UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DISEÑO E IMPLEMENTACIÓN DE UNA METODOLOGÍA DE PREDICCIÓN DE FUGA DE CLIENTES EN UNA COMPAÑÍA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

FRANCISCO JAVIER BARRIENTOS INOSTROZA

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
GASTÓN L'HUILLIER CHAPARRO
SANDRA TERESA MOLINA MENA
ANDRÉS FELIPE CHACÓN SANDOVAL

SANTIAGO DE CHILE
DICIEMBRE 2011

Resumen Ejecutivo

La minería de datos es una nueva tecnología que está cobrando relevancia en la actualidad, su utilidad para resolver complejos problemas a lo que se enfrentan las empresas (de múltiples variables y casos) ha dado entrada a la aplicación e investigación sobre la misma. Sin embargo, esta tecnología no es una heurística cualquiera, se fundamenta en la rama de las ciencias de la computación denominada inteligencia artificial y las matemáticas mediante la estadística.

En un comienzo, las empresas sólo se preocupaban por el almacenamiento de los datos, datos históricos que permitían cálculos matemáticos simples con una finalidad, la generación de reportes. De esta manera, se buscaba responder las preguntas referentes al control del negocio. Posteriormente se profundizaron estas preguntas de control hasta llegar a la creación de un repositorio consolidado, expresado en la tecnología de *data warehouse*. En la actualidad Las exigencias de los consumidores cada día aumentan más, puesto que la competencia comienza a ser más dinámica, por ende, para establecer una ventaja competitiva, las empresas requieren responder preguntas que van más allá de los datos históricos, es decir, necesitan extraer información que pueda ser útil para el futuro, y de esta manera, dejar el paradigma de una empresa reactiva y pasar a ser una entidad proactiva y preventiva. En este nuevo desafío aparece la tecnología de minería de datos, la cual va inserta en un procedimiento Knowledge Discovery on Databases (KDD), puesto que para obtener información del futuro se debe estar seguro del presente.

Esta tecnología se aplica actualmente en variadas empresas, sin embargo, no se vislumbra explícitamente. Las personas son afectadas por ella como parte de un paradigma de consumismo, cuando compran un producto y se le hace un descuento, un aviso publicitario mencionando la promoción de un nuevo producto, cuando se les ofrece un crédito bancario o se les llama telefónicamente para mejorar un servicio que ya tienen contratado, e incluso cuando ingresan a Internet para navegar en sus redes sociales o buscar información. También se ve en los avances biológicos como un diagnóstico rápido y efectivo, una cura basada en la ingeniería genética, entre otros.

Actualmente la minería de datos se ha subdivido en múltiples ramas según su aplicación, es así, como se pueden encontrar distintos tipos de minería: Web, de Texto, de Procesos. Estos solamente generan la diferencia en la perspectiva en que se ejecuta el KDD, siendo el último tipo el más reciente. Cabe mencionar que los principales algoritmos de han adaptado según su uso y día a día se implementan mejoras sobre los mismos. Análogamente, también, se desarrollan nuevas formas de valorización sobre sus resultados.

Esta memoria busca investigar sobre el KDD y las distintas técnicas que pueden ser utilizadas, para luego aplicarlas a un producto particular en una empresa determinada. En ella se describen todos los procesos por los cuales se transcurrió cada uno visto desde el punto de vista del KDD, por

lo que su estructura es como realizar un KDD a un documento de esta índole.

Sin embargo, no todo fue la aplicación, puesto que se refinan los modelos y algoritmos tanto de transformaciones como de imputaciones de datos, lo que converge en un aprendizaje incremental, en el que cada intento es expresado como relevante puesto que destaca una etapa particular del KDD.

Además, de describir la aplicación del KDD se añade una evaluación comercial utilizando recursos de la compañía y bajo el soporte del área de Aseguramiento de Ingresos y la Vicepresidencia Comercial. En base a esta evaluación comercial, se tiene la evaluación técnica de cada modelo y las peculiaridades que se forman al efectuar el contraste entre ambas. Adicionalmente se evalúa monetariamente los resultados obtenidos desde dos puntos de vista, lo que conlleva al establecimiento de propuestas futuras.

Agregado a lo anterior, se presentan problemáticas no documentadas, debido a que su acontecer es propio dentro de lo que es desarrollar un proyecto que tiene al KDD como eje articulador. A su vez, se muestran soluciones y planteamientos para ingresar un proyecto a un área determinada, en otras palabras, se presentan herramientas que ayudan a generar confianza al interior de una empresa para que origine un cambio a nivel organizacional respecto a esta tecnología

Finalmente se concluyen los aprendizajes y las acciones correctivas que debiesen ejecutarse en caso de implementar el piloto a nivel operacional.

Agradecimientos

Esta etapa universitaria se cierra en esta memoria, mas es sólo el comienzo de mi vida como profesional. Sin embargo, nunca hay que olvidar a aquellos que fueron pilares fundamentales tanto en toda la vida, como en un punto en particular. Por ende, agradezco a todas las personas que me ayudaron en este camino, que cada vez se refina más y más.

Quiero agradecer a mi familia por todo el apoyo que me brindó desde niño, alentando cada uno de mis pasos para ser un profesional aunque en realidad fuese para tener más herramientas frente a la vida. Mi modo de ser, mi inteligencia, mi sabiduría la heredé o aprendí de ellos. En particular, esta memoria va dirigida a mis padres, Juan Barrientos y Norma Inostroza, quienes han sido el pilar más sólido en mi vida y las personas más importantes en ella. Por lo mismo, este documento, es uno de los regalos más preciados que espero poder hacerles. Gracias por todo lo que me han dado y lo que me dan. Sin ustedes esto no sería lo mismo.

Respecto al negocio y todo lo que fue la facilitación de información quiero agradecer enormemente la tutoría de don Andrés Chacón, quien fue el mejor de todos los instructores que he tenido dentro de una empresa, gracias por enseñarme a vivir dentro de una compañía, por todo el soporte que me entregaste, por dejarme trabajar sin presionar, por alentar mi creatividad en este proyecto, por aguantar mis desesperaciones producto de que el modelo no daba, simplemente te pasaste, un 30 % de esta memoria es tuya. Además, agradezco a Gonzalo Sepúlveda quien al igual que Andrés, siempre tuvo una disposición excepcionalmente abierta para ayudar en el desarrollo de esta memoria, entregando todo lo que pudo. También he de agradecer a la persona que desarrolló la tesis base en la compañía la Sra. Sandra Molina, puesto que gracias a ella, la memoria tomó un giro más interesante y a quien me introdujo a esta empresa de forma externa Eduardo Duran, muchas gracias. Adicionalmente quiero agradecer a quien facilitó las bases de datos correspondientes con una paciencia de oro, Solange Pastén, Carmen Gloria Insunza y María Elena Olmos.

Académicamente estoy muy agradecido con gran parte de los profesores de la Universidad, en especial con tres personas. La primera de ellas fue la que por primera vez me habló de análisis de datos y me incentivó en sus clases a seguir averiguando del tema, esta persona es Manuel Reyes, gracias profesor, sus clases prácticas reveló una perspectiva más profunda acerca de lo que parece tan superfluo. La segunda persona es mi profesor guía, Sebastián Ríos, ahora bien, generalmente las personas dicen que para que un alumno trabaje se le debe presionar constantemente, sin embargo, si mi guía hubiese hecho tal cosa esta memoria no hubiese sido lo mismo, al igual que Andrés usted me dejó explorar todas mis ideas (aún las más locas) sobre este proyecto, me dió responsabilidad por él, en otras palabras, terminó mi enseñanza de cómo ser un profesional, le agradezco también su apoyo frente a situaciones límites, de hecho gracias a usted entendí que: “No es mejor que te presionen,

puesto que si lo hacen limitan tu creatividad", y esa es la clave para generar un proyecto entretenido, creatividad. La tercera persona, diría yo que fue mi maestro del Mining, don Gastón L'Huillier, esta persona fue mi co-guía y ya había hecho su labor hace un año atrás en el curso de Minería de Datos, todo lo que sabía previo a esta memoria se lo debo a él, es una persona extraordinaria, un excelente docente al igual que Sebastián, ambos solamente frenan la idea cuando el tiempo lo amerita, muchísimas gracias Gastón, por lo que me enseñaste, por lo que me ayudaste a aprender (al ser uno de los gestores en la introducción a la compañía) y quizás qué termine aprendiendo de ti pero sé que será entretenido. A mis dos profesores guías, les agradezco mucho (incluso la ayuda en latex).

En la universidad aparte de profesores, conocí varias personales especiales a quienes está memoria presenta sus agradecimientos, puesto que fueron gestores de la misma: Hector Álvarez y Eduardo Merlo, respecto a la primera persona, agradezco su ayuda para aprender a escribir en latex y sus lecciones sobre no ser un alumno tan protestante que me llevarán a ser un buen profesional. La segunda persona, es un alumno ejemplar que también me ayudó en el diseño de varias presentaciones y entregó su aporte en la programación de esta memoria, te agradezco tu forma de ser Edu, eres una de las pocas buenas personas que quedan. Como mención honrosa a quienes me apoyaron en esta etapa universitaria se encuentran todos mis conocidos, de los que destaco: Carlos Arancibia, Carlos Villa, Sandra Jeraldo, Pablo Junyent, Nabor Erices, mil gracias.

También he de agradecer a Roberto Jaramillo quien es un ser genial, desde la adolescencia hasta la universidad agradezco todo tu apoyo, las discusiones que sostuvimos, los juegos de pool, etc. Un gran amigo. Adicionalmente agradezco a Evelyn Araya quien también me envío todas sus buenas energías para que esto fuese posible. Gracias.

Agradezco a la familia de mi pareja quien me entregó mucho apoyo y sus rezos por las noches, también ellos, han de quedar escritos en esta melodía denominada memoria, pues ellos también facultaron algunas notas.

Finalmente quiero agradece a mi pareja Valeria Zapata, quien aparte de ayudar en un 100 % en la corrección de este documento, me entregó todo su apoyo y amor, uno de los seres más geniales con los que he tenido la oportunidad de compartir grandes extensiones de tiempo. Agradezco todas sus enseñanzas y la larga espera, en la que tomé como ejemplo su esfuerzo y sacrificio, una profesional de alta calidad. Agradezco las vivencias que me hizo pasar y el apoyo que me entregó en esta etapa.

Gracias a todos por compartir este momento conmigo!

Tabla de contenido

Índice de figuras

Índice de cuadros

Capítulo 1

Contextualización Memoria

1.1. Introducción

En esta memoria se presenta la descripción de actividades a seguir para poder predecir de manera óptima el churn de los clientes del servicio NGN (Next Generation Networks) de una compañía de Telecomunicaciones Nacional. Para ello, se utilizará el procedimiento denominado KDD (Knowledge Discovery on Databases), generalmente usado para solucionar problemas de esta índole. Dicho procedimiento posee 5 etapas: integración, preprocesamiento, transformación, modelamiento y evaluación. El resultado de cada una de ellas está sujeto a la salida de la etapa anterior. Bajo estas etapas, se pretende integrar los datos de la muestra de clientes pedida por la empresa, preprocesar sus datos, originar transformaciones, implementar un modelo y generar conclusiones de fácil interpretación para las áreas clientes puedan tomar medidas efectivas respecto a la retención de clientes.

Además, se plantean herramientas adicionales que fueron usadas como complemento este trabajo tales como la segmentación, clusterización, análisis factorial. En términos aplicados se utilizan tres softwares disponibles académicamente: Rapidminer (versión 5), Microsoft Office y SPSS. Cada uno de ellos estuvo abocado a un aspecto particular del KDD.

Este trabajo se fundamenta en una tesis de magister de la Sra. Sandra Molina [77]. La cual establece las bases conceptuales de la memoria. Dicha experiencia, fue efectuada en el mismo producto, por lo que también se usa tal y como se explica en los capítulos siguientes.

En un mercado como el de las telecomunicaciones, esta tecnología se hace necesaria, puesto que su constante variabilidad sugiere un conocimiento más avanzado sobre el comportamiento de los clientes. Además, se ha descubierto la presencia de áreas dedicadas a ella a nivel nacional, por lo que el interés de esta empresa particular facilitó la obtención de datos y ejecución de entrevistas telefónicas.

Este documento presenta la valorización monetaria de los resultados de cada experimento desde dos perspectivas, una basada en el ciclo de vida del cliente y la segunda representando la fuga de ingresos. De esta misma manera, se ahonda en varias estrategias para la aplicación del KDD.

Cabe destacar que el comportamiento de los clientes estudiado en este documento muestra una peculiaridad que lo hace distinto a la mayoría de los problemas tratados generalmente. Dicha particularidad alude a la detección de un número reducido de clientes que se encuentran en una base extensa de datos.

1.2. Antecedentes Generales

1.2.1. Antecedentes Técnicos

La idea que avala el uso del procedimiento KDD en la compañía, es utilizar la herramienta de modelamiento más adecuada para el negocio en cuestión, de manera tal que pueda ser aplicada a futuro por la empresa. Tema que en el interior de la misma es inexistente a nivel operacional en toda área.

A su vez, se analiza una experiencia previa que existe acerca de esta materia. Esta fue confeccionada y defendida por una persona perteneciente a la compañía. En base a esta experiencia, se puede reducir el tiempo de la etapas de preprocesamiento y transformaciones de datos, además, del hecho de que gran parte de la etapa de integración de datos haya sido mejorada bajo un *data warehouse* implementado por la compañía.

El estudio del churn en el área de las telecomunicaciones se ha hecho cada vez más necesario producto de la alta competitividad que se está desarrollando en Chile. Prueba de esto, son los altos niveles de fuga de clientes (alrededor de 30 %) que se presentan en diversas empresas del medio [57, 77]. A partir de lo anterior, se hace necesaria la aplicación de herramientas avanzadas que permitan tomar acciones proactivas frente a la fuga de clientes. Es aquí donde se requiere el conocimiento de la minería de datos o data mining, la que contiene en sí una metodología cuasi-estándar a seguir para llegar a la predicción de los clientes que sean más proclives de terminar su contrato con la compañía para migrar a la competencia.

Las líneas de negocio presentes en la empresa de telecomunicaciones son: tráficos de larga distancia, telefonía local, internet, cargos de accesos, servicios privados, facturas y cuentas corrientes, atención de clientes (referente a las solicitudes de atención y reclamos), clientes y contratos, modelos de operaciones, participación de mercado y suscriptores [12]. El diseño de la metodología va enfocado específicamente al servicio NGN que, a modo de síntesis, se define, en este informe, como un "pack" de servicios con un solo acceso para las Pymes, que consiste en un agrupamiento de los servicios de telefonía local, servicios de banda ancha y servicios de larga distancia. El principal atributo de este "paquete.es que se adapta a las necesidades de los clientes y que cuenta con la flexibilidad de poder incluir nuevas herramientas (más avanzadas) dentro de él [77].

Además de los antecedentes previamente descritos, en el transcurso del proyecto la empresa estuvo sujeta a una fusión interna lo que produjo múltiples cambios tanto de personal como de direccionamiento estratégico de la misma.

La base de datos de la compañía acepta el uso de consultas SQL, cubos OLAP, repositorio *warehouse* y, además, presenta un prototipo de modelo de data mining ya confeccionado en la compañía. No obstante, el acceso a ellas está restringido para personas externas a la compañía, por lo tanto, en pos de su obtención se debió acudir a terceros.

1.2.2. Antecedentes de mercado

El rubro de las telecomunicaciones se refiere a todo lo que tiene que ver con 4 servicios que actualmente se están presentando a las personas y estos son: telefonía local, telefonía móvil, internet y televisión de prepago, siendo este último el más novedoso dentro de la industria[6, 7].

A nivel mundial, el tamaño de esta industria es de 4,03 trillones de dólares, y se pronostica un

crecimiento de un 6 % anual hasta el año 2013[1]. Dentro de este mercado se puede apreciar el tamaño de la industria que exceptuando el mercado estadounidense llega a los 3,033 trillones de dólares, siendo los principales actores China, Japón, Alemania, Reino Unido, Brasil, Italia, Francia, India, México, España, Corea del Sur, Rusia que componen cerca del 47 % del gasto internacional en el año 2009 [6].

Ahora bien, para el sector de Latinoamérica esta cifra de gastos en Equipos y servicios es de 398,6 Billones donde se proyecta un crecimiento del 8,6 % en el 2010 y un 10,7 % para el 2011 mientras la economía se expanda [6].

Desde una perspectiva más general y de tendencias, *"a nivel mundial...la telefonía local continúa perdiendo protagonismo con tasas de crecimiento cada vez menores en la mayoría de los países salvo algunas excepciones en países en vía de desarrollo...las grandes empresas telefónicas siguen evolucionando hacia la incorporación de..IPTV, la TV por ADSL, ADSL2+, acceso a internet con un mayor ancho de banda"* [7]. Por lo que en este ámbito resulta relevante el estudio del churn para el caso de la telefonía local, como medida reactiva a la pérdida de clientes a nivel mundial.

Sin embargo, la crisis financiera del años 2008 golpeó a nivel mundial a este sector, de manera que ha quedado con ciertas pérdidas. Por ejemplo *"en EEUU, la industria de la telefonía local no ha podido reducir los costos en proporción a la caída de ingresos...existió una caída en las nuevas conexiones de banda ancha, provocada por la disminución de la confianza del consumidor...la inversión de capital de operadores estadounidenses fue a la baja durante el año 2008"* [7], además, existió una caída de ingresos a nivel global de 4,02 a 3,87 trillones de dólares entre los años 2008 y 2009 [1].

En un contexto de país, el PIB del sector de las telecomunicaciones *"siguió creciendo aunque a una tasa menor a la presentada el año anterior"* [7], esto se puede apoyar en el siguiente gráfico:

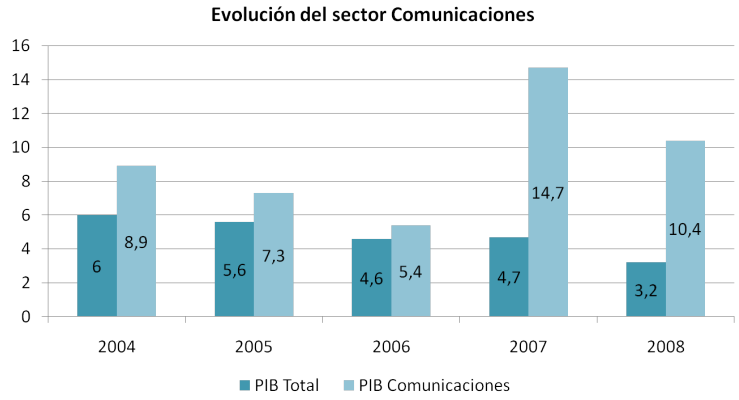

Figura 1.1: Evolución sector de Comunicaciones

Cabe señalar que la crisis financiera mundial no impactó a Chile el 2008 debido a *"la ley de bancos que sustenta a Chile...ante el otorgamiento de un crédito hipotecario, puede seguir otros bienes del patrimonio del deudor para recuperar el crédito"* [7].

También, se puede visualizar el tamaño en el país del sector por las inversiones reales (no financieras), tal como se aprecia en la figura 1.2 de inversión y empleo [3]:

Los supuestos de los valores de esta tabla son los siguientes:

Empresa	2004 [Millones de dólares]	2005 [Millones de dólares]	2006 [Millones de dólares]	2007 [Millones de dólares]	2008 [Millones de dólares]	2009 [Millones de dólares]
AT&T Chile	8	11	8	63	1.844	1.014
Carrier 121	2.475	2.171	2.639	4.513	4.215	3.048
Entel-Chile	32.146	36.396	41.223	58.163	56.966	63.305
Chilesat (Telmex)	9.378	4.054	3.274	19.767	66.937	23.648
Movistar	81.832	106.980	117.086	157.421	170.715	143.304
Claro Chile	14.591	25.474	129.977	63.154	91.695	61.761
Entel PCS	69.748	69.250	116.744	140.999	197.377	189.460
CMET	984	1.221	4.406	3.197	2.944	112
Telefonica del Sur	12.157	7.016	8.942	18.927	25.106	15.690
Telefónica Chile	65.320	72.393	84.052	103.859	118.516	111.355
CTR	1.351	813	920	852	2.069	906
Entelphone	1.558	2.234	1.039	844	1.113	916
GTD Manquehue	3.057	3.123	2.243	3.545	3.067	2.816
VTR	44.757	59.419	68.376	86.889	81.406	94.794
Telmex	-	12.118	20.274	14.938	16.237	21.180
Telmex TV	-	-	-	3.240	35.316	10.801
Directv Chile	2.907	7.390	2.474	2.827	2.905	3.844
Telefónica Multimedia	-	-	22.550	33.389	20.216	11.422
Nextel	4.897	226	5.157	5.905	13.375	10.557
Total	**351.735**	**411.788**	**636.045**	**727.951**	**916.420**	**778.153**

Figura 1.2: Inversiones en Comunicaciones

1. Las empresas informan la inversión real (no financiera) realizada durante el año respectivo, expresada en Millones de pesos corrientes [3].

2. La información se refiere a aquellas inversiones relacionadas con los servicios de telecomunicaciones que se prestan y referidas al RUT que informa[3].

3. Sólo se incluyen aquellas empresas que informan en el STI(Dirección de Servicios de tecnologías de información)[3].

4. Sólo se colocan las inversiones superiores a 1.000 millones de pesos.

A lo que, además, se agregan los niveles de reclamos por empresa [9], ya que la principal razón para que un cliente deje de comprar los productos de una compañía según Rossat en [90] son la disconformidad y la falta de políticas de retención efectivas expresadas en un mejor trato hacia ellos. Para posicionar lo anterior en un contexto nacional se presenta la figura 1.3.

Se puede apreciar en la figura 1.3 que el nivel promedio de reclamos cada 10 mil líneas es de 3,5. Un punto relevante es que la compañía de telecomunicaciones a la cual se le está ofreciendo este proyecto churn, presenta una tasa menor de reclamos en la SUBTEL. Esto puede significar que los mecanismos de reclamos se han dado a conocer, lo cual es el primer paso para la retención de los clientes. Dentro de este marco de reclamos, las 10 razones principales detectadas por la SUBTEL (Subsecretaría de telecomunicaciones) son [4]:

Donde se puede destacar la dificultad del término de contrato, que alude directamente a intentos de retención mediante burocracia, que al final, conllevan a empeorar la imagen de la compañía y la disconformidad con la suscripción o continuidad del servicio, lo que requiere usualmente proactividad del ejecutivo a cargo. No obstante, al no contar con alguna alarma, queda imposibilitado a realizar acciones de retención efectiva.

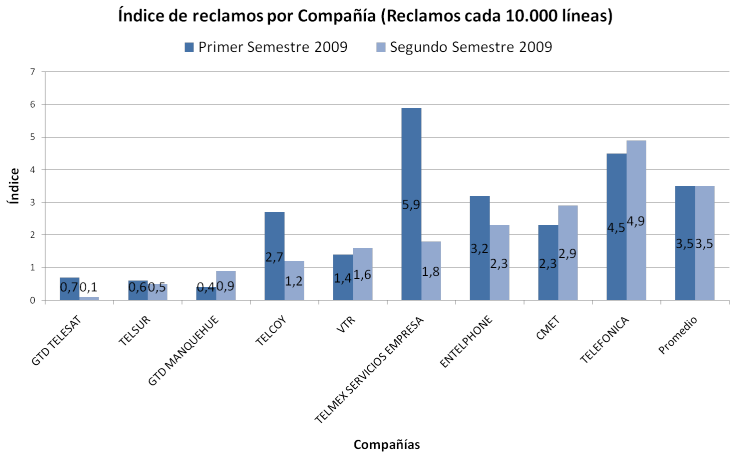

Figura 1.3: Indice reclamos compañía

Motivo	Cantidad
Disconformidad con la suscripción del servicio	1.652
Disconformidad con la continuidad del servicio	1.029
Disconformidad con cumplimiento del contrato o promoción	844
Dificultad con el término de contrato	648
Disconformidad con la calidad del servicio prestado	460
Disconformidad con cobro de saldo anterior	324
Los distintos cobros de llamadas	179
Disconformidad con cobros de la larga distancia internacional	268
Disconformidad con cobro de servicios complementarios	228
Motivo se definirá durante la revisión (reclamo en línea)	228

Figura 1.4: Motivos que causan los reclamos

Benchmarking de minería de datos en telecomunicaciones

Como el rubro de las telecomunicaciones ha crecido de forma sostenida desde hace un tiempo, la cartera de clientes y la información de los mismos ha aumentado, generando una característica casi propia del grado de información que manejan las telecomunicaciones, citando a [17] *"Uno de los rasgos que caracterizan al sector de las telecomunicaciones es la cantidad de información que generan y almacenan sus empresas"*. Por ende, nuevas tendencias e ideas han aparecido para hacer uso de la minería de datos en el rubro, donde su uso más común se aplica en *"tres ámbitos fundamentales: marketing, detección de fraudes y control de calidad"* [17].

En lo que se refiere al marketing, el almacenamiento de los registros de llamadas y la información demográfica de los clientes, la minería de datos puede generar perfiles de clientes, pues *"si conseguimos segmentar nuestra clientela en grupos con características similares, será mucho más sencillo emprender una campaña de promoción al conocer cómo es éste segmento y qué es lo que busca"* [17]. No obstante, el segmentar la cartera de clientes no es el único beneficio que plantea la minería de datos para el marketing en las telecomunicaciones. También se puede apreciar que *"otra de las utilidades de este profundo estudio de la clientela es su fidelización* [17]. Las empresas del sector están especialmente sensibilizadas con la pérdida de clientes que escogen una compañía de la competencia, puesto que es un proceso fácil y está demostrado que *"el coste de conseguir*

un cliente nuevo es sustancialmente más costoso que mantener al antiguo" [17, 57, 67, 93]. Esta última afirmación se justifica en que *"la estimación del costo anual en las telecomunicaciones es de 4 mil millones de dólares* [57]. Dentro de esta categoría se puede mencionar el caso de *"Vodafone que aplicó una segmentación y catalogación de los clientes"* [17]. Cabe destacar es en este ámbito donde existe el término de churn de clientes (el cual se explicará posteriormente en la sección 2.4.1), en particular en el problema de la fidelización de los mismos. Además, se agrega que aún en los casos de pack de servicios por empresas esto suele suceder, por lo que el modus operandi de la industria en general ha cambiado desde la adquisición de clientes a la retención de los mismos [53].

Otro ámbito que puede verse afectado por la minería de datos es el de control de gestión, mediante la detección de fraude, donde el fraude en las telecomunicaciones es *"prestar un servicio sin obtener contraprestación económica alguna"* [17], sin embargo, este se puede dividir en dos considerando las principales causantes de este fenómeno: el primero es *"fraude suscriptivo, donde el cliente contrata el uso de una línea pero no paga sus facturas al día...el adecuado uso de la minería de datos puede elaborar modelos que predigan qué clientes son más susceptibles de dejar de pagar sus facturas"* [17], en cambio el segundo, es *"el fraude superimpuesto, donde una persona no registrada accede a la línea de un cliente para hacer un uso fraudulento de ella...es posible su detección utilizando minería de datos para analizar los cambios en los comportamientos de los clientes"* [17]. Respecto a lo anterior, el primer fraude resulta fácil de detectar con una segmentación, no así el segundo, en donde se deben analizar los valores fuera de rango de los atributos que tengan los registros correspondientes, en este caso los clientes. Para este ámbito existe el ejemplo de la aplicación de minería de datos en la compañía de telecomunicaciones *"Brasil Telecom que implementó un sistema para la detección de fraudes"* [17].

El último ámbito detectado, en donde la minería de datos pudiese entregar información de apoyo, es el área técnica de la empresa, en particular en el proceso que se refiere al control de calidad, pues *"dada la complejidad de las redes desplegadas actualmente, el operador necesita ser capaz de solucionar cualquier error en el menor tiempo posible...Ante esto, la minería de datos puede ayudar a interpretar esta información permitiendo al operador saber cuando y dónde se localiza un problema"* [17]. El problema anterior, se sintetiza en la segmentación y predicción de fallas de cada componente que el cliente tenga, utilizando los atributos propios de cada producto. Un ejemplo para este ámbito fue el de la creación de *"TASA, un sistema que descubre regularidades en las alarmas y puede localizar episodios frecuentes de alarmas presentándolos como normas"* [17].

Todos estos ejemplos son, en su mayoría, casos externos al país, sin embargo, se descubrió que a nivel nacional existe una empresa inserta en el mercado de las telecomunicaciones. Esta empresa ha llevado el KDD a otra escala, diferente de la conslutoría o proyecto singular, se trata de una implementación del KDD del orden organizacional. Esta empresa pasó por una etapa de cambios organizacionales lo que permitió la entrada del KDD como área (subgerencia) de la compañía cuyo rol fuese facilitar información al resto. Por lo tanto, sus roles como identidad son de análisis, sistemas y operación. Consecuentemente, su jerarquía de cargos presentada en 1.5 muestra estos roles agrupados por trabajo. Esta área se encuentra actualmente operando y sus principales resultados en un período de 4 meses han sido los siguientes:

Donde el grupo de blindaje tiene su explicación en la figura 1.6, lo que puede ser descrito como el grupo de clientes que probablemente se fugarán y que se le aplicarán campañas de retención, por consiguiente, el grupo de control es aquel al que no se le aplican estas acciones.

Cuadro 1.1: Resultados de aplicación del KDD organizacional

Mes	Blindaje [%]	Grupo de control [%]
1	1,34	3,92
2	0,67	3,89
3	0,76	3,95
4	0,85	4,01

Figura 1.5: Jerarquía del área respecto a sus roles

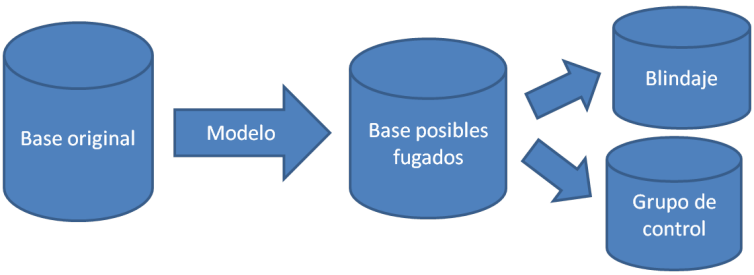

Figura 1.6: Procedimiento para la obtención del grupo de blindajes

Un punto relevante es que el tiempo que tardó esta empresa en implementar el KDD a este nivel fue de 3 años.

1.3. Descripción del Proyecto, Planteamiento del Problema y Justificación

La finalidad del proyecto es introducir el proceso del KDD dentro de la empresa de telecomunicaciones para que pueda ser aplicado en todas sus áreas. Particular en la detección de fuga de clientes. Por lo mismo, el problema tratado consiste en predecir la fuga de clientes en esta empresa para un producto particular (NGN) cuyo segmento objetivo son las pequeñas y medianas empresas (PYMES), siendo lo anterior representado en el objetivo general del estudio.

Un punto relevante relacionado con el churn es que puede ser clasificado en involuntario, aquel que se refiere a cuando la empresa es quien decide terminar el contrato con el cliente por asuntos de que el cliente no se ha comportado de la forma esperada; o bien, Voluntario, donde el cliente es quien decide cambiar de proveedor de servicios [77]. Esto aclarará más adelante la parte específica

donde se concentrará el estudio de la predicción de churn para NGN.

El servicio NGN, visto de aquí en adelante como el producto en cuestión, se presenta al cliente, en forma de versiones donde cada modificación a realizar mejora el producto. Las modificaciones referidas pueden ser: oferta comercial, herramientas de apoyo y redes disponibles.

El problema de predicción de churn es de aprendizaje supervisado, descrito de la siguiente forma: *"Dado un horizonte de pronóstico determinado, el objetivo es predecir los futuros fugados sobre ese horizonte, dado los datos asociado a cada uno de los clientes en la red de clientes"* [86].

La definición misma de churn es incierta, no obstante, en el sector de las telecomunicaciones el término *"es usado para describir colectivamente el cese de servicios de la suscripción de un cliente...donde el cliente es alguien que se ha unido a la compañía por al menos un período de tiempo...un churner o fugado es un cliente que ha dejado la compañía"* [43].

Ahora bien, la razón de que el churn sea un problema en las compañías es que *"causan una gran pérdida de servicios en línea"* [43], además, implica una pérdida en el sentido de que el cliente se retira y, por consecuencia, la empresa deja de percibir ese ingreso. En otras palabras, la pérdida completa está relacionada tanto con el costo que significaba mantener esa línea sin uso, como en el ingreso perdido al momento de que el cliente se retira oficialmente de la compañía.

El horizonte de estudio planteado para el proyecto fue de seis meses, para generar resultados. Sin embargo, mediante entrevista informales con el área comercial relacionada directamente con el producto, se obtuvo el conocimiento de que el período de vida de un cliente en la empresa era de 17 meses, lo cual discrepó frente a la mayoría de las telecomunicaciones donde le período estándar de un cliente es de 6 meses.

Otra característica del problema es que el churn es bastante pequeño (1 % aproximadamente) por lo que entra en el problema de rareza de clases (explicado en [107]), que en pocas palabras, señala que existe una clase mayoritaria sobre la otra, lo que tiene como consecuencia el hecho de que el algoritmo o modelo a usar tiene menos cliente o registros para "aprender" a predecir. Ahora bien, la razón para estudiar los churns pequeños es cuando ocurre la situación de que la tasa de ingreso es equivalente a la tasa de churn, por lo que se presenta un decaimiento del producto que converge al término de su comercialización. Esto es aplicable a la empresa tal y como se bosqueja en el gráfico 1.11, cuando el ingreso es menor a la tasa de fuga, entonces la tarea de mantener a los clientes cobra relevancia y por lo tanto el estudio de este tipo de churns.

1.3.1. Problema en el interior de la empresa

Para obtener una mejor visualización del proyecto en la empresa se procede a describir la inserción del mismo a partir de una perspectiva de negocios. Por consiguiente, se muestran las fortalezas y debilidades que presentaba la empresa para la implementación del piloto de KDD, así como también, las áreas que sustentaron el proyecto completo.

Fortalezas y Debilidades

En el siguiente recuadro extraído de [37], se visualizan las fortalezas y debilidades desde un punto de vista de evaluación de proyectos informáticos. Al cual se añade la fortaleza de que existió una experiencia previa al proyecto de KDD en la empresa para el mismo proyecto, lo cual facilitó las etapas del KDD.

Cuadro 1.2: Fortalezas y debilidades detectadas en la empresa para el piloto de KDD

			Fortaleza	Debilidad
Recursos Humanos	Habilidades	Aspectos técnicos		x
		Cultura Organizacional		x
	Capacidades	Número de personas	x	
		Disponibilidad de tiempo	x	
Recursos Técnicos	Hardware		x	
	Software			x
	Comunicaciones			x
Recursos Económicos	Existentes		x	
	Futuros		x	
Administrativa	Infraestructura		x	
	Procedimientos administrativos			x
	Compromiso y soporte organizacional			x

Áreas involucradas en el proyecto

Las áreas que apoyaron la creación e implementación del piloto se muestran en la figura 1.7 encerradas por un círculo, cabe señalar que este organigrama es representativo, debido a que a la fecha en la fue confeccionado no existía una clara estructura del mismo:

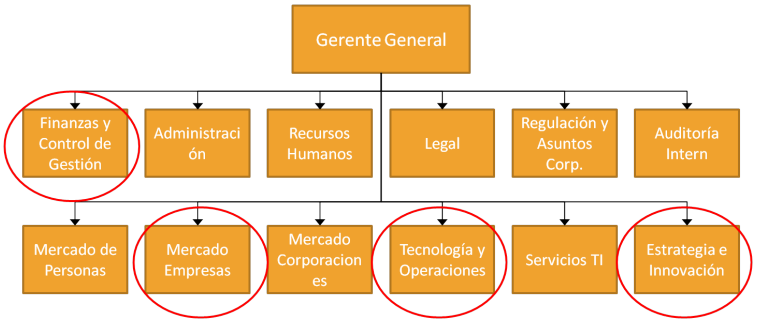

Figura 1.7: Organigrama de la empresa

En el mismo se puede destacar que el problema se ubicó esencialmente en el área de Mercado Empresas, lugar en el que se encuentra actualmente el producto NGN.

Período de estudio y aplicación para cada experimento

En la siguiente tabla se muestran, los períodos usados para estudiar cada situación del KDD y elaborar un experimento respectivo, no obstante, se debe enfatizar que estos no son los períodos de estudio de las bases de datos (los cuales son 6 meses para cada experimento), si no que se refieren a los meses utilizados para generar el experimento correspondiente:

Cuadro 1.3: Períodos de generación de experimentos

Experimento	Período de estudio y aplicación
1	Marzo-Abril 2010
2	Mayo-Junio 2010
3	Julio-Octubre 2010
4	Diciembre 2010
5	Diciembre 2010
6	Enero-Febrero 2011
7	Marzo-Abril 2011

1.4. Objetivos

1.4.1. Objetivo General

Dadas las secciones anteriores, el objetivo principal de este trabajo de titulo es: "Diseñar e implementar una metodología para la predicción del churn de los clientes del servicio NGN (Next Generation Network) de una compañía de telecomunicaciones y evaluarla empíricamente contactando a los clientes".

1.4.2. Objetivos Específicos

- Replicación y reconstrucción de experiencias anteriores documentadas.
- Lectura y recopilación de antecedentes de predicción de churn en telecomunicaciones y otras empresas.
- Caracterización del churn en la compañía estudiada. Búsqueda y selección de fuentes de información.
- Establecimiento de métricas de evaluación, para los modelos planteados y las predicciones.
- Diseño de un repositorio para almacenar los resultados obtenidos históricamente.
- Evaluación los resultados obtenidos.
- Validación con encuestas telefónicas los resultados obtenidos.
- Implementación un procedimiento de predicción de churn.

1.5. Descripción de Metodología a utilizar

Para una mejor comprensión de la realización del estudio, el siguiente esquema sintetiza la metodología de trabajo ejecutada en esta memoria:

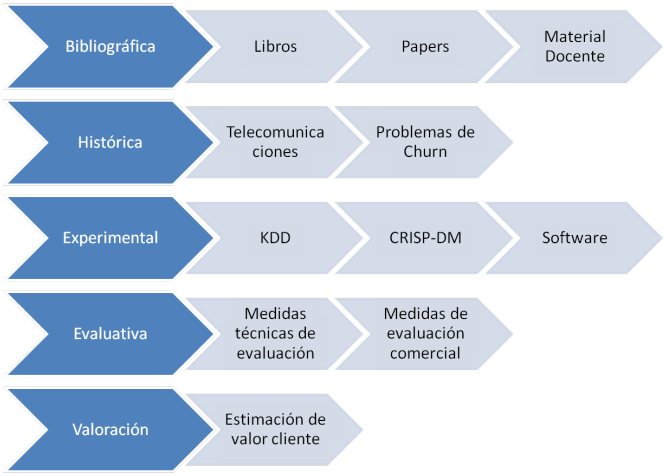

Figura 1.8: Metodología de trabajo

La metodología que se pretende utilizar en este estudio, considera el procedimiento KDD como forma de apoyar la realización del modelo predictivo. Por ende, abarcará un poco más allá de los que es el KDD y los resultados que entrega, es decir, se desea agregar una etapa ausente en dicho procedimiento, ya que la última etapa del KDD consiste en traspasar los resultados del modelo ocupado, lo que no equivale a valorizarlos. Por esto, el estudio pretende agregar esta etapa de valorización del nuevo conocimiento y establecer acciones correctivas que puedan generar mayor impacto en la empresa de telecomunicaciones. Una vez dicho esto, se procede a presentar la metodología en los siguientes pasos:

- Antecedentes de experiencia previa acerca del producto NGN.
- Especialización con Software Escogido.
- Descripción de datos del producto.
- Inspección principales sistemas y responsables de los mismos.
- Prueba de resultados antiguos con base de experiencia anterior.
- Representación del churn actual con el KDD.
- Estudio de modelo adecuado para la situación actual.
- Discusión de resultados.
- Establecimiento del estudio como prototipo.
    - Actualización del procedimiento KDD
    - Valorización de churn y Resultados
    - Prueba histórica sobre el modelo escogido

    - Generación de prototipo completo de prueba

- Mejoramiento del KDD

- Establecimiento de acciones correctivas

Con estos puntos expresados anteriormente, se detalla cada uno de los mismos, en las páginas siguientes, para esclarecer el proceder en cada uno de ellos y los posibles obstáculos que se puedan encontrar en ellos.

1.5.1. Antecedentes de experiencia anterior acerca del producto NGN

Considerando la experiencia previa se procedió a generar un completo análisis de esta experiencia, destacando sus punto débiles y que podrían reforzarse en el estudio actual. Es así, como se extrajo una muestra de la principales bases de datos usadas y sus respectivas variables, constatando el preprocesamiento realizado y los modelos ejecutados. Con dicho análisis se muestra a continuación las propuestas que surgieron:

- Se sugiere un nuevo preprocesamiento de los valores perdidos y los valores fuera de rango para observar el impacto de este tipo de valores en el KDD.

- Se propone modelar el problema como vista general en dos líneas: una supervisada y una no supervisada, partiendo con la primera mencionada para ver el mejor rendimiento posible (debido a que se posee información adicional sobre la variable objetivo) y, en base a dicho rendimiento, se decidirá si proseguir al nivel de no supervisado.

- Se contemplan los modelos tipo multiclasificador usando diferentes modelos, entre ellos: árboles de decisión, naive bayes, SVM, redes neuronales, regresión logística, y en el caso de técnicas no supervisadas se puede recurrir a algoritmos de clustering, como por ejemplo, el *fuzzy C-means* o una variante de éste.

- Se realizará un clustering conceptual para vislumbrar la distribución de los clientes dentro de la base de datos, y para entender la relación preliminar entre la variable a predecir y las predictoras.

- Como este es un problema de clasificación en una vista preliminar se sugieren las siguientes técnicas para usar en el modelamiento [24]: análisis discriminante, métodos de regla inductiva o de asociación, árboles de decisión, redes neuronales, algoritmos genéticos, K vecino más cercano.

- Para el aprendizaje de los modelos se usará la misma técnica de evaluación (cross validation) y, además, se pretende establecer un período de 2 a 3 meses de prueba para su validación.

1.5.2. Especialización del software escogido

Aparte del análisis de la experiencia anterior, se procedió a investigar los distintos softwares que permiten entregar una solución al problema señalado. Dentro de los modelos investigados se encuentran:

- **Clementine**: Es un software creado por la compañía SPSS Inc. (de la cual IBM es la actual dueña [5]), que implementó la metodología CRISP-DM. Su manejo es relativamente sencillo aunque está descontinuado debido a que cuando IBM adquirió SPSS y dejó de producirlo, no obstante, fue uno de los primeros softwares con interfaz que implementaba el KDD expresado en la metodología.
- **R**: Es el software más básico en cuanto al KDD, en el sentido de que carece de una interfaz y la mayor parte de los algoritmos se ejecutan mediante línea de comando, por ende, es ideal para la etapa de minería de datos del KDD, en especial porque se puede adaptar el algoritmo de los modelos.
- **Rapidminer**: En la última versión de este software (versión 5 y derivados), se implementó una interfaz gráfica, además, es uno de los programas con más algoritmos implementados, y posee la mayoría de los pertenecientes al software WEKA. En las últimas versiones también se incluyen algoritmos del software R. Su lenguaje de programación JAVA, se puede llevar a una aplicación confeccionada con este lenguaje y llamar la librería de Rapidminer desde la misma. Posee la particularidad de ser Open Source para efectos comerciales con su edición de comunidad, en particular está licenciado por la GPL. Sin embargo, su desventaja radica en la incapacidad para modificar los algoritmos directamente.
- **PASW Modeler**: Este software es la actualización que le hizo IBM al software Clementine. Dentro de sus ventajas se encuentran: el fácil manejo de los modelos y de la interfaz. Además, tiene implementados varios tests que permiten validar los resultados, mas la desventaja de este software radica en que no es de uso libre y, por lo tanto, su licencia se debe ir renovando.
- **SAP Business Objects Predictive Workbench**: Esta aplicación es un módulo adaptativo del sistema SAP Business Objects, en el que se incorporan herramientas de minería de datos al ERP(Enterprise Resource Planning) original. En caso de que éste se posea la aplicación podría generar reportes con facilidad relacionados con la predicción, no obstante, su desventaja radica en su alto costo debido a que es un módulo de un ERP.
- **Statistica**: Un software no tan conocido, pero es bastante completo, sobre todo en el área gráfica. Dentro de sus ventajas se encuentra la interfaz familiar y la implementación de una gran gama de algoritmos y test estadísticos, al igual que el SAP y el PASW. Sin embargo, sus desventajas se refieren a la gran cantidad de íconos de rápido acceso y el requerimiento de licenciamiento.
- **WEKA**: Uno de los software más usados en el procedimiento KDD, sobre todo por ser gratuito y programable. Posee la particularidad de modificar los algoritmos, así como también utilizarlos desde una interfaz. No obstante, su implementación en Rapidminer lo hace prescindible.

- **SAS**: Sin duda alguna uno de los software más eficiente, altamente usado en el mercado desde la perspectiva de la predicción, debido a que tiene implementado una gran cantidad de algoritmos de toda índole, a diferencia de Rapidminer que se concentra en problemas de clasificación, por ende la variedad de sus algoritmos es amplia en este sentido, SAS posee gran variedad de algoritmos en clasificación y predicción. Sin embargo, su costo por licencia es bastante alto.

Luego del análisis de softwares se escoge Rapidminer por ser aquel que tiene licencia Open Source, mayor potencia y adaptabilidad en su aplicación, una comparativa se puede observar en la siguiente tabla:

Criterios	R	Clementine	PASW Modeler	RapidMiner	SAP Workbench	Rattle
Baja Complejidad				✓		
Bajo Costo	✓			✓		✓
Alta Adaptabilidad					✓	
Alto Manejo de Grandes Volúmenes (más de 20 mil instancias)	✓	✓	✓		✓	✓
Alta Variedad de modelos	✓		✓	✓		

Figura 1.9: Criterios para elección de software

Para la especialización con el software escogido, se consultó los videos existentes en la Internet [13]. Además, se agrega la existencia de conocimiento previo de la herramienta en su versión anterior (4.4). Sin embargo, la nueva versión(5.0) agrega mejoras, iniciando con la nueva interfaz. Cabe destacar que en la fecha en que se necesitó el aprendizaje rápido del software, no existían manuales disponibles, debido a que recién venía saliendo esta nueva versión. No obstante, en el transcurso de esta memoria han salido nuevos manuales para el aprendizaje efectivo de este software.

1.5.3. Descripción de datos del producto

El producto en concreto se denomina NGN o Next Generation Network, que traducido al español significa redes de próxima generación, el cual utiliza la telefonía IP para agrupar varios servicios en un solo acceso. Por lo cual, se definirá la telefonía IP para complementar el concepto general del NGN.

Producto Telefonía IP

Para definir el producto de las redes de la próxima generación, se debe ir al origen del mismo producto, el cual se encuentra en el servicio de Telefonía, éste *"permite comunicaciones locales,*

móviles y de larga distancia...así como servicios comerciales(por ejemplo, central virtual) y administrativos propios del servicio telefónico" [12]. Sin embargo, este servicio se puede dividir según el tipo de tecnología usada, una de ellas es la telefonía IP, que *"reúne la transmisión de datos y de voz, posibilitando la utilización de las redes informáticas para efectuar llamadas telefónicas"* [8].

Producto NGN

Tras el concepto de la telefonía IP y el rol de esta en las telecomunicaciones, se puede definir el producto NGN como *"NGN un modelo de red destinado a entregar todos los servicios de comunicación a través de un sólo acceso, estos son telefonía local, servicios de banda ancha, servicios de larga distancia, y que, como principal atributo, tiene la potencialidad de adaptarse y crecer sin límites, para satisfacer cada vez de mejor forma las necesidades de los usuarios"* [77]. Además, al ser usar la tecnología de la telefonía IP, tiene una *"Central Virtual que permite llamadas gratis e ilimitadas entre anexos, además, incluye minutos SLM"* [11], estos últimos se definen como los minutos referentes al servicio local medido. Esto quiere decir que el "producto" NGN es un paquete de servicios que combina los distintos productos de la empresa de telecomunicaciones como la internet, la comunicación interna y externa. Aún siendo un producto adaptable para la microempresa, puede presentar exceso de fallas, dado que en un comienzo estaba *"orientado a la pequeña y mediana empresa"* [77], sin embargo, con el tiempo, esas mismas empresas que iniciaron con el producto, crecieron y se debió crear un nuevo producto para su satisfacción (Trunk IP) que viene a ser un NGN evolucionado. Por ende, se puede decir que *"el producto se presenta en forma de versiones, dónde cada modificación que se realiza es para perfeccionar el producto, ya sea en términos de oferta comercial, herramientas de apoyo, o redes disponibles. Estas versiones se materializan en los diferentes tipos de planes que adquiere cada cliente"* [11, 77].

Ahora bien, los principales beneficios que muestra el producto en los clientes son los siguiente[11]:

- Línea telefónica disponible de forma simultánea.
- Tarifa plana y menor debido a la integración de servicios
- Conexión múltiple con los ordenadores de la empresa
- Flexibilidad de configuración de servicios

Un punto relevante es el hecho de que el producto se vende en formato de contrato, el cual se divide en uno o varios planes, donde cada plan consta de uno o más accesos. Cada uno de ellos es descrito posteriormente [11]:

- Plan Access NGN
  - Desde 2 líneas telefónicas con 500 minutos de SLM.
  - Banda ancha con WIFI desde 2 MB/600 Kbps.
  - Desde 5 casillas de correo.
  - Administración web de tus servicios (telefonía, Internet, facturación, etc.).
- Plan Business Office I

  - Desde 2 líneas telefónicas con 500 min. de SLM y 100 min LDN.
  - Central virtual IP (comunicación gratuita entre anexos).
  - Administración web de tus servicios (telefonía, Internet, facturación, etc.).
- Plan Business I
  - Desde 4 líneas telefónicas con 1.000 min. de SLM y 200 min. LDN.
  - Central virtual IP (comunicación gratuita entre anexos).
  - Administración web de tus servicios de telefonía y facturación.
- Plan Business Office II
  - Desde 2 líneas telefónicas con 500 min. de SLM y 100 min. LDN.
  - Central virtual IP (comunicación gratuita entre anexos).
  - Banda ancha con WIFI desde 2 MB/600 Kbps.
  - Desde 5 casillas de correo.
  - Administración web de tus servicios (telefonía, Internet, facturación, etc.).
- Plan Business II
  - Desde 4 líneas telefónicas con 1.000 min. de SLM y 200 min. LDN.
  - Central virtual IP (comunicación gratuita entre anexos).
  - Banda ancha con WIFI desde 2 MB/600 Kbps.
  - Desde 5 casillas de correo.
  - Administración web de tus servicios (telefonía, Internet, facturación, etc.).

1.5.4. Inspección principales sistemas

Se puede apreciar el flujo de datos respecto al servicio de telefonía en la figura 1.10.
Una breve descripción de los principales sistemas se muestra a continuación [12]:

- **VENUS**: Sistema único de atención a clientes vía call center y ejecutivo CACE, permite la consulta de datos de clientes y el registro de solicitud y reclamos.
- **Intranet preventa y venta**: Aplicación interna inter-ejecutivos con funciones de gestión de planes.
- **Sistema de venta y post venta SAP (CCSAP)**: Plataforma única que maneja las relaciones cliente-contrato y catálogo de productos.
- **BUN**: Sistema cliente-contrato para negocio de larga distancia.
- **USV**: Sistema actual para gestionar los contratos desde el punto de vista de ventas.
- **Kenan**: Plataforma general de facturación.
- **OWF**: Sistema actual que gestiona requerimientos post venta.
- **OTC-AR**: Workflow de provisión técnica de servicios.

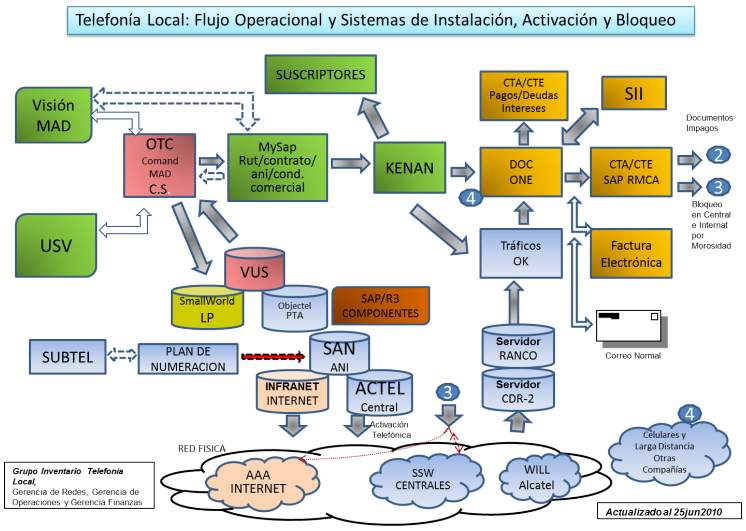

Figura 1.10: Flujo en sistemas de datos de telefonía

1.5.5. Prueba de resultados antiguos con base de experiencia anterior

Se solicitó la base antigua para probar la validez de la experiencia anterior. Además, se considerarán otros modelos para construir futuros prospectos que puedan entregar mejores resultados que el modelo J48 (utilizado en la experiencia anterior). Al describir la situación del churn se puede apreciar un 10 % de clientes fugados y un 90 % de clientes vigentes, todo ello acorde a la experiencia anterior. Cabe mencionar que las fuentes de información para generar este modelo, tuvieron origen en las área comercial y de operaciones, a lo que se debe agregar la consolidación de datos debido a la inexistencia de un *data warehouse* en esa fecha. En el momento del inicio de esta memoria se cuenta con un *data warehouse* en el área comercial.

1.5.6. Representación del churn actual con el KDD

En esta etapa se buscará modelar el nuevo comportamiento del churn, siguiendo los lineamientos de cálculo de la empresa para su obtención. Para ello, se contactarán las áreas responsables tanto del producto y como de su fuga, debido a que son dos áreas distintas. Esto implica que se debió decidir con que persona tratar realmente, que para efectos inmediatos fue la persona responsable del producto, dado que esta persona recibe una estadística del área en general.

Se opta por tomar los contratos renunciados de los clientes del sistema general, y posteriormente tomar la fuga combinando la plataforma OWF y la base de datos “NGN instalados” (ver sección 5.1.1), y luego solamente la plataforma OWF.

La situación actual carecía de similitud con respecto a la de hace dos años, en la que el producto no se había establecido como tal dentro del mercado, donde no se habían estudiado más opciones del mismo producto. El churn cambió y se redujo considerablemente de un 10 % a un 1 %, debido a que el producto aumentó su cartera de clientes y con la fuga se hizo despreciable. Sin embargo, esto sigue siendo un problema puesto que al comparar el churn actual con la tasa de captación del

producto, no se obtiene una diferencia significativa, lo que quiere decir que están fugándose tantos cliente como están ingresando. El siguiente gráfico expone lo mencionado anteriormente:

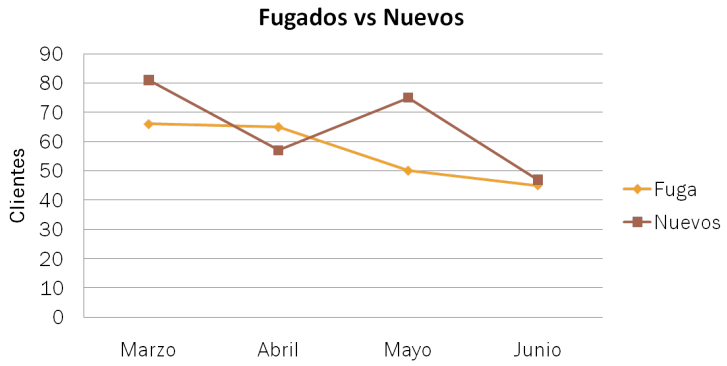

Figura 1.11: Gráfico de clientes fugados versus clientes nuevos

1.5.7. Estudio de modelo adecuado para la situación actual

A pedido del personal, se incluyen nuevas variables como es el caso de los reclamos técnicos y comerciales, pues intuyen que existe una relación entre el churn actual y el comportamiento de reclamos de cada cliente. Esto fue obtenido en entrevistas formales con el encargado del producto y su respectivo jefe para mantener una concordancia y dar a conocer el KDD. También fueron útiles para determinar que las variables usadas fuesen válidas para los futuros experimentos.

Se opta por mantener las transformaciones de la experiencia anterior sobre las variables de la base de facturación, así como también en las variables nominales consideradas. Sin embargo, se proponen transformaciones para las nuevas variables que no estaban descritas en dicho documento. Posteriormente en el proceso se mejoraron las transformaciones en facturación que culminaron con el uso del análisis factorial. En el caso de variables nominales se llegó a usar técnicas de clusterización como parte de su transformación.

Durante todo el estudio se ejecutaron modelos acordes a las distintas teorías de predicción (predicción directa, posterior a la clasificación y con distintos subconjuntos de datos). Para cada modelo propuesto se estudiaron distintas configuraciones para intentar adecuarlo al problema, lo cual se profundizará en el capítulo 3.

Finalmente se a evaluará cada modelo mediante métricas válidas en el KDD, las cuales dan una perspectiva técnica al procedimiento. Cabe señalar que las métricas contempladas fueron evaluadas en todas las situaciones, es decir, en el entrenamiento, prueba y validación del modelo.

1.5.8. Establecimiento del estudio como prototipo

En esta etapa se propuso el procedimiento a modo de prototipo, para lo cual el procedimiento KDD debió ser valorizado y que converge en cuantificar el churn monetariamente. Para temas de costo en el proceso, el acceso a las bases resultó ser directo, lo que significó un costo bajo. Esto se debió principalmente a que se utilizaron bases de datos que se encuentran en sistemas de acceso

sencillo para el usuario común, así como también reportes que ya se generaban previo a la aparición del estudio.

Es así como en esta etapa se puede dividir en los siguientes hitos relevantes:

- Actualización del procedimiento KDD

  Para actualizar el procedimiento KDD se revisó cada una de las etapas del procedimiento previamente ejecutado. En la etapa de integración de bases se verificó la validez de cada una de las variables dependiendo de su origen, así como también del grado de información que entregaban. En la etapa de preprocesamiento, se replanteó el comportamiento de los valores perdidos y su relación con la variable objetivo. En base a esto se propusieron distintos caminos del tratamiento de los mismos para, posteriormente, tomar en cuenta el efecto de los valores fuera de rango y mejorar la predicción. En la etapa de transformación, se propusieron cambios de índole comercial y técnica de manera que el resultado permitiese la interpretación instantánea. En la etapa de modelamiento, se abarcaron más modelos para aumentar certeza de las predicciones, incluyendo la experimentación con multiclasificadores y distintas bases para verificar la existencia de estacionalidad en el comportamiento de los clientes. En la etapa de evaluación, se descartaron algunas medidas por su difícil cálculo en la validación. Sin embargo, se consideraron en el entrenamiento para observar la estabilidad de los modelos. A su vez, se tomó en cuenta el hecho de que el modelo pudiese entregar variables claves en la predicción y deducir las causales de la fuga.

- Valorización de churn y resultados

  En esta fase se procedió a llevar el estudio fuera de los números técnicos y traspasar el problema a ingresos no percibidos. En un inicio se valorizó la pérdida inmediata, es decir, la cantidad que la compañía dejaba de percibir al mes siguiente de la fuga. Esta valorización subestimaba el problema, por lo que en valorizaciones posteriores se tomaron en cuenta las consideraciones de tiempo correspondientes a que el cliente no vuelve.

  A su vez, se valorizó en base a los ingresos no percibidos y se consideró el costo, tanto del software como de la capacitación en el mismo, junto con un espacio para la creación del set de datos necesario en el *data warehouse*.

- Generación de prototipo completo de prueba

  Una vez valorizado el churn y el procedimiento implementado, y habiendo revisado el KDD previo, se puede proceder a ejecutar el prototipo en el mercado real, con clientes actuales. Todo ello para que el estudio efectuado agregue valor a la compañía. Ahora bien, este valor quedará expresado en una propuesta de posibles fugados, de los cuales se hará un muestreo y se les contactará telefónicamente a los pertenecientes de dicha muestra con el objeto de verificar la predicción. Dicha muestra será tomada por el personal responsable del producto y los resultados se compartirán posteriormente para su análisis, con ello se busca encontrar una valorización real en el sentido comercial, así como también, la efectividad y eficiencia de la metodología usada del KDD.

- Prueba histórica sobre el modelo escogido

Para que el procedimiento KDD generase confianza dentro de la empresa se realizó una prueba histórica del modelo en consideración, en un lapso de 11 meses, obteniéndose una alta predicción, en los primeros meses, que posteriormente declinó. Cabe señalar que se tomó una base particular como entrenamiento y se validó en los meses posteriores, lo cual volvió a plasmar la idea de la estacionalidad presente en el comportamiento de los clientes.

Otro punto relevante de esta prueba histórica es que se realizó el KDD 11 veces, 1 por base.

1.5.9. Mejoramiento del KDD

Posterior a la ejecución del prototipo, se prosigue a la modificación del modelo tal que se pueda refinar predecir desde la perspectiva técnica, dado que en resultados comerciales se descubrió que, a pesar de que la predicción técnica fuese baja, la predicción resultó ser efectiva. Esto planteó serias dudas sobre la variable fuga contemplada, debido a que la evaluación comercial de la predicción arrojó un resultado mucho mejor que la evaluación técnica. Con esto se procedió a actualizar el KDD. No se efectuaron cambios en la etapa de integración, pues gran parte de las variables ya habían sido consideradas. En la etapa de preprocesamiento se trataron los valores fuera de rango con reglas de exclusión, que se dedujeron a partir de tablas de comparación de variables. En la etapa de transformación se cambiaron las variables de facturación, puesto que se detectó la poca efectividad de propuestas anteriores. En la etapa de modelamiento se agrupó y predijo en base a los generados previamente. Finalmente en la etapa de evaluación se efectuó la misma técnica que en el prototipo inicial.

1.5.10. Establecimiento de acciones correctivas

Si bien, la ejecución del prototipo llevó al proceso del KDD más cerca del cliente, no presentó mejoras respecto al churn, ya que en el muestreo solo se buscaba determinar si existía la intención de fuga (para lo cual se contrató a una persona que detectase dicha intención). no se llevaron a cabo acciones correctivas, debido a que el área comercial solamente puede controlar el precio del producto por cliente, sin modificar características del producto en el corto plazo, pues no existe una relación estrecha con el área operacional. Por ende, el establecimiento de acciones correctivas resulta ser un conjunto de propuestas factibles, formuladas en base a los resultados paramétricos del modelo. Con ello, se busca la inserción del procedimiento del KDD en la compañía y la ejecución del mismo.

Un punto a destacar es el tema de las relaciones que existen entre el área comercial y operacional, las cuales no se pueden mejorar debido a que actualmente se está llevando a cabo una fusión a nivel de la empresa.

1.6. Alcances del trabajo

El alcance de la memoria comprende la realización de varios experimentos en un ambiente sistémico de múltiples plataformas, con datos históricos y la evaluación de la implementación del procedimiento KDD para uno o varios meses de prueba con datos reales, consecuentemente, se ejecutan llamadas a clientes actuales de la compañía así como también, se verifican sus estados

reales de vigencia. De este modo es posible cuantificar empíricamente el beneficio de la aplicación de los modelos desarrollados desde dos puntos de vista, el primero de una mirada marginal, es decir, la fuga de ingresos del mes siguiente debido a que un cliente determinado se fuga en el mes actual; el segundo, en cambio, se ve desde la vista de pérdida, la cual sólo puede ser calculada en base al ciclo de vida del cliente, siendo todos los ingresos no obtenidos, debido a que el cliente deja la compañía antes de terminar su ciclo de vida. Dentro de la empresa abarcan distintas áreas: aseguramiento de ingresos, comercial, sistemas y planificación como actores del proyecto mismo. Cabe señalar que el principal beneficiado del proyecto es el área comercial.

Desde la visión de servicios involucrados, el alcance principal serán los servicios enfocados a las pequeñas y medianas empresas que son clientes del servicio NGN.

El alcance desde una perspectiva técnica abordará conceptos básicos de minería de datos y manejo de un software facilitador, así como también, los algoritmos de aprendizaje supervisado y no supervisado. Abarcando desde su expresión inicial como algoritmo hasta las medidas de distancia que ocupa. Lo mismo aplica para los métodos de imputación de datos, detección de valores fuera de rango y transformaciones.

Dentro de los límites del proyecto se encuentra la predicción de clientes NGN, y no los clientes de otro servicio. Es probable que la predicción pueda expandirse hacia los clientes indirectos de forma completa y no sólo avocada al servicio NGN. Además, se pretende valorizar a los mismos clientes y encontrar las variables más relevantes dentro de la base de datos, para generar respecto a ellas, acciones correctivas correspondientes en conjunto con el área comercial.

Como alcance a futuro, la idea es que con este proyecto de churn, genere la integración e implementación del procedimiento KDD dado que hasta este momento no está incorporado, pues la competencia sí lo tiene implementado, al menos para los productos de telefonía móvil.

1.7. Resultados Esperados

Objetivos Específico: Estudio y Reconstrucción de experiencia anterior de la empresa.

- Resultado 1. Reporte resumen sobre documento de experiencia anterior, detallando métodos, variables utilizadas, dificultades encontradas y fuentes de información.
- Resultado 2. Detección de mejoras al modelo propuesto en la experiencia base, analizando todas las etapas del modelo según el procedimiento del KDD, desde un punto de vista técnico-teórico.
- Resultado 3. Fase de reuniones con personal de la compañía para la obtención de datos necesarios en la reconstrucción de la experiencia anterior.
- Resultado 4. Creación de una base de datos con estos necesarios consolidados.
- Resultado 5. Reconstrucción del modelo planteado en la experiencia base.
- Resultado 6. Comparación y selección de procesamiento de datos en cuanto a mejoras de resultados en el modelo.

Objetivos Específico: Lectura y recopilación de antecedentes de predicción de churn en telecomunicaciones y otras empresas.

- Resultado 1. Reporte comparativo sobre técnicas de minería de datos ad-hoc al problema planteado, con registro de la fuente correspondiente.
- Resultado 2. Resumen sobre el churn en el entorno en las telecomunicaciones del churn.
- Resultado 3.Fase de reuniones con personal de la compañía encargado directamente del cálculo del churn NGN.
- Resultado 4. Reporte de valorización del churn actual.

Objetivos Específico: Caracterización del churn en la empresa, búsqueda y selección de fuentes de información.

- Resultado 1. Fase de reuniones con personal de la compañía para la declaración de los objetivos buscados en la metodología.
- Resultado 2. Lista de selección de variables a utilizar en la nueva metodología.
- Resultado 3. Integración de fuentes de información relevantes como base de datos del modelo.
- Resultado 4. Informe detallado con el preprocesamiento efectuado a las nuevas variables.

Objetivos Específico: Planteamiento y selección de modelos acordes al problema señalado.

- Resultado 1. Reporte con principales modelos a utilizar según el estilo del problema.
- Resultado 2. Generar lista con ventajas y desventajas de cada modelo y de cómo impactan en el negocio de las telecomunicaciones. Así como también, en sus resultados.

Objetivos Específico: Establecimiento de métricas de evaluación para el desempeño de los modelos planteados y de la bondad de las predicciones.

- Resultado 1. Informe que describe la selección del modelo en base a las métricas presentadas y evaluadas.
- Resultado 2. Detección de mejoras al modelo propuesto inicialmente en la experiencia base respecto a las distintas métricas propuestas.
- Resultado 3. Fase de reuniones con personal del área comercial para evaluación del desempeño del modelo en el negocio.

Objetivos Específico: Evaluación de los resultados obtenidos

- Resultado 1. Valorización del impacto del estudio realizado.
- Resultado 2. Informe de establecimiento y valorización de acciones correctivas, en el mes de holgura.
- Resultado 3. Valorización de la automatización y almacenamiento de la metodología, en el largo plazo.

Capítulo 2

Marco Conceptual

Dentro del estudio, una parte fundamental resultan los conceptos que sustentan el valor mismo, por lo que a continuación, se presentará el marco conceptual. Este especifica cada concepto presente en esta memoria, incluyendo los sistemas que se encuentran en la empresa, la arquitectura, los softwares estudiados y utilizados, apoyando cada definición con una imagen cuando esto amerite.

Dado que el estudio presente en este informe trata acerca de la minería de datos y su desenvolvimiento, se partirá con dicha definición ya que es el eje articulador de esta memoria.

2.1. Definición de minería de datos, fundamentos, evolución

2.1.1. Definición de minería de datos

En el mercado, el término minería de datos, hace referencia al área que estudia los hallazgos de información en los datos propiamente tal de un negocio determinado, o bien, es el proceso de análisis de bases de datos que busca encontrar relaciones inesperadas que son de interés o valor para el poseedor de dicha base de datos [39].

Otra definición relevante es que la minería de datos es *"una capacidad sofisticada de búsqueda de datos que usa algoritmos estadísticos para descubrir patrones y correlaciones en los datos"* [93]. En términos simples sería *"una manera de encontrarle significado a los datos"* [93].

Por otro lado, hay que destacar que la minería de datos es *"parte de un largo procedimiento denominado Descubrimiento del Conocimiento (Knowledge Discovery)"* [93] que, aplicado a bases de datos, se denomina Knowledge Discovery in Databases (KDD).

2.1.2. Fundamentos de la minería de datos

Los principales fundamentos de la minería de datos residen en las ramas de las ciencias de la computación, por la inteligencia artificial (en particular el aprendizaje sistemático), y de las matemáticas, por la estadística. Debido a que el área o campo de la minería de datos implica modelar problemas, analizar información e interpretar información bajo los supuestos de validez, es clara la base estadística que requiere. Ahora bien, la necesidad de la minería de datos se debió a que la clásica aproximación para el análisis de datos solía ser que un conjunto de analistas se familiarizase con

los datos y sirviesen de interface entre las bases de datos y los usuarios y productos. No obstante, a medida que el tiempo ha avanzado, los volúmenes de información han crecido considerablemente, debido a que el detalle de cada información requerido para entregar mejores soluciones en los distintos negocios es bastante mayor. Es por ello, que bases de datos con el orden de 100 a 1000 variables comenzaron a tomar relevancia y sobre todo la información que se escondía en ellas.

Debido a las altas dimensiones de las bases de datos actuales, la forma clásica de *"prueba manual de un conjunto de datos es lenta, cara y altamente subjetiva y a medida que el volumen de datos crece, el análisis manual se vuelve impracticable en ciertos dominios"* [33]. En el problema anterior, es donde el machine learning; que se define como una rama de la inteligencia artificial(que a su vez, es una rama de las ciencias de la computación) se encarga del diseño y aplicación de algoritmos de aprendizaje [6]; entra a jugar el papel de implementación dentro de lo que es la minería de datos. Por ello, el fundamento de este campo son el machine learning y la estadística, la cual es definida como *"el estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas"* [10].

En pocas palabras, la inteligencia artificial resuelve el problema de la cantidad y capacidad de análisis, mientras que la estadística se preocupa de la calidad de dichos análisis, ya sean estos a resultados o a variables.

2.1.3. Estilos de aprendizajes en la minería de datos

Para enfatizar más el concepto de la minería de datos, se deben agregar los estilos de aprendizajes que existen en ésta, los cuales corresponden a: clasificación, asociación, *clustering* y predicción numérica.

El primero de los mencionados aparece *"con un set de ejemplos clasificados de los cuales se espera aprender una manera de clasificar ejemplos no observados"* [108]. En otras palabras, se refiere a clasificar a nuevos clientes o servicios partiendo de una agrupación ya originada.

El segundo estilo, asociación, señala que *"cualquier asociación entre características es buscada, no solo aquellas que predicen un valor particular de una clase(término a definir posteriormente)"* [108]. Esto tiene alude a la forma en que se relacionan las variables.

El tercer estilo, se define como *"conjuntos de ejemplos que pertenecen a un mismo grupo son buscados"* [108]. Este estilo de aprendizaje se recomienda para casos en que no se tenga una noción del negocio, o bien para complementar cuantitativamente la visión sobre los registros que una persona partícipe del negocio tenga.

Finalmente, el cuarto estilo se da a conocer cuando *"el resultado a predecir no es una categoría discreta sino una cantidad numérica"* [108]. Este estilo está presente en predicciones que requieran algún tipo de regresión sobre los datos.

2.1.4. Evolución histórica

La evolución temporal, se presenta como un esquema y las tecnologías en cada época [93] en la figura 2.1.

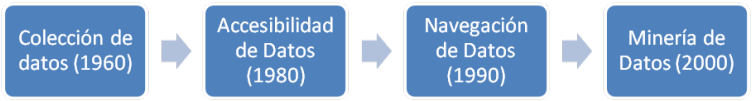

Figura 2.1: Evolución histórica del Data Mining

Época	Tecnologías
Colección de Datos	Computadores, cintas, discos.
Accesibilidad de Datos	Bases de datos relacionales, Lenguaje de consulta estructurado (SQL), ODBC.
Navegación de Datos	Procesamiento analítico en línea(OLAP), Bases de datos multidimensionales.
Minería de Datos	Algoritmos avanzados, Bases de datos masivas.

Figura 2.2: Tabla de evolución histórica del Data Mining

2.1.5. Knowledge Discovery in Databases

El Knowledge Discovery in Databases o KDD, es término ingresado por Usama Fayyad en los años 90's, el cual, se define como *"...el proceso no trivial de identificar patrones novedosos, válidos, potencialmente útiles y descifrables en el conjunto de datos"* [33], siendo uno de los procesos más utilizados a nivel global, dada su generalidad en sus aplicable a distintas áreas, por ejemplo, *...marketing finanzas (inversiones específicamente), detección de fraude, manufactura, telecomunicaciones y agentes de internet"* [33], donde cada una tiene distinta connotación.

Este proceso es donde más se usa la minería de datos como parte de una etapa completa del proceso mismo, donde el negocio, si bien es parte importante, sólo está presente efectivamente en la última fase de este procedimiento. Por ende, a nivel técnico resulta muy útil, puesto que se enfoca más en el entendimiento de los patrones más que en el negocio mismo, por lo tanto, *"propone la intersección de diferentes campos de investigación como lo son inteligencia artificial, estadística, visualización de los datos"* [33].

Este proceso consiste en 5 etapas que se describen a continuación [33]:

1. **Integración o Selección**: En esta etapa se escogen las variables a considerar en el proceso completo, por lo que incluye la identificación de los objetivos del estudio de minería de datos, desde el punto de vista del cliente como referencia fundamental del negocio, así como también, incluye la creación del conjunto de datos objetivos sobre los cuales se procede a integrarlos como la base de estudio en el proceso.

2. **Preprocesamiento**: En esta etapa, el análisis y limpieza de los datos, son las líneas principales a seguir. Es aquí donde se produce el tratamiento de valores sin información (datos incompletos en el caso de que no se pueda extraer su valor original) o ausentes (*missing*), los valores fuera de rango u *outliers* (donde se incluye el caso de valores incompleto en que sí se puede determinar el valor original). Para ello, se emplean distintas técnicas de imputación de datos que van desde un reemplazo simple (*simple imputation*) hasta un reemplazo múltiple (*multiple imputation*). Todos los tipos de valores presentados en esta etapa son definidos, posteriormente, en la sección tipos de datos.

3. **Transformación**: Aquí se generan nuevas variables (definiéndose ésta como un conjunto de datos que describen una característica determinada, lo que quiere decir que es un atributo de un producto o columna de una base de datos) utilizando diferentes técnicas, por ejemplo, el traspaso de una variable continua a una nominal (conceptos definidos en la sección tipos de datos), o de una variable nominal a una variable binomial, entre otras, para las cuales se pueden usar funciones discretas y continuas.

4. **Minería de datos**: Este paso en el proceso de KDD, consiste en la aplicación de análisis de datos para descubrir un algoritmo ad-hoc que *"bajo limitaciones computacionales aceptables, produzca una particular enumeración de patrones"* [33]. En esta etapa se selecciona el modelo a ocupar, bajo los supuestos que mantienen los objetivos primarios del estudio. Además, es en esta etapa en donde los algoritmos "aprenden" a partir de los datos, por lo que se ejecuta múltiples veces el "entrenamiento" del modelo [108].

5. **Interpretación y Evaluación**: Esta última fase implica la selección de medidas de evaluación, que permitan decidir la confiabilidad y validez del modelo, así como también, el traspaso de los resultados de dichas medidas a conocimiento y acciones correctivas en el negocio, que permitan la solución al fenómeno estudiado. Respecto a la evaluación, ésta se puede aplicar técnica y comercialmente, en donde la primera se subdivide acorde al tipo de validación que se le realiza al modelo, mientras que la evaluación comercial no se encuentra estandarizada. La subdivisión de la evaluación técnica se basa según los algoritmos que permitan más solidez en los resultados. En este documento se describen dos: el *holdout* y la validación cruzada. El primero consiste en: Dado un conjunto de datos preprocesados y transformados, *"se retiene una cantidad para probar el modelo y el resto es usado para entrenar"* [108], entonces puede darse el caso de que el conjunto de datos de prueba no sean los más adecuado ni representativos del conjunto completo, por lo que se implementa la validación cruzada. En esta técnica estadística *"Se decide un número fijo de particiones del conjunto de datos n, luego se separa el conjunto en n particiones iguales y en cada iteración se utiliza cada una de ellas para probar mientras que el resto se usa para entrenar el modelo* [108], en la que se puede tomar cada partición de forma estratificada de tal manera que cada partición represente el conjunto de datos original.

El bosquejo de estas etapas, su sucesión y la característica cíclica que muestra, se presentan en la figura 5.17[1].

2.2. Principales metodologías del KDD

2.2.1. Cross Industry Standard Process for Data Mining

Una de las metodologías que adaptan el KDD a los negocios es el Cross Industry Standard Process for Data Mining (CRISP- DM), el cual surge en Daimler Chrysler en el año y posteriormente es

[1]Dibujo extraído, con el permiso de Eduardo Merlo, de su memoria: "Identificación de redes sociales de copia en instituciones educacionales mediante el uso de minería de datos sobre documentos digitales" confeccionada en la primavera del 2010

formalizada. Particular, *"se describe en términos de un modelo de procesos jerárquicos, consistente en un conjunto de tareas descritas en 4 niveles de abstracción"* [24]. Estos son: fases o áreas, tarea genérica, tarea específica e instancia de proceso. En el siguiente cuadro se explicará brevemente cada uno de estos niveles:

Nivel de abstracción	Definición
Fase o Área	Es la descripción el contexto o entorno en el que se va efectuar la minería de datos.
Tareas genéricas	Es la búsqueda de todas las situaciones donde se pueda aplicar minería de datos.
Tareas específicas	Es la descripción de cómo se deberían llevar a cabo las acciones en las tareas genéricas para casos peculiares.
Instancia de Proceso	Es el conjunto de acciones, decisiones y resultados de un acuerdo de minería de datos.

Figura 2.3: Niveles de la Metodología CRISP-DM

El proceso de CRISP-DM puede variar dentro de estos niveles de abstracción, es por ello que se tiene un modelo referencial, el cual, se separa en distintas etapas que no siguen un orden secuencial. Las etapas con las que cuenta este proceso son 6:

- **Entendimiento del negocio**: Se enfoca en establecer los objetivos y requerimientos del negocio.
- **Comprensión de los datos**: Tiene estrecha relación con el tema de la familiarización con los datos, es decir, conocer las distintas bases de datos del negocio y comprender su trasfondo.
- **Preparación de los datos**: Esta etapa contempla toda la preparación de la base final de datos a utilizar
- **Modelamiento**: Se refiere a modelar el problema con los distintos algoritmos existentes y con la base de datos final obtenida en la etapa anterior.
- **Evaluación**: Es evaluar los resultados del modelo acorde a una visión complementaria con el negocio en el que se está ejecutando.
- **Despliegue**: Es la etapa que contiene las acciones correctivas a generar e implementar a partir de los resultados finales entregados por el modelo.

2.2.2. Sample, Explore, Modify, Model, Asess

La segunda metodología considerada es denominada *Sample, Explore, Modify, Model, Asess* (SEMMA); desarrollada por el Instituto Statistical Analysis System (SAS) [14]. Considera las partes más relevantes de la minería de datos, por lo que no es una metodología de minería de datos [15]

propiamente tal, sino una organización de pasos prácticos, incluídos en un software elaborada por SAS (SAS Enterprise Miner). No obstante, puede ser de utilidad como guía a seguir, por lo que se procede a describir sus etapas:

- **Muestreo**: Como bien lo dice su nombre, se refiere a extraer una fracción de una base extensa de datos.
- **Exploración**: En esta etapa se efectúan análisis para detectar anomalías y tendencias que permitan la familiarización con el conocimiento de la base de datos.
- **Modificación**: En esta sección se aplican los criterios respectivos de selección de variables e instancias, así como también, la creación y transformación de las variables. En otras palabras es un procesamiento de las variables.
- **Modelamiento**: Tal como se indica, en esta etapa se aplican algoritmos de aprendizaje a la muestra.
- **Evaluación**: Consiste en la revisión de los resultados, en pos de encontrar hallazgos relevantes y útiles para el problema a solucionar.

Solamente se consideraron estas dos metodologías debido a que son las usadas o creadas por las dos empresas más grandes dedicadas al KDD, SPSS y SAS.

2.3. Problemas resueltos por la minería de datos

Acorde a la metodología CRISP-DM, se pueden sintetizar la mayoría de los problemas en los que es aplicable la minería de datos, particularmente se mencionan 6 tipos de problemas, los cuales se describirán a continuación:

- **Descripción de datos**: Cuando se tiene una base de datos en la que se sabe muy poco acerca de su contenido, se puede utilizar la minería de datos para detectar relaciones que *"apuntan a la descripción concisa de características de los datos"* [24]. Cabe señalar que este tipo de problemas son *"de baja escala en minería de datos"* [24]. No obstante, suelen aparecer *"en combinación con otros tipos de problemas de minería de datos"* [24], lo que conlleva a saber como requisito básico el trato adecuado de éste.
- **Segmentación**: Se puede sintetizar su significado en la búsqueda de agrupaciones en un conjunto de datos, por lo que este problema *"apunta a la separación de los datos en subgrupos y clases interesantes y con algún significado para el negocio"* [24]. Por lo tanto, su solución va ligada a un experto del negocio y, además, posee dos miradas: una técnica y una semántica propia del negocio.
- **Descripción de conceptos**: El problema de la descripción de conceptos abarca la región de la semántica, la cual puede ser obtenida mediante medidas cuantitativas y algoritmos que la minería de datos provee, lo que a fin de cuentas, llevará a la observación del comportamiento de una o más variables involucradas en la base de datos. En otras palabras, *"apunta a una*

descripción comprensible de conceptos o clases. El propósito no es desarrollar completos modelos..., sino ganar puntos de vista" [24]. Cabe destacar que este tipo de problema *"tiene una conexión con los problemas de clasificación y segmentación"* [24], sobre todo para el tema de asignación de nombres a los segmentos encontrados o validar alguna clasificación preexistente al análisis.

- **Clasificación**: Este tipo de problemas es bastante similar a la segmentación. Su diferencia radica en que la clasificación *"asume que existe un conjunto de objetos caracterizados por un grupo de variables o características que pertenecen a diferentes clases o agrupaciones"* [24], mientras que la segmentación genera dichas agrupaciones caracterizadas. Sintetizando el punto anterior, en la clasificación existe un resultado esperado a llegar, mas en la segmentación no, por ello, la minería de datos puede entregar la validación de una clasificación.

- **Predicción**: Uno de los problemas más importante que puede resolver la minería de datos es el requerimiento de efectuar acciones proactivas y preventivas, para lo cual se necesita información o puntos de referencia sobre el futuro. Es así como la predicción *"apunta a encontrar el valor numérico de una característica objetivo para objetos donde aún no se le ha observado"* [24]. Este problema suele aparecer frecuentemente en predicciones de valores futuros (series de tiempo), aunque también puede aplicarse a la segmentación.

- **Análisis de Dependencias**: El análisis de dependencias aparece cuando se requiere averiguar la relación existente entre dos o más variables o características, es decir, *"consiste en encontrar un modelo que describa dependencias significativas entre items de datos o eventos"* [24]. El problema se puede expander al negocio en cuanto al diseño de productos y las características óptimas para éste.

Dentro de estos problemas y los estilos previamente mencionados, se puede hacer una división en lo que se refiere al estado del problema. Estos pueden presentarse de manera supervisada, que indica que *"el algoritmo a usar opera bajo la supervisión, la cual es facilitada por un resultado actual para cada uno de los ejemplos de entrenamiento"* [108]. En otras palabras, la supervisión es cuando se posee un resultado de la variable a predecir, el caso contrario se denomina no supervisado.

Antes de proseguir, en esta memoria, el término "significativo" tiene una connotación matemática o propia del lenguaje, para ello, se define el p-valor y el grado de significancia:

- P-Valor: *"Probabilidad de que el resultado obtenido sea debido al azar, éste se obtiene a partir de los datos que se ingresen a la prueba estadística"* [92].

- Grado de Significancia: *"Probabilidad de rechazar de manera incorrecta la hipótesis nula cuando sea cierta"* [92], usualmente se asignan valores de 5 %, 1 %, entre otros, dependiendo a la validez estadística que se desee (mientras menor, mayor su validez, pero también, mayor su costo de ejecución, además, es una convención establecer el 5 % como válido estadísticamente).

Considerando todo lo anterior se puede hablar de una dependencia significativa cuando su test de hipótesis da por resultado un p-valor mayor al grado de significancia.

2.4. Churn

2.4.1. Concepto y tipos

Concepto de churn

El churn, dentro de las telecomunicaciones, es *"la acción de cancelar el servicio prestado por la compañía"* [67]. En dicha cancelación, el cliente puede decidir renunciar a la empresa (voluntaria), o bien, la empresa puede expulsarlo(involuntaria). En particular, la connotación de churn hace referencia la fuga de los clientes, por lo que, para efectos de este estudio, se cuenta el churn en base a la decisión del cliente en abandonar la empresa. Otra definición del churn, en el sector de las telecomunicaciones, es aquel término *"usado para describir colectivamente el cese de servicios de la suscripción de un cliente...donde el cliente es alguien que se ha unido a la compañía por al menos un período de tiempo...un churner o fugado es un cliente que ha dejado la compañía"* [43].

Una definición adicional se presenta como: *"La propensión de clientes a efectuar el cese de los negocios que tenga con una compañía en un período de tiempo determinado"* [23].

Con lo anterior, el manejo del churn *"consiste en desarrollar técnicas que permitan a las firmas mantener a su clientes rentables y apuntar al incremento de la lealtad de los clientes"* [57].

Tipos de churn

La clasificación del churn se puede ver desde dos aristas, la primera que va en el sentido de grado de fuga de un cliente, mientras que la segunda va asociada directamente al negocio de una empresa. Es así como la primera divide al churn en [85]:

- **Absoluto**: suscriptores que se han desligado sobre la base total en un período
- **De línea o servicio**: Este tipo de churn el número de servicios discontinuados sobre la base total
- **Primario**: referente al número de fallas
- **Secundario**: Decenso en el volumen de tráfico

En cambio en la segunda, la categorización va acorde al negocio, es decir [43]:

- **Churn de paquete**: Este churn se caracteriza por el hecho de que *"los clientes se mueven entre los paquetes ofrecidos por la compañía"* [43], en otras palabras, no es una fuga completa de la compañía sino que es una fuga de un paquete de servicios determinado, por ende ese churn comprende una serie de servicios.
- **Churn del servicio**: En este churn, existe solamente fuga a nivel de servicio.
- **Churn de la compañía**: Sin lugar a dudas el más costoso, en este churn *el cliente se fuga hacia la competencia*, por ende, no solamente se pierde el ingreso no percibido, sino que también el prestigio de la compañía expresado en el *market share* de la competencia.

En general, el estudio del churn presenta atractivos para la compañía por los beneficios en juego, cada uno de ellos va asociado a reducir las "pérdidas" monetarias en la empresa. Así se puede decir que los principales beneficios que aporta el conocer el churn son [85]:

- Menor inversión en adquirir a un cliente.
- Alta eficiencia en el uso de la red.
- Incremento en el valor agregado de las ventas, gracias a los clientes fieles en el largo plazo.
- Incremento en ventas, por el "boca a boca"(este tipo de comunicación *"es un factor determinante entre el 20 y el 50 por ciento de todas las decisiones de compra"* [16]) de los clientes fieles.
- Reducción de gastos en mesa de ayuda o en servicio al cliente.
- Mayor confianza en los inversionistas.
- Reducción de exposición a fraudes y malos débitos.

2.5. Tipos de datos

2.5.1. Conceptos generales y escalas de datos

Para formar un marco conceptual desde su punto más básico se necesita la definición de algunos conceptos previos y de uso común en el procedimiento del KDD. Generalmente, estos conceptos son nombres utilizados para identificar variables o características de las mismas.

Es así como es que lo primero que hay que definir es el término "Concepto", entendido en esta área como *"el objeto a ser aprendido por el algoritmo"* [108] y la descripición de conceptos como *"el resultado producido por un esquema o estilo de aprendizaje"* [108]. Estas definiciones hacen referencia a la o las variables, objetivo del estudio.

Los ejemplos, son aquellos objetos o individuos caracterizados en una base de datos, generalmente las filas del conjunto de datos. No obstante, en términos del KDD, los ejemplos se denominan instancias que *"son las cosas que serán clasificadas, asociadas, o clusterizadas"* [108]. También puede denominárseles registros, aunque puede tender a la confusión.

Como existen filas denominadas "instancias" en una base de datos, también están los atributos que son *"las características asociadas a las instancias"* [108]. Usualmente, estas características son las variables del modelo, y los valores que contiene cada atributo en cada instancia son *"medidas de la cantidad a la cual cada atributo se refiere"* [108]. Otra definición, alude a que el atributo es el *"objetivo del instrumento de medida o característica que se utiliza para describir objetos(personas, estímulos, respuestas)"* [58]. Con esto se puede definir la base de datos como *"una matriz de instancias versus atributos"* [108]. Dos conceptos apartados de las bases de datos en sí, son los de confiabilidad y validez. El primero *"indica que algo o alguien tiene la capacidad de llevar a cabo y mantener sus funciones en circunstancias ordinarias pero también en las extraordinarias"* [20]. Esto quiere decir que es la medida en que un instrumento mantiene sus resultados, independiente de la situación. En cambio, la validez es el hecho de *"que las observaciones que realizamos con el instrumento corresponden a los fenómenos reales que queremos observar"* [20], es decir, que el instrumento sea efectivo y coherente con la realidad.

Las variables del modelo, poseen propiedades relacionadas directamente con su escala, la cual se define como *"los valores numéricos asignados a cada variable según ciertas reglas"* [58], definición que no tiene relación con la terminología de encuestas.

Los tipos de escala convencionalmente usados en estadística son aquellos pertenecientes a la propuesta de Stevens(1946) [58], que diferencia cuatro principales: la escala nominal, ordinal, intervalo y de razón. Las dos iniciales son *"escalas no métricas o cualitativas, puesto que reconocen en cada instancia una determinada cualidad o propiedad"* [58], en cambio las dos finales, son *"escalas métricas o cuantitativas, capaces de reflejar diferencias de grado o cantidad"* [58]. Se debe señalar que *"los atributos numéricos a veces son denominados atributos continuos"* [108].

2.5.2. Tipos de escala

- **Nominal**: Es aquella escala que *"permite identificar categorías, de ahí que las variables medidas en escala nominal reciban el nombre de categóricas-mutuamente excluyentes y exhaustivas. Los números que se asignan a cada categoría no encierran ningún significado concreto ni relación de orden alguna"* [58], o también, *"tiene valores que son distintos símbolos"* [108]. En particular posee propiedades únicas del resto de las escalas, como por ejemplo, *"la única operación permitida es contar...sólo las estadísticas que consisten en contar son válidas: porcentajes, moda, etc."* [20]. Además, tiene una subescala incluída como caso especial, que es *"la dicotomía, la cual tiene sólo dos miembros usualmente designados como verdadero o falso, o sí y no...tales atributos también reciben el nombre de boolean"* [108].

- **Ordinal**: Esta escala es similar a la nominal, pero *"los números asignados a cada una (de las categorías) sí guardan una relación de orden"* [58]. No obstante, *"si bien existe relación de orden, no hay noción de distancia"* [108]. Esta última es vital para el desarrollo del análisis y preprocesamiento de datos, puesto que para las variables con esta escala no se pueden utilizar estadísticos que involucren a la distancia, como por ejemplo, distancia euclídea, en caso de que se efectuase dicho estadístico, se estaría forzando un resultado e incurriendo en un error. Otra particularidad de esta escala es que permite *"además de las operaciones basadas en conteo, las estadísticas basadas en el ordenamiento de los datos...percentiles, cuartiles, mediana"* [20].

- **Intervalo**: Esta escala tiene *"las mismas características que la ordinal, pero la diferencia entre dos niveles o categorías consecutivas es constante a lo largo de toda la escala. Sin embargo, esta escala carece de un cero absoluto, lo que impide afirmar que un valor determinado de ella sea...un múltiplo de otro valor de la propia escala"* [58]. Entre las peculiaridades que posee esta escala se encuentran la permisión de *"transformaciones lineales del tipo y=a +bx...además de los estadísticos de las escalas anteriores...permiten calcular aquellos más usuales: media aritmética, desviación estándar"* [20].

- **Razón**: es aquella escala que *"presenta un cero absoluto...que hace posible realizar todas las operaciones matemáticas"* [58], por ende, existe una noción de distancia, lo que resulta de gran importancia para la mayoría de los algoritmos de la minería de datos que trabajan con la continuidad de las instancias.

Generalmente, todas las variables tienen alguna de estas escalas asociadas de forma natural, sin embargo, existe una característica respecto a las variables no métricas, y esta es que *"una variable no métrica puede ser convertida en métrica a través de variables ficticias binarias. Sería necesario contar con un número de ellas igual al número de categorías, de la variable no métrica, menos uno"* [58]. La característica anterior hace referencia a las denominadas variables *dummy*, y la condición puesta al final de la cita se debe a que si se asigna un número de variables *dummy* igual a la cantidad de categorías se pierde robustez en los modelos por redundancia de información. En otras palabras, se pierde la validez del modelo. Otra definición de este tipo de variable se muestra como *"una variable artificial creada para representar un atributo con una o más categorías"* [98], así como también, una conceptualización adicional señala que *"la variable dummy es un simple y útil método de introducir, en un análisis de regresión, información contenida en variables que no son convencionalmente medidas en una escala numérica"* [99]. La importancia de esta variable radica en la incorporación o desglose de información agregada en las variables con escala no métrica.

A modo de resumen, se bosqueja la siguiente tabla que agrega ejemplos de las principales escalas [71]:

Escala	Ejemplos Comunes	Ejemplos de Marketing	Estadísticas permitidas	
			Descriptivas	Inferenciales
Nominal	Enumeración de jugadores de fútbol.	Marca, Tipos de tienda.	Porcentaje, moda.	Chi-Cuadrado, pruebas binominales.
Ordinal	Rankings de calidad.	Rankings de preferencias, posición de mercado, clase social.	Percentiles, mediana.	Correlación de Ranking-orden, ANOVA de Friedman.
Intervalo	Temperatura (Farenheit).	Actitudes, opiniones, índices.	Rango, media, desviación.	Producto-momento
Razón	Peso, Altura.	Edad, ventas, ingresos, costos	Media Geométrica, Media harmónica.	Coeficiente de variación

Figura 2.4: Ejemplos de tipos de escala

2.5.3. Valores Ausentes o Missings

Se definen como aquellos valores que no se encuentran de forma explícita. La idea de estudiar este tipo de valores fue el quiebre que tuvo Rubin en su momento, puesto que *"el proceso que causa los valores perdidos es ignorado después de haber asumido como un accidente en algún sentido u otro"* [30]. Además, *"la realidad indica que existen situaciones en que debido a los objetivos de la investigación, deliberadamente se omite información de personas que no forman parte de la población de estudio"* [36]. Estos valores omitidos suelen denominarse valores ausentes, pues no existe un registro que indique una pista del valor real de la instancia en el caso de que se hubiese llenado.

Estos valores se dividen en tres tipos [91]:

- **Missing Completely At Random (MCAR)**: Son aquellos datos perdidos que son completamente al azar, es decir, no poseen ningún tipo de relación con los datos presentes en otras variables. En este escenario, las distribuciones de probabilidad de los datos faltantes y de todos los datos son idénticas, debido a que es una variable independiente y, por ende, sólo posee una distribución de probabilidad. Como detalle a la definición, su traducción sería "datos ausentes completamente al azar".
- **Missing At Random (MAR)**: Estos datos o valores a diferencia de los anteriores, presentan relaciones con el resto de las variables, por lo que la distribución de probabilidad de los valores perdidos puede tener una distribución distinta a la variable en general, puesto que depende de la relación existente entre los datos de la variables con otros pertenecientes a otra variable.
- **Not Missing At Random (NMAR)**: Son valores perdidos que tienen significado, es decir, el hecho de que esté ausente no es un error, sino información relevante para la variable.

Cabe señalar, que *"para los valores MCAR, existe el supuesto de que la distribución de los valores perdidos y los completos es la misma, mientras que en el caso de los valores de tipo MAR son diferentes por lo que los valores de este tipo pueden ser predichos usando datos completos"* [30, 62].

2.5.4. Valores fuera de rango

Los valores fuera de rango u *outliers* son valores anormales que se pueden definir como *"una entrada en el conjunto de datos que es anómalo con respecto al comportamiento observado en la mayoría de las entradas de la base de datos"* [18]. De esa definición se concluye que un outlier engloba tanto a los valores fuera de rango como también a un tipo de valores incompletos (aquellos de los que se pueda deducir su valor original). Los valores fuera de rango pueden, mediante una medida de distancia, ser ignorados, para que su comportamiento distinto a los valores (a nivel global) de la variable no parcialice el comportamiento global de la variable, es decir, que no afecte el análisis global de la misma. En el caso de los valores incompletos, estos son errores del sistema o humanos (escribir mal el valor de la instancia en un atributo), por lo que, dependiendo del grado de error en el valor (puesto que puede presentar parte de un valor admisible dentro de la variable, generando así, un reemplazo fácil y asumible) y del flujo de dicho atributo en el negocio, se proceden a transformar a valores ausentes, o bien, reemplazar por un valor (que incluso puede ser un valor que indique la falta del campo por error).

La diferencia entre los valores fuera de rango y los ausentes, es que los primeros contienen información y su problema radica en que representan una posible amenaza para el desarrollo de un modelo predictivo o de clasificación, debido a que se refiere a valores incompletos, o bien, a valores que escapan de la tendencia general a un nivel que imposibilita un análisis correcto de la situación. En cambio los valores ausentes, en caso de ser MCAR o MAR, no indican información alguna.

2.5.5. Variables Temporales

Las variables temporales se observan en el problema de series de tiempo y se definen como *"variables en las que los casos o sujetos son diferentes momentos en el tiempo"* [24]. Existen

también lo que se denominan secuencias temporales, éstas *"se forman con los datos recopilados en una base sobre la evolución en el tiempo de un conjunto de características"* [33], por lo que se concluye que es el mismo concepto.

Con respecto a la temporalidad ya sea de la secuencia o variable respectiva, éstas pueden ser estacionarias, lo que significa que su *"valor central y variabilidad no cambian"* [24]; o bien, no estacionarias, es decir, su *"media y variabilidad cambian a lo largo del tiempo, no oscila en torno a un valor constante"* [24].

El análisis de las series depende del tipo correspondiente, si es estacionaria, tendrá sentido describir el comportamiento de la variable mediante *"histograma, media, desviación típica"* [24]. En caso contrario, si se trata de series no estacionarias, estos estadísticos no tienen sentido pues, para este caso, éstos *"ignoran la característica principal de una variable no estacionaria: su cambio discontinuo con el tiempo"* [24].

Cabe agregar que en la no estacionalidad, se agregan los principales componentes en los que se puede dividir la variable, estos son: *"la tendencia, que es el movimiento suave de la serie a largo plazo...la estacionalidad, que el movimiento de oscilación y la componente irregular que se refiere a variaciones aleatorias alrededor de los componentes anteriores"* [24].

Otro concepto relacionado con este tipo de variables es la concordancia entre secuencias de tiempo, la cual puede ser completa o parcial, donde la completa se define como: *"dada una colección de N secuencias de números reales S_i, y otra secuencia Q a consultar, se quiere encontrar aquella secuencia que se encuentras a una distancia ϵ de Q. Todas las secuencias deben tener a misma longitud"* [33], mientras que la parcial difiere en que el *"tamaño de las secuencias es arbitrario"* [33], es decir, las secuencias no necesariamente deben tener la misma longitud.

2.6. Imputación de datos

2.6.1. Estrategias de imputación de datos

Para el tratamiento de dicho valores perdidos o ausentes existen distintas alternativas, dentro de las cuales, las más relevantes son [30]:

- **Descarte de los registros con datos faltantes**: Esta alternativa suele utilizarse cuando la información que aporta la variable es baja, o bien, la cantidad de valores perdidos es baja y la variable tiene poca varianza, no obstante, siempre se debe cuidar el hecho de que al descartar los valores perdidos no se produzca ningún tipo de sesgo.

- **Reemplazo de los datos faltantes con otro valor**: Esta alternativa suele usarse para identificar al valor perdido; en temas de inferencia del valor, no añade ningún tipo de información; sin embargo, al usarlo en la clusterización de los registros puede ser muy útil y entregar información relevante acerca de la naturaleza del valor perdido.

- **Imputación de los datos faltantes**: Esta alternativa es factible cuando la cantidad de atributos con datos faltantes es relativamente pequeña en relación al número de registros que presentan dicha condición. Este método, sí influye en la información de la variable, en otras palabras, puede producir error dentro de la misma. No obstante, todo depende de la experiencia de la

persona que está tratando los valores perdidos, debido a que con la alternativa anterior, no puede llevar a cabo estadísticos descriptivos sin targiversar la información.

Otra perspectiva o alternativa de solución, frente a los datos perdidos, la presentan Roderick Little et al.(2002), que establece las siguientes técnicas para tratar los valores perdidos [62]:

1. **Procedimientos basados en instancias completas**: Esto se refiere a que *"cuando algunas variables no están guardadas para determinadas instancias...se sugiere simplemente eliminar éstas y analizar solamente las instancias con atributos completos...Generalmente es fácil de efectuar y quizás satisfactoria en pequeñas cantidades de valores perdidos. Puede guiar a tendencias equívocas, mas usualmente es bastante eficiente"* [62]

   - **Listwise deletion**: Este análisis es usado por la mayoría de los paquetes (o softwares) estadísticos y sugiere el ignorar el problema de la información completa para su estudio, es decir, se da prioridad a las instancias completas para la realización de los análisis. Sin embargo, esta estrategia es reconocida por no ser adecuada, pues *"genera error en los coeficientes de asociación y de correlación"* [36]. Particularmente el *listwise* se refiere a que, para efectuar los análisis de datos y preprocesamiento e incluso al momento de realizar la predicción, *"se trabaje únicamente con las observaciones que disponen de información completa para todas las variables"* [36]. Lo anterior, es efectuado bajo el supuesto de que *"los datos faltantes siguen un patrón MCAR"* [36], lo que quiere decir que se presume que los datos perdidos o incompletos presentan dicho estado solamente por el azar.

     Cabe destacar que al eliminar las instancias que contengan valores perdidos se está asumiendo que la base completa de datos sigue el mismo comportamiento que la muestra tomada, lo cual es equivalente a tomar un muestreo personalizado basándose en la completitud de las instancias. En consecuencia, resulta menos eficiente que un muestreo estratificado (es decir, tomar la misma proporción de instancias con comportamiento similar), pues puede darse el caso de que el valor perdido en una variable sea un indicador suficiente para concluir el comportamiento del cliente. Esto último, se avala en el hecho de que en *"una muestra probabilística, eliminar observaciones no es correcto ya que se debe tener presente que las unidades fueron elegidas con un procedimiento aleatorio y con probabilidad de selección, conocida y distinta de cero...si la eliminación de registros no se acompaña de un ajuste apropiado de los factores de expansión...se obtendrán estimadores sesgados de los parámetros poblacionales, lo que podría invalidar las conclusiones"* [36].

   - **Pairwise deletion**: Este análisis considera la información completa, pero usando distintos tamaños de muestra. A diferencia del análisis *listwise*, solamente se eliminan aquellas observaciones que no poseen ningún dato, y *"los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados"* [36], siendo está la principal dificultad de este tipo de análisis. Por lo mismo, no se recomienda esta metodología, pues esta incapacitada para contrastar resultados y no se puede dilucidar si el camino a seguir es el correcto o no.

Si bien ninguno de los métodos descritos anteriormente es recomendable, resultan eficientes en cuanto al tiempo dedicado al preprocesamiento de los datos, es decir, se obtienen resultados en un período de tiempo reducido, por lo que, se sugiere su práctica para casos que presenten una baja cantidad de valores incompletos y perdidos, así como también, en los análisis exploratorios iniciales de una base de datos.

2. **Procedimientos basados en la imputación**: En estos procedimientos *"los valores perdidos son llenados y la base de datos completada es analizada por métodos estandarizados. Los métodos comúnmente usados incluyen Hot Deck...Imputación por promedio e imputación por regresión"* [62].

   - **Cold Deck**: Un método no muy conocido en el mercado de la minería de datos, pues requiere de información adicional fuera de la base de datos de estudio. Éste consiste en *"seleccionar los valores o relaciones de uso obtenidos de fuentes distintas al conjunto de datos actual"* [104], es así como este método *"es el opuesto de la imputación hot deck, pues lo que se hace es imputar los valores perdidos de una variable Y con datos de otras variables auxiliares"* [96]. Sea s una muestra seleccionada de una población finita P, y sea r el conjunto que no contiene valores perdidos en la variable señalada. Además, sea x_i una base de datos auxiliar que tiene valores completos para todo $i \in s$. Si el estimador Horvitz-Thompson para la población total de los valores completos y_i es

   $$\hat{Y_R} = \sum_k \sum_{i \in r_k} w_i y_i + \sum_{i \in r_k} w_i \hat{\beta_k} x_i \tag{2.1}$$

   Donde $\hat{\beta_k}$ es un estimador del parámetro de la siguiente ecuación,

   $$y_i = \beta_k x_i + x_i^{\frac{1}{2}} e_i \tag{2.2}$$

   y dicha estimación se realiza con la siguiente fórmula:

   $$\hat{\beta_k} = \frac{\sum_{i \in r_k} w_i y_i}{\sum_{i \in r_k} w_i x_i} \tag{2.3}$$

   Entonces, la estimación total de los valores perdidos de la variable Y se realiza para todo $i \in s - r$ donde s es el conjunto total y r es el subconjunto sin valores perdidos:

   $$\hat{Y_C} = \sum_{i \in r} w_i y_i + \sum_{i \in s-r} w_i x_i \tag{2.4}$$

   Donde esta última forma se denomina Cold Deck simple [96]. En vista de lo anterior, se puede apreciar que la efectividad de este método depende exclusivamente de la existencia de datos pasados o históricos. Además, se requiere de un conocimiento mucho mayor del negocio y de las bases de datos a tratar, así como de un estudio exhaustivo de las mismas para no llegar a conclusiones erróneas

   - **Hot deck**: Es aquel método de imputación en el cual *"para cada ejemplo que contenga un valor perdido, se encuentra el ejemplo más similar y los valores perdidos son impu-*

tados de dicho ejemplo" [30]. Este método supone que la instancia que tiene la máxima similitud con la instancia del valor perdido, no posee el mismo en estado ausente.

Otra definición de este método se refiere al *"reemplazo de un valor perdido con un valor seleccionado de una distribución estimada a partir de las respuestas similares para cada valor perdido"* [104], es decir, se estudia las instancias que no contenga un valor perdido en dicha variables para luego, reemplazar los valores perdidos. Dentro de sus principales ventajas se encuentran *"la complejidad conceptual, el nivel de mantenimiento y medición apropiada de las variables, y la disponibilidad de un conjunto completo de datos al final del proceso de imputación"* [104], esto permite que se requieran de escasos conocimientos de estadística para su ejecución, no obstante, dentro de sus desventajas están *"la dificultad en definir qué es similar"* [104].

- **Media**: En este método, *"los valores perdidos son imputados por la media para datos continuos o, en el caso de categorías, por el valor más frecuente (moda)"* [30], por ende, se sugiere utilizar este método cuando el porcentaje de valores perdidos a reemplazar sea cuasi despreciable(menor al 1 %) respecto a los valores completos en la variable.
- **Regresión**: Este método se sugiere en el caso de que los datos faltantes o ausentes sean del tipo MAR, *"imputando la información de una variable Y a partir de un grupo de covariables* $X_1, X_2, ..., X_p$*"* [58]. Para que no existan casos en que no se pueda predecir una variable, se eliminan las observaciones con datos incompletos (*pairwise* o *listwise*) y las instancias que no contenga alguna de las covariables. El nombre alternativo para este tipo de imputación es de media condicionada.

3. **Procedimiento de asignación de pesos**: Este método es usado cuando se desea adquirir una probabilidad de selección, pues *"las inferencias aleatorias de la muestra de una encuesta sin respuesta comúnmente están basadas en pesos de diseño que suelen ser inversamente proporcionales a la probabilidad de selección"* [62]. Usualmente se usa para estimar la población promedio a partir de una muestra.

    - **Reponderación**: Este tipo de imputación adquiere importancia cuando se tienen muy pocas respuestas o valores de alguna categoría de interés. Las ponderaciones se interpretan como *"el número de unidades de la población"* [36], cuando se detecta la ausencia de información *"los ponderadores de las unidades que respondieron se utilizan para ajustar los factores de expansión...para que la submuestra observada genere estimaciones compatibles con la población"* [36]. El uso de este tipo de imputación se recomienda para casos de problemas de rareza (explicados en [107] o con alto porcentaje de valores ausentes.

4. **Procedimientos basados en modelos**: Este procedimiento se aplica cuando se intuye una relación, lineal o no lineal, entre un subconjunto de variables. No obstante, el procedimiento completo se sintetiza en *"definir un modelo para los valores parcialmente perdidos y basando las inferencias en la similitud del modelo, con parámetros estimados bajos procedimientos tales como el de máxima verosimilitud"* [62]. Dentro de las ventajas de este procedimiento se encuentra la detección de relaciones complejas, además de ser más flexible que los anteriores, en el sentido de la alta diversidad de parámetros que se pueden asignar a los distintos modelos.

2.7. Transformación de datos

2.7.1. Transformaciones para variables temporales

Las variables temporales suponen un problema respecto a la consistencia de las bases de datos debido a que contienen relaciones entre sí y generalmente, contienen una escala muy distinta al resto de los datos. Estas variables son referenciadas como secuencias temporales, las cuales *"se forman con los datos recopilados en una base sobre la evolución en el tiempo de un conjunto de características: evolución de precios, sistemas de producción, datos meteorológicos, medidas de sensores, evolución de sistemas dinámicos, etc."* [82]. Si bien las series temporales tratan este tema, en el procedimiento KDD, en su etapa de transformación se puede recurrir a ciertas técnicas pertenecientes al área de series temporales, pues el objetivo que se busca en esta etapa es reducir la dimensionalidad de variables. A continuación se bosquejan las transformaciones más a usadas:

- **Índices**: Los índices son un *"procedimiento para describir la evolución de la realidad a través del tiempo"* [74]. La fórmula general para un conjunto de datos de facturación es:

$$I_t = \frac{\text{Transacción en mes t}}{\text{Transacción en mes base}} * 100 \tag{2.5}$$

  Aparte de la fórmula anterior, es posible ponderar dichos índices de manera tal que se forme una suma o promedio ponderado, sin embargo, sus mayores desventajas son el hecho de que la dimensionalidad de los datos solamente disminuye en una unidad, por ende *"no trabajan bien con datos de espacios de grandes dimensiones"* [82].

- **Transformación por funciones**: Una alternativa para representar la secuencia temporal es una transformación que preserve la ortonormalidad de la distancia, la cual *"es aconsejable que sea independiente de los datos"* [82]. La función más sugerida para efectuar la transformación de estos datos es la Transformada Discreta de Fourier (TDF), pues *"conserva la energía de sus primeros coeficientes"* [82]. La fórmula de la TDF es:

$$X_k = \sum_{n=0}^{N-1} X_n e^{\frac{-2\pi i}{N} kn} \tag{2.6}$$

  Con k=0,....,N-1. Al momento de trabajar con variables reales, los resultados satisfacen la siguiente ecuación:

$$X_{N-K} = X_K \tag{2.7}$$

  Otra alternativa a la transformación por funciones, es la de promedios ponderados, si bien la información que se pierde es bastante, resulta fácil de interpretar y no contamina el desarrollo del modelo a la magnitud que lo hacen los índices, debido a la correlación entre los mismos. Esto también indica que se pueden efectuar mezclas de promedios para resumir este tipo de variables.

2.7.2. Análisis factorial

Este análisis tiene el propósito de *"simplificar las numerosas y complejas relaciones que se puedan encontrar en un conjunto de variables cuantitativas observadas"* [58]. Esto quiere decir que no se encarga de reducir las variables, sino que busca encontrar el significado de los nuevos factores generados producto del análisis de componentes principales. Por ende, su definición converge a *"un procedimiento matemático mediante el cual se pretende reducir la dimensión de un conjunto de p variables obteniendo un nuevo conjunto de variables más reducido, pero capaz de explicar la variabilidad común encontrada en un grupo de individuos sobre los cuales se han observado las p variables originales"* [58], siendo estas nuevas variables reducidas, los factores obtenidos de dicho análisis.

Este procedimiento posee el siguiente modelo matemático [58]:
Sean

- $X_1, X_2, ..., X_p \equiv$ Variables observadas.
- $F_1, F_2, ..., F_p \equiv$ Factores comunes.
- $e_1, e_2, ..., e_p \equiv$ Factores específicos.

Entonces el modelo consiste en:

$$\begin{aligned} X_1 &= l_{11}F_1 + ... + l_{1m}F_m + e_1 \\ X_2 &= l_{21}F_1 + ... + l_{2m}F_m + e_2 \\ &\dots \\ X_p &= l_{p1}F_1 + ... + l_{pm}F_m + e_p \end{aligned}$$

Donde l_{hj} es el factor h de la variable j y el nombre asignado a dichos coeficiente es de cargas factoriales. Luego siguiendo la lógica del algoritmo, se escriben las variables observadas como una combinación lineal entre factores comunes y específicos. Esto se puede expresar matricialmente como procede:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} l_{11}, ..., l_{1m} \\ l_{11}, ..., l_{1m} \\ \vdots \\ l_{11}, ..., l_{1m} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \tag{2.8}$$

No obstante, el modelo para que se pueda ejecutar de manera adecuada, requiere de determinadas hipótesis sobre los factores comunes, específicos y las relaciones existentes entre estos dos últimos:

Hipótesis sobre factores comunes

- $E[F] = 0$.
- La matriz varianzas-covarianzas es la identidad.

Hipótesis sobre factores específicos

- $E[e] = 0$.
- La matriz varianzas-covarianzas es una matriz diagonal.

Hipótesis sobre la relación entre factores comunes y específicos

- La matriz varianzas-covarianzas coincide con la matriz de correlaciones.
- La matriz de correlación poblacional se puede descomponer en 2 matrices, aquella debida a los factores comunes y otra coincidente con la matriz varianzas-covarianzas de los factores únicos.
- La varianza de la variable poblacional X_j se descompone como: $l = h_j^2 + w_j^2$ Donde h_j es la comunalidad, definida como *"la parte de la varianza que es debida a los factores comunes"* [58], en otras palabras es el grado de extracción de la varianza de los datos originales, que efectuó el modelo del análisis factorial y w_j es la parte de la varianza debido a los factores específicos.
- La correlación entre las variables observadas se pueden reproducir a partir de las cargas factoriales.

En pocas palabras, el análisis factorial busca estimar los coeficientes l_{jk}, es decir, las cargas factoriales, que pueden ser determinadas una vez conocidos los factores comunes. Ahora bien, existe un surtido de métodos para estimar los factores, dentro de los cuales el que resulta más relevante es el de componentes principales debido a que *"las componentes son factores reales porque se derivan directamente de la matriz de correlación"* [58], la es de rápida obtención.

Métodos de elección de factores

Análisis de componentes principales (ACP) Este método consiste en combinaciones lineales de las variables originales incorreladas (covarianza entre e y F igual a 0) entre sí. La primera componente principal es la combinación lineal de variables originales con varianza máxima. El resto de las componentes principales se obtienen de forma equivalente a la primera. Su utilidad para la etapa de preprocesamiento del procedimiento KDD es *"representar los datos iniciales en un espacio con menos dimensiones con una pérdida de información mínima"* [72]. Está herramienta en sí añade información a las variables, pero es en combinación con el análisis factorial que se puede interpretar completamente.

La formulación del problema de los componentes principales contempla las variables tipificadas originalmente para evitar problemas de escala. Supongamos que la matriz de variables originales es Y. Entonces toda combinación lineal h se puede expresar como

$$h = Yv \tag{2.9}$$

con v como vector que permite la combinación lineal

Posteriormente se busca $\|v\| = 1$ tal que la varianza de la componente h_1 sea máxima. Es así que la varianza de las componentes principales es:

$$S_h^2 = v^t V_y v \tag{2.10}$$

Luego se busca resolver el siguiente problema de maximización:

$$\max_v v^t V_y v \tag{2.11}$$

$$\text{sujeto a} v^t v = 1 \tag{2.12}$$

Utilizando el lagrangeano respectivo se obtiene:

$$L = v^t V_y v - \lambda \left(v^t v - 1 \right) \tag{2.13}$$

Lo cual, al derivarlo, se obtiene

$$\frac{\partial L}{\partial v} = 2V_y v - 2\lambda = 0 \tag{2.14}$$

Y esto implica,

$$\left(V_y - \lambda I \right) v = 0 \tag{2.15}$$

Donde v es vector propio de la matriz de varianzas-covarianzas de los datos originales tipificados. Es decir que la primera componente $h_1 = Yv_1$, así como también, λ es valor propio de la mencionada matriz.

La interpretación de las componentes principales se efectúa mediante el análisis de correlaciones entre las componentes principales y las variables originales. Dichos coeficientes de correlación tienen asociada la siguiente expresión para su cálculo:

$$r_{kj}^* = h_{kj} \sqrt{\lambda_k} \tag{2.16}$$

Donde que r_{kj} es el coeficiente de correlación lineal entre la k-ésima componente principal (h_k) y la j-ésima variable inicial tipificada (Y^j). Además, h_{kj} es la j-ésima coordenada de la k-ésima componente principal.

Finalmente se puede obtener la puntuación factorial, una vez calculados los coeficientes h_{jk} de la siguiente manera:

$$C_{mi} = h_{m1}X_{1i} + h_{m2}X_{2i} + h_{m3}X_{3i} + ... + h_{mp}X_{pi} h = 1, 2, ..., pi = 1, 2, ..., n \tag{2.17}$$

Máxima verosimilitud Este método estima un conjunto factores por medio de factorizaciones sucesivas, donde cada uno explica su máximo de varianza en la matriz de correlaciones poblacional, expresando dicho valor en forma de estimaciones sobre la matriz de correlación muestral. El mayor argumento que acredita su uso es que la robustez del análisis se refleja en tests estadísticos que señalan el grado de significancia de cada factor extraído.

Si L es la matriz de cargas factoriales, μ es el vector $E(X)$ con E siendo la esperanza de las variables observadas, Σ es la matriz $V(X)$ donde V es la matriz de varianza-covarianza, F es el

vector de factores comunes y e el vector de factores específicos, entonces, el método se modela de la siguiente forma [83]:

Sea X $N_p(\mu, \Sigma)$ donde $X - \mu = LF + e$ y $\Sigma = LL' + \Psi$. Entonces

$$f(x_1, x_2, ..., x_n) = \prod_{k=1}^{k=n} \frac{1}{(2\pi)^{\frac{p}{2}} + |\Sigma|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x_i - \mu)' \Sigma^{-1}(x_i - \mu)\right\} \tag{2.18}$$

El objetivo de la ecuación anterior es encontrar $\hat{L}$, $\hat{\Psi}$ y $\hat{\mu}$ (donde el sombrero significa un estimador para el vector señalado) que maximizan $f(x_1, x_2, ..., x_n)$

Además, cuenta con las siguientes propiedades:

- No hay óptimo único por lo que se requiere que $L'\psi^{-1}L = \Delta$ siendo este último símbolo el símil de una matriz diagonal.
- La solución se obtiene computacionalmente y las comunalidades se calculan como la suma de las cargas factoriales.
- La proporción de varianza explicada por el factor j-ésimo es:

$$Varexp_j = \frac{\hat{h_j}}{\text{Varianza total muestral}} \tag{2.19}$$

Ejes Principales Es un método similar al ACP, mas su diferencia radica en el supuesto de un procedimiento iterativo, el cual *"en lugar de encontrar 1 en la diagonal de la matriz de correlación, las comunalidades son estimadas"* [58]. Por lo tanto, se reduce la matriz de correlación para calcular los autovalores, con los que se determina el número de factores a considerar. Las comunalidades se estiman como la suma de los cuadrados de las cargas factoriales.

Los requisitos para llevar a cabo este método son [83]:

- E[F]=0 y Varianza(F)=Identidad
- E[e]=0 y Varianza(e)=Ψ
- F y e son incorrelados, es decir, cov(e,F)=0

Rotación de factores

Para poder interpretar los factores resultantes del análisis factorial se deben rotar las soluciones, pues *"los factores con muchas cargas factoriales son difíciles de interpretar y no existe una solución perfecta o ideal en el análisis factorial"* [58]. Por lo tanto, se utiliza el hecho de que toda rotación ortogonal de los ejes de solución, también es solución. Sin embargo, hay dos tipos de rotaciones, las ortogonales y las oblicuas. La ventaja principal de las ortogonales es la simplicidad, ya que los pesos representan las correlaciones entre los factores y las variables, lo cual no es aplicable para el caso de las rotaciones oblicuas [73]. Es por ello que a continuación solamente se describen las rotaciones ortogonales [58]:

- Varimax: Su objetivo es minimizar el número de variables con altas cargas en un solo factor, por lo que se realiza una maximización de la suma de las varianzas de las cargas factoriales dentro de dicho factor, resultando por columna valores cercanos a 1 o 0.

- Quartimax: Este método busca conseguir que una variable tenga una carga alta en un factor y baja en los demás, es decir, que se asocie a una variable representativa. Sin embargo, resulta ineficiente cuando se está en presencia de múltiples variables.

- Equimax: Solución mixta que toma en consideración las habilidades de los dos métodos anteriores, donde además se pueden asignar pesos a cada criterio.

Medidas de calidad del Análisis factorial

Si bien es necesario reducir la dimensionalidad de las variables en el procedimiento KDD, deben existir los estadísticos suficientes como para concluir que el análisis efectuado ha sido un éxito, he aquí las principales medidas encargadas de dar tal apoyo [58]:

- **Determinante**: Indica el grado de intercorrelaciones, en el caso de que sea muy bajo, indica que las correlaciones entre los factores son altas.

- **Estadístico Kaiser-Meyer-Olkin (KMO)**: Es una de las medidas más usadas para evaluar la calidad de los factores, es un coeficiente de correlación parcial que mide la correlación entre dos variables una vez que se han descontado los efectos lineales adversos referentes a los factores comunes. En consecuencia, el coeficiente de correlación parcial entre dos variables es equivalente al coeficiente de correlación entre factores únicos de dos variables. En particular su forma matemática se representa como:

$$KMO = \frac{\sum\sum_{h\neq j} r_{jh}^2}{\sum\sum_{h\neq j} r_{jh}^2 + \sum\sum_{h\neq j} a_{jh}^2} \tag{2.20}$$

  Donde $r_{jh} \equiv$ coeficientes de correlación observados entre variables originales y $a_{jh} \equiv$ coeficientes de correlación parcial entre variables originales

  En el caso de que se tenga una adecuación correcta de los datos a un modelo de análisis factorial, a_{ij} será pequeño, por ende el KMO tenderá a estar cercano a 1, lo cual es el equivalente de un buen modelo factorial, es decir, los datos pueden ser representados por un modelo de factores. Cabe destacar, que se recomienda el uso de variables continuas para la aplicación de esta técnica, debido a que trabaja con nociones de división y multiplicación, las cuales no son aplicables a las variables nominales.

- **Contraste de esfericidad de Bartlett**: Busca comprobar que la matriz de correlaciones es significativamente distinta a la matriz identidad. De ser el caso contrario, indicaría que no existiría correlación entre variables, lo cual implicaría un absurdo en la aplicación de la técnica. Por lo tanto, este test permite realizar tal comparación y consiste en una transformación de la χ^2, suponiendo una transformación normal multivariante, lo que presupone que los datos

provienen de una distribución normal multivariante. Siendo así, su expresión matemática es la siguiente:

$$Test = \chi^2 \left[0,5 \left(p^2 - p\right)\right] = -n \left[n - 1 - \frac{1}{6}(2p + 5)\right] Ln|(R|) \tag{2.21}$$

Donde

- $n \equiv$ dimensión de la muestra
- $p \equiv$ número de variables observadas
- $R \equiv$ matriz de correlación observada

2.7.3. Segmentación

Diferencia entre segmentación y clusterización

Existen múltiples definiciones acerca de la segmentación, dentro de las cuales, se destacan las siguientes:

- La segmentación es *"el proceso de agrupar personas u organizaciones dentro de un mercado acorde a necesidades, características o comportamientos similares"* [19].
- *"La segmentación es la subdivisión de un mercado en distintos tipos de clientes. En estos segmentos, los miembros son diferentes entre segmentos pero similares dentro de uno mismo"* [78].

Equivalentemente para el término de clusterización se da una problemática similar, cuyas definiciones destacadas son:

- La clusterización es *"un proceso de machine learning no supervisado que crea clústers de tal forma que los puntos de datos dentro de un clúster están cercanos unos con otros y además, alejados de los puntos pertenecientes a otros clústers"* [94].
- La clusterización se puede ser señalada como que *"agrupará los datos en conjuntos de observaciones relacionadas o clústers. Observaciones entre instancias del mismo grupo son más similares que otras observaciones de externas al grupo. Clusterización es un método no supervisado de aprendizajes"* [79].

A partir de estas definiciones tanto de segmentación como de clusterización se puede notar la similitud entre estos términos. Para aclarar en qué se diferencian se simplificarán las definiciones de ambos, lo que para el caso de segmentación se traducirá en un método de *"dividir algo en piezas acorde a algún criterio y cada pieza se denominará segmento"* [2], y la clusterización será una técnica para *"encontrar regiones en un espacio con diferente densidad de objetos que el resto de las regiones"* [2]. Dicha diferencia, puede ser apreciada en el siguiente gráfico:

En base a la figura 2.5 se puede establecer que *"clusterizar es encontrar bordes significativos entre grupos, mientras que la segmentación es usar bordes para formar grupos"* [2], lo que en otras palabras indica que en un conjunto de datos siempre se podrá segmentar para poder interpretar los

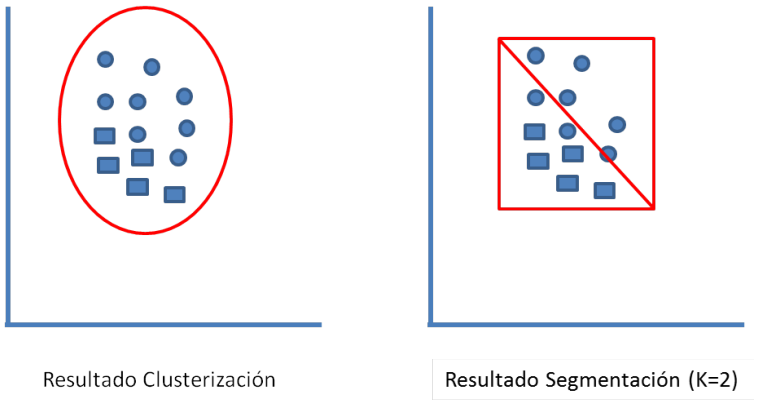

Figura 2.5: Gráfico de diferencia entre clusterización y segmentación

deseado, en cambio la clusterización solamente será útil si el conjunto de datos es no homogéneo, de lo contrario, entregará un único grupo. Por lo tanto, estos términos son equivalente en el caso de que la clusterización entregue un resultados que pueda ser interpretado, puesto que en una definición informal, la clusterización es una segmentación válida desde el punto de vista técnico.

Una vez comprendida la diferencia, se detalla la segmentación, pues posee un criterio a asignar.

Tipos de segmentación

La segmentación se puede dividir en dos tipos:

- Segmentación a priori: Es aquella en la que *"tanto el número de segmentos como su tamaño o su descripción se establece antes de que el estudio se lleve a cabo"* [58]. Para esa elección, se requiere un conocimiento del mercado bastante alto, usado como validador de una intuición en el negocio. Dicha intuición puede poseer fundamento técnico, así como también, puede haber sido creado por un experto del negocio.

- Segmentación post hoc: A diferencia de la anterior, en ésta *"se desconocen las características de mercado y sus reacciones ante un nuevo producto"* [58], por ende, *"en este modelo, el número de segmentos, su tamaño y su descripción se conocen tras el análisis, no antes"* [58]. Para averiguar el número de segmentos o los otros elementos faltantes, generalmente, *"se aplica una técnica estadística conocida como análisis de conglomerados"* [58], que *"también se denomina análisis clúster"* [58]. Para ambos casos se debe anticipar *"una exploración cualitativa para conocer en profundidad el mercado"* [58].

Beneficios de la segmentación

Los motivos que suelen originar la necesidad de una segmentación son [84]:

- Encontrar grupos en que sus elementos sean
    - Lo más heterogéneos posible respecto a los elementos pertenecientes a los otros grupos.

  - Lo más homogéneos posible respecto a los elementos que pertenecen al mismo grupo.

- Encontrar grupos

  - Significativos (tamaño justificable).
  - Alcanzables (accesibles para la compañía de acuerdo a sus recursos y experiencia).
  - Identificables (interpretables).

El más común de los motivos es *"la elaboración de clasificaciones o tipologías"* [58]. Este último es la razón principal por el que se utilizó la segmentación en este estudio.

Ahora bien, existen muchas otras áreas donde la segmentación resulta útil, puesto que entrega más beneficios que los técnicamente enunciados anteriormente. Sin embargo, para que estos beneficios se vislumbren, se necesita la aplicación de medidas a la información entregadas por la segmentación. Es así, como a continuación, se muestran los principales alcances que posee [78, 110]:

- Firma o Compañía: Identificación de clientes valiosos, así como también, clientes destructivo para la compañía. Promociones y publicidad más enfocada hacia los clientes. Mayor valor del cliente en el tiempo. Esto en el mediano-largo plazo converge a un crecimiento de rentabilidad sustentable.

- Cliente: Productos y servicios personalizados, por ende, una experiencia más personal con la empresa, que conlleva a una mayor satisfacción. Esto, en el mediano-largo plazo converge a una mayor retención de clientes y la aparición del efecto lealtad sobre los mismos.

- Según área de aplicación [67]:

  - Ingeniería: Síntesis de información, reducción de errores en los datos.
  - Ciencias de la computación: Clasificación de sitios web, de accesos vía web, descubrimiento de conceptos implícitos en foros.
  - Medicina: Definición de taxonomías, segmentación de genes.
  - Astronomía: Clasificación de sistemas y planetas.
  - Ciencias Sociales: Segmentación de personas según patrones de comportamiento.
  - Economía: Segmentación de clientes, servicios, productos.

Sin embargo, para la obtención de la segmentación como tal, se puede acudir a distintas herramientas de análisis, las cuales pueden llegar a resultados distintos debido a que se sostienen bajo diferentes supuestos acerca de la composición de la base de datos. Estas herramientas posibles son [78]:

- Análisis de clúster (o clusterización): Este método es uno de los más usados a nivel académico y de investigación, debido al nacimiento de la minería de datos que reúne los conocimientos de estadística con los de *machine learning*.

- Análisis de panel [70]: Se denomina "estudio de panel" a la recolección de información sobre una pluralidad de unidades de análisis en varios instantes del tiempo.

- Análisis regresión: Este método, sugiere la utilización de funciones regresivas para la agrupación de clientes en base a una puntuación entregada por las mencionadas funciones.
- Segmentación a base de juicio: Este método sugiere una heurística particular originada por el criterio de una persona del negocio, respecto a una o más variables o atributos que caractericen al cliente, producto o servicio.

Adicionalmente existe un procedimiento para la segmentación y otro para la clusterización:
El primero se divide en cuatro etapas, las cuales se describen brevemente a continuación [58]:

1. Obtención de la matriz de datos: Se refiere a poseer una base de datos cuasi estandarizada, es decir, sin valores perdidos.
2. Estandarización de la matriz de datos: Esta etapa es opcional y es donde se normalizan las variables para que sean comparables, sin embargo, *"la estandarización elimina la identidad de cada variable y transforma su valor numérico en unidades adimensionales"* [58], pero su uso se justifica en el hecho de que *"elimina la arbitrariedad"* [58].
3. Cálculo de la matriz de semejanzas: Para definir esta etapa se requiere el concepto de medida de similaridad, el cual se presenta como sigue: la medida de semejanza es un coeficiente, que mide la semejanza global entre cada par de entidades. Existen dos tipos de medidas de semejanza, las medidas de disimilaridad y las medidas de similaridad, la diferencia entre ambas radica en la afirmación *"cuanto más pequeño es el valor de una medida de disimilaridad, más semejantes son dos objetos, mientras que cuanto más grande es el valor de una medida de similaridad, más semejantes son"* [58].
4. Ejecución del método de agrupamiento.

En cambio, cuando se refiere a la clusterización, se debe ser más riguroso en cuanto a la evaluación que valida de los grupos, es decir, no se tiene una clasificación a partir de un criterio, sino que se busca una métrica que represente la validez técnica de los clústers obtenidos. Es así, como el procedimiento que deben seguir en la aplicación de la clusterización es el mostrado a continuación [110]:

- **Selección y extracción**: En esta etapa se extraen las variables relevantes, ya sea en el mismo formato que viene en la base original, o bien, con alguna transformación que entregue mayor información.
- **Diseño y selección de algoritmo de clusterización**: Es aquella etapa en la que se buscan y se seleccionan medidas de proximidad, las cuales se utilizarán en la construcción de un algoritmo de clusterización, dichas medidas afectan directamente los resultados del algoritmo usado.
- **Validación de clústers**: Existen criterios y medidas que dan cuenta de la certeza de la clusterización, muchas de ellas se basan en la agrupación que el algoritmo hace respecto a la agrupación inherente en la base de datos sobre la cual se aplica el algoritmo o bien, buscan el número de clústers o grupos apropiados para la base de datos. En síntesis, existen

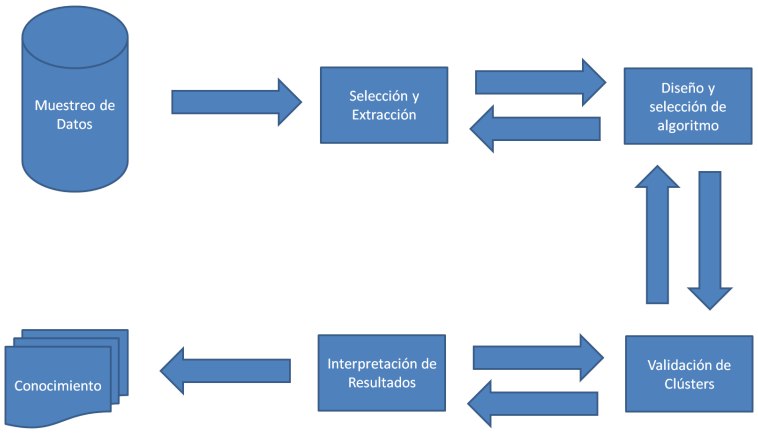

Figura 2.6: Procedimiento para efectuar la clusterización

tres categorías generales de criterios de validación: índices externos, índices internos e índices relativos, cada uno de ellos son definidas para tres tipos de estructuras de clusterización: particional, jerárquica e individual. Los índices externos se basan en una estructura pre especificada en la base de datos, es decir, toman en consideración una posible clasificación que se presentaba previo al análisis. Los índices internos no dependen de información externa, sino que buscan evaluar la estructura más similar a la que está implícita en la base de datos. Por último, los índices relativos, busca comparar diferentes estructuras de clusterización, para decidir la más apropiada para la base de datos a utilizar. Ahora bien, las medidas más importantes se pueden encontrar en [100]. Cabe destacar que la verdadera calidad se mide en la aplicación a la realidad de los clústers, o bien, con una validación conjunta con el experto en el negocio.

- **Interpretación de resultados**: En esta etapa se utilizan los valores representativos entregados por el algoritmo para cada grupo originado, en base a estos, y en conjunto con expertos en el negocio, se buscan los atributos que más caracterizan el producto o servicio y se etiqueta a los grupos, basándose en los valores que tiene cada "representante".

Medidas de semejanza

Estas medidas de semejanza o distancia se definen como *"un coeficiente que mide la semejanza global entre cada par de entidades...consiste en una fórmula matemática que sirve para calcular lo semejantes que son dos entidades"* [58]. Esta definición, señala la medición matemática de la similitud entre dos instancias o registros pertenecientes a una misma base de datos. No obstante, dichas medidas se dividen en dos tipos dependiendo de la perspectiva o dirección que plantee la medición. Esta división es: medidas de disimilaridad o medidas de similaridad, su diferencia es que *"cuánto más pequeño es el valor de una medida de disimilaridad, más similares son dos objetos; en cambio cuanto más grande es el valor de una medida de similaridad, más similares son"* [58]. Sin embargo, dependiendo del tipo de escala se tendrá otra división para las medidas de semejanza debido a que la configuración matemática será distinta. Por ende, las medidas de distancia serán

distintas si las variables son de intervalo, binarias, nominales o mezcla de varias variables. Las principales medidas utilizadas son:

- Medida Euclidiana [26]

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{N} \left(x_{ip} - x_{jp}\right)^2} \tag{2.22}$$

- Medida Minkowski [26]

$$d(x_i, x_j) = \left\{ \sum_{p=1}^{N} \left(x_{ip} - x_{jp}\right)^q \right\}^{\frac{1}{q}} \tag{2.23}$$

- K. Pearson [26]

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{N} \left(\frac{x_{ip} - x_{jp}}{s_p}\right)^2} \tag{2.24}$$

- Canberra [26]

$$d(x_i, x_j) = \sum_{p=1}^{N} \frac{|x_{ip} - x_{jp}|}{|x_{ip}| + |x_{jp}|} \tag{2.25}$$

- Clark [26]

$$d(x_i, x_j) = \left\{ \frac{1}{n} \sum_{p=1}^{N} \frac{\left(x_{ip} - x_{jp}\right)^2}{\left(x_{ip} + x_{jp}\right)^2} \right\}^{\frac{1}{2}} \tag{2.26}$$

- Mahalanobis [58]

$$D^2 = \left(\bar{X}_i - \bar{X}_j\right)' V^{-1} \left(\bar{X}_i - \bar{X}_j\right) \tag{2.27}$$

  Donde

  - $\bar{X}_i$ = Vector de medias del grupo i
  - $\bar{X}_j$ = Vector de medias del grupo j
  - V^{-1} = Inversa de la matriz de varianzas-covarianzas intragrupo

En cambio, si se trata de variables con escala binaria(que son un subtipo de la escala nominales), las medidas; que son para las variables de escala de intervalo; no aplican, por lo que se requiere medidas que describan la distribución de concordancia entre dos instancias. Para tener un mayor entendimiento de estas medidas, los resultados se presentan en la siguiente matriz [58]:

Acorde a esta matriz, se aclara que los valores “a” y “d” hacen referencia a que ambas instancias poseen el mismo valor en el atributo de escala binaria. Una particularidad es que $a + b + c + d = p = total$ de variables de escala binaria. Por último, se muestran algunas de las fórmulas para medidas de similaridad [26, 58]:

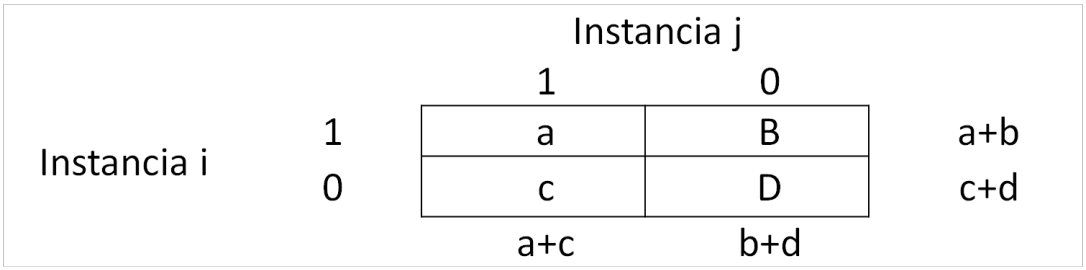

Figura 2.7: Esquema para medidas nominales

- Concordancia Simple

$$CS = \frac{a+d}{p} \tag{2.28}$$

- Sorenson

$$Sor = \frac{2a}{2a+2b+c} \tag{2.29}$$

- Jaccard

$$Jac = \frac{a}{a+b+c} \tag{2.30}$$

- Sokal y Sneath

$$Sok = \frac{2(a+d)}{2(a+d)+b+c} \tag{2.31}$$

- Russell y Rao

$$Rao = \frac{a}{a+b+c+d} \tag{2.32}$$

Algunas consideraciones antes de aplicar estas medidas de distancia son el hecho de que existan *"casos en los que puede ser totalmente inapropiado considerar dos entidades como similares simplemente porque ambas carezcan de algo, así como también, en otras situaciones puede resultar inapropiado olvidarse de las ausencias de una característica a la hora de estimar la similaridad"* [58]. Lo cual indica que para afirmar si una instancia es más semejante a una que a otra, se debe estudiar la importancia de las variables en las cuales las instancias son semejantes, de lo contrario, se incurrirá en el error de tener dos instancias igual de semejantes a una tercera sin poder concluir concretamente a cual se parece más. Además, *"antes de decidirse por un determinado coeficiente debemos observar si las variables son simétricas o asimétricas, siendo esta última aquella en la que existe una gran desproporción entre los ceros y unos, o viceversa"* [58].

Para las variables con escala ordinal, nominal o de razón, se pueden aplicar transformaciones que las dejen con una escala de intervalo, o bien, generando variables binarias (dummies) para cada categoría en el caso de las variables de tipo nominal. En particular, Jean-Pierre Lévy et al. sugiere que para las variables nominales, *"se convierta a datos binarios, o , analizar los datos tal cual, utilizando la medida de concordancia simple"* [58]. En el caso de las variables ordinales, *"se reco-*

mienda trata la escala como si fuese de intervalo, aplicando las fórmulas correspondientes" [58]. Por último, para las variables de razón, se puede *"tratarlas como si fuesen de intervalo, realizar primero una transformación logarítmica sobre las variables y luego tratarlas como de intervalo, tratarlas como si fuese ordinales y tratar los rangos como si fuesen de intervalo"* [58].

Todo lo anterior tiene su justificación completa en el caso de que la base de datos en la que se trabaje solamente cuente con variables de un tipo de naturaleza. No obstante, si se tiene una base de datos con variables de distinto tipo de escala, entonces se sugieren 4 soluciones [58]

- Ignorar el problema: Solución más simple, pero usualmente conlleva mayor error.
- Llevar a cabo diferentes análisis: Solución que contrasta ambos tipos de distancias (binarias y de intervalo).
- Transformar los datos cuantitativos en cualitativos: Utilizando variables dummy para las variables de tipo intervalo, conlleva a pérdida de información.
- Utilizar el coeficiente de Gower

Un detalle matemático para el lector experto es que la disimilaridad no es lo mismo que la distancia, pese a que en [58] se trate como tal, *"la distancia δ sobre un conjunto Ω es una aplicación de Ω x Ω en IR, tal que a cada par (i,j) hace corresponder un número real $\delta(i, j) = \delta_{ij}$, cumpliendo algunas de las siguientes propiedades"* [26]:

- Propiedad 1: $\delta_{ij} \geq 0$
- Propiedad 2: $\delta_{ij} = 0$
- Propiedad 3: $\delta_{ij} = \delta ji$
- Propiedad 4: $\delta_{ij} \geq \delta_{ik} + \delta jk$
- Propiedad 5: $\delta_{ij} = 0 \Longleftrightarrow i = j$
- Propiedad 6: $\delta_{ij} \leq max\left\{\delta_{ik}, \delta_{jk}\right\}$ (desigualdad ultramétrica)
- Propiedad 7: $\delta_{ij} + \delta_{kl} \leq max\left\{\delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk}\right\}$ (desigualdad aditiva)
- Propiedad 8: δ_{ij} es aditiva
- Propiedad 9: δ_{ij} es Riemanniana
- Propiedad 10: δ_{ij} es una divergencia

Por ende una similaridad es un subconjunto del concepto de distancia, es así, como *en general, dada una similaridad cualquiera (no necesariamente comprendida entre O y 1), podemos definir la distancia* $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$*"* [26], Una clasificación de las distancias en base a sus propiedades, se muestra a continuación [26]:

1) Disimilaridad: P.1, P.2, P.3

2) Distancia métrica: P.1, P.2, P.3, P.4, P.5

3) Distancia ultramétrica: P.1, P.2, P.3, P.6

4) Distancia euclídea: P.1, P.2, P.3, P.4, P.8

5) Distancia aditiva: P.1, P.2, P.3, P.7

6) Divergencia: P.1, P.2, P.1 O

Donde P significa propiedad que debe cumplir.

Criterios y Algoritmos de segmentación

Los algoritmos de segmentación son aquel esquema matemático implícito que permite la agrupación de elementos similares entre sí. Utilizan características innatas de las escalas de las variables, o bien, usan la instancia como un vector y efectúan una comparación más global. Dentro de los algoritmos de segmentación existen distintas clasificaciones, de las cuales sólo se darán a conocer dos. La primera categorización es respecto al tipo de problema(clasificación o de clusterización), si es de clasificación, el algoritmo se denomina supervisado. En cambio, en caso de que el problema sea de clusterización, el algoritmo a utilizar está en la categoría de no supervisado. Cabe enfatizar el hecho de que la estructura del algoritmo se basa tanto en el tipo de problema a resolver como en la estructura misma. Esto último muestra la existencia de la segunda categorización: algoritmos jerárquicos y algoritmos no jerárquicos.

Los algoritmos de asignación jerárquica buscan *"determinar un criterio de cercanía o método de unión entre cada par de objetos...bastante explicativa, pero poco eficiente cuando se trabaja con muchos casos"* [84]. Estos métodos se subdividen en dos tipos principales [84]: Los métodos aglomerativos (cada objeto es un grupo y en cada iteración de aglutinan los más similares) y los métodos divisivos (se tiene un gran grupo general y se efectúan divisiones hasta llegar a un criterio de corte). El otro tipo de métodos son los de asignación no jerárquica. Estos *"determinan un número predefinido de grupos en donde deben ser asignados todos los casos..., menos explicativa, pero eficiente con muchos casos"* [84]. Es así como en los de métodos de asignación no jerárquica existe una subdivisión, la cual es [84]: Métodos discretos (donde cada objeto sólo puede pertenecer a un único grupo, por ejemplo K-means, Two-Step Clúster, Árboles de decisión) y los métodos difusos (donde cada objeto tiene un grado de pertenecía a cada uno de los grupos, por ejemplo, Fuzzy C-Means, clase latente).

Como se habrá observado, la asignación a los grupos depende exclusivamente del criterio que se escoja, así como también, la conglomeración en los métodos jerárquicos depende de las medidas de similitud que se utilicen.

No obstante, a continuación se detallan los criterios de asignación en los métodos jerárquicos, debido a que las medidas de similitud ya han sido descritas.

Criterios de asignación Los criterios de asignación basan su validez en la visión de un experto.

Los principales criterio generalizados que existen son [58, 84]:

- **Vecino más cercano**: En este criterio *"se juntan los grupos que presentan la mínima distancia entre objetos* [84].*"Este método no resulta apropiado en aquellas situaciones en las que dos conglomerados se acercan demasiado, ya que inmediatamente pasarían a estar unidos. Dicho fenómeno que se denomina encadenamiento, da como resultado conglomerados alargados, en los que algunos miembros se encuentran muy alejados de otros"* [58]. Para ayudar en la comprensión de este método se esquematiza la idea general, a continuación:

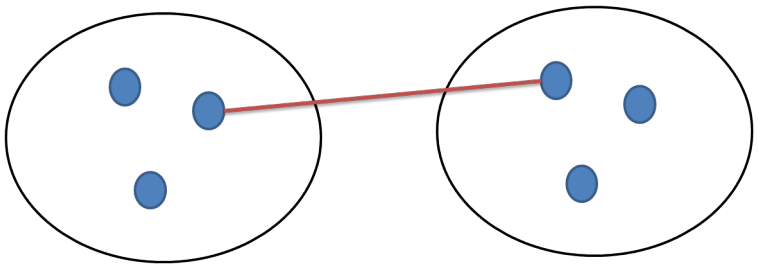

Figura 2.8: Esquema del método del Vecino más cercano

- **Vecino más lejano**: *"Se juntan los grupos que presentan la mínima distancia entre los objetos más distantes del grupo"* [84]. *"Lógicamente esta técnica menos es proclive a producir encadenamientos, no obstante, produce conglomerados compactos, de pequeño diámetro"* [58].

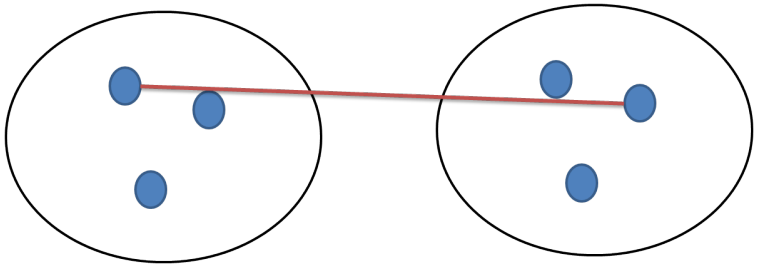

Figura 2.9: Esquema del método del Vecino más lejano

- **Vecino Promedio**: Se juntan los grupos que presentan la mínima distancia promedio entre grupos [84]. La disimilaridad entre los conglomerados es la media de todas las disimilaridades. *"resulta una técnica apropiada para encontrar conglomerados de forma más o menos esférica"* [58].

- **Método del centroide**: Se juntan los grupos que presentan la mínima distancia entre sus centroides (medias para todas las variables) [84]. Un punto importante es que, para usar este método, el algoritmo utilizado debe generar centroides o "representantes" de cada grupo.

- **Método de Ward**: También llamado método de la varianza mínima. Lo que hace es buscar dos conglomerados cuya unión conlleve el menor incremento de la varianza. Esto significa que en cada paso debe probar con todas las combinaciones posibles de dos conglomerados,

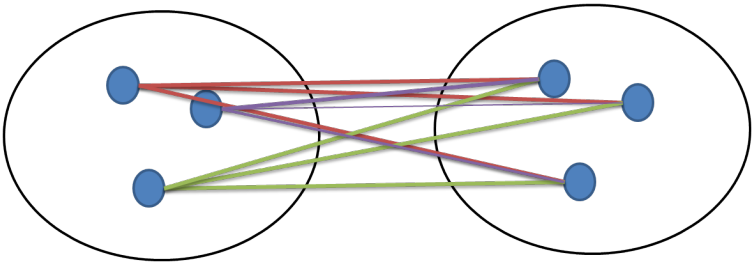

Figura 2.10: Esquema del método del Vecino promedio

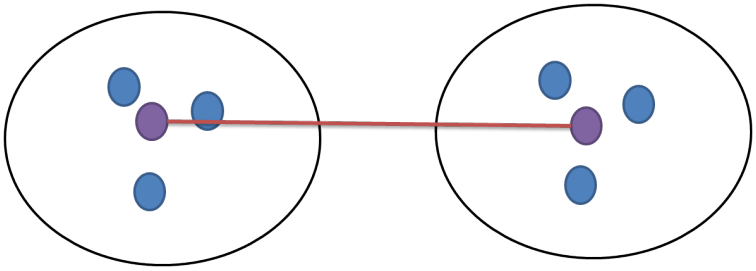

Figura 2.11: Esquema del método del Centroide

calcular el valor del índice de la suma de cuadrados (el índice que cuantifica la varianza) y seleccionar aquel con el menor valor [58]. Esto se puede apreciar en la figura 2.12.

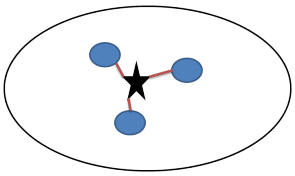

Figura 2.12: Esquema del método de Ward

Algoritmos de Segmentación Los algoritmos de segmentación como se ha mencionado anteriormente, pertenecen a una clasificación entre jerárquicos y no jerárquicos, no obstante, existen ocasiones en las que se pueden complementar ambos algoritmos, utilizando aquellos jerárquicos para obtener una noción a sobre la cantidad de clústers o grupos que tiene el conjunto de datos estudiado, por ende, se presentan los algoritmos más usados en lo que se refiere a la segmentación de variables:

Algoritmos No Jerárquicos

- **K-means**: Este algoritmos busca una partición de X (como conjunto de datos o instancias) en K grupos a través de un criterio particular, siendo *lejos el método de asignación más utilizado en la segmentación de mercados* [84]. Su formulación matemática es [64, 84]:

Sea $X = \{x_1, x_2, ..., x_m\}$ la base de datos con M instancias. Además, sea la función criterio (la más usada, suma de errores al cuadrado):

$$SSE = \sum_{j=1}^{K} \sum_{x_k \in C_j} |\{x_k - \mu_j|\}^2 \tag{2.33}$$

Donde μ_j es el centroide de las instancias del clúster C_j y K es el número de clústers seleccionado. Esto declara que el K-means es un algoritmo iterativo que minimiza de forma local la suma de errores cuadráticos. Finalmete el proceso que se itera es:

- Inicialización
- t=0
- Elegir arbitrariamente $\mu_j(t)$
- Asignación y actualización de centros.
- Asignar X_i al grupo más cercano para todo $i = 1...N$.
- Recalcular $\mu_j j = 1...K$
- t=t+1
- Criterio de parada: Si $\mu_i(t) - \mu_i(t+1) \leq e \forall i$ parar.

El algoritmo estándar usa la distancia euclídea como medida de disimilaridad [64], por lo que requiere del uso de variables continuas para entregar resultados válidos y confiables.

- **Fuzzy C-Means**: Este algoritmo es similar al K-means, su diferencia radica en la utilización de la lógica difusa, la cual, sugiere que las instancias no tienen un valor absoluto de pertenencia, sino que grados de pertenencias para los distintos grupos a formar. Además, en este algoritmo se requiere de un conjunto difuso Y, el cual se define como [84]:

$$Y = \left\{ \left(x, u_y(x)\right) ; x \in X \right\} \tag{2.34}$$

Donde $u_y(x)$ es el grado de pertenencia del elemento x al conjunto difuso Y. Sin embargo, para el caso del algoritmo en sí, se tiene el siguiente procedimiento:

- Antes de la primera iteración se elige arbitrariamente el grado de pertenencia de cada objeto a cada grupo, $u_{ij} \in [0, 1]$, u_{ij} = grado de pertenencia de objeto i a grupo j, donde $\sum_j u_{ij} = 1$
- En cada iteración se determinan los centros de los grupos y se actualizan los grados de pertenencia, utilizando una función con un grado de difusividad asociado.
- Iterar hasta que los cambios en los centros de cada grupo no sean significativos.

- **Two-Step Cluster**: Este algoritmo tiene una ventaja fundamental sobre el K-Medias debido a que permite el uso de variables categóricas y nominales sin afectar el resultado en cuanto a su robustez y validez. El esquema de pasos se divide en dos etapas, en donde la primera consiste en aplicar iterativamente un algoritmo jerárquico, contemplando únicamente las variables

categóricas o nominales, hasta formar clústers previos. En la etapa posterior, el algoritmo cambia en una dirección no jerárquica, tomando como elementos de entrada los clústers previos, provenientes de la etapa anterior, luego utiliza las variables continuas para entregar los grupos finales. Formalmente el algoritmo sigue la secuencia [65, 84]:

- Primera etapa: Aplicación de CFTree
  - Escaneo instancia por instancia.
  - Decide si fusionar la instancia con los clúster previos o formar uno nuevo basado en el criterio de distancia escogido.
  - Aplicación el CFTree, el cual, consiste en un árbol de altura balanceada con dos parámetros, el factor de raminifcación B y el corte de decisión T, además, cada nodo hoja o hijo debe tener a lo más L entradas, tal y como se representa en la figura
  - Identificación de la hoja apropiada descendiendo recursivamente el CFTree hasta encontrar el nodo hijo más cercano.
  - Si no hay espacio en la hoja, se separa escogiendo el par de instancias más alejado como semillas y redistribuyendo el resto de las instancias.
  - Actualización de la información CF para cada hoja falsa en el camino del árbol, siendo la hoja falsa aquella que no se puede separar más.

  Un ejemplo de cómo se mueve el CFTree es el siguiente: Cabe señalar que CF se refiere

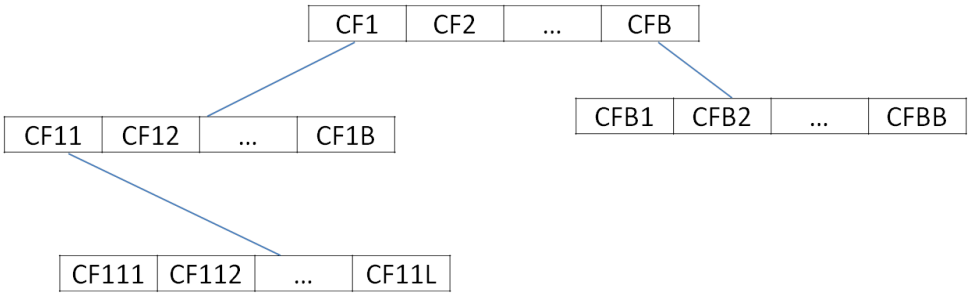

Figura 2.13: Esquema de CFTree

  al atributo del Clúster, la cual se define como CF=(N,M,V,K) con
  - N= Número de instancias en el clúster
  - M= La s medias de cada variable continua de las N instancias
  - V= Varianza de cada variable continua de las N instancias
  - K= Frecuencia de cada categoría de cada variable nominal.
- Segunda etapa: Agrupación de datos
  - Cálculo del criterio (AIC o BIC) para cada número de clústers en el rango especificado y usarlo para descubrir el número inicial de grupos. Donde AIC se refiere al criterio de Akaike y BIC se refiere al criterio de Schwartz o inferencia bayesiana, donde sus fórmulas se explicitan en las ecuaciones 2.38 y 2.39.

- Refinamiento de la estimación inicial encontrando los mayores cambios de distancias entre dos grupos cercanos en cada etapa de la clusterización jerárquica.

- **Algoritmo EM**: Este algoritmo busca la distribución que cada instancia de entrenamiento tiene y los parámetros a estimar. Para ello, lo que hace este algoritmo es adivinar los cinco parámetros para calcular la probabilidad de pertenencia, luego con esa probabilidad de pertenencia intenta re-estimar los parámetros e iterar sucesivamente. Su formulación en esquema de pasos es [108]:

  - Cálculo de las probabilidades de los grupos, las cuales son los valores esperados del grupo.
  - Cálculo de los parámetos de distribución, lo cual es la maximización de la verosimilitud de las distribuciones del conjunto de datos entregado
  - Ajuste de las ecuaciones de estimación de parámetros para considerar las probabilidades de los grupos, dichas probabilidad actúan como pesos. Sea w_i la probabilidad de que una instancia i pertenezca al grupo A, entonces la media y la varianza para este grupo son:

$$\mu_A = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n} \tag{2.35}$$

$$\sigma_A^2 = \frac{w_1 (x_1 - \mu)^2}{w_1 + w_2 + ... + w_n} \tag{2.36}$$

  Luego el algoritmo converge hacia un punto fijo pero nunca llega a él, sin embargo, se puede observar cuán lejos queda la solución del punto ideal calculando la similitud total original del conjunto de datos, de la siguiente manera:

$$\prod_i p_A Pr\left[x_i|A\right] + p_B Pr\left[x_i|B\right] \tag{2.37}$$

Cabe señalar que las probabilidades dadas de los grupos A y B son determinadas desde la función de distribución normal $f(x; \mu, \sigma)$. Esta verosimilitud total es una medida de cuán bueno es el clúster y aumenta con cada iteración del algoritmo.

Algoritmos Jerárquicos

- **Clúster Jerárquico**: Este método de clusterización no requiere ningún parámetro, por ello, resulta de gran utilidad al momento de buscar una estructura de clúster en un conjunto de datos. Sin embargo, su resultado puede verse altamente influenciado por la presencia de valores fuera de rango. No obstante, se muestra la forma general de este algoritmo, la cual es [97]:

  - Entrada: Matriz de disimilaridad D^1.
  - Salida: El clúster jerárquico en el conjunto de objetos S, donde $|S| = n$;
  - Método: k=1
  - Iniciar clusterización: $\pi^1 = \alpha_S$

- **while** (π^k contenga más de un bloque **do**
- fusionar un par de dos grupos cercanos;
- sale un nuevo grupo;
- k++;
- calcular la matriz de disimilaridad D^k
- **endwhile**

Donde la matriz de disimilaridades es escogida mediante los criterios de cercanía explicitados anteriormente. Cabe destacar, que *"los algoritmos jerárquicos tienen un coste O(N2) donde N es el núumero de puntos de entrada, y, además realizan gran cantidad de accesos de entrada/salida. Se trata de un problema grave cuando se trabaja con grandes volúmenes de datos"* [82].

Número de grupos

El primer problema de un modelo de conglomerados es la definición del parámetro correspondiente al número de conglomerados. Para este problema existen muchas soluciones que se dividen en índices y criterios ya sea usando solamente la naturaleza de los datos o los juicios de un experto en el negocio. Los índices de validación de grupos no se utilizan en esta memoria, por lo que si se desea averiguar acerca de ellos se puede consultar a Milligan-Cooper en [75] quienes comparan 30 índices, o bien, unos índices más actuales se muestran por Thilagamani-Shanthi en [100]. Por otro lado, dentro de los múltiples criterios se describen 4 [94]: Validación cruzada, Estimación de verosimilitud penalizada, Criterio la rodilla en la curva de error, Criterio del codo en la curva F y Criterio jerárquico.

- **Validación cruzada**: Esta técnica generan modelos *"que intentan ajustar el conjunto de datos lo más preciso posible"* [94] mediante varias particiones. La idea general al aplicarla para clusterizar es *"dividir la muestra total en V particiones, o submuestras disjuntas aleatorias...luego el mismo tipo de análisis es ejecutado en forma iterativa a las observaciones de las V-1 particiones (denominado instancias de entrenamiento) y los resultados son aplicados a la muestra V para computar algún índice de validación predictiva...los resultados de la V iteraciones son promediadas para obtener una media singular de la estabilidad del modelo"* [80], en otras palabras, a partir de una muestra general se le divide en V partes, posteriormente se utiliza las V-1 particiones para ingresarlas como entradas para el aprendizaje del modelo, luego en la partición restante prueba una configuración determinada del mismo. En un paso siguiente se itera el algoritmo para todas las particiones generadas en un principio y se calcula una medida agregada para observar la validación del modelo frente a nuevas observaciones.

- **Estimación de verosimilitud penalizada**: Como la técnica anterior intenta ajustar el modelo a un conjunto de datos de la forma más acertada posible, pero como agregado *"minimiza la complejidad del modelo"* [94]. Sus principales criterios de corte (es decir, de decisión de la cantidad de clústers) son: BIC, AIC, SIC, entre otros. De los cuales se procede a describir el

AIC (*Akaike Information Criterion*) y el BIC (*Bayesian Inference Criterion*) en las ecuaciones 2.38 y 2.39 respectivamente [110]:

$$AIC(K) = \frac{-2(N - 1 - N_k - \frac{K}{2})l(\hat{\theta})}{N} + 3N_p \tag{2.38}$$

$$BIC(K) = l(\hat{\theta}) - \left(\frac{N_p}{2}\right) log(N) \tag{2.39}$$

Donde los símbolos se traducen como:

- N: Total de patrones
- N_k: Número de parámetros para cada clúster
- N_p: Número de parámetros estimados
- $l(\hat{\theta})$: Máxima log-verosimilitud
- K: Selección de cantidad de clústers con el valor mínimo de AIC o BIC.

- **Criterio la rodilla en la curva de error**: Es un criterio *"definido como el punto de máxima curvatura"* [94], el cual se puede obtener mediante los siguientes métodos en un grafo de grupos versus métrica de evaluación [94]:
  - Máxima diferencia de magnitudes entre dos puntos.
  - Máxima diferencia radial entre dos puntos.
  - El punto en la curva que está más alejado de una línea ajustada a la curva completa.

- **Criterio del codo en la curva F** [84]: Para este criterio se necesita realizar el gráfico Información F versus Número de clústers. Con él se opta por aquel punto de inflexión que se asemeja de aun codo, en términos matemáticos se puede aludir a aquel punto en que la derivada de función intrínseca al gráfico cambia de signo.

- **Criterio Jerárquico** [84]: Este criterio usa el modelo particular de clusterización jerárquica, donde se pueda observar el grafo propio denominado dendrograma y se efectúa un corte que mejor acomode los grupos, como en el caso de la figura 2.14:

  En esta figura se puede optar por cualquier distancia para hacer el corte definitivo, puesto que también es a criterio.

Como criterio adicional está el caso de que el personal relacionado con el negocio conozca gran parte del mercado y plantee una heurística en base a su propio juicio. No obstante, este último criterio se puede combinar de manera efectiva con cualquiera de los otros criterios, en donde concretamente se puede "negociar" la cantidad de grupos a considerar, aunque en este último caso se converge a una segmentación.

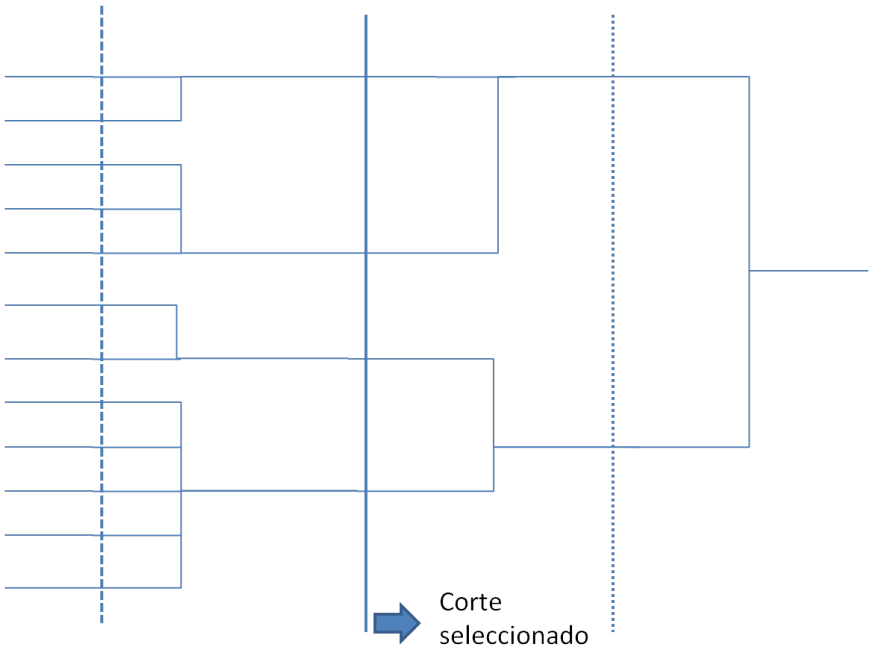

Figura 2.14: Gráfico sintetizado análisis jerárquico

Etiquetación de clústers

Dado que este problema de segmentación no tiene una clasificación a priori (lo que lo hace un problema sin una variable obtenida a comprar, es decir, un problema no supervisado), se requiere la necesidad de etiquetar los grupos obtenidos considerando las características del negocio que han sido incorporadas en el análisis. Para poder asignar un nombre a algún grupo se requiere de algún tipo de referencia, es así, como se utilizará un mapa de diagnóstico de lealtad (MADIL) que usualmente se usa para evaluar la lealtad del cliente en base a cuestionarios estructurados, no obstante, la minería de datos por sí sola no puede entregar un resultado similar a un cuestionario, el procedimiento del KDD (Knowledge Discovery on Databases) sí puede, debido a que existe la evaluación de un experto detrás del analista, es por ello, que se puede considerar de forma análoga.

El mapa de lealtad MADIL muestra dos dimensiones: El encanto del cliente con el producto y el Compromiso a Futuro con el este productos, bosquejados de la siguiente manera [109]:

De esta matriz se puede visualizar claramente la etiqueta apropiada para cada conglomerado basándose en la perspectiva que tiene cada variable respecto al compromiso y al encanto. Por ejemplo, la variable facturación, está asociada al compromiso, debido a que mientras más factura, significa que el producto realmente soluciona su problemática total o parcialmente, en cambio, otra variable como la competencia o los reclamos tienen asociada la arista del encanto con el producto.

Por lo anterior, se señalan unos perfiles genéricos de los nombres mencionados en la matriz [109]:

- **Los Apóstoles**: Son los incondicionales de la empresa. Son aquellos que harán el boca a boca acerca de la empresa, en sus círculos cercanos.

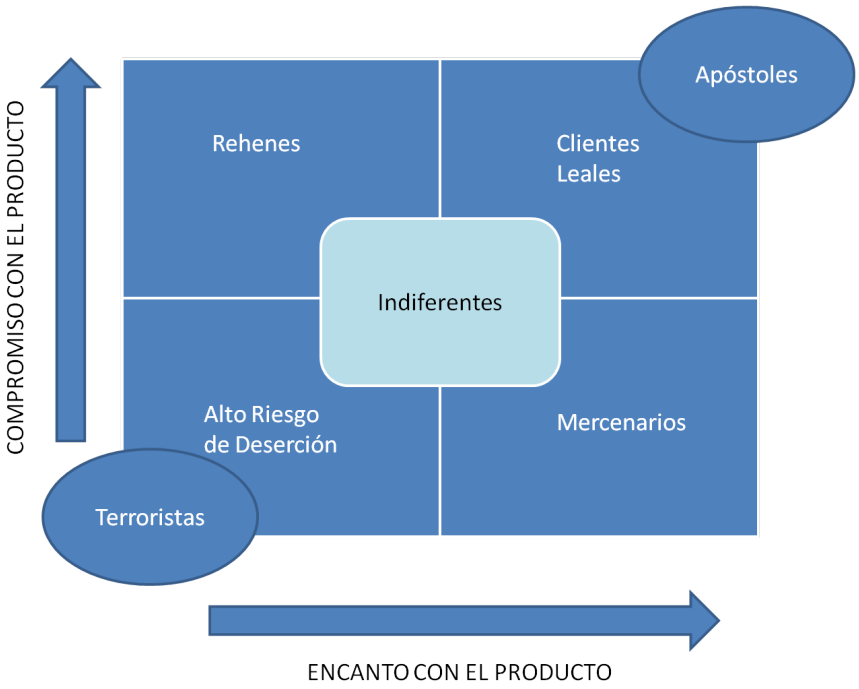

Figura 2.15: Matriz de clasificación MADIL

- **Los Clientes Leales**: Iguales a los anteriores pero con menor intensidad.
- **Los Terroristas**: Frecuentemente han tenido una o varias experiencias malas con los productos de la empresa, a diferencia de los apóstoles, este tipo de clientes trata de destruir la imagen de la empresa.
- Potenciales desertores: Similares a los terroristas pero con menor intensidad, en ocasiones no muestran el patrón de destruir la imagen.
- **Indiferentes**: Son aquellos que no muestran señales de que el producto de la compañía este bien o mal.
- **Rehenes**: Son los que a pesar de no estar contentos con el producto permanecen con él. Dentro de las razones por las cuales lo hacen es por su alto costo de salida, o también, por el hecho de que el mercado es semimonopólico.
- **Los Mercenarios**: Estos clientes usualmente consideran el producto en cuestión como un commodity, es decir, el precio dicta su comportamiento. No obstante, existen distintos tipos de mercenarios:
- **Los Switchers**: Aquellos que tienen pocas marcas favoritas y brincan entre éstas según esté o no en promoción.
- **Los Negociadores**: Cambian regularmente dentro de un abanico más amplio de marcas aceptables para ellos.

- **Los Sensibles al precio**: Sistemáticamente compran la marca del menor precio, sin importar cuál sea ésta.

Cabe destacar que esta clasificación está fundamentada en el artículo de Jones y Sasser expuesto en [50].

2.7.4. Modelo Recency, Frequency, Mount

El modelo Recency, Frequency, Mount (RFM) data antes del año 2000 en el marketing cualitativo, como forma de medir el comportamiento del consumidor. Esta medición se hace desde tres perspectivas [51]. La primera es *Recency*, que *"indica hace cuánto el cliente respondió"* [51], que en otras palabras, significa el tiempo transcurrido desde la última vez que el cliente registró un accionar con la compañía. La segunda perspectiva es la Frecuencia, que *"provee una métrica de cuán seguido, el cliente ha respondido a recibir mails"* [51], que pretende indicar, el tiempo transcurrido entre interacciones del cliente con la compañía. Finalmente, la tercera perspectiva es el valor monetario, que *"mide el monto en dinero o el número de productos que el cliente ha gastado o consumido en respuesta a los mails enviados"* [51], lo que expresado de forma distinta, indica el valor monetario o cantidad que el cliente gasta, emplea o consume en cada accionar con la compañía. De esta manera se puede formular, para el caso de los reclamos en las telecomunicaciones las siguientes variables:

- Recency (R): La última vez o mes que el cliente emprendió un reclamo hacia la compañía
- Frecuencia (F): El número de meses en las que el cliente reclamó.
- Monto (M): El número de reclamos promedio involucrado en cada ocasión.

Generalmente M va asociado a un valor monetario, no obstante, acorde con J.-J. Jonker en [51], también se puede ocupar como una variable de monto de acciones. Estas variables se pueden combinar en cuocientes para determinar KPI's útiles para la compañía, mas no poseen una fórmula general aplicable, por lo que el concepto trasciende dependiendo de los contextos, por ende, se puede implementar este modelo ya sea hablando en términos de compra, de reclamos, de consumo, etc.

2.8. Problema de rareza o desbalanceo

De vez en cuando, dependiendo del tipo de mercado, aparece este problema expresado en la base de datos respecto a las clases, es decir, a las categorías o valores de la variable objetivo. Esta rareza de clases o desbalanceo de clases se da cuando existe escasez de una de las clases, esta escasez puede ser de dos tipos [107]:

- Rareza de clases: se define como la ocasión en que un valor de la variable objetivo se encuentra fuera del común denominador o en los extremos de la distribución de la variable objetivo. En el caso de que la variable objetivo sea nominal, se refiere a aquella categoría con una frecuencia despreciable respecto a las otras categorías.

- Rareza de casos: el segundo tipo *corresponde a un conjunto de datos significativo pero a su vez pequeño"* [107], en otras palabras, son aquellas instancias que escapan al común, en cuanto a su comportamiento, por ejemplo, si se tienen una población que tiene 1.5 mts de estatura y se desea establecer una clasificación de personas altas y bajas y se encuentra una pequeña subpoblación de personas cuya altura promedio es de 2 mts, entonces esta subpoblación entra a ser un problema de rareza de casos. Generalmente son *definidos por el dominio y compartirán características comunes"* [107]

Ambos tipos de escasez son consideradas una desbalanceo interno de las bases de datos.

Además, de la clasificación anterior, se puede notar el desbalanceo de datos respecto a si la rareza es relativa o absoluta [107], donde la última se define como el caso en el que *el número de instancias asociadas a la clase rara es pequeño en un sentido absoluto"* [107]. Sin embargo, en esta clasificación no se posee una regla tangible que permita estandarizar las instancias o valores que pertenecen a la rareza. Por ejemplo, en el caso de churn predictivo mostrado en esta memoria, se tienen un churn del 1 % respecto a la base general(o sea, es una clase rara absoluta), no obstante, no puede etiquetarse la categoría de fuga efectiva como rareza, debido a que es el comportamiento que se desea estudiar, ni tampoco se pueden dejar de lados los casos que se comportan similarmente dentro de ese 1 %(en esta última ocasión se hablaría de rareza de casos).

Este tipo de problema conlleva a dificultar la labor de la implementación del KDD, debido a que existen consecuencias asociadas a la ignorancia de dicha problemática, entre las cuales destacan [21, 107]:

- Métricas de evaluación inapropiadas: En ocasiones, la métrica que ayuda a construir el modelo se basa en obtener una certeza adecuada, sin embargo, en el caso de que existe una rareza de clases del 1 %, estos modelos se construirán para obtener el 99 % de certeza, dejando de lado la clase rara que puede ser aquella de interés para el proyecto. En otras palabras, esto indica que *las clases raras tienen menor impacto en el accuracy (o certeza) que las clases comunes"* [107].

- Escasez de datos (Rareza absoluta y relativa): Esta consecuencia se da en las bases de datos en donde la cantidad de instancias que pertenecen a la clase rara es mucho menor con respecto al resto de las clases. No obstante, esta consecuencia también aplica para la rareza de casos, puesto que las instancias raras tienen una tasa de error de clasificación mucho mayor que los casos comunes [107], donde este problema degenera en otra consecuencia que afecta al aprendizaje del algoritmo, siendo este último denominado *problem with small disjuncts* [107], que se refiere a que el algoritmo no puede distinguir bien el patrón en el caso de las instancias raras o que pertenecen a una clase rara. Si bien, el abordamiento usual que se efectúa sobre este problema es la eliminación de registros o clases, *el efecto neto de la eliminación total de los pequeños conjuntos disjuntos es difícil de predecir, porque depende del destino de las instancias emancipadas , instancias que fueron clasificadas por los conjuntos disjuntos que han sido eliminados"* [42]. Un punto relevante es que los resultados en [42] indican que la eliminación de los conjuntos disjuntos (traducción de *small disjuncts*) *resulta en un incremento de la tasa de error total y, por ello, no es una buena estrategia"* [107]. Respecto a la rareza relativa, es cuando las instancias son raras respecto a otras, generalmente agregan ruido innecesario, que conllevan a resultados erróneos cuya causa es difícil de encontrar.

- Fragmentación del conjunto de datos: Cuando los algoritmos de minería de datos, se da esta consecuencia, el cual, *es un problema porque las regularidades pueden ser solamente encontradas en particiones individuales que contienen menos datos"* [107], esto quiere decir que los patrones finales terminan bajo la influencia de los patrones internos de cada partición.

- Tendencia inducida: En la minería de datos para comprender el patrón general subyacente en el problema, se tiende a inducir tendencias, de hecho, *muchos algoritmos de aprendizaje usan una tendencia general de manera de encontrar la generalización y evitar el sobreajuste, por lo tanto, la tendencia puede impactar la habilidad de aprender de casos o clases raras"* [21].

- Ruido: El ruido, cuando es consecuencia de rareza, *tiene un mayor impacto sobre los casos raros que sobre los casos comunes, puesto que los casos raros tienen menos instancias para empezar, por lo tanto, requerirán menos ejemplos ruidosos para impactar el subconcepto aprendido"* [107]. Esto quiere decir que si se tiene ruido en las clases raras, el patrón aprendido será mucho más difuso e inexacto respecto al patrón real.

No obstante, [107] facilita los métodos usuales para manejar dicha rareza, los cuales se expresan en la siguiente tabla:

Cuadro 2.1: Métodos para manejar rarezas

Métodos para manejar rareza	Descripción
Métricas de evaluación apropiadas	AUC, o alguna que no se vea influenciada por el corte
Técnicas de búsqueda no ambiciosa	No árboles, mejor reglas de asociación
Interacción conocimiento humano	Puede proveer una explicación a la rareza
Aprender solo de la clase rara	Considerar solo las instancias de la clase rara para predecirla
Segmentar los datos	Permite reducir las rarezas o agruparlas, identificándolas
Aprendizaje sensitivo al costo	Asignando un mayor costo a las rarezas se puede instigar el aprendizaje
Muestreo	Submuestrear los datos comunes y sobremuestrear las rarezas
Otro métodos	Boosting, Inducción en dos fases

2.9. Medidas de Evaluación

Una vez escogido el software, y el modelamiento acorde al KDD, se sugiere medir la efectividad del mismo a través de métricas de evaluación. Sin embargo, previo a su definición se debe enmarcar el concepto de la matriz de confusión que es la que *"contiene información acerca de las clasificaciones actuales y las predichas, realizadas por un sistema de clasificación"* [44]. Un ejemplo del esquema de esta matriz se muestra a continuación:

En base a este ejemplo, que describe una matriz de confusión para un problema de clasificación binaria, se definen las siguientes medidas:

$$TasaFP = \frac{FP}{N} \tag{2.40}$$

Cuadro 2.2: Ejemplo de tabla de confusión

		Clase Actual	
Categorías		0	1
Clase hipotética	0	TN	FN
	1	FP	TP
Columnas Totales		N=FP+TN	P=TP+FN

$$TasaTP = \frac{TP}{P} \tag{2.41}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.42}$$

$$Recall = \frac{TP}{P} \tag{2.43}$$

A continuación se procede a definir medidas más avanzadas que las anteriores, las cuales representan de forma más realista los resultados obtenidos en la etapa de modelamiento. Estas se muestran a continuación [32]:

- **Accuracy**: Medición que se refiere al nivel de certeza del modelo dentro del universo total del problema. Su fórmula es:

$$Accuracy = \frac{TP + TN}{P + N} \tag{2.44}$$

- **F-Measure**: Esta medida es una media geométrica entre dos cocientes relativos a la precisión y al recall. Sirve para obtener una medida más cercana a la realidad en cuanto a la certeza del modelo. La fórmula que la describe, se muestra a continuación:

$$F - Measure = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{2.45}$$

- **Curvas Roc**: Son curvas que muestran la habilidad del clasificador para posicionar las instancias verdaderas respecto a las falsas [32]. En una definición más acertada se puede decir que las curvas ROC son las que miden la relación de la tasa de verdaderos positivos (predicciones acertadas) versus la tasa de falsos positivos (predicciones erradas). Siendo el *positive* el referente a la clase de fuga cuando se trata de un problema de clasificación binario. Estas curvas no tienen una fórmula asociada. No obstante, sí tienen una métrica, la cual es el *Area Under the Curve*(AUC), que se define como el rea bajo la curva ROC, además, tiene la siguiente propiedad de la estadística: *"La AUC de un clasificador es equivalente a la probabilidad que el clasificador posicionará una instancia aleatoria positiva más alto que una instancia aleatoria negativa"* [32].

Algorithm 1 Conceptual method for calculating an ROC curve. See algorithm 2 for a practical method.

Inputs: L, the set of test instances; $f(i)$, the probabilistic classifier's estimate that instance i is positive; min and max, the smallest and largest values returned by f; $increment$, the smallest difference between any two f values.

```
for t = min to max by increment do
  FP ← 0
  TP ← 0
  for i ∈ L do
    if f(i) ≥ t then                    /* This example is over threshold */
      if i is a positive example then
        TP ← TP + 1
      else      /* i is a negative example, so this is a false positive */
        FP ← FP + 1
      end if
    end if
  end for
  Add point (FP/N, TP/P) to ROC curve
end for
end
```

Figura 2.16: Algoritmo Conceptual de las Curvas ROC

Un punto relevante a destacar es el hecho de que las curvas ROC *"no dependen de un parámetro de corte, por ello es mejor métrica de evaluación que la accuracy"* [21].

- **Lift**: Es una medida que *"compara la precisión con la tasa de churn total en la base de prueba"* [21]. Su fórmula respectiva es:

$$Lift = \frac{Precision}{\frac{P}{P+N}} \tag{2.46}$$

2.10. Algoritmos de minería de datos

2.10.1. Descripción principales algoritmos

Dentro del procedimiento KDD, la etapa que más ha destacado en este último tiempo es la de minería de datos, tanto así que se ha llegado a confundir con la aplicación del KDD. Esto se debe principalmente a que esta etapa ofrece el diferenciamiento o valor agregado a los procedimientos estándar ya existentes en la mayoría de las empresas, producto de la generación de reportes para la facilitación de la gestión del negocio en el que se encuentren inmersas. Es por ello, que los algoritmos de minería de datos entran a ser la parte más relevante del KDD para el mercado. A continuación se describen los principales algoritmos usados en la actualidad, gran parte de ellos ya vienen incluídos en los paquetes estadísticos comerciales dedicados al descubrimiento de patrones desconocidos:

K-Nearest Neighboor (KNN)

El algoritmo del K vecino más cercano o KNN es uno de los más simples. Este algoritmo no requiere de ningún parámetro fuera del número de vecinos a considerar. En pocas palabras, el algoritmo puede resumirse en que *"reúne los K vecinos más cercanos y los hace votar, la clase con más vecinos gana, ..., mientras más vecinos consideramos, menor la tasa de error"* [49]. Dicha cercanía ,generalmente se mide en base a alguna distancia, por lo que se pueden obtener distinto resultados dependiendo de la distancia escogida, pues *"diferentes métricas definirán diferentes regiones"* [55]. Ahora bien, su formulación matemática, se describe de la siguiente forma:

Sea $V = \{v_1, ..., v_v\}$ un conjunto etiquetado de referencia, el cual se desea usar como variable a predecir, y también denominado prototipo, donde un prototipo es *"un elemento representativo de una clase"* [55]. Si los prototipos son etiquetados en las c clases, es decir, para cada v_i conocemos su etiqueta l(v_i). Entonces para clasificar una instancia x, los k-prototipos más cercanos son llamados a votar. El resultado de dicha votación es la clase que se asignará a la instancia x.

Cuando se trabaja con un problema de clasificación, específicamente binaria, el modelo requiere de dos parámetros, K y p, donde K es *"el tamaño del vecindario"* [34], o la cantidad de vecinos a considerar y p es *la probabilidad de corte para elección"* [34]. Sin embargo, en el caso de que la clasificación no sea binaria, la probabilidad de elección de una instancia se puede expresar como:

$$p\left(x|w_i\right) \approx \frac{k_i}{N_i V_R} \tag{2.47}$$

Donde R es la región que contiene exactamente k elementos del conjunto de referencia V, k_i es el número de elementos en R de la clase w_i y N_i es la cantidad de elementos de la clase w_i en el conjunto original de datos. De esta forma la probabilidad posterior se obtiene:

$$P\left(w_i|x\right) \approx \frac{k_i}{k} \tag{2.48}$$

Se debe destacar que para la formulación anterior, se asumió la distancia euclídea y que para el caso de tener variables nominales, se debe trabajar con similaridades en vez de distancias.

El algoritmo previamente descrito se puede resumir como sigue [47]:

- Escoger la función de distancia en las instancias
- Tener en cuenta el conjunto de datos de entrenamiento consistente en los pares (a_1, c_1),....(a_N, c_N), donde a son las instancias y c corresponde a la variable que se desea predecir.
- Por lo que al clasificar una nueva instancia a se ejecutan los siguientes pasos:
  - Sea (a_{j1}, c_{j1}),....(a_{jK}, c_{jK}) las K instancias, cuyos atributos se encuentran cercanos a "a".
  - Etiquetar "a" con la clase que más aparezca dentro de los $c_{j1}, ..., c_{jK}$ vecinos.

Como se ha podido apreciar, existen tres formas de cambiar el KNN [55]:

1) Escoger distintas distancias.
2) Elegir diferentes valores para la cantidad de vecinos a considerar.
3) Tomar una parte del conjunto original como prototipos.

Naive Bayes

En general los algoritmos de clasificación que utilizan el aprendizaje bayesiano resultan complejos en el sentido de la cantidad de parámetros. Sin embargo, el método de naive bayes convierte dicha complejidad en una simpleza factible, *"debido a que hace un supuesto de independencia condicional que reduce el número de parámetros a estimar, cuando se modela P(x|y)"* [76]. De forma cuantitativa, si la variable a predecir tiene dos valores pasa de estimar $2\left(2^n - 1\right)$ parámetros a $2n$. La utilidad de los algoritmos de aprendizaje bayesiano es que *"da una medida probabilística de la importancia de esas variables en el problema, y, por lo tanto, una probabilidad explícita de las hipótesis que se formulan"* [68].

La independencia condicional se refiere a que *"Dadas X, Y y Z variables aleatorias, sea X condicionalmente independiente de Y dado Z, si y sólo si la distribución de probabilidad que rige X es independiente del valor de Y dado Z"* [76], es decir:

$$\left(\forall i, j, k\right), P\left(X = x_i | Y = y_j, Z = z_k\right) = P\left(X = x_i | Z = z_k\right) \tag{2.49}$$

Los algoritmos más reconocidos, que usan aprendizaje bayesiano, son las redes bayesianas, que poseen alta complejidad en su desarrollo pero son adaptables a una gran variedad de problemas, por lo que es *"bastante costoso y usualmente se aproxima por métodos como cadenas de Markov o Montecarlo...aún así poseen una convergencia difícil de diagnosticar, ya sea por la ocurrencia de esta o por la validez de la misma"* [66]. Cabe destacar que los modelos de naive bayes *"pueden ser vistos como redes bayesianas en las cuales cada instancia X_i tiene una variable a predecir C, como el único padre, que no tiene padres"* [66]

Naive Bayes, dentro de los algoritmos de aprendizaje bayesiano se llaman "naive"(ingenuo) por el *"supuesto que todas las variables X_i son mutuamente independientes condicionalmente dada una variable especial C"* [66]. Por ejemplo, en un problema de dos dimensiones, con 2 atributos y una variable a predecir, se tiene que [76]:

$$P\left(X|Y\right) = P\left(X_1, X_2|Y\right) = P\left(X_1|X_2, Y\right) P\left(X_2|Y\right) \tag{2.50}$$

Lo que gracias al supuesto mencionado anteriormente se simplifica a:

$$P\left(X|Y\right) = P\left(X_1|Y\right) P\left(X_2|Y\right) \tag{2.51}$$

De esta manera, la probabilidad para múltiples atributos puede expresarse de la siguiente forma:

$$P\left(X_1, ..., X_n|Y\right) = \prod_{i=1}^{N} P\left(X_i|Y\right) \tag{2.52}$$

Una vez aclarado lo anterior, la explicación del algoritmo naive bayes se simplifica: sea Y la variable a predecir y X_i los atributos considerados para el entrenamiento del modelo, entonces, la probabilidad para tomar el k-ésimo valor es, bajo la regla bayesiana:

$$P\left(Y = y_k | X_1 ... X_n\right) = \frac{P\left(Y = y_k\right) P\left(X_1 ... X_n | Y = y_k\right)}{\sum_j P\left(Y = y_j\right) P\left(X_1 ... X_n | Y = y_j\right)} \tag{2.53}$$

Que al aplicar el supuesto de Naive Bayes, queda:

$$P\left(Y = y_k | X_1 ... X_n\right) = \frac{P\left(Y = y_k\right) \prod_i P\left(X_i | Y = y_k\right)}{\sum_j P\left(Y = y_j\right) \prod_i P\left(X_i | Y = y_k\right)} \tag{2.54}$$

Entonces, sea X una nueva instancia, y tanto P(Y) como $P\left(X_{ij}|Y\right)$ dadas producto de la estimación del conjunto de datos de entrenamiento. El valor más probable de Y con Naive Bayes será:

$$Y \leftarrow argmaxy_k \frac{P\left(Y = y_k\right) \prod_i P\left(X_i | Y = y_k\right)}{\sum_j P\left(Y = y_j\right) \prod_i P\left(X_i | Y = y_k\right)} \tag{2.55}$$

No obstante, la estimaciones internas que hace Naive Bayes de los parámetros que no se le entregan, cambian según la naturaleza de la variable, es decir, si es continua o discreta. Para el caso discreto, son tres los parámetros a estimar. Suponiendo n atributos X_i donde cada uno tiene j valores discretos e Y es la variable a predecir tomando K valores posibles, el primer parámetro es [76]:

$$\theta_{ijk} \equiv P\left(X_i = x_{ij} | Y = y_k\right) \tag{2.56}$$

Para cada atributo X_i, de los cuales pueden tener x_{ij} valores posibles e y_k valores de Y. El segundo parámetro viene dado por la probabilidad a priori de Y:

$$\pi_k \equiv P\left(Y = Y_k\right) \tag{2.57}$$

Estos dos parámetros son los fundamentales, sus fórmulas de estimación se presentan a continuación:

$$\hat{\theta_{ijk}} = \frac{\#D\left\{X_i = x_{ij} \wedge Y = y_k\right\} + l}{\#D\left\{Y = y_k\right\} + lJ} \tag{2.58}$$

En la ecuación anterior, J es el número de valores distintos que X_i y l es un factor para suavizar la estimación, esta suavización se conoce como la suavización de Laplace y l determina la fuerza de la suavización. El segundo parámetro, se puede estimar de múltiples formas, en particular se puede usar el estimador de máxima verosimilitud:

$$\hat{\pi_k} = \frac{\#D\left\{Y = y_k\right\}}{|D|} \tag{2.59}$$

Donde $|D|$ es el número de elementos del conjunto de datos de entrenamiento D. Sin embargo, para el caso de los valores datos o variables continuas, naive bayes puede utilizar la ecuación 2.52, así como también, pueden escogerse otras formas de representar P($X_i|Y$) en tal caso se deben estimar la media y la varianza de la distribución, usando el estimador de máxima verosimilitud. De esta manera, la media se estima:

$$\hat{\mu_{ik}} = \frac{\sum_j X_i^j \delta\left(Y^j = y_k\right)}{\sum_j \delta\left(Y^j = y_k\right)} \tag{2.60}$$

Y la varianza, la cual es estimada con el método de mínima varianza sin tendencia(MVUE):

$$\hat{\sigma_{ik}}^2 = \frac{\sum_j \left(X_i^j - \hat{\mu_{ik}}\right)^2 \delta\left(Y^j = y_k\right)}{\sum_j \delta\left(Y^j = y_k\right) - 1} \tag{2.61}$$

Es así como, finalmente, se muestra el algoritmo Naive Bayes en su formato secuencial:

Table 1. The NBE learning algorithm.

INPUT: training set T, hold-out set H, initial number of components k_0, and convergence thresholds δ_{EM} and δ_{Add}.

Initialize M with one component.
$k \leftarrow k_0$
repeat
  Add k new mixture components to M, initialized using k random examples from T.
  Remove the k initialization examples from T.
  repeat
    E-step: Fractionally assign examples in T to mixture components, using M.
    M-step: Compute maximum likelihood parameters for M, using the filled-in data.
    If $\log P(H|M)$ is best so far, save M in M_{best}.
    Every 5 cycles, prune low-weight components of M.
  until $\log P(H|M)$ fails to improve by ratio δ_{EM}.
  $M \leftarrow M_{best}$
  Prune low weight components of M.
  $k \leftarrow 2k$
until $\log P(H|M)$ fails to improve by ratio δ_{Add}.
Execute E-step and M-step twice more on M_{best}, using examples from both H and T.
Return M_{best}.

Figura 2.17: Algoritmo de Naive Bayes

Aparte de Naive Bayes, existen otros clasificadores Bayesianos, los cuales están implementados en el software de Weka, así como también vienen incorporados en la última versión del software Rapidminer como extensión del mismo o upgrade. He aquí un breve descripción de cada uno de ellos [108]:

1) **NaiveBayesSimple**: Usa una distribución normal para modelar los atributos numéricos.

2) **NaivesBayes (kernel)**: Esta variación permite usar estimadores de densidad kernel, los cuales

permiten relajar el supuesto de naive bayes. También acepta atributos numéricos, no obstante, los discretiza.

3) **NaiveBayesUpdateable**: es una versión que procesa una instancia por iteración, puede usar estimadores kernel pero no discretiza las variables numéricas

4) **AODE**: promedia sobre un espacio de modelos bayesianos que tenga supuestos de independencia más débiles que naive bayes.

5) **BayesNet**: Es una red bayesiana bajo los supuestos de que solamente se ingresen atributos nominales(en caso de que se tenga atributos continuos se deben discretizar previamente) y sin valores perdidos.

Árboles de Decisión

Los árboles de decisión son modelos que usualmente se representan en forma de grafos. Es *"un modelo predictivo que puede ser usado para representar tanto modelos regresivos como aquellos de clasificación...se refiere a un modelo jerárquico de decisiones y sus consecuencias"* [88]. Dentro de un esquema general, el árbol de decisión consiste en un grafo donde existe un nodo único o parental, el cual, contiene las instancias a contemplar en el modelo.

Posteriormente, mediante un test o prueba sobre los atributos se seleccionan las mejores divisiones para dichas instancias, lo que provoca que el nodo parental se separe en uno o más nodos hijos, lo cuales se siguen subdividiendo tal como se efectuó con el nodo parental. En el caso de que un nodo hijo ya no se pueda separar, es decir, en el caso donde todas las instancias que contenga dicho nodo hijo pertenecen a una misma clase, o bien, el test estadístico o prueba deja de ser significativo, se denomina nodo Hoja o terminal. En otras palabras, *"cada nodo interno (o nodo hijo) es un atributo, cada rama corresponde a un valor del atributo"* [59] y *"las hojas representan la respuesta, la clase o la propiedad buscada sobre el dato"* [87] donde la "clase" se refiere a una de las categorías o valores de la variable a predecir en caso de que sea categórica. Gráficamente:

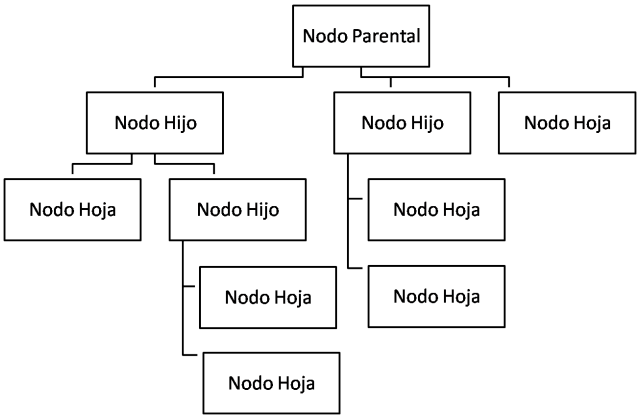

Figura 2.18: Grafo Árbol de Decisión Estándar

Ahora bien la division o *split* de un nodo es *"una variable más una lista de condiciones sobre dicha variable"* [87], por ejemplo, Variable: Clima, Condiciones: Clima=Lluvioso, Clima=Despejado,

Clima= Nubosidad Parcial. Es decir, *"un split significa que una parte separada del conjunto de datos (del nodo parental) es tomada para cada nodo hijo"* [55]. Estos *splits* se efectúan en base a criterios, donde en su forma más simple, cada criterio alude a los valores de un atributo en particular, lo cual, para variables del tipo numérico o continuo suele asociarse a rangos [88]. Generalmente, un nodo interno o hijo, *"se divide acorde al valor de un atributo singular"* [88] esto quiere decir que son univariantes. Existe una gran variedad de criterios en base a este tipo de divisiones, las cuales se pueden sintetizar en dos tipos de criterios [88]:

- Origen de la medida: Este tipo incluye lo referente a la teoría de cantidad de información, dependencia entre variables y distancia.
- Estructura de la medida: Es el tipo de criterio que hace alusión a la pureza de la variable, respecto a su distribución propiamente tal.

Los criterios más usados en el modelo árbol de decisión son aquellos en base a la estructura de la medida. Sin embargo, previo a ahondar en este tipo, suele asociarse la división o *split* de los árboles de decisión, a una bifurcación, en donde el nodo se separa en dos. No obstante, esto no siempre es así, lo anterior, es recomendado debido al costo computacional y a la interpretación posterior del modelo. En el caso de que se tenga variables con múltiples categorías se toma en consideración cada valor de dicha variable como una categoría a separar, así como también se puede categorizar por rangos en el caso de las variables continuas.

Tipos de Splits

Criterio basado en Impurezas Antes de describir los distintos criterios que existen, se definirá el término de medida impureza, la cual, se presenta como *"dada una variable aleatoria x con k valores discretos, distribuidos como* $P = (p_1, p_2, ..., p_k,)$*, la impureza es una función* $\phi : [0, 1]^k \rightarrow R$ *que satisface las siguientes condiciones"* [88]:

- $\phi(P) \geq 0$.
- $\phi(P)$ es mínimo si $\exists$ un componente $p_i = 1$.
- $\phi(P)$ es máximo si $\forall i$, 1¡i¡k, $p_i = \frac{1}{k}$.
- $\phi(P)$ es simétrica respecto a las componentes de P.
- $\phi(P)$ es diferenciable en cualquier punto.

En otras palabras, si el vector de probabilidades contiene un componente de 1, entonces dicha variable se define como pura, puesto que el valor a entregar será uno solamente, en cambio, si tiene las componentes son iguales entre sí, lo que equivale a decir que el nivel de impureza es máximo, debido a que no se podrá decidir entre algún valor sin perder altos márgenes de información de la variable.

Luego, para un conjunto de datos S, el vector de probabilidades de una variable "y", es [81]:

$$P_y(S) = \left(\frac{|frec\left(y = c_j, S\right)|}{|S|}, ..., \frac{|frec\left(y = c \in domy, S\right)|}{|S|} \right) \tag{2.62}$$

Donde $frec\left(y = c_j, S\right)$ se refiere a la frecuencia de la categoría c_j en la variable "y", $|S|$ es la cardinalidad del conjunto de datos considerado, por ende, $P_y(S)$ no es más que *"la probabilidad de ocurrencia de la clase x en el nodo n"* [59], donde el nodo n es la partición S considerada. Ahora bien, la bondad del *split* puede ser conceptualizada como una función que minimiza la impureza de una variable "y", tal y como se describe a continuación [88]:

$$\Delta\Phi(ai, S) = \phi\left(P_y(S)\right) - \sum_{j=1}^{|dom(y)|} \frac{|frec\left(y = c_j, S\right)|}{|S|} \phi\left(P_y(S)\right). \tag{2.63}$$

Como se habrá observado *"lo ideal es encontrar los splits que generen nuevas ramas cuyos nodos sean lo más limpios posibles, es decir, los nodos con menor impureza"* [59]. Por ello, en base a esta última función se generan los distintos criterios existentes que buscan la mínima impureza. Entre los principales se encuentran [81, 87, 88]:

- **Ganancia de Información**:este criterio utiliza la función de entropía como medida de impureza (que en este contexto es la función $\phi()$), esta función se define como:

$$Entropía = \sum_{c_j \in dom(y)} -\frac{|frec\left(y = c_j, S\right)|}{|S|} log_2 \left(\frac{|frec\left(y = c_j, S\right)|}{|S|} \right) \tag{2.64}$$

$$Entropía = \sum_{x \in X} -P(n, x)\, log_2 P(n, x) \tag{2.65}$$

  La entropía, dentro de este tema, es una función que *"mide la impureza de los datos S promedio necesaria para codificar las clases de los datos en el nodo"* [59]. Interpretado de otra forma, mide la complejidad o el caos que posee la variable en cuanto a la cantidad de categorías que tiene, en este sentido una variable continua resulta altamente caótica debido a su multiplicidad de valores.

  Usando esta función, la ganancia de información se mide por la siguiente fórmula:

$$GananciadeInformación(y, S) = Entropía(y, S) - sum_{c_j \in dom(y)} \frac{|frec\left(y = c_j, S\right)|}{|S|} Entropía(y, S) \tag{2.66}$$

- **Índice de Gini**: Cuando los criterios referentes a la teoría de la información son tajantes en los resultados entregados por el modelo de árbol escogido, se puede recurrir al índice de Gini, el cual presenta una perspectiva distinta, basado en la impureza, siendo definido como *"un criterio que mide la divergencia entre las distribuciones de probabilidad de los valores de los atributos objetivos"* [88], lo que en palabras más sencillas se traduce en un nodo como *"la suma de los cuadradados de la proporciones de las clases"* [27]. Otra concepción del mismo

índice es como aquel que *"mide el grado de impureza del nodo en cuestión con respecto a las clases"* [45]. La fórmula general del índice es [59, 87]:

$$Gini\,(n) = 1 - \sum_{x \in X} P\,(n, x)^2 \tag{2.67}$$

Donde P(n,x) es la probabilidad de ocurrencia de la clase x en el nodo n, Gini(n) es la probabilidad de no sacar dos registros de la misma clase del mismo nodo. No obstante, el *split* o división de nodos se efectúa en base al GiniSplit, el cual, dado un Split S=$s_1, ..., s_k$ del nodo n

$$GiniS\,plit\,(n, S\,) = \sum_{s \in S} \frac{|s|}{|n|} Gini\,(s) \tag{2.68}$$

Cabe señalar que $|s|$ hace referencia a la cantidad de elementos de la clase deseada a predecir en la partición, así como también, $|n|$ es la cantidad de instancias que se encuentran en el nodo n. Finalmente la mejor división será aquella que entregue un GiniSplit menor, lo que significará una mayor pureza.

- **Radio de Ganancia**: consigue normalizar la ganancia de información, no obstante, debido a que el denominador puede ser cero desfavorece a aquellos atributos con escasa entropía, por ende, el procedimiento que se suele efectuar es calcular la ganancia de información de todos los atributos, y luego se calcula el radio de ganancia solamente a aquellos atributos que posean una ganancia de información mayor al promedio de todos. A continuación se expresa la fórmula que lo define como tal:

$$RadiodeGanancia = \frac{GananciadeInformación\,(y, S\,)}{Entropía\,(y, S\,)} \tag{2.69}$$

Un último alcance es la implicancia que conlleva este radio la cual es que a mayor radio, mejor división.

Dentro de los dos principales splits explicitados, entre el de Gini y Ganancia de información se pueden destacar sus diferencias más observables [87]

- **Gini**
  - Selecciona divisiones que aislan una clase mayoritaria en un nodo.
  - Crea splits desbalanceados.
  - Tiende a aislar clases numerosas de otras clases.
- **Entropía**
  - Mayores divisiones balanceadas en cuanto a número de instancias.
  - Tiende a encontrar grupos de clases que suman más del 50 por ciento de los datos.

En un aspecto más amplio se puede ver que el número de divisiones posibles varía según la naturaleza de la variable objetivo (aquella que se busca predecir), es decir [46, 55]:

- Variable Continua o numérica: n-1 divisiones posibles. (n número de instancias).
- Variable Nominal no ordenada: k-1 divisiones posibles. (k es el número de categorías por variable nominal).
- Variable Nominal no ordenada: $2^k - 1$ divisiones posibles.

Para el caso continuo, es decir, para los casos en los que el árbol de decisión juega un papel de modelo de regresión, la división requiere de la utilización de la suma de los cuadrados

$$D = \sum_{instancia\, j} \left(y_j - \mu_{[j]}\right)^2 \tag{2.70}$$

Criterio de Detención No siempre se llegará a nodos hoja con una clase dominante, sobre todo en aquellos conjuntos de datos, donde las instancias sean muy similares, es en estos casos, donde se aplican los criterios de detención. Otros casos donde se puede necesitar esta herramienta, son aquellos donde se den árboles con clases dominantes en todos sus nodos hojas, pero con una validación imprecisa, lo que conlleva a la conclusión de que se ha sobreentrenado el modelo. En esta caso, la herramienta de detención ayuda a evitar este tipo de errores [55]. Sin embargo, en este criterio de detención debe ser lo suficientemente válido para no generar un subentrenamiento del modelo, es decir, que no considere toda la información contenida en el conjunto de datos. Por ello, las siguientes alternativas pueden utilizarse sin dañar el modelo [55]:

- Usar un conjunto de prueba fuera del conjunto de entrenamiento, cuando el error del conjunto de prueba aumente detener divisiones.
- Considerar una reducción de corte "b", de tal manera de que si se tiene la máxima reducción de impureza en un nodo, detener y declarar que el nodo hoja es del tipo de la clase (categoría de la variable objetivo) mayoritaria. No obstante, esta alternativa depende del parámetro "b" escogido por el analista en cuestión
- Detener por el número de instancias contenidas en un nodo, modificando este parámetro.
- Penalizar la complejidad del árbol, tomando en cuenta alguna de sus medidas principales como la profundidad, cantidad de nodos hojas, etc.
- Usar tests estadísticos que permitan decidir si la división futura es conveniente o agrega información.

Poda Teniendo en cuenta las herramientas anteriores, se puede construir un modelo de árbol de decisión convincente, no obstante, se ha optimizado el aprendizaje del modelo, mediante la poda del árbol, la cual se efectúa para tratar el fenómeno del efecto horizonte, el que consiste en las veces donde el criterio de detención dicta una paralización temprana impidiendo mayor aprendizaje [55]. Es así, como la poda mide el *"balance entre el incremento del error del modelo en el conjunto de entrenamiento y la profundidad del árbol"* [55], o bien, se puede definir como *"es un mecanismo para obtener árboles con menor error de predicción"* [87]. Se suele clasificar en dos tipos [87, 59]:

- Pre-poda: Detener la construcción del árbol en algunas nodos que no generen mayor información o sobreajusten el modelo. Se suelen utilizar los criterios de detención mencionados anteriormente para este tipo de poda.

- Post-poda: Construir un árbol complejo (probablemente sobreajustado) y ejecutar la poda posteriormente. Nótese que *"al plantear el problema como un modelo de optimización, por el momento sólo se deben considerar heurísticas dado que es un problema NP-hard"* [59].

Dentro de los principales criterios para podar un árbol se encuentran [55, 88]:

- **Poda de Reducción de error**: Es el método más simple, haciendo uso de un conjunto extra de entrenamiento, como conjunto de subprueba o poda (actúa como tal pero no pertenece al conjunto de datos donde se probará el modelo). Comienza desde las hojas y poda hasta el nodo parental, calculando el error para los nodos que no son esquinas, términos o inicios del árbol. Esto último se efectúa reemplazando cada nodo por una hoja, cuya etiqueta es la clase mayoritaria que contenga el nodo. Posteriormente se calcula el error del nuevo árbol en el conjunto de poda y se compara con el error del árbol original. Si es menor entonces se considera el árbol poda. Este procedimiento se repite por cada nodo no parental, ni hoja del árbol original, hasta llegar al árbol de menor error, es decir, se poda hasta el punto en que la certeza (*accuracy* comience a decrecer). Cabe destacar que cada nodo se revisa una sola vez.

- **Poda del error pesimista**: Es similar a la poda de reducción de error, no obstante, la diferencia de este criterio radica en que utiliza el mismo conjunto de datos para construir el árbol y para podarlo, por ende, no se puede usar el error directamente. Es así, como se recorre el árbol del nodo parental hacia las hojas. Su cálculo corresponde a:

  Sea n el número de instancias en el nodo t, y e(t) el número de errores si t fue reemplazado por una hoja. Además, sea T_t el sub-árbol emparentado (es decir, cuyo nodo parental es) t, L_t el conjunto de hojas de T_t y $e,(T_t)$ el número de error(o clasificaciones erradas) de T_t con la corrección de complejidad. Entonces el nodo t es reemplazado por un nodo hoja si:

$$e(t) \leq e,(T_t) + \sqrt{\frac{e,(T_t)\,[n - e,(T_t)]}{n}} - \frac{1}{2} \tag{2.71}$$

$$e,(T_t) = \sum_{l \in L_t} e(l) + \frac{|L_t|}{2} \tag{2.72}$$

- **Poda del valor crítico**: En este criterio se fija un parámetro como valor crítico. Luego el árbol es revisado partiendo desde los nodos intermedios sobre las hojas y terminando en el nodo parental. Los nodos donde la ganancia en la tasa de error sea menor que el valor crítico son reemplazados por hojas. Sin embargo, no posee una garantía de encontrar el mejor árbol debido a que exige la fijación de un parámetro.

- **Error-Based Pruning**: En este tipo de poda los nodos pueden ser reemplazados por nodos hijos junto con el resto de los subárboles, además de las opciones anteriores. Por lo tanto se puede transformar en la poda. Para decidir el reemplazo, se usa el estadístico del intervalo

de confianza para el error en el nodo, donde se está haciendo la decisión de mantenerlo o reemplazarlo. El límite superior se calcula tratando las instancias como una muestra estadística y asumiendo que los errores se distribuyen como una variable aleatoria binomial. De esta forma, la decisión sobre el nodo se efectúa en base al error esperado, los errores de los nodos hijos y del nodo con mayor cantidad de instancias.

Finalmente si el árbol es muy complejo impactará sobre la certeza del modelo, así como la extensión del mismo lo tendrá sobre la validez del modelo. Por ello, la evaluación de un modelo de árbol de decisión vendrá dada por *"el total de nodos, el total de hojas, la profundidad del árbol y el número de atributos que usa...cabe señalar que esto va dado por el criterio de detención y el método de poda"* [88].

En síntesis, el algoritmo descrito presenta las siguientes características [55]

1) Si todos los objetos son distinguibles, es decir, que no hay elementos idénticos con diferentes categorías en la variable objetivo, entonces se puede formar un árbol de decisión con error de resustitución igual cero. En otras palabras, pueden percatarse de pequeños cambios en el conjunto de datos de entrenamiento.

2) Permiten el seguimiento de las decisiones que conllevan a la categorización respectiva de la variable a predecir.

3) Pueden aprender independiente de la naturaleza de las variables, ya cualitativas (o nominales) como cuantitativas(continuas).

4) No requiere de una distancia o espacio cartesiano.

Support Vector Machines (SVM)

A diferencia de los algoritmos anteriores, la máquina de soporte vectorial, o bien, Support Vector Machines, utilizan planos complejos para encontrar la mejor división de las instancias que permita clasificarlas de manera óptima. Por ello, se representa como *"un problema de minimización cuadrático con un número de variables igual al número de objetos del entrenamiento"* [69]. Por lo tanto para grandes números de datos, se debe usar un equipo de gran capacidad. Además, se puede destacar su diferencia con la formulación del árbol de decisión y del modelo bayes, pues el primero hace uso de la inteligencia artificial más tests (estadísticos o no) para la división de los mismos; el segundo usa probabilidades y las SVM utilizan la rama de optimización de la matemática. Una peculiaridad de este problema es que *"involucra optimización de una función convexa"* [105], esto quiere decir que no contiene mínimos locales (y su máximos se encuentran en el infinito). Otra particularidad que comprende este algoritmo es que no requiere información acerca de la distribución del conjunto de datos, lo cual, implica la realización de un aprendizaje libre de distribuciones, es por ello que las SVM se suelen llamar modelo no parámetricos, lo cual, *"no significa que los modelos SVM no tenga ningún parámetro...sino que los parámetros son los que definen la capacidad de que el modelo se ajuste a los datos y a su complejidad"* [105]. Es decir, se busca el balance entre certeza y cantidad de datos a aceptar. Es esta combinación la que induce el origen del problema de minimización del riesgo estructural.

Minimización del Riesgo Estructural Para definir el problema de la minimización del Riesgo Estructural, se deben describir los pasos que llevaron prácticamente a su formulación. Este riesgo *"nace como necesidad de incorporar la capacidad de generalización de manera explícita en la construcción de un modelo predictivo y prevenir, de esta forma, el problema de sobreajuste"* [69], el cual, en líneas generales consiste en que el modelo aumente su aprendizaje sobre los datos de entrenamiento.

Para comprender este principio se requieren algunos conceptos básicos:

- **Funciones Q(z,A)**: Es la función de pérdida, donde el argumento z está asociado a la dimensión de los datos, en otras palabras, está asociado a la cantidad de atributos que posee el conjunto de datos de entrenamiento y prueba. Por otro lado, el argumento A va asociado al conjunto de objetos que la función Q usará internamente, por consiguiente, *"los elementos pertenecientes a A no son necesariamente vectores, pueden ser parámetros abstractos"* [102].

- **Dimensión VC**: Se refiere al número máximo de vectores $(z_1, ..., z_h)$que pueden ser divididos en dos clases, en 2^h caminos usando funciones del conjunto Q, es decir, es el número máximo de vectores que pueden ser encerrados por un conjunto de funciones [102].

- **Funciones Indicatrices**: Son aquellas funciones que pueden tomar solamente dos valores, uno o cero.

- **Intervalo de confianza (ϵ)**: Es aquel intervalo donde es posible argumentar estadísticamente que la función presentará el comportamiento que se desea demostrar.

Una vez explicado el principio, se procede a describir sus principales características expresadas a continuación [102]:

Sea S un conjunto de funciones Q(z,α) con $\alpha \in A$, y sea este conjunto provisto de una estructura consistente de subconjuntos anidados de funciones $S_k = \{Q(I, \alpha), \alpha \in A_k\}$ tal que $S_1 \subset S_2 \subset \ldots \subset S_n \ldots$, y los elementos de la estructura satisfacen:

- La dimensión VC: h_k de cada conjunto $S_k$de funciones es finita.

- La estructura del conjunto es admisible.

Como la dimensión VC, está asociada a la cantidad de vectores que aguantará la función Q y la admisibilidad de la estructura va asociado a que la función este limitada superiormente. De esta manera, se puede llegar a la conclusión de que *"maximizando el margen de separación de entre clases es posible generar funciones de clasificación que logren ese balance"* [69].

Representación del algoritmo El algoritmo en palabras concretas consiste en *"encontrar un hiperplano de separación que divida el espacio de entrada en dos regiones...estas regiones corresponderán a las clases definidas"* [60, 69], además, ocupa la perspectiva de *"mantener el valor del riesgo empírico ajustado (es decir, que sea igual a cero) y minimizar el intervalo de confianza"* [102]. Desde una vista general, la idea de que ejecutan las SVM es esquematizar los vectores de entrada x en espacio de alta dimensionalidad Z a través de un mapeo no lineal escogido a priori [102], la cual, se bosqueja a continuación [69]:

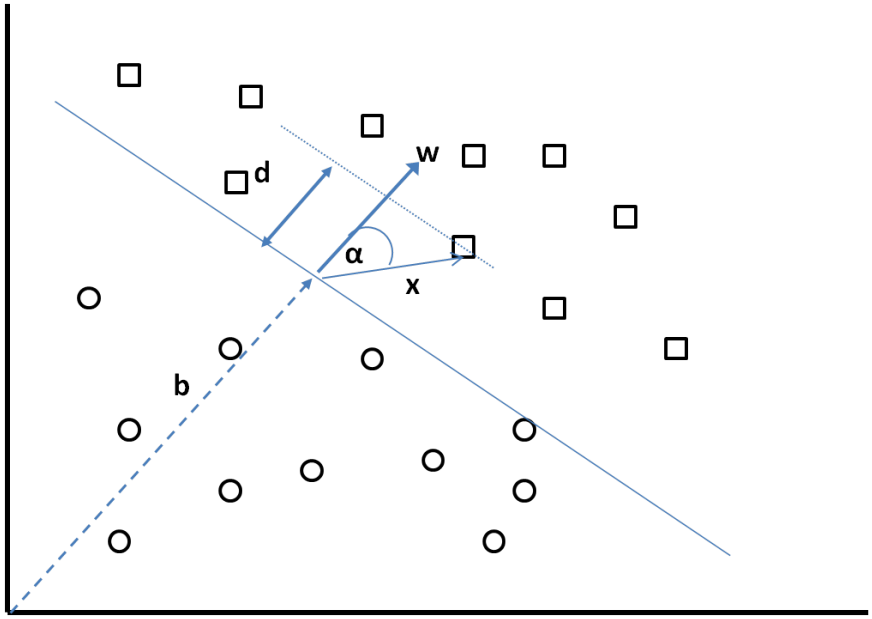

Figura 2.19: Representación Hiperplano de clasificación para dos dimensiones

Siendo $\vec{v}$ el vector que va desde el origen al hiperplano y es perpendicular a este último, y b el valor de su magnitud. Además, E es el punto de intersección entre el vector v y el hiperplano separador. Considerando a $\vec{w}$ como el vector normal al hiperplano, $\vec{x_i}$ el vector que va desde E hasta la instancia i y α el ángulo entre el vector $\vec{w}$ y el vector $\vec{x_i}$, se puede describir la implementación del modelo SVM para la ocasión en que se tengan dos clases.

Separabilidad de las SVM En el capítulo anterior se mencionó el hiperplano separador, éste es aquel que genera las fronteras, para las regiones que representarán las clases, ahora bien, consecuente al objetivo de las SVM, *"un hiperplano de separación ideal debe maximizar el margen de separación y minimizar el error de clasificación"* [69]. Ahora bien, su fórmula matemática es $\vec{w} \cdot \vec{x} + b = 0$. Para encontrar el mejor hiperplano que separe las clases se define el siguiente problema de minimización [22, 69, 102]:

$$\min_{w} \frac{1}{2}\vec{w}^T \cdot \vec{w} \tag{2.73}$$

$$\text{Sujeto a} \quad y_i \left(\vec{x_i} \cdot \vec{w} + b\right) - 1 \geq 0 \tag{2.74}$$

$$\forall i = 1, ...l \tag{2.75}$$

Cuya formulación Dual de Wolfe es:

$$\max_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x_i} \cdot \vec{x_j} \tag{2.76}$$

$$\text{Sujeto a} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.77}$$

$$\alpha_i \geq 0 \quad \forall i = 1, ..., l \tag{2.78}$$

Esta formulación no agrega el hecho de que se puede aprender sobre las malas clasificaciones de los datos de entrenamiento, y por ende supone que al trazar un hiperplano lineal el error será relativamente pequeño, por ello se usa este planteamiento del problema para el denominado caso linealmente separable. Una representación de ambos casos descritos previamente se puede observar en la figura 2.20 [105].

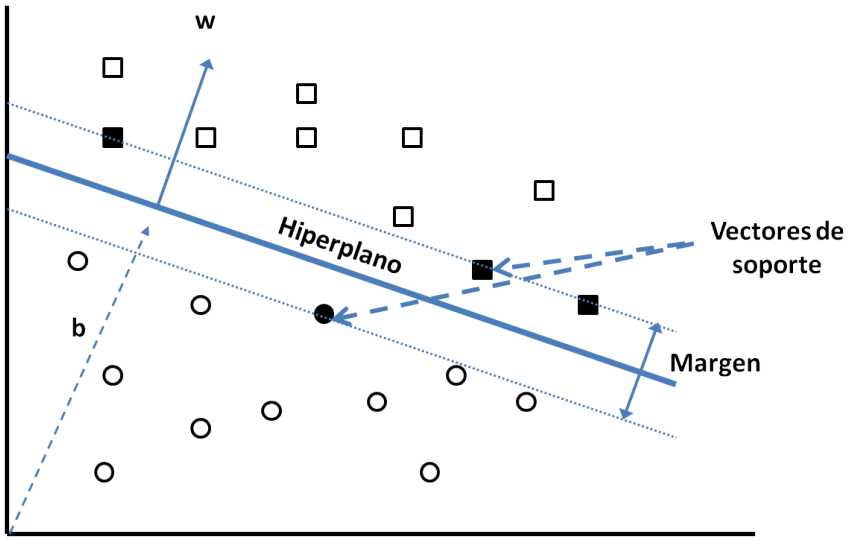

Figura 2.20: Solución caso linealmente separable

En los casos en que el comportamiento del conjunto de datos de entrenamiento no se pueda dividir en clase de forma factible por un hiperplano lineal, se estará en presencia de un caso no linealmente separable, lo cual es producido principalmente por la inexistencia de un hiperplano que cumpla con la restricción 2.74. La forma en que se adapta el problema inicial a la nueva disyuntiva mencionada es agregando una variable de holgura a la formulación original, la cual permite el manejo de la penalización sobre las malas clasificaciones, esto es una modificación que se le efectúa en el caso en que se requiera un plano no linealmente separable. De esta forma, el nuevo planteamiento del problema, que ayuda a encontrar el mejor hiperplano, será [22, 60, 69]:

$$\min_{\vec{w},b} \frac{1}{2}\vec{w}^T \cdot \vec{w} + C \cdot \sum_{i=1}^{l} \epsilon_i \tag{2.79}$$

$$\text{Sujeto a} \quad y_i \left(\vec{x_i} \cdot \vec{w} + b\right) - 1 \geq 1 - \epsilon_i \quad \text{para} \quad i = 1, ...l \tag{2.80}$$

$$\epsilon_i \geq 0 \forall i \tag{2.81}$$

Cuya formulación Dual de Wolfe, respectiva es:

$$\max_{\alpha_i} \sum_{i=l}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x_i} \cdot \vec{x_j} \tag{2.82}$$

$$\text{Sujeto a} \quad 0 < \alpha_i \leq C \tag{2.83}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.84}$$

No obstante, para que la clasificación posea el menor error(y cuyo margen se máximo) se debe definir un hiperplano óptimo que sea canónico, pues el margen máximo indicará que el modelo clasificará mejor los nuevos datos que sean ingresados, además, se le añade el hecho de que sea canónico para *"facilitar la búsqueda de patrones significativos, denominados vectores de soporte o support vectors"* [105], es decir, simplificará los cálculos. Estos vectores de soporte son aquellos vectores (usualmente paralelos) al hiperplano separador que declaran la frontera real de la clase en cuestión, siendo una especie de sentinelas. Formalmente son *"aquellos vectores en donde la desigualdad 2.74 es alcanzada"* [105]. Se aclara entonces, que el margen tratado en este apartado se refiere a la distancia entre vectores de soporte, es decir, la distancia entre las fronteras de las clases. Un punto importante, es que la solución de w, para ambos casos es [22]:

$$w = \sum_{i=1}^{U} \alpha_i y_i x_i \tag{2.85}$$

Donde U es el número de vectores de soporte y, además, se utiliza la siguiente expresión para su cálculo:

$$|w|^2 = \alpha^T H \alpha \text{ con H siendo el Hessiano del problema Dual} \tag{2.86}$$

Técnica de funciones Kernel para SVM Los casos en que un hiperplano es lineal o no linealmente separable, no comprende a aquellas situaciones en que la división de clases es no lineal, es decir, *"cuando la función de decisión (indicatriz Q) no es lineal en los datos"* [69], lo que se traduce en aquellas ocasiones en donde la línea divisoria no puede ser recta, debido a que se comete un error bastante grande. Para estos casos se requiere usar una técnica adicional al modelo, cuya idea principal es: *"proyectar los objetos en otro espacio euclidiano H de mayor dimensión en el cual sean linealmente separables"* [69], posteriormente se ejecuta el procedimiento anterior, es decir, se encuentra el hiperplano en H resolviendo el problema de minimización, y finalmente se vuelve al espacio original con el hiperplano encontrado. No obstante, este nuevo hiperplano, *"ya no será un*

hiperplano en el espacio original sino una hipersuperficie no lineal" [69], tal y como se puede apreciar en la figura 2.21 [105]:

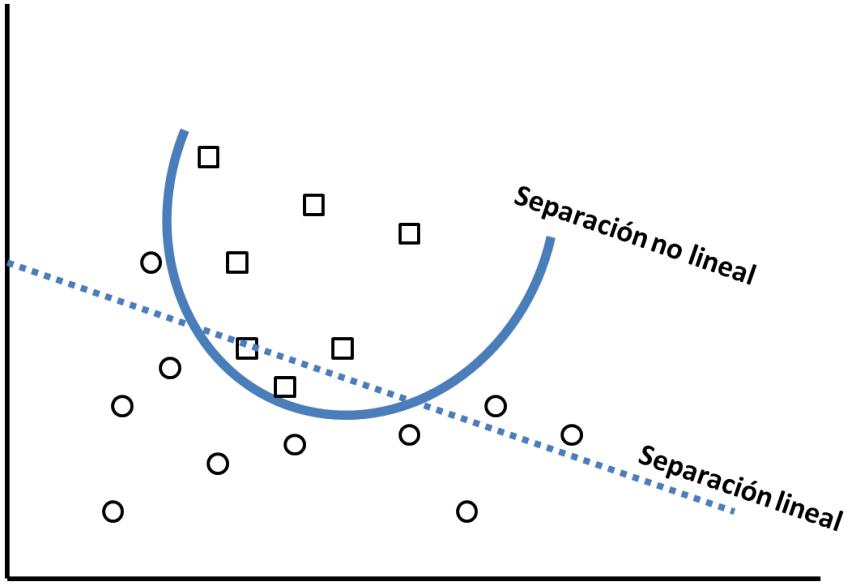

Figura 2.21: Situación no lineal

Por lo tanto, se requiere un función de mapeo entre espacios, en forma de producto punto, la cual se define como $\phi : IR^n \rightarrow H$. De esta manera, la nueva técnica sugiere que una vez mapeado el conjunto de datos por separado, se determine el hiperplano. Por consiguiente la aparición de los datos (ya mapeado) en el problema del hiperplano, en el espacio H, tendrá la forma $\phi(x_i) \cdot \phi(j)$. Ahora bien, con esto el problema se vuelve simple, en el sentido de que no se necesita saber la función $\phi(\cdot)$ para llevar a cabo el modelo, puesto que se busca una función kernel de la forma: $K\left(x_i, x_j\right) = \phi(x_i) \cdot \phi(j)$[22]. Sin embargo, esto no soluciona el tema de que el vector $\vec{w}$ estará en el espacio H, mas eso no es un problema debido a que en la fase de prueba o testeo, las SVM requieren *"calcular los productos punto de un punto de prueba dado x con w"* [22], lo que se traduce en el cálculo del signo de:

$$f(x) = \sum_{i=1}^{U} \alpha_i y_i \phi(s_i) \cdot \phi(x) + b = \sum_{i=1}^{U} \alpha_i y_i K\left(x_i, x_j\right) \tag{2.87}$$

En donde U es el número de vectores de soporte y s_i son los vectores de soporte.

Agregado a lo anterior, se menciona la definición de las funciones kernel como aquellas que permiten *"transformar el espacio de entrada, mediante una función no lineal de mapeo, a un espacio de mayor dimensionalidad"* [25], lo que de forma matemática se traduce en: *"Sea X un conjunto de datos no vacío. Una función k en X cruz X, que* $\forall m$*, con* $m \in N$ *y* $\forall x_1, ..., x_m \in X$ *da lugar a una matriz Gram Kernel definida positiva, es llamada una kernel definida positiva"* [95], donde esta última matriz, se define como: *"dada una función* $k : X^2 \rightarrow K$ *(con K=IR), y sean los patrones* $x_1, ..., x_m \in X$*, la matriz K de m x m con elementos* $K := k(x_i, x_j)$ *es denominada Matriz Gram (o*

Cuadro 2.3: Principales kernels asociados a clasificadores

Funciones Kernel	Tipo de clasificador
$K(x,y) = (x^T y)$	Linear, producto punto, Kernel, CPD.
$K(x,y) = \left[(x^T y) + 1\right]^d$	Polinomio completo de grado d.
$K(x,y) = e^{\frac{1}{2}\left[(x-y)^T \Sigma^{-1}(x-y)\right]}$	RBF Gaussiano.
$K(x,y) = tanh\left[(x^T y) + b\right]$	Multilayer(multicapa) Perceptron.
$K(x,y) = \frac{1}{\sqrt{\|x-y\|^2+\beta}}$	Función inversa multicuadrática.

matriz Kernel)" [95].

El problema ahora es ¿Cuáles funciones kernels se pueden usar?, la respuesta a esta disyuntiva reside en el par (Espacio H y Función de mapeo $\phi(\cdot)$) el cual, tendrá que cumplir la condición de Mercer la que consiste en [22, 102]:

$$\exists \phi(\cdot), K(x_i, x_j) = \phi(x_i) \cdot \phi(j) \Leftrightarrow \forall g(x) \text{ tal que} \tag{2.88}$$

$$\int g(x)dx \quad \text{es finita entonces se tiene que} \quad \int K(x_i, x_j)g(x_i)g(x_j) \geq 0 \tag{2.89}$$

Esta condición *"permite asegurar que el Hessiano de la formulación Dual esté definido y el problema cuadrático tenga solución"* [69]. Finalmente se muestran las funciones kernels más usadas en la actualidad en la tabla 2.3 [105].

Redes Neuronales

Este modelo de minería de datos suele ser uno de los más potentes, desde un punto de vista de capacidad, en la actualidad, puesto que emula el funcionamiento de un cerebro humano a un nivel superior, entregando buenos resultados en su mayoría. Sin embargo, debido a la complejidad que posee, no se puede saber con exactitud de dónde provino ese resultado, lo que es una dificultad a la hora de explicar su funcionamiento. Se puede aplicar a numerosos tipos de problemas como, por ejemplo, de clasificación, reconocimiento, agrupación y mapeo de datos para descubrir patrones. En un sentido directo, una red neuronal artificial (o denominada simplemente red neuronal, o ANN) *"consiste en procesar elementos (llamados neuronas) y las conexiones entre ellos con coeficientes(pesos) ligados a las conexiones, las cuales constituyen una estructura neuronal, y un entrenamiento y algoritmos recordatorios adjuntos a la estructura"* [52], lo que en palabras simples puede ser descrito como *"una piscina de unidades simples de procesamiento que se comunican enviando señales entre ellas sobre un gran número de conexiones ponderadas"* [54]. En base a estas definiciones, se establece una serie de supuestos que explican el funcionamiento básico de un red neuronal [31]:

- El procesamiento de información ocurre en la unidad más pequeña llamada neurona.
- Las señales se pasan entre neuronas mediante enlaces de conexión.

- Cada enlace de conexión tiene un peso asociado.
- Cada neurona aplica una función de activación para procesar la información

Estos supuestos caracterizan las principales componentes de una red neuronal, así como también, los términos asociados al modelo, puesto que describe que existe una red de neuronas conectadadas, lo cual tiene asociado un patrón en el sentido en cómo se forman las conexiones de la red. Esto último se denomina la arquitectura de la red neuronal. También señala la existencia de pesos relacionados con estas conexiones, lo que alude a la necesidad de calcularlos, el método que se utiliza para calcularlo tiene como nombre entrenamiento, aprendizaje o algoritmo de la red. La última característica de este modelo es la función que procesa la información en cada neurona.

De esta manera, la red neuronal, puede ser representada matemáticamente, enfocándose en la actividad de una neurona, la que consta de los siguientes parámetros [31, 52]:

- **Conexiones de entrada**: Se representan como los datos de entrada, es decir, $x_1, ..., x_n$ siendo n la cantidad de instancias. Los pesos asignados a estas conexiones se denotan como $w_1, ..., w_n$, sin embargo, existe una conexión de entrada que es constante, llamada tendencia y se escribe x_0
- **Función de entrada**: Es aquella encargada de efectuar el cálculo acumulado de la red, por ende, tiende a tener la forma de una sumatoria, es decir, $u = u(x, w) = \sum_{i=1}^{n} x_i \cdot w_i$.
- **Función de activación**: También denominada señal, se encarga de generar el nivel de activación de la neurona, su notación es $a = a(u)$. Nótese que el argumento de esta función es la función de entrada de la neurona, la cual considera las observaciones y los pesos. Las funciones más usadas de este tipo son:
  - Función de corte difícilmente limitada: Toma dos valores como una función indicatriz (es decir, 0 ó 1).
  - Función de corte lineal: Esta función va aumentando linealmente su valor acorde al aumento en la función de entrada u, lo que tras un determinado corte se satura a un solo valor.
  - Sigmoidal: En esta categoría de funciones caen todas las funciones con forma de S, y que sean acotadas, con monotonía creciente, continua y diferenciable en todo el dominio.
- **Función de salida**: Esta función entrega el resultado de la neurona, su notación será $o = o(u)$, sin embargo, se tiende a asumir que es igual a la función de activación, o sea, o=a.
- **Tasa de aprendizaje**: Es la tasa que se utiliza para ajustar la cantidad de modificación de los pesos en cada iteración del entrenamiento.
- **Parámetro de corte** θ: Es aquel parámetro bajo el cual la neurona decide (mediante la función de activación) cuando una neurona es activada o no.

Un bosquejo de estos parámetros se observa en la figura 2.22.

Ahora bien, una de las clasificaciones más generales de las redes neuronales está relacionada con la arquitectura de la red, la cual, eventualmente, va enlazada con la presencia o ausencia de conexiones con retroalimentación [52, 54]:

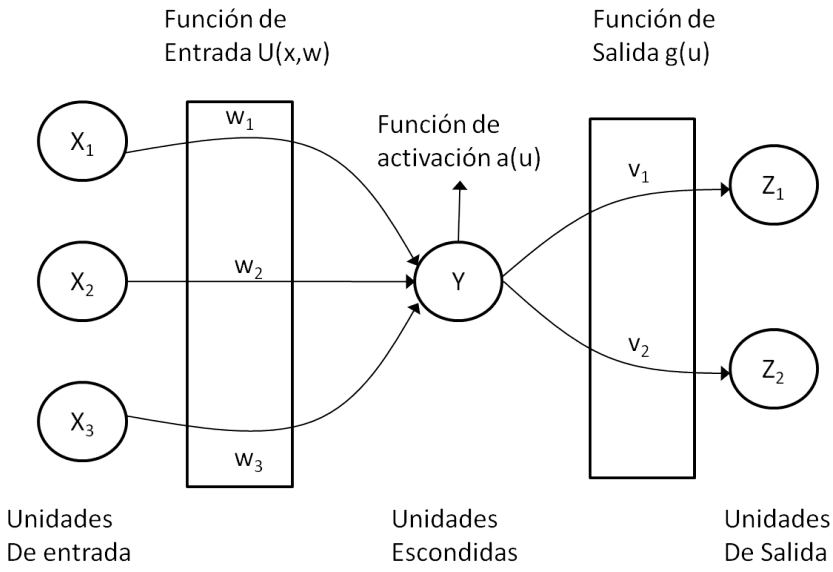

Figura 2.22: Esquema de red neuronal simple

- **Arquitectura FeedForward**: Esta arquitectura se presenta cuando *"no hay conexiones inversas desde los resultados hacia las neuronas de entrada...la red no mantiene una memoria de sus resultados anteriores y los estados de activación se quedan en las neuronas"* [52]. Esto quiere decir, que el modelo no estudia en base a sus resultados ni a cómo se comportó en las iteraciones, por lo que no penaliza sus resultados. Una definición alterna a esta arquitectura sería aquella *"donde los datos fluyen de las unidades de entrada a las unidades de salida solamente con retroalimentación hacia adelante...no presenta conexiones que se extiendan desde las unidades de salida a las unidades de entrada en la misma capa o en capas diferentes"* [54], donde se define capa como una agrupación de neuronas bajo un mismo proceso. Un ejemplo de estas redes son las que efectúa el modelo Perceptrón.

- **Arquitectura FeedBack o Recurrente**: Esta arquitectura mantiene conexiones *"desde las unidades de salida hacia las neuronas de entrada...manteniendo memoria de sus estados previos y su estado siguiente no depende sólo de las funciones de activación de entrada sino de los estados previos de la red"* [52], es por ello que relaciona todos los datos ya sea en forma acumulada o bien hacia atrás, con esto, a diferencia de la arquitectura anterior, la red puede penalizar las predicciones erróneas realizadas en cada capa, cabe señalar que *"en algunos casos los valores de la función de activación de las unidades pueden relajarse para llegar a un estado estable donde estas funciones no cambien más"* [54]

Una vez considerados los aspectos anteriores, se puede establecer un algoritmo general de aprendizaje para las redes neuronales, en los casos que se tenga un problema de aprendizaje supervisado [52]:

- Paso 1: Ajustar la arquitectura de la red neuronal y los pesos iniciales para las conexiones (a lo menos una neurona de entrada por variable(instancia en este caso).

- Paso 2: Entregar vectores de entrenamiento a la red.
- Paso 3: Calcular el vector resultado.
- Paso 4: Compara el vector resultado deseado y, con el obtenido por la red, evaluar el error.
- Paso 5: Corregir los pesos de las conexiones, tal que el vector resultado se ajuste al vector deseado.
- Paso 6: Repetir desde 2 a 5, hasta la convergencia.

Es así como se pueden presentar los algoritmos más usados de las redes neuronales:

1) **Perceptrón**: Este algoritmo fue uno de los primeros en salir a la luz en lo que respecta a las redes neuronales, buscaba modelar la percepción visual del ser humano, específicamente fue *"diseñado para ilustrar algunas de las propiedades fundamentales de los sistemas inteligentes, en general, sin llegar a ser demasiado profundo en condiciones especiales para organismos"* [89]. La arquitectura que posee este algoritmo es *"feedforward y de tres capas,...la primera capa es donde se almacenan los datos, la segunda combina la información con diferentes configuraciones, no obstante,... los pesos entre la primera capa y la segunda quedan fijos, razón por la cual sólo son visibles gráficamente dos capas"* [52]. Luego el algoritmo, que *"no es sensitivo a los valores iniciales tanto de sus pesos como de la tasa de aprendizaje"* [31], procede como se muestra a continuación:

   - Paso 1: Inicializar los pesos y la tendencia (x_0) (generalmente estos valores son 0). Ajustar la tasa de aprendizaje α que debe estar entre 0 y 1.
   - Paso 2: While condición de detención (usualmente relacionada con la cantidad de error a aceptar) hacer pasos 3-7.
   - Paso 3: Para cada par de entrenamiento s:t (siendo t la variable que se busca a predecir y s la instancia) hacer pasos 4-7.
   - Paso 4: Configurar las unidades de entrada, es decir, $x_i = s_i$.
   - Paso 5: Calcular resultado para la unidad de salida, considerando el parámetro de corte θ:

$$y' = x_0 + \sum_i x_i w_i \tag{2.90}$$

$$y = \begin{cases} 1 & \text{si } y' > \theta \\ 0 & \text{si } -\theta \leq y' \leq \theta \\ -1 & \text{si } y' < -\theta \end{cases}$$

   - Paso 6: Modificar la tendencia y los pesos de las conexiones si ocurre un error para el

patrón entregado, es decir, si $y \neq t$ entonces,

$$w_i(nuevo) = w_i(viejo) + \alpha t x_i \tag{2.91}$$
$$x_0(nuevo) = x_0(viejo) + \alpha t \tag{2.92}$$
$$\text{si y=t entonces} \tag{2.93}$$
$$w_i(nuevo) = w_i(viejo) \tag{2.94}$$
$$x_0(nuevo) = x_0(viejo) \tag{2.95}$$
$$\tag{2.96}$$

- Paso 7: Si los pesos no varían en el paso 7, entonces detener, de lo contrario volver al paso 5

Un punto relevante frente a este algoritmo es que existe un teorema de convergencia asociado, el cual asegura la misma para un algoritmo del área de inteligencia artificial. Éste señala que:

Teorema: *"Si existe un conjunto de pesos de conexiones w_i en el cual es posible efectuar una transformación y=d(x), la regla de aprendizaje del perceptrón convergerá a una solución, en número finito de pasos para cualquier elección inicial de los pesos"* [54].

2) **Adaptative linear element (Adaline)**:Este algoritmo fue descubierto al resolver el problema de *"controlar un conjunto de resistores en un circuito el cual pudiese sumar las corrientes causadas por las señales de voltaje de entrada"* [54]. Por lo general, tiene funciones de activación bipolares (-1 o 1) para las funciones de activación de entrada y *"los pesos de las conexiones de las unidades de entradas, en este caso, son ajustables...,admemás, la función de activación es la función identidad"* [31], esto último permite que esta red prosiga en su aprendizaje sobre patrones de entrenamiento. Respecto a la arquitectura, este algoritmo *"es una unida singular (neurona) que recibe entradas desde varias unidades, también recibe una entrada peculiar cuya señal(o resultado de la función de activación) siempre es +1, para que la tendencia(o peso tendencia x_0) sea entrenada en el mismo proceso como los otros pesos (los de las conexiones)"* [31], en otras palabras, es para que el valor de la tendencia sea tratado como los otros pesos a lo largo del algoritmo. Si bien, la descripción no entrega mucho detalle, su algoritmo sí lo hace, dejando entrever una clara diferencia con el Perceptrón, en especial, al momento de modificar los pesos. Esto puede ser apreciado al mostrar el algoritmo del Adaline:

- Paso 1: Inicializar los pesos, tendencia y tasa de aprendizaje α.
- Paso 2: While condición de detención (usualmente relacionada con la cantidad de error a aceptar) hacer pasos 3-7.
- Paso 3: Para cada par de entrenamiento s:t (siendo t la variable que se busca a predecir y s la instancia) hacer pasos 4-7.
- Paso 4: Configurar las unidades de entrada, es decir, $x_i = s_i$.

- Paso 5: Calcular resultado para la unidad de salida, considerando el parámetro de corte θ:

$$y' = x_0 + \sum_i x_i w_i \tag{2.97}$$

- Paso 6: Modificar la tendencia y los pesos de las conexiones si ocurre un error para el patrón entregado, es decir, si y≠t entonces,

$$w_i(nuevo) = w_i(viejo) + \alpha \left(t - y'\right) x_i \tag{2.98}$$

$$x_0(nuevo) = x_0(viejo) + \alpha \left(t - y'\right) \tag{2.99}$$

$$\tag{2.100}$$

- Paso 7: Si el cambio más grande entre los pesos que ocurrió en el paso 2 es más pequeño que una tolerancia especificada, detener, de lo contrario volver al paso 5

Cabe señalar, que este algoritmo, sí posee sensibilidad frente a los parámetros iniciales asignados a diferencia del Perceptrón. Dicha sensibilidad queda plasmada, en especial, en la tasa de aprendizaje, por lo que se sugiere ajustar valores pequeños para la tasa de aprendizaje en este método, del orden de 0,1.

Se menciona la regla delta, que permite la convergencia de este algoritmo, esta regla también se denomina la regla de los mínimos cuadrados y consiste en modificar los pesos tal que el error cuadrático medio entre el resultado real y el obtenido por el modelo sea mínimo. Lo que implica un proceder en el que se minimiza tal error, se deriva y se llega a que el cambio en los pesos por iteración debe ser:

$$\Delta w_I = \alpha \left(t - y'\right) x_I \tag{2.101}$$

Una demostración detallada de esta regla se puede encontrar en [31, 54].

3) **Perceptrón multicapa**:La principal desventaja del perceptrón es que sólo puede clasificar problemas que son linealmente separables, característica compartida con la mayoría de las redes neuronales de dos capas. Para problemas que no son linealmente separables, las éstas requieren más capas, las cuales se ubican entre las de entrada y las de salida. Estas capas intermedias se denominan escondidas, porque contienen neuronas escondidas y se asocia dicho nombre a que los procesos que ocurren dentro de ellas, son desconocidos y no pueden visualizarse a través de los resultados. El perceptrón es un modelo de dos capas, por lo que no puede resolver el problema XOR, que en otra palabras, es el problema del "ó" exclusivo, el que puede observarse en la figura 2.23. Dicho problema fue el origen del modelo de Perceptrón multicapa o *multilayer perceptron (MLP)*, el cual consiste en añadir las capas escondidas de manera de efectuar cálculos que permitan resolver problemas que no son linealmente separables. Esta red *"está completamente conectada porque cada nodo en una capa está conectado con todos los nodos de la siguiente capa. Si algunas conexiones están ausentes, la red está parcialmente conectada"* [48], esto quiere decir, que su resultado depende de la calidad de los datos, en el sentido de que un alto grado de imputación (sea 30

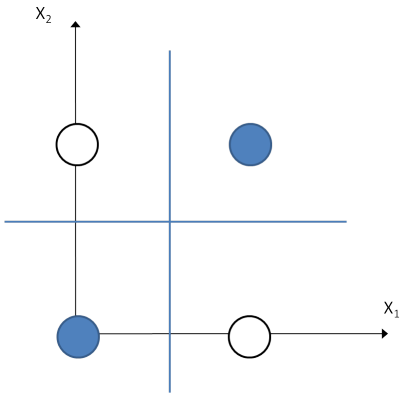

Figura 2.23: Problema de ó exclusivo

por ciento, por ejemplo) afectará gravemente los resultados. Dentro de las características del MLP se pueden mencionar que *"las neuronas tienen valores de entrada y salida continuos, una función de entrada de suma y una función no lineal de activación"* [52], ahora bien, el objetivo en este tipo de modelo es encontrar los pesos óptimos que minimicen el error global E.

Una característica relevante de este modelo es que *"un MLP con una capa escondida puede aproximar cualquier función continua a cualquier certeza deseada, sujeto a un número suficiente de nodos escondidos"* [52].

El modelo multicapa es representado principalmente por el algoritmo de propragación inversa o *back propagation* que también se conoce como la regla delta generalizada. Éste se muestra a continuación [31]:

- Paso 1: Inicializar los pesos(valores aleatorios pequeños).
- Paso 2: While la condición de detención sea falsa ejecutar pasos 3-10
- Paso 3: Para cada par de entrenamiento (s:t), ejecutar pasos 4-10. cabe señalar, que los pasos de 4 al 6, son pasos relacionados con el aprendizaje hacia adelante. Mientras que los pasos 7 y 8 hablan de la propagación a la inversa del error.
- Paso 4: Cada neurona de entrada (sea X_i, $i = 1, ..., n$) recibe una señal x_i y la transmite a todas las unidades de la capa posterior(capa escondida).
- Paso 5: Cada unidad escondida (sea Z_j, $j = 1, ..., p$) suma los pesos de las señales de entrada, es decir

$$z'_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij} \tag{2.102}$$

  Donde v_{0j} es la tendencia de la unidad escondida j y v_{ij} son los pesos de la capa escondida. Además, se aplica la función de activación para calcular la señal de salida:

$$z_j = f(z'_j) \tag{2.103}$$

Y se envía esta señal a todas las unidades de la capa posterior(unidades de salida).

- Paso 6: Cada unidad de salida (Y_k,k=1,...,m) suma su señales ponderadas de entrada.

$$y'_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk} \tag{2.104}$$

Y, correspondientemente, se calcula la función de activación para calcular la señal de salida

$$y_k = f\left(y'_k\right) \tag{2.105}$$

- Paso 7: Cada unidad de salida (Y_k, k=1,...,m) recepciona un patrón objetivo(patrón de entrenamiento) y calcula el término de error en información, lo que expresado en otras palabras, corresponde a

$$\delta_k = (t - k - y_k) f'\left(y'_k\right) \tag{2.106}$$

El cual es utilizado para calcular la correción de los pesos y la tendencia, de la siguiente manera

$$\textbf{Pesos}: \Delta w_{jk} = \alpha \delta_k z_j \tag{2.107}$$

$$\textbf{Tendencia}: \Delta w_{0k} = \alpha \delta_k \tag{2.108}$$

Además la presencia en las ecuaciones anteriores, δ_k es enviado a las unidades en la capa anterior para dar el efecto de propagación a la inversa del modelo

- Paso 8: Cada unidad escondida Z_j, suma las entradas delta para entregar un efecto acumulado.

$$\delta'_j = \sum_{k=1}^{m} d_k w_{jk} \tag{2.109}$$

Luego, se calculan las correcciones de los pesos y la tendencia de la capa escondida, es decir,

$$\Delta v_{ij} = \alpha \delta_j x_i \tag{2.110}$$

$$\Delta v_{0j} = \alpha \delta_j \tag{2.111}$$

Finalmente se procede con la actualización de los pesos y tendencia.

- Paso 9: Cada unidad de salida (Y_k, k=1,...,m) y unidad escondida actualiza sus tendencias y pesos (j=0,...,p):

$$w_{jk}(nuevo) = w_{jk}(viejo) + \Delta w_{jk} \tag{2.112}$$

$$v_{ij}(nuevo) = v_{ij}(viejo) + \Delta v_{ij} \tag{2.113}$$

- Paso 10: Probar la condición de detención

Cabe destacar que *"implementando este algoritmo, deberían usarse arreglos separados para los deltas en las unidades de salida"* [31], lo expresado anteriormente solamente es para

entrenar la red, el detalle de cómo aplicar la red posterior al entrenamiento se puede observar en [31].

En el modelo MLP, debido a que tiene capas escondida, se introduce un nuevo concepto llamado época, el cual se define como *"un ciclo a través de todo el conjunto de vectores de entrenamiento...muchas épocas son necesarias para entrenar una red neural de propagación a la inversa"* [31]

La base matemática de este algoritmo es el gradiente descendente, pues el gradiente de una función entrega tanto el crecimiento como el decrecimiento de una función, por lo que hace más eficiente la dirección en la que se actualizan los pesos de la red.

A pesar de que el algoritmo resulta ser atractivo, y, en ocasiones adecuado y eficiente, posee dos desventajas: la primera es que la red se puede paralizar mientras se entrena debido a que los pesos pueden terminar en grandes valores, por consiguiente, la cantidad de unidades escondidas será muy alta y la función de activación colapsará. La segunda desventaja es el tema del mínimo local, pues *"la superficie del error de una red compleja está lleno de cerros y valles...por el gradiente descendente, la red puede caer en un mínimo local"* [54].

Regresión

El término regresión partió por un estudio de Francis Galton acerca de la tendencia presente en que los padre de estatura alta tenían hijos de estatura alta y los padres de estatura baja tenían hijos de estatura baja, por ende, Galton estudió el efecto en la estatura promedio de la población total y observó como se movía dicha tendencia. Esta ley, posteriormente fue validada por Karl Pearson.

La regresión, en la actualidad, consiste en *"el estudio de la dependencia de la variable dependiente, respecto a una o más variables (las variables explicativas), con el objetivo de estimar y/o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las últimas"* [38]. Por lo que este modelo sirve para predecir y clasificar, donde su uso típico es el de predecir la demanda o el inventario futuro de una empresa. En el caso de la clasificación, la regresión que se utiliza comúnmente no resulta muy efectiva, puesto que la variable a predecir posee una connotación nominal, sin embargo, existe un tipo de regresión que se encarga de predecir variables nominales y se denomina regresión logística.

La regresión propiamente tal se asocia a la linealidad, mas existen modelos de regresión no lineales. Se define lo que es la linealidad en la variables mediante dos significados: el primero señala que existe linealidad cuando *"la esperanza condicional de Y(variable a predecir) es una función lineal de X_i, es decir, $E(Y|X_i) = \beta_1 + sum_i\beta_i x_i$"* [38]. Nótese que si el exponente de alguna de las variables explicativas x_i se encuentra elevada a un exponente mayor a 1 entonces la función ya no es lineal. El segundo significado aparece *"cuando la esperanza condicional de Y, $E(Y|X_i)$ es una función lineal de los parámetros β"* [38], independiente de si es lineal o no con las variables explicativas. Siendo está última la más usada en la teoría de la regresión.

Al modelo original se debe agregar una componente más, denominada perturbación estocástica la cual no se introduce en el modelo explícitamente, pero ayuda a abarcar la información que no puede ser descifrada por el modelo original, todo lo anterior en pos de evitar una pérdida de información.

Generalmente los parámetros β deben ser estimados, el método más utilizado es el de la máxima verosimilitud, y además, se debe agregar un test estadístico que permita evaluar dicha estimación, dentro de los cuales se suelen usar los t-student y las F de Fischer. Cuando la variable a predecir es nominal, los modelos de regresión comienzan a llamarse modelos probabilísticos, pues su *"objetivo es encontrar la probabilidad de que un acontecimiento suceda"* [38]. Siendo uno de los modelos más usados es el de la regresión logística, el que en palabras sencillas *"provee de una forma funciónal f y un vector de parámetro α para expresar P(Y—X) como* $P(Y|X) = f(X, \alpha)$*"* [29], dichos parámetros son determinados mediante la estimación de máxima verosimilitud aplicada sobre el conjunto de datos general.

En los problemas de clasificación binaria se asume una función sigmoidal, debido a que generalmente la proporción de respuesta se comporta de manera tal que se genera una curva de tipo S. La función sigmoidal más usada para predecir la probabilidad respectiva es:

$$P_i = E(Y = 1|X_i) = \frac{1}{1 + e^{-(\beta_1 + \sum_{j=2}^{n} \beta_j x_{ij})}} \tag{2.114}$$

Con esta función se genera una razón de probabilidad $\frac{P_i}{1-P_i}$, de tal manera que se llegue a una relación lineal utilizando el logaritmo, es decir:

$$L_i = ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 + \sum_{j=2}^{n} \beta_j x_{ij} \tag{2.115}$$

Lo que argumenta que *"el logaritmo de la razón de probabilidades no es solamente lineal en X, sino también lineal en los parámetros...*L_i *es llamado logit"* [38]. Por ende, hablar tanto del modelo logit como de la regresión logística sería equivalente para el caso binario, sin embargo, la regresión logística contempla una extensión a que variables que puede predecir que son multinomiales, mas para los casos en que se tengan observaciones o instancias repetidas, la estimación de parámetros no se puede hacer de manera usual(debido a la heteroscedasticidad; que se define como el efecto de que las varianzas condicionales de la variable predictora son distintas por instancia; de la perturbación). Se usa el método de los mínimos cuadrados ponderados(MCP), lo que conlleva a que la estimación original 2.115, cambie a:

$$\sqrt{w_i} L_i = \sqrt{w_i}\beta_1 + \sum_{i=2}^{n} \beta_j \sqrt{w_i} X_{ij} \tag{2.116}$$

Pues *"se puede calcular la ecuación 2.116 mediante MCO (Mínimos cuadrados ordinarios), lo cual es un MCP, donde* w_i *son las ponderaciones"* [38]. Entonces se puede sintetizar lo anterior en un algoritmo como el siguiente:

- Paso 1: Ver si la varianza de la variable a explicar puede ser sintetizada en una constante.
- Paso 2: Considerar una función sigmoidal y calcular la forma de regresión (es decir, buscar la linealidad con los parámetros y la variable a predecir).
- Paso 3: Calcular los parámetros β en base a los pasos 1 y 2, utilizando si es posible la máxima verosimilitud.

- Paso 4: Ajustar parámetros o la manera en que evolucionan acorde a los resultados

Un punto relevante es el hecho de que la medida de error o bondad de ajuste R^2, posee valores cuestionables para casos de respuesta dicótoma, pues entrega valores de dispersión no acordes al cálculo, por ende, *"el uso del coeficiente de determinación como estadístico de resumen debe evitarse en modelos con variable dependiente cualitativa"* [38].

Multiclasificadores

Los multiclasificadores, a diferencia de los modelos anteriores, busca encontrar formas o combinaciones de volver una predicción o clasificación efectiva en una predicción eficiente. Para ello se pretende explorar la mayor cantidad de caminos abordables, abarcando toda la información posible. Sin embargo, al buscar esta eficiencia, este tipo de modelos suele caer en algoritmos de gran complejidad, además, puede suceder que el modelo no sea válido a un nivel de fundamentos matemáticos. Por lo que se muestran algunas razones que acreditan las ventajas de su uso [55]:

1) **Razón estadística**: Relacionada con el hecho de que *"en vez de tomar sólo un predictor, una opción más segura sería usarlos todos y promediar sus resultados...tal vez el nuevo predictor no sería mejor que el predictor singular, pero disminuiría o eliminaría el riesgo de tomar un predictor singular inadecuado"* [55]. Esto señala el hecho de que se evitan caminos o predictores que no cumplen las condiciones suficientes, respecto al conjunto de datos, para entregar una predicción válida.

2) **Razón computacional**: Se refiere a que algunos modelos tienen el problema de los mínimos locales, por lo que una ,mala selección de parámetros iniciales, o bien, una mala elección del modelo, puede llevar a predicciones erróneas, que en el caso de que se ocupase un multiclasificador (o multipredictor) se podrían evitar.

3) **Razón representativa**: *"es posible que para el espacio de predictores considerado para un problema, no contenga un predictor óptimo"* [55], en ese caso, el multipredictor o multiclasificador, puede utilizar combinaciones de los predictores de un mismo espacio para solucionar el problema, por ejemplo, si se tiene un problema en donde el comportamiento de los datos es no lineal; usar modelos lineales sería una decisión incorrecta respecto a los resultados que entregaría, no obstante, mediante la combinación de modelos lineales, se puede llegar a una separación óptima incluso para casos no lineales.

Un diagrama de cualquier multipredictor o multiclasificador se observa en la figura 2.24. En donde, los distintos tipos de algoritmos se distinguen en la forma en que abordan los 4 niveles expresados en la figura mencionada.

Uno de los problemas que enfrentará un multiclasificador corresponde a la cantidad de clases que requiera predecir. En este sentido, se puede categorizar un multiclasificador, según como estén clasificadas las instancias, de la siguiente manera [64]: será de una categoría singular *"donde cada caso es conocido por pertenecer exactamente a una de las n clases"* [64] y será de multicategoría, si *"cada caso puede pertenecer a varias, ninguna o incluso todas las clases"* [64]. Eventualmente, cuando se trata del segundo tipo, se puede descomponer el problema en distintos subproblemas

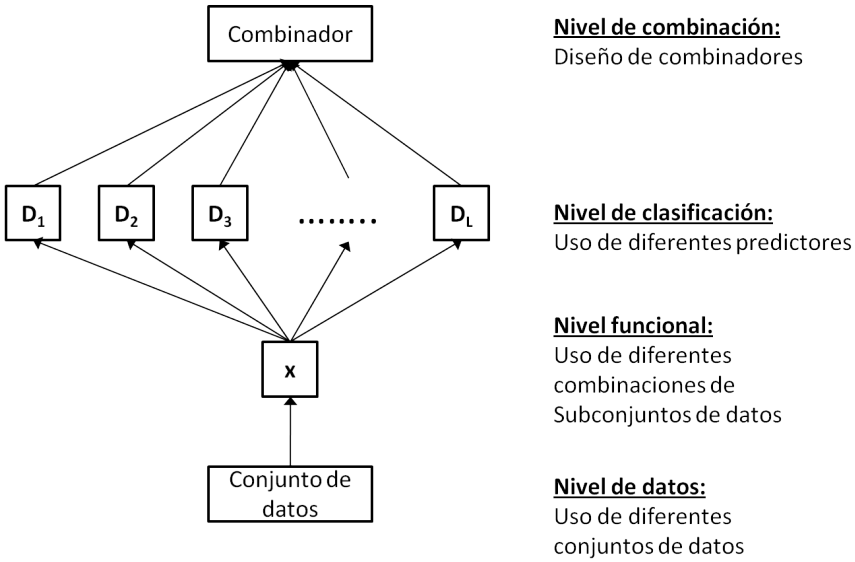

Figura 2.24: Esquema de un modelo multipredictor cualquiera

donde la variable a predecir sea una variable binaria y, de esta forma, proseguir a resolver un problema de clasificación binaria. Dentro de los distintos algoritmos que pertenecen a esta familia de modelos, se pueden apreciar los siguientes 5:

1) **Votación**: El algoritmo de votación consiste en una elección por votos, donde los entes que votan son los clasificadores y sus resultados. Mediante distintas técnicas se puede entregar mayor importancia al voto de un clasificador, o bien, se puede optar por el hecho de que todos posean la misma relevancia al momento de votar, lo que es equivale a señalar que *"provee un método base para combinar clasificadores mediante el promedio de sus probabilidades estimadas (en caso de clasificación) o predicciones numéricas"* [108]. Generalmente, este algoritmo se asocia a un panel de expertos, lo que degenera en un nombre adyacente a "votación de expertos", donde cada experto es un clasificador. Este multiclasificador se puede dividir en dos tipos:

   - **Mayoría de votos**: Este algoritmo consiste en tomar los resultados de todos los expertos y ver la clase ganadora, la cual es aquella clase un número k de las predicciones coinciden entre los expertos. Como se mencionó anteriormente, en el caso de que se tenga un problema de múltiples clases, lo más conveniente es separar este problema en distintos subproblemas. Además, si ocurre que existe un número igual de mayorías, es decir, el 50 por ciento de los expertos dice que un cliente se fuga y el otro 50 por ciento de los expertos totales señala que no, se tiende a asumir una distribución binomial. Ahora bien, aquel número k, es calculado mediante la siguiente fórmula, donde n es el número de expertos [56]:

$$k = \begin{cases} n/2 + 1 & \text{si n es par} \\ \frac{n+1}{2} & \text{si n es impar} \end{cases}$$

En caso de que no se tenga al menos k-expertos, la instancia es rechazada por el algoritmo. Esto puede pasar en el caso de que un experto se niegue a votar, lo que ocurre si el experto no fue capaz de predecir ningún valor.

El problema clásico de votación supone las siguientes tres afirmaciones [56]:

- El número de votantes es impar(de esa forma no se tiene el problema de decidir en caso de que haya un número equitativo k de votos para dos categorías.
- Cada experto tiene la misma probabilidad de votar por una categoría.
- Las decisiones individuales son independientes.

Sin embargo, las dos primeras pueden romperse con algunas restricciones y la tercera puede relajarse. Debe denotarse que *"combinar las decisiones de los expertos no es un sustituto para diseñar mejores clasificadores"* [56].

- **Votación ponderada**: El procedimiento general de la votación ponderada es similar al de la votación por mayoría, sin embargo, su diferencia radica en que se efectúa una aprendizaje mayor a la simpleza de contar votos, en otras palabras, su procedimiento consiste en que *"la instancia se entrega al conjunto de expertos. Cada algoritmo genera una predicción, las cuales son agrupadas en un algoritmo maestro...que recibe la variable objetivo y lo entrega al conjunto de expertos"* [63]. Es decir, la votación ponderada efectúa un aprendizaje sobre la votación de los expertos de manera tal que se pueda determinar la validez de cada experto así como también, la relevancia de su resultados. Su forma básica es:

  - Paso 1: Se asocia un peso positivo a cada algoritmo del conjunto de expertos(Generalmente igual a 1).
  - Paso 2: El algoritmo votación ponderada forma su predicción mediante la comparación del peso total q_0 de los algoritmos del conjunto de expertos que predigan una clase (0, por ejemplo) del peso total q_1 de los algoritmos que predijeron otra clase (1, para el caso binario, nótese que para casos multinominales se puede dicotomizar por cada categoría y separar en distintos problemas de predicción binaria).
  - Paso 3: El algoritmo de votación ponderada predice acorde a la extensión total (en caso de que se tenga una votación empatada). Cuando el algoritmo de votación ponderada se equivoca los pesos de los expertos, cuyo resultado no fue igual al valor de la variable objetivo (en el entrenamiento), son multiplicados por un parámetro β tal que $0 \leq \beta < 1$.

  Cabe destacar que este tipo de votación ponderada solamente modifica los pesos de los expertos cuando un error es cometido(es decir, en los juicios en que se discrepa con la realidad), por ende, se originan otros algoritmos en los que la actualización de los pesos de los expertos puede ser modificado en cada iteración de entrenamiento(juicio), sin embargo, *"mantiene el mismo error de límite que en el caso de que se modifique solamente en las iteraciones que llevan un error"* [63]. Entonces sea la actualización del juicio j, aquella actualización que ocurre cuando se discrepa, si $x_i^{(j)}$ es el resultado del experto i, $w_i^{(j)}$ es el peso del experto i, en el juicio j, y $s^{(j)} = \sum_{i=1}^{n} w_i^{(j)}$, entonces la

predicción del algoritmo maestro en el juicio j $\lambda^{(j)}$ vendrá dada por:

$$\lambda^{(j)} = \frac{\sum_{i=1}^{n} w_i^{(j)} x_i^{(j)}}{s^j} \tag{2.117}$$

El factor para actualizar los pesos va dado por una cota que es una función de x_i, β, dicha cota argumenta que este factor F existe, es decir:

$$w_i^{j+1} = F w_i^j \tag{2.118}$$

Donde F puede ser cualquier factor que satisfaga:

$$\beta^{|x_i^j - \rho^j|} \leq F \leq 1 - (1-\beta)|x_i^j - \rho^j| \tag{2.119}$$

2) **Bagging (Bootstrap Aggregating)**: Es una agrupación de expertos sujetos a un método simple denominado *bootstrap* cuyos resultados son combinados en un voto plural(es decir, cada uno vota por su preferencia). La técnica denominada *bootstrap* se aplica en los casos en que la estimación usual por intervalos de confianza (usando la t-student) o la estimación por máxima verosimilitud no son apropiadas debido a que se margina parte de los datos en el problema inicial. Bajo esta perspectiva, *"el bootstrap usa los datos y el poder computacional para estimar la distribución muestral desconocida"* [28]. Basado en el algoritmo presentado por [28], sea un conjunto de observaciones independientes e idénticamente distribuidas X_i con i=1,...,n; se define un parámetro como una función de los datos poblacionales $\theta = T(x)$ y un estimador cuya función es idéntica, o sea, $\hat{\theta} = T(x)$, entonces el *bootstrap*, estimará la distribución muestra $F_x(x)$ de dicha función T usando los datos. Por consiguiente, las muestras *bootstrap* son elaboradas en forma iterativa sobre la población estimada, así, la función poblacional es evaluada para cada muestra entregando resultados en el formato $\hat{\theta_i^B}$, luego la distribución empírica de dichos valores $\hat{F_B(x)}$ se usa para estimar la distribución teórica $F_\theta(x)$. Finalmente la distribución *bootstrap* es usada para estimar la tendencia (o valor fijo), el error estándar, o bien, para construir el intervalo de confianza. Por ello, el bagging ayuda a evitar el sobreajuste puesto que toma la opinión de uno o más predictores o expertos, los cuales estudian el problema bajo distintas variables evitando caer en estacionalidades. Su algoritmo, expresa dicha cualidad, el cual se muestra a continuación [55]:

- **Fase de entrenamiento**
  - Iniciar párametros. Conjunto de fusión $D = \emptyset$ y L= número de clasificadores o expertos a entrenar.
  - Para k=1,...,L
    - Tomar una muestra *bootstrap* S_k de Z (siendo Z el conjunto de entrenamiento con variable objetivo visible).
    - Construir un clasificador o experto D_k usando S_k como conjunto de entrenamiento.
    - Añadir el clasificador al conjunto fusión, es decir, $D = D \bigcup D_k$
  - Retornar D.

- **Fase de clasificación**
  - Correr $D_1, ..., D_L$ en los datos de entrada x.
  - La clase con mayor votos para una instancia x_i es seleccionada como su categoría objetivo.

Siendo así, el bagging, debe unir los resultados usando diferentes subconjuntos del conjunto general de datos de entrenamiento, los que son escogidos por la técnica *bootstrap*. Para que este modelo tenga alguna utilidad extra, es decir, para que las variaciones en los distintos conjuntos sean usadas, *"el clasificador base debe ser inestable, esto quiere decir que pequeños cambios en el conjunto de entrenamiento conlleven a extensos cambios en los resultados"* [55]. Si bien el bagging *"apunta a desarrollar clasificadores independiente pos la técnica bootstrap... puede que no entreguen resultados independientes"* [55]

Una de las variantes más conocidas del bagging se denomina Random Forest o bosque aleatorio. Un bosque es un conjunto de árboles de decisión, por ende, el Random Forest es un modelo que por cada subconjunto de entrenamiento generado por el bootstrap, y por selección de atributos aleatoria, genera un árbol de decisión, los que posteriormente son llevados a una votación de la misma forma que señala el bagging, para decidir el resultado para una instancia determinada. Lo anterior es la idea general que envuelve al random forest, no obstante, su definición formal es la siguiente: *"un random forest es un clasificador consistente en una colección de clasificadores con estructura de árbol, donde cada árbol crece respecto a un vector aleatorio Q_k, donde Q_k con k=1,...,L, son independientes e idénticamente distribuidos. cada árbol arroja una unidad de voto para la clase más popular de la entrada x"* [55].

3) **Boosting**: El origen del Boosting se encuentra en el algoritmo denominado Hedge, el cual, *"ubica pesos a un conjunto de estrategias para predecir el resultado de un evento determinado...dicho peso s_i se puede interpretar como la probabilidad de que la estrategia asociada a ese peso es la mejor estrategia predictora del conjunto...donde las estrategias con la predicción correcta reciben más pesos que las estrategias con predicciones incorrectas"* [55]. Esta es la esencia del boosting, generar iteraciones de uno o varios modelos sobre un conjunto de datos, castigando las predicciones incorrectas y premiando las correctas, mas esta parte solamente es ejecutable en el entrenamiento para adquirir los pesos más óptimos. De esta manera, puede definirse el Boosting como *"el problema general de producir una regla de predicción bastante certera, mediante la combinación bruta y moderando las reglas que generen incerteza"* [55]. Ahora bien, la forma en que se toman las muestras inicia bajo la consideración de la distribución uniforme, para posteriormente abarcar aquellas instancias de difícil acceso. Cabe señalar, que la distribución es modificada en cada paso.

Si bien no existe un algoritmo claro para el Boosting(aunque su representativo sería el Hedge), generalmente se hace referencia a él por el algoritmo llamado Adaboost, el cual *"tiene dos implementaciones: con iteración de pesos o con iteración de muestreo"* [55], se muestra entonces, a continuación el algoritmo con la implementación de iteración de muestreo [55]:

- **Fase de entrenamiento**
  - Inicializar los parámetros.

- Ajustar los parámetros $w^1 = [w_1, ..., w_N]$,$w_j^1 \in [0,1]$, $\sum_{j=1}^{N} w_j^1$ (usualmente $w_j^1 = \frac{1}{N}$)
- Inicializar conjunto de expertos. $D = \emptyset$.
- Escoger L, el número de expertos o clasificadores a entrenar.

- Para k=1,...,L
  - Tomar un muestreo S_k de Z (conjunto de entrenamiento con variable objetivo), usando la distribución w_k.
  - Construir un clasificador D_k usando S_k como instancia de entrenamiento.
  - Calcula el error ponderado conjunto en el paso k mediante la fórmula:

$$\epsilon_k = \sum_{j=1}^{N} w_j^k l_k^j$$

  - Donde l_k^j se define como:

$$l_k^j = \begin{cases} 1 & \text{si } D_k \text{ clasifica erróneamente a la instancia de entrenamiento } z_j \\ 0 & \text{si no} \end{cases} \tag{2.120}$$

  - Si $\epsilon_k = 0$ o $\epsilon \geq 0{,}5$, ignorar D_k, reinicializar w_j^k a $1/N$ y continua.
  - Si no, calcular

$$\beta_k = \frac{\epsilon}{1-\epsilon}, \quad \text{donde } \epsilon_k \in (0, 0{,}5) \tag{2.121}$$

  - Actualizar los pesos individuales

$$w_j^{k+1} = \frac{w_j^k \beta_k^{1-l_k^j}}{\sum_{i=1}^{N} w_i^k \beta_k^{(1-l_k^i)}} \tag{2.122}$$

- Retornar el conjunto D y $\beta_1, ..., \beta_L$.

- **Fase de clasificación**
  - Calcular el soporte para la clase w_t mediante

$$\mu_t(x) = \sum_{D_k(x)=w_t} Ln\left(\frac{1}{\beta_k}\right) \tag{2.123}$$

  - La clase con el máximo soporte es escogida como valor de la variable objetivo para x.

Cabe destacar que la diferencia entre el boosting y el bagging, es que *"el bagging reduce de forma general la varianza, mientras que el adaboost reduce tanto la tendencia o variable constante y la varianza...también hay evidencia que el bagging es más efectivo que el Adaboost en reducir la varianza"* [106]. Sin embargo, siempre dependerá del problema y el comportamiento en los datos.

2.10.2. Comparativa de principales algoritmos

A continuación se describen brevemente los modelos básicos con sus ventajas y desventajas [29]:

- **SVM (Support Vector Machines)** Usando las funciones Kernel se pueden incluir distintos grados de NO linealidad y de esta manera, flexibilizar el modelo.

  La desventaja de este modelo es que el resultado de la clasificación es puramente dicotómico y no hay probabilidad de pertenencia, además, no se tiene una idea clara de explicación dada la complejidad del modelo en sí.

- **K- vecino más cercano (K-Nearest neighbor)** La ventaja es que los vecinos pueden dar una explicación de los resultados de clasificación. La desventaja de este modelo es que requiere definir una métrica que mida la distancia entre los datos, puesto que no es clara como definir dicha métrica.

- **Árboles de decisión** La desventaja de este modelo está en que las variables continuas o numéricas son implícitamente discretizadas en el proceso de separación, perdiendo información en el camino. No obstante, previo al análisis anterior, los árboles son tolerantes al ruido, a los atributos no significativos y a los valores faltantes; son escalables a grandes volúmenes. Dentro de sus desventajas están la de su imprecisión y su debilidad en el sentido de que 2 muestras distintas sobre la misma distribución pueden llevar a 2 árboles muy diferentes. La ventaja es que establece una explicación clara acerca de la predicción, es decir, su utilidad va directamente relacionada con el hecho de que puede aportar explicaciones asociadas a los clientes fugados, con lo que se puede conocer el perfil del cliente y su comportamiento previo a la fuga de manera práctica y fácil.

- **Regresión Logística** Este modelo es flexible por el hecho de que se pueden incluir términos de interacción, es decir, productos que hagan el modelo no lineal,. La desventaja es que no posee alto grado como las redes neuronales ya que la alta flexibilidad conlleva un alto riesgo de sobreajuste u *overfitting*, lo que reduce el accuracy.

Capítulo 3

Análisis y Resultados

La idea de este capítulo es expresar la experimentación completa del KDD dentro de la empresa de telecomunicaciones, mediante la explicación de cada experimento aplicado, para luego converger al resultado implementado final. Dentro de dicha explicación se observará que cada etapa del KDD sufrió cambios en cada experimento, por ende, los resultados no son del todo comparables, mas las estrategias ejecutadas sí lo son, lo que se debe primordialmente a la complejidad de configuraciones usadas en las distintas etapas del KDD.

Otro alcance frente a este capítulo es el hecho de que solamente dos modelos fueron probados de forma efectiva en el negocio. A esto se agrega el hecho de que existe un estudio histórico de un modelo en particular, lo cual es considerado como tal debido a la múltiple aplicación del KDD para cada mes. Con ello, se pretende mostrar si la hipótesis, de que el modelo estudiado sea acanónico, es cierta. Este último concepto se basa en el hecho de que un modelo de datos no sea canónico, lo que se define como *"una relación matemática, donde un atributo es una relación de equivalencia y el valor de un atributo es una clase de equivalencia exceptuando la variable objetivo"* [61]. A partir de esto, se puede definir los modelos acanónicos como aquellos que varían según el tiempo en el que se aplican. Otra perspectiva del concepto es contemplando las bases de datos con las siguientes características: un modelo de datos canónico es aquel que *"está normalizado (en quinta forma normal), no tiene datos redundantes, y se preservan las dependencias de los datos"* [103], por ende, retoma lo anterior que implica que un modelo acanónico variará las relaciones de los datos acorde al tiempo. Cuando se presentan circunstancias sujetas a externalidades, como se dan en la lógica de los negocios, la tercera característica queda incumplida. Por ejemplo, el terremoto que afectó la zona donde reside la compañía de telecomunicaciones, influyó los resultados finales obtenidos en los experimentos que contemplan en su horizonte de datos dicho efecto, mientras que en aquellos cuyo horizonte de datos no se acerca a dicha externalidad no es perceptible. Nótese que todos los experimentos tienen un horizonte de datos igual a 6 meses y que la distinción horizonte de datos se conceptualiza como aquel período contemplado en la integración de datos en el KDD.

El objetivo de la descripción individual de cada experimento es poder expresar la evolución de un proyecto KDD de principio a fin. Finalmente se presentan los resultados y las principales ventajas y desventajas de cada experimento.

Además, se recuerda que este problema está relacionado con el aprendizaje supervisado y con una variable objetivo de carácter binario. A lo cual se añade la existencia de una experiencia anterior

o tesis base en la compañía la cual plantea un modelo de árbol de decisión con un conjunto de datos determinado.

Una vista general al procedimiento implementado en esta memoria se puede apreciar en la figura 3.1:

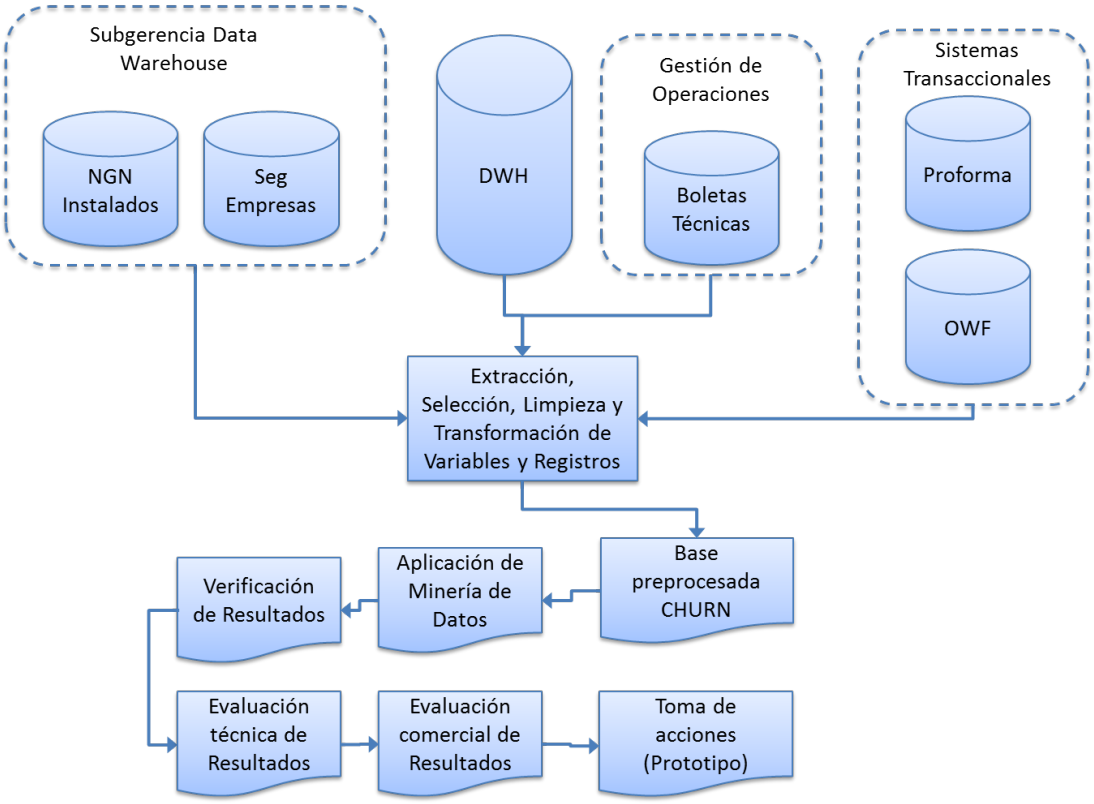

Figura 3.1: Procedimiento KDD en el piloto

Donde la vista general de las variables por experimento, así como su descripción respectiva, se puede apreciar en la tabla 5.1. Cabe destacar que la nomenclatura a usar a continuación para distinguir entre bases de datos y variables será la siguiente: para las variables se aplicará letra cursiva, es decir, variable *Fuga*. En cambio, las palabras referentes a las bases de datos se destacarán en negrita, por ejemplo, base de datos **NGN_Instalados**.

3.1. Fase 1: Exploración

El primer experimento se resume en la palabra exploración, esto se debe a que se asumieron ciertos supuestos teóricos que son válidos al momento de decidir la creación de un proyecto que contenga el procedimiento KDD como elemento fundamental. Además, como se mencionó anteriormente, se comparan los distintos modelos válidos en la tesis que se tiene como experiencia anterior. Con ello se hacen los descubrimientos iniciales y se comienza la fabricación de lo que

será el primer modelo. En esta etapa no ocurre ningún tipo de valorización de la fuga de los clientes, puesto que recién existe un inicio en la integración con el negocio. No obstante, se comienza a visualizar las principales bases a integrar en los experimentos. Sin embargo, el primer experimento debe dividirse en dos: la primera división trata acerca de la experimentación de modelos sobre el conjunto de datos utilizado en la tesis [77], es decir, un conjunto que data del año 2008. La segunda división del experimento, busca actualizar la tesis base y verificar si su validez se mantiene en el período actual.

3.1.1. Experiencia anterior a partir de la tesis base

El experimento de la tesis base consiste en tomar el conjunto de datos del año 2008, usado como entrada final para el modelo y comparar los distintos algoritmos que pudieron haber sido utilizados para dicho propósito. Un punto relevante es el hecho de que este modelo no poseyó una etapa de validación, es decir, no se probó si la predicción era acorde con el mercado, por lo tanto, no se pudo averiguar si existieron problemas como el sobreajuste o el subajuste. Sin embargo, resultó positivo, el hecho de contar con un piso teórico-práctico, en cuanto al conocimiento de las bases, la maduración de perspectiva en las estrategias a ejecutar y al abordaje del problema. La duración *"para la extracción de la muestra del estudio fue de un período de ocho meses"* [77], cuando el producto en cuestión recién se encontraba en etapa de maduración. Ahora bien, no existió un proyecto ni una iniciativa como empresa al momento de la elaboración de la tesis base, por ende, existieron validaciones generales que no pudieron concretarse, ni tampoco poner en marcha una serie de acciones correctivas en base a esto. No obstante, *"se contó con el apoyo de la Dirección de Data Warehouse"* [77], cuya dirección tiene a cargo *"gran parte de la información de negocio de la compañía"* [77], es decir, contiene la consolidación del negocio en cuanto a la información. Un punto a destacar, es el hecho de que en la época en la que se ejecutó este experimento la dirección de Data Warehouse recién estaba asentándose en la empresa, por lo que las bases de datos presentaron algunas inconsistencias que debieron ser resueltas por la tesista a través de los sistemas transaccionales.

Una breve descripción del procedimiento KDD implementado en dicha ocasión, se describe 5.18 basado en [77].

Integración

En esta etapa, se acude a las distintas áreas encargadas de ciertos datos, debido a que en esa época la empresa de telecomunicaciones no contaba con un Data Warehouse establecido, por lo tanto, la consolidación de datos solamente pudo ser efectuada mediante entrevistas informales, es decir, confiar en lo que el personal de dicha área señalaba. Las principales bases que se usaron en la experiencia corresponden a:

- **Segmentación de empresas**
- **NGN Instalados**
- **Proforma**
- **Boletas Técnicas**

Cuyas descripciones se encuentran en la sección 5.1.1 en los Anexos.

Cuadro 3.1: Categorización Variable fuga en base a variable Estado_Cliente

Categorías Variable Estado_Cliente	Categorías Variable Fuga
Vigente	No
Voluntario	Sí
Involuntario	No
Terminado	No

Preprocesamiento

- Se eliminaron 11 casos duplicados (2 con instancias duplicadas iguales, 9 instancias con valores distintos en un atributo e igual en el resto).
- 6 Registros con estado de fuga vigente, pero con ID no válido, a esto se le agrega el hecho de que los 6 registros presentan al menos un 45 % de información incompleta. Cabe señalar que la tesista consultó con los expertos del negocio antes de eliminar dichos registros.
- Se crea la variable *Fuga* en base a los valores de la variable *Estado_Cliente* (cuya descripción se encuentra en la tabla 5.1), como se muestra a continuación:

  La lógica de esta categorización es que sólo se considerarán los valores VIGENTE y VOLUNTARIO, pues en el INVOLUNTARIO la empresa decide acabar el contrato y en el caso TERMINADO ya no hay una relación con el cliente. Acorde a esto se eliminan 1056 casos. Quedando en 6538 registros.
- Los valores incompletos se cambian todos a valores perdidos, menos el caso de la categoría NO IDENTIFICADO que se presenta en el atributo Canal, que tiene una interpretación válida.
- Se revisan los datos en forma de gráficos de las distribuciones de las variables.
- Se efectúa un análisis descriptivo de las variables numéricas y nominales.
- Se elabora tabla de contingencia para ver la completitud de los datos.
- Se detectan 60 casos anómalos, cuyo error está en el número de identificación del cliente. Estas instancias son eliminadas debido a que no pertenecen al producto NGN, lo que deja la cuenta de los datos de estudio en 6478.
- La variable *Segmento* adquiere varianza nula, ya que no aporta información a la investigación, por lo que no se considerará en los análisis posteriores.
- A pesar de que la variable *GIRO* tenga un porcentaje de valores perdidos igual a 58 % se decide no eliminar, ni modificar dado que es una variable externa dependiente del servicio de impuestos internos. No se recomienda dejar en blanco para el caso actual, dado que se debe disminuir al máximo los valores perdidos para evitar el ruido en el modelo a usar.
- Se eliminan registros por alta cantidad de datos ausentes, dejando en una muestra total de 6473 casos.

Transformación

- Se cambia la variable numérica *Sucursales* a nominal, debido a la presencia de altos valores considerados como fuera de rango (*outliers*) desde la perspectiva del cliente objetivo del producto. A su vez, se discretiza según intervalos, los cuales se describen a continuación [77]:

Cuadro 3.2: Tabla de Transformación de la variable *Sucursales*

Categoría Sucursales	Rangos en Cantidad de Sucursales
SUC 1	≤ 1
SUC 2	> 1 y ≤ 6
SUC 3	> 6 y ≤ 15
SUC 4	> 15 y ≤ 25
SUC 5	> 25 y ≤ 50
SUC 6	> 50

- Las variables asociadas al consumo y facturación se tratan con el propósito de reducir la dimensionalidad de este grupo de variables. Es así como se determina llevar las 12 variables consideradas (6 de consumo, 6 de facturación) al promedio respectivo y a una variable que es formada de la siguiente manera: En un principio se crea una variable "Var_{ij}" que nominaliza la variación entre el mes i- y el mes i, donde j está asociado a dos valores Facturación, Consumo, esta variable *Var*, se define como [77]:

$$Var_{ij} = \begin{cases} 0 & \text{si la variable j disminuyó entre } mes_{i-1} \text{ y el } mes_i \\ 1 & \text{si la variable j se mantuvo entre el } mes_{i-1} \text{ y el } mes_i \\ 2 & \text{si la variable j aumentó entre el } mes_{i-1} \text{ y el } mes_i \end{cases}$$

  Finalmente todo lo anterior se comprime en:

$$IDCURVAj = \sum_{i=1}^{5} Var_{ij} * 3^{5-i} \tag{3.1}$$

- Se crea una variable extra denominada *INGRESO CONT POST INST* cuyos valores son "SI" o "NO". Esta variable *"representa si el cliente tuvo una instalación del servicio con anterioridad al ingreso del contrato"* [77] y su formulación respectiva es:

$$ICPI^{1} = \begin{cases} SI & \text{si fecha variable } \textit{PRIMERA INSTALACION} < \text{fecha variable } \textit{INGRESO CONTRATO SAP} \\ NO & \text{si fecha variable } \textit{PRIMERA INSTALACION} \geq \text{fecha variable } \textit{INGRESO CONTRATO SAP} \end{cases}$$

- Análoga a la variable anterior, *DIAS INST CONT* indica *"el número de días transcurrido entre el ingreso del contrato y la instalación del servicio"* [77]

[1] INGRESO CONT POST INST

- Agregado a las variables *ID CURVA*, se crean las variables *Promedio consumo* y *Promedio facturación*.

- La última transformación empleada, es la aplicada a la variable *ICP* (Índice de comportamiento de pago), la cual consiste en una discretización de la misma. Su definición práctica es:

$$TIPO\_ICP = \begin{cases} \text{Riesgo Alto, si } ICP \text{ tiene un valor menor a 0,98} \\ \text{Riesgo Bajo, caso contrario} \end{cases}$$

Modelo

El modelo que se usa en la tesis base corresponde al de Árboles de decisión J4.8 que es la implementación práctica en el software WEKA del algoritmo teórico C4.5, creado por Quinlan en [81], que tiene las siguientes peculiaridades:

- Usa medidas de información para efectuar los *splits*, es decir, *Gain Information ratio*, Entropía, *Gini*, entre otras.

- A diferencia del ID3 [2] que usa un método *listwise* para los valores perdidos, maneja valores continuos y valores ausentes reemplazando los valores por métodos de preprocesamiento, dicha forma se encuentra descrita en los capítulos 2 y 3 de [81].

- Para la poda del árbol usa el Error-Based Prunning y el Cost-Complexity(este último consiste en evaluar cada poda según el costo en términos de información que significa la pérdida de un nodo hoja, para mayor información consultar [55, 88]).

- Asume que las clases (de la clasificación) son disjuntas, lo cual puede traer problemas si se está en presencia de instancias muy cercanas difíciles de clasificar.

Generalmente, el modelo de árboles utiliza particiones por lo que una vez elegida alguna (el subconjunto y el o los atributos escogidos que separaran a los nodos parentales) no se vuelve a evaluar la alternativa, es decir, que el criterio de partición juega un papel rígido que impide el ingreso de datos muy aleatorios o variables (de varianza) . No obstante, la particularidad del J4.8 es que las particiones no son rígidas, esto quiere decir que puede volver a evaluar otros atributos. Respecto al criterio de poda, el J4.8 utiliza el método de reducción de error al igual que el C4.5. Dentro de los criterios de partición del árbol se usan el error esperado, Gini, la entropía, entre otros [77].

El modelo J4.8 se prueba con un set de 5.795 clientes vigentes y 678 clientes fugados voluntarios, con 34 atributos (incluído aquel que representa la fuga), para un horizonte de 8 meses. A su vez también se testea el modelo Naive Bayes para tener una comparativa.

Posteriormente se usan ténicas que vienen incorporadas en el software WEKA para efectuar una selección de atributos e instancias, dentro de estas técnicas *"existen de dos tipos: evaluadores de subconjuntos o selectores (SubsetEval), los que necesitan elegir un método de búsqueda (SearchMethod) de los subconjuntos, y los prorrateadores de atributos (AtributteEval) que se combinan con*

[2] Un modelo de árbol de decisión para problemas supervisados binarios, que usa la entropía como medida para los *splits*

un "Ranker", ya que no seleccionan atributos, sino los ordenan por relevancia" [77]. Se seleccionan tres opciones: *GainRatioAttributeEval*, *CfsSubsetEval* y *ConsistencySubsetEval*, denominados J4.8 GR, J4.8 CFS y J4.8 CS respectivamente en la tabla 3.3).

Evaluación

En la etapa, la experiencia anterior efectúa una de las tres fases de validación, la técnica, es decir, se entrena con 8 meses para verificar la predicción de la fuga del noveno que ya viene etiquetada. Para ello, el modelo se entrena y prueba con la técnica de *Cross-Validation* o validación cruzada.

De esta manera, se muestran los resultados finales con las medidas de evaluación correspondientes en la tabla 3.3:

Cuadro 3.3: Tabla de resultados: Experiencia anterior

Criterios Técnicos	Modelos				
	J4.8	Naive Bayes	J4.8 GR	J4.8 CFS	J4.8 CS
Accuracy	90.95 %	31.07 %	90.638 %	90.561 %	90.746 %
Medida F	70.10 %	55 %	68.151 %	67.639 %	68.601 %
TP	184	608	139	125	136
TN	5703	1400	5728	5737	5738
FP	92	4385	67	58	57
FN	494	70	539	553	542

Se puede apreciar que el modelo J4.8 original es superior al resto de configuraciones y modelos utilizados, pues el *accuracy* es superior, además, se acepta el modelo dado que su medida ROC (que no depende del criterio de corte) es mayor a 0,70, lo cual es aceptable en el mercado del KDD.

Cabe señalar que ninguno de los J4.8 modificados supera al original en cuanto a la precisión de la predicción, no obstante, igual se tiene un 90 % de precisión y de *accuracy*. Esta tesis base contiene un apartado adicional de un modelamiento del churn considerando el tema de los reclamos solamente, estableciendo una estrategia base para esta memoria.

3.1.2. Experimento Análogo

En un inicio, para analizar la posibilidad de aplicar el KDD al producto en la actualidad (es decir, dos años después de la experiencia anterior), se efectúa; con la misma base preprocesada y transformada de la tesis base; una serie de pruebas con otros modelos que vienen incorporados en Rapidminer, como redes neuronales, árboles de decisión como el LADTree, que es una expansión del ADTree, el cual toma un conjunto de condiciones y sus negaciones respectivas y va dividiendo el árbol afectando la puntuación final de la predicción. Sin embargo, su diferencia radica en que la clase que el árbol de decisión estándar asigna es la mayoría de votos en el nodo hoja en el que termina una instancia determinada, en cambio, el ADTree considera que cada condición o regla tenga una puntuación propia, de modo, que al sumar todas las puntuaciones de la reglas (más una puntuación base propia de la instancia) se pueda decidir mediante una función (que para el caso binario es el

signo), la clase a asignar para una instancia en general [35]. Por consiguiente, el LADTree es definido para problemas con multiclase (más de una categoría en la variable objetivo) como problema binario al cual se le aplica el algoritmo LogitBoost, es decir, si se tienen J clases, se propone la siguiente fórmula para el cálculo de la clase probable:

$$p_j(x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}, \quad \sum_{k=1}^{J} F_k(x) = 0 \tag{3.2}$$

Ahora bien, el LogitBoost se puede aplicar al ADTree de dos formas, *"la primera...llamada LT1PC construye árboles (ADTree) separados para cada clase en paralelo. La segunda, solamente se construye un árbol prediciendo todas las clases probables simultáneamente"* [41], cada una es explicada con profundidad en [41]. En la siguiente tabla muestra los principales resultados:

Cuadro 3.4: Tabla de resultados: Experimento análogo

Criterios Técnicos	Modelos				
	J48	LADTree	Random Forest	Naive Bayes	Red Neuronal
Accuracy	90.66	90.46	89.27	89.27	89.2
Medida F	33.16	39.13	Desconocida	Desconocida	3.52
AUC	0.672	0.76	0.5	0.5	0.676
Lift	681.28	589.01	Desconocida	Desconocida	388.48
TN	2239	2215	2264	2264	2257
TP	60	79	0	0	5
FN	212	193	272	272	267
FP	25	49	0	0	7
Total muestra	2536				

En base a la tabla 3.4 se puede apreciar que existen dos modelos superiores al J4.8 en base a la medida AUC: el LADTree y las redes neuronales lo que constata que el modelo anterior se puede mejorar y da cabida a que se efectúe un segundo estudio.

3.2. Fase 2: Estableciendo Estrategia

Una vez explorada la experiencia anterior existente, se procede a analizar las bases de datos, donde se visualiza un cambio en el atributo fuga. Dentro de la experiencia anterior, se detallaba que esta variable le fue entregada directamente en la base **NGN_Instalados**, sin embargo, no se había estudiado su procedencia. A esto, se agrega el hecho de que el producto se había establecido en los dos años transcurridos, lo que se expresó en una caída en la cantidad de fugados. Si bien en el 2008, cuando el producto recién estaba ingresando al mercado tenía un churn de aproximadamente 10 %, en el 2010 ese mismo churn se encontraba en un 1 %, según estadísticas internas a la compañía. Por ende, la forma en que se aborda el problema es diferente a la experiencia anterior, en términos

de estrategia; puesto que en el 2010 existe un problema de desbalanceo o rarezas en las clases[3]. Además, se comienza a dilucidar el tipo de churn que se calcula en la compañía, el cual es a nivel de productos de servicios, en particular de accesos caídos. Esto quiere decir que no se mide si el cliente deja el servicio, y por ende, tampoco se mide si el cliente deja la compañía completamente.

En términos del problema abarcado por la memoria, solamente se puede llegar al nivel de medir el churn del servicio, esto se debe a que el cliente en la compañía tiene más de un producto, incluso tratándose de un cliente pequeño.

3.2.1. Experimento 1: Actualización de experiencia anterior

Este experimento consta como parte de la exploración de las bases y del producto, así como también, busca observar la efectividad del procedimiento antiguo si se aplica tal cual. El período de tiempo asignado a este experimento es de dos meses para su realización contemplando 6 meses como período de tiempo de estudio en las bases(y no 8 como se efectuó en la experiencia anterior) tal como se muestra en la tabla 1.3. Esta cifra se decide en conjunto con el personal de la compañía y su razón principal es la disponibilidad de las bases según fecha. No obstante, a pesar de que este experimento es una actualización del procedimiento anterior, también corresponde a una puerta de exploración hacia posibles transformaciones y tratamientos de datos para futuros experimentos.

Integración y Selección

La primera idea al momento de intentar refinar el modelo de la experiencia anterior es partir por las fuentes de bases. En particular volver a calcular la variable *Fuga*, debido a que mediante entrevistas informales con el personal, se descubre que no hay información consistente sobre la existencia de un atributo que describió la fuga previa. Por ello, se comienza a investigar sistema por sistema, base por base, esta variable, llegando así, a dos bases que posiblemente pueden contenerla. La primera es un extracto de 172.977 registros proveniente de la base de la plataforma mySap (ver figura 1.10), la segunda es una base trabajada directamente por el product Manager de NGN, con 9111 registros. Ambas presentan una cantidad distinta a la experiencia anterior en donde se tomaron aproximadamente 7000 registros inicialmente.

Una vez escogida la base de datos, se empieza la búsqueda de las variables especificadas en la experiencia anterior, no obstante, se descubre una separación entre clientes no mencionada anteriormente. Dicha categorización separa a los clientes del producto en 2 fases (0 y 1), esto se debe a una migración de los datos de una plataforma antigua a una nueva. Los clientes de fase 0, poseen gran parte de sus datos distribuidos en plataformas diferentes de la compañía, mientras los datos de los clientes fase 1 se encuentran en la plataforma MySap. A pesar de esto, se decide considerar ambos casos. Para ello, se toman las variables de *Giro, Comuna, Planes (la descripción de los mismo)*, el estado del contrato en el sistema (en la variable *estado_contrato_sap*), la cantidad de contratos vigentes, anulados, renunciados y nuevos que cada cliente posee, y el rut, una descripción detallada de estas variables se puede encontrar en la tabla 5.1. Sin embargo, una particularidad de algunos clientes se dispuso en esta ocasión, la cual consiste en que el rut no está constituido de un solo

[3]Definido como la ocasión en que un valor de la variable objetivo se encuentra fuera del común denominador o en los extremos de la distribución de la variable objetivo. En el caso de que la variable objetivo sea nominal, se refiere a aquella categoría con una frecuencia despreciable respecto a las otras categorías.

campo (rut completo), tampoco de dos (cuerpo del rut y rut verificador), si no que se encuentra sólo el cuerpo del rut pero no el dígito verificador, esto conlleva a usar una transformación interna a la compañía para obtener el dígito verificador a partir del cuerpo del rut.

Otra base de datos considerada es la de las **boletas técnicas** (explicada en la sección 5.1.1) . En ella, se tomó el mismo campo que en la experiencia anterior, en donde se consideraba la variable *Resp_falla* (que indica si el cliente o la compañía es responsable de la falla, tal como se puede apreciar en la tabla 5.1), y el rut como llave foránea.

La base **Proforma** (caracterizada en la sección 5.1.1), que contiene los registros de facturación mes a mes, viene en un formato muy adecuado para la extracción, es decir, presenta variables como el rut, el monto de facturación del producto y los segundos consumidos por el usuario, por lo tanto las variables seleccionadas en esta base son: *Facturación(monto de facturación total), la cantidad de minutos totales hablados(Consumo)* y el rut como llave foránea.

La base **segmentación de empresas (o seg_empresas)** (cuya definición se encuentra en la sección 5.1.1), es usada indirectamente en este experimento, dado que sus variables se obtienen a partir de los datos de la plataforma MySap, que es el origen de la base previamente mencionada, las variables seleccionadas se pueden apreciar en la tabla 5.1.

La base **NGN_Instalados**, que es la base que usa el product manager del producto NGN, se usa en este experimento para determinar las variables relacionadas con las sucursales, no para establecer la cantidad de ruts vigentes. Dentro de estas variables se encuentran: *Sucursales, sucursales con y sin banda ancha, cantidad de accesos, y cantidad de centros primarios* (que señala si el cliente tiene sucursales fuera de la región metropolitana), entre otras.

La variable *ICP* (descrita en la tabla 5.1) para mayor detalle, fue obtenida de forma externa, por petición específica a un miembro del área Data Warehouse, encargado de manejar dicha variable.

Sin embargo, a la mitad del experimento, se descubre que la base de datos de clientes, aquella que contiene la llave principal, proveniente de la plataforma MySap posee los registros duplicados, de esta manera, todo rut tiene asociado más de un contrato (cuando en la práctica no es así), por lo que se opta por utilizar una nueva base de datos limpiada, extraída nuevamente de la plataforma MySap, en pos de efectuar un reemplazo de la anterior. Con esta última se efectúa la totalidad del experimento. Esta nueva base al ser combinada con los clientes de la fase 0, da un total de 11439 instancias.

Preprocesamiento

Registros En el preprocesamiento de los registros se consideran aquellos clientes que están vigentes (es decir, están activados y facturando) o que están finalizados(o sea, que todos sus contratos están finalizados), en otras palabras se discriminan aquellos clientes que solamente tienen contratos anulados o bien, que son clientes nuevos. De esta manera, las 11439 instancias se redujeron a 8955, no obstante, también se considera la opinión de un experto y trabajador actual de la compañía que sugiere que el producto está directamente relacionado con las empresas PYME (Pequeñas y medianas empresas), por ende, las grandes empresas, corporaciones y mayoristas no deben ser contempladas. En base a lo anterior y dada la obtención de una base denominada Ruts *Friendly User*[4], corpora-

[4]Friendly User: es un tipo de rut ficticio utilizado generalmente para probar el comportamiento de las plataformas y los sistemas en la introducción de nuevos productos [12]

ciones y mayoristas; que se describe en el apartado de bases de datos de los anexos; se discriminan aquellos ruts catalogados como *Friendly User*, obteniéndose 8924 registros para el trabajo.

Variables Para el preprocesamiento de las variables se debe observar la naturaleza de la misma, puesto que de esta característica depende el tipo de reemplazo a efectuar. Para explorar el estado actual del comportamiento de los clientes, se opta por realizar los mismos pasos efectuados por la experiencia anterior. Nótese que en esta sección solamente se analizan los reemplazos de los valores ausentes, esto último se debe a que los valores fuera de rango no son considerados en el experimento anterior, además, se debe adquirir un mayor conocimiento acerca del producto para elaborar un criterio válido en la etiquetación de determinados casos como *outliers*. Se agrega que para los valores incompletos se opta por transformarlos a valores perdidos o *missing* para simplificar la utilización de criterios disponibles para estos últimos. De esta manera, se procede a describir el preprocesamiento efectuado en este experimento:

- Se reemplazan los datos *missing*(o valores perdidos) en la variable *Consumo*, a cero. Se hace lo mismo para los valores "DIV0". Este tratamiento fue análogo para la variable *facturación*. La justificación de este reemplazo radica en que el indicador presentado en el preprocesamiento de la experiencia anterior sólo busca explicar la intensidad de facturación o consumo y su movilidad en el tiempo, es decir, si se mantiene, aumenta o disminuye, además, esta variable presenta valores perdidos del tipo NMAR [5] respecto a que la ausencia solamente implica que el cliente no factura o consumo. Cabe decir, que se hace una modificación leve al tema del indicador, es decir, en la experiencia anterior se señalaba que si la variable se mantenía, el indicador Var, exponente del indicador original, tenía el valor 1. Para que este valor no se tome como un valor sin sentido (donde pocas instancias iban a adquirirlo) se establece una vecindad en cuanto a un porcentaje de variación, es decir, si las ventas del mes actual tuvieron una variación menor al 1 % de las ventas del mes pasado, entonces se tomaba como si el consumo o facturación se mantenía para un cliente determinado. Ahora bien, una tabla con la cantidad de valores perdidos para cada variable de la base **Proforma** se muestra a continuación:

  Esta tabla bosqueja la cantidad de valores ausentes o perdidos para las 12 variables originales extraídas de la base **Proforma**, en donde, se nombran como $Fact_i$ o $Cons_i$ con el índice i correspondiente al mes número x antes del tiempo contemplado en el experimento, es decir, si el experimento 1 se efectúa en base a Marzo del año 2010, entonces el mes número 5 será Octubre del año 2009 y Marzo del 2010 sería el mes número 1. Además, Fact se refiere a facturación y Cons, es el consumo reflejado en minutos ocupados (para lo cual simplemente se divide el consumo obtenido directamente de la **Proforma** por 60).

  Se puede observar que los vacíos de estas dos variables presentan una sincronía en cantidad y en las instancias en las que se dan, es decir, cuando alguna de estas dos variables para un mes determinado, está ausente; su análogo para el mes correspondiente también se encuentra ausente, en otras palabras, si $Fact_5$ está ausente para una instancia determinada, entonces $Cons_5$ estará ausente para esa misma instancia, por ende, las cantidades de valores perdidos entre meses coinciden.

[5] *Not missing at random* descrito previamente en la sección 2.5.3

Cuadro 3.5: Tabla de resultados: Valores perdidos de variables provenientes de la base de datos **Proforma**

Variables	Valores Perdidos	Estrategia
Fact6	3796	Reemplazo por valor 0
Fact5	3815	Reemplazo por valor 0
Fact4	3841	Reemplazo por valor 0
Fact3	3801	Reemplazo por valor 0
Fact2	3795	Reemplazo por valor 0
Fact1	3823	Reemplazo por valor 0
Cons6	3796	Reemplazo por valor 0
Cons5	3815	Reemplazo por valor 0
Cons4	3841	Reemplazo por valor 0
Cons3	3801	Reemplazo por valor 0
Cons2	3795	Reemplazo por valor 0
Cons1	3823	Reemplazo por valor 0

- Se agregan las variables de reclamos técnicos con responsabilidades respectivas dentro del período considerado, es decir, desde el mes de Octubre del año 2009 hasta el mes de Marzo del año 2010. Estas variables poseen gran cantidad de valores perdidos, sin embargo, se realiza el mismo reemplazo anterior por el valor 0, debido a que si el cliente no tiene valor en las variables dedicadas al reclamo técnico, implica que el cliente simplemente no reclamó. La tabla que describe la cantidad de valores perdidos y la estrategia aplicada es bosquejada en la tabla 3.6.

Cuadro 3.6: Tabla de resultados: Valores perdidos de variables relacionadas con los reclamos técnicos

Variables	Valores Perdidos	Estrategia
Ent_Oct09	7940	Reemplazo por valor 0
Ent_Nov09	8031	Reemplazo por valor 1
Ent_Dic09	7998	Reemplazo por valor 2
Ent_Ene10	8070	Reemplazo por valor 3
Ent_Feb10	8132	Reemplazo por valor 4
Ent_Mar10	7575	Reemplazo por valor 5
Cl_Oct09	8540	Reemplazo por valor 6
Cl_Nov09	8531	Reemplazo por valor 7
Cl_Dic09	8502	Reemplazo por valor 8
Cl_Ene10	8493	Reemplazo por valor 9
Cl_Feb10	8562	Reemplazo por valor 10
Cl_Mar10	8477	Reemplazo por valor 11

- Se corre un análisis anova, tomando como factor la variable *Sucursales* y aplicando sobre el resto de las variables, donde el resultado más relevante se muestra en la tabla 3.7, la cual significa que la variable *ICP*, depende o es muy similar a la variable *Sucursales*. Cabe señalar, que ambas variables son continuas y que, a pesar de lo anterior, se utilizan ambas variables para el experimento 1.

Cuadro 3.7: Análisis ANOVA para la variable *ICP*

ANOVA con factor Sucursales		Suma de cuadrados	Gl	Media cuadrática	F	Sig.
ICP	Inter-grupos	1.731.904	50	34.638	0,123	1,00
	Intra-grupos	2.093.113.447	7.45	280.955		
	Total	2.094.845.351	7.5			

- Ahora bien, para el caso de la variable *ICP*, se cuenta con 1423 valores perdidos, los cuales son desconocidos de tipo MCAR[6]. Se observan las correlaciones en las tablas 5.3, 5.4, 5.5 entregando como producto que ninguna variable que relacionaba linealmente con el *ICP*, lo que indica que esta variable si bien tiene un cálculo a partir de otras variables no incluídas en el estudio, no se relaciona con ninguna a nivel lineal (justificando la etiquetación de sus valores perdidos como MCAR) . Por consiguiente, se observa la naturaleza estrátegica de dicha variable, es decir, el hecho de que no se sepa el valor del ICP ya presenta un alto riesgo, por lo que lo más aplicable es que se reemplace por el promedio de los valores que efectivamente son riesgo alto. Esto da un total de 61,16, el cual fue reemplazado en los 1423 registros ausentes en la variable *ICP*.

- También se analiza el caso de las correlaciones (en valor absoluto) válidas según los test Pearson, cuyos resultados completos también se encuentran en las tablas 5.3, 5.4, 5.5. A continuación, se presenta la tabla de correlaciones entre la variable objetivo y las variables que presentaba un valor de correlación significativo mayor al 50 %:

Cuadro 3.8: Tabla de Correlaciones con variable *Fuga*

Variables	FUGA_NUM
ID_CURVA_CONSUMO	0,530
ID_CURVA_FACTURACION	0,538
Nom_CAN	0,502

- Cabe señalar que estas correlaciones puede que potencien algunos modelos como la regresión, no obstante, el valor que tienen no es suficiente para considerarlas como amenaza. De todos modos, se muestra a continuación la variable que presentó mayores correlaciones entre sus pares, la cual fue la variable *Sucursales*:

[6] *Missing completely at random* descrito previamente en la sección 2.5.3

Cuadro 3.9: Tabla de variables con correlaciones relevantes respecto a la variable *Sucursales*

Variables	Sucursales
CANT_CPO	0,708
PROMEDIO_CONSUMO	0,619
Q_ACCESOS_BA	0,807
Q_ANIS	0,804
Q_RECL_TEC_RESP_EMPRESA	0,636
Q_SUCURSALES_CON_BA	0,807
Q_SUCURSALES_SIN_BA	0,694

La alta correlación que existe entre *Sucursales* y los *Q_SUCURSALES* se refiere a que el valor de *Sucursales* puede ser obtenido a partir de las dos variables que poseen el nombre señalado anteriormente como prefijo. No obstante, se opta por dejar aparte una de las variables extras, *Q_SUCURSALES_CON_BA*, puesto que de esta forma no se pierde tanta información (expresada en la varianza), además, puede ser determinada a partir de otras dos presentes en el negocio. Esto se diferencia con el preprocesamiento de la experiencia anterior, donde se dejaba de lado la variable *Q_Accesos_BA* por tener una correlación de 1 con la variable *Q_SUCURSALES_CON_BA*.

- Se descarta la variable *GIRO* debido a su excesivo ruido (más del 60 % de datos incompletos), además, al realizar la prueba del modelo J4.8 en la experiencia anterior, se detecta que esta variable no era necesaria para producir resultados similares al expresado en la tesis base, sin embargo, en entrevistas informales con el personal encargado del producto, se declara desde un punto de vista subjetivo que esta variable resulta relevante dado que describe el rubro o la actividad a la cual se dedica el cliente.

- Para la variable *PLANES*, el tratamiento es describirlos por los atributos que contienen, no obstante, dado el tiempo disponible en el proyecto a efectuar, se opta simplemente por considerar el plan "más contratado" por el cliente, es decir, el plan con mayor presencia en los contratos pertenecientes a las transacciones del cliente. Posterior a ello, se detecta la presencia de valores nulos, esto quiere decir que no existe un plan presente en ningún contrato registrado por el cliente (una anomalía del sistema o bien el cliente tiene dos planes en igual cantidad), el trato para estos valores anómalos fue simplemente borrar el registro de valor 0 en la variable de *Nominalización de planes*(o *Nom_Plan_real*), la cual consiste en una categorización de los distintos planes que la compañía vende de forma estandarizada, su glosa puede observarse en la tabla 3.11. Por lo que el valor 0 en esta variable indica un valor missing de la instancia en análisis. Finalmente se reemplazó por la moda de los planes.

- La variable *CANAL* se transforma mediante una nominalización de sus categorían que vienen en formato de cadenas de texto. En este cambio quedan 4 instancias con valores fuera de rango, distribuidas en dos categorías con una presencia de menos del 0,05 %. El tratamiento

Cuadro 3.10: Tabla de Transformación de la variable *USV Canal*

Canal	
Glosa	Valor
DCACE	1
DDEALER	2
Director de Cuentas	3
Ejecutivo COMPANY	4
DEVI	5
No definido	6

que se le da a estos valores es reemplazar con la moda de las categorías de la variable *CANAL*, debido a que existían muy pocos casos en esta situación lo que se puede visualizar en la tabla 5.6. El detalle de dicha transformación se encuentra en la tabla 3.10.

- Para el resto de las variables nominales, se realiza el reemplazo de categorías incompletas en su significado, por ejemplo, si un rut x tiene el siguiente valor "Masdrjig" en la variable *Clasificación*, se reemplaza ese valor por una categoría "0" o "-99", dicha categoría se establece para mantener un control sobre los valores ausentes.

- El restante de variable se mantiene igual, debido a su escasez de valores perdidos.

Transformación

En la etapa de transformación se dejan todas aquellas nominalizaciones o funciones que en cierto sentido resumen una variable sin quitarle información vital, no obstante, se debe destacar que en este experimento se pretende actualizar la experiencia anterior. Por ello, se tomaron las mismas transformaciones existentes en la experiencia anterior y se añadieron otras, como parte de la exploración del nuevo comportamiento del cliente con el producto. Un punto importante, es señalar que no se transforman los registros, debido a que se puede sobreajustar el modelo, así como también se puede ensuciar la información y su comportamiento subyacente en la base. Es así como se procede a describir las transformaciones en el mismo formato en que se describió el preprocesamiento.

- En el caso de la variable *CANAL*, se cuenta con todos los valores, para cada uno de los ruts, en otras palabras, no existen valores ausentes, sin embargo, a modo de descripción se presenta la tabla de frecuencias de esta variable ya nominalizada (cuyo nombre es *Nom_CAN*) en la tabla5.6, cuya nominalización se puede ver en detalle en la tabla 3.10. Cabe destacar que en esta variable el valor *missing* no existe, debido a que la categoría NO DEFINIDO se refiere a un significado válido en el negocio.

- Respecto a la variable relacionada con planes (es decir, la variable *PLANES* se realiza una transformación de tipo discretización, mediante distintas características de los planes. Se utiliza como criterio de elección (dado que un cliente tiene varios contratos y cada contrato puede tener muchos planes) el plan mayoritario a nivel vigente o anulado/finalizado, donde

Cuadro 3.11: Tabla de Transformación de la variables *PLANES*

Plan	Valor nominal asignado
Access	1
Access + Algo	2
Business I	3
Business II	4
Business II + Algo	5
Business Office I	6
Business Office II	7
Business Office II + Algo	8
Otros	10

se prioriza el vigente, es decir, si el cliente tiene un plan vigente y 5 planes renunciados o anulados, el plan que se escoge es el vigente. Todo lo anterior converge a la siguiente nominalización a la variable *Planes* expresada en la tabla 3.11. Donde las frecuencias respectivas se encuentran en la tabla 5.7 , además, se puede observar a simple vista un valor fuera el plan 4, y un posible *outlier* en la categoría 5 debido a su escasa cantidad. El tratamiento que se procede a hacer respecto a esto, es cambiar el *outlier* del plan 4 a un valor ausente, lo cual se debe a la baja cantidad perteneciente a dicha categoría. En este caso los valores fuera de rango se reemplazaron por la moda.

- La variable *ICP*, o Índice de comportamiento de pago, se transforma mediante la creación de una variable categórica para diferenciar el riesgo asociado al valor del ICP, dado que internamente en la compañía se cuenta con un criterio de corte que divide a los clientes, en categorías de alto riesgo y bajo riesgo, tomando en cuenta el valor de la variable *ICP*. Dicha transformación, fue la misma que se aplicó en la experiencia anterior, es decir, aquella representada en la expresión 3.1.1. Las frecuencias de las categorías de dicha variables pueden ser apreciadas en la tabla 5.8.

- La variable de *INGRESO CONTRATO SAP*, posee un problema en su origen dado que proviene de dos fuentes distintas: el área comercial y el área de sistemas, no obstante, mediante entrevista informales con el personal de la compañía, se opta por utilizar aquella variable de origen comercial, sin embargo, al observar los datos rigurosamente se detecta una mayor presencia de valores perdidos en la variable entregada por el área comercial. Por lo que se concluye que lo más óptimo es combinar ambas variables, de manera de contener la mayor cantidad de información referente a la variable en cuestión. La forma de combinar las dos variables es la siguiente:

$$ICSV = \begin{cases} ICSC & \text{si ICSC es no nulo e ICSS es nulo} \\ ICSS & \text{si ICSS es no nulo e ICSC es nulo} \end{cases}$$

En donde las siglas son descritas según la tabla 3.12. Cuando el valor se encuentra en ambas

variables, da la coincidencia que son iguales. De esta manera, los valores perdidos disminuyen, los cuales se pueden apreciar en la tabla 5.9, perteneciente a los anexos.

Cuadro 3.12: Nomenclatura para la variable *INGRESO CONTRATO*

Sigla	Significado
ICSV	Ingreso Contrato Sap Variable
ICSC	Ingreso Contrato Sap (Comercial)
ICSS	Ingreso Contrato Sap (Sistemas)

- Posteriormente se procede a recrear la variable *INGRESO CONT POST INST*, ya presente en la experiencia anterior. Su formato es similar a la expresión 3.1.1, no obstante, en este caso, la variable señala si el contrato fue ingresado posterior a la instalación del servicio. Es así, como se muestran sus frecuencias, en este experimento, en la tabla 5.10, en donde se denota una fuerte tendencia hacia la categoría "NO". Esto indica que se han cometido una menor cantidad de errores en cuanto al proceso ingreso-instalación. Cabe destacar, que la intención de la recreación de esta variable en el estudio era, efectivamente, la detección de fallas en el sistema. Sin embargo, a pesar de presentar una tendencia muy marcada, se opta por mantenerla sin alteración para poder replicar de forma exacta el modelo anterior. Por otro lado, para un posible tratamiento a la variable y sus valores ausentes o *missings*, se propone reemplazarlos por la categoría "SI" para que adquiera un mayor peso y se pueda detectar un patrón menos predominante (además, el hecho de que esté ausente ya es símbolo de error) . En otras palabras, la propuesta va en dirección a aumentar la varianza de esta variable y que con ello aporte más información al estudio.

- La variable *Sucursales* es categorizada en 6 valores distintos debido a su gran distribución de frecuencias, dicha categorización es equivalente a la efectuada en la experiencia anterior, la cual se encuentra en la expresión 3.2, donde las frecuencias respectivas, para este experimento, se muestran en la tabla 5.11.

- Para recrear la variable *Fuga* (cuyo origen en la experiencia anterior fue directo), se considera la base directa de MySap, en la parte que se refiere a los planes. Se debe señalar que los campos extraídos de MySap se dan también en los clientes de fase 0, siendo la variable *Planes* (equivalente a la variable *SAP_DESCRIPCION_ITEM*) uno de esos campos. Con el mismo, añadido al estado del plan(que también se encuentra para ambos tipos de clientes), se contabilizan los planes renunciados y los planes vigentes, así, si el cliente no tiene ningún plan activo, entonces se considera como fugado. Una descripción de sus frecuencias está en la tabla 5.12 donde la fuga efectiva se relaciona con el valor 1 y el caso de que no se fugue es indicado por el valor -1.

  Un punto importante es el hecho de que este problema de predicción de churn está un poco alejado del problema de clases desbalanceadas o raras de Weiss en [107], debido a que tiene más de un 30 % de fugados, contrario a lo que el personal de la compañía ha señalado respecto a la cifra del churn (la cual bordea el 1 %), no obstante, la tabla 5.12 contradice dicha

afirmación, esto conlleva a suponer que este resultado se puede deber a que el área comercial considera un porcentaje de fuga a modo de provisión dentro de los gastos declarados como empresa.

Modelamiento

En la parte de modelamiento se procede a considerar los principales algoritmos previamente descritos, esto se debe tanto a la escasez de plazo como también al propósito de este experimento, el cual busca actualizar el procedimiento plasmado en la experiencia anterior. De esta manera, se muestran en la tabla 3.14 los resultados de los principales modelos utilizados en este experimento, todos ellos con sus medidas respectivas.

Por otro lado, la fuga se modela respecto a la base incremental entregada por la plataforma MySap (cuya descripción se encuentra en la sección de bases de datos de los anexos) . No obstante, se realiza un muestreo estratificado sobre la misma, la principal razón que justificar este procedimiento es el hecho de que el programa utilizado (Rapidminer 5) no tiene la capacidad de analizar los 8 mil registros en un tiempo reducido, por consiguiente, el muestreo consta de 2677 clientes para generar el modelamiento final. En concreto se presenta la distribución de frecuencias, tanto para el muestreo como para la base principal en la tabla 3.13:

Cuadro 3.13: Tabla de frecuencias variable FUGA

Variable FUGA	Base total		Muestreo	
	Frecuencia	Porcentaje [%]	Frecuencia	Porcentaje [%]
SI	3130	35.07 %	940	35.11 %
NO	5794	64.93 %	1737	64.89 %
Total Instancias	8924	100 %	2677	100 %

Posteriormente se usan los mismos modelos utilizados en el experimento análogo, todo ello, para poder establecer una comparación concreta y válida, la única diferencia radica en el conjunto de datos usado.

Evaluación

Para la evaluación se usa una validación cruzada, pues esta técnica es lo bastante robusta en sus resultados para generar una interpretación confiable. Además de esto, se utilizan medidas adicionales a las expuestas en el experimento análogo, dentro de las medidas usuales se agrega el Lift y AUC, contemplando el muestreo estratificado, de esta forma se llega a la tabla 3.14.

Todos estos modelos son evaluados en las etapas de entrenamiento y prueba con la validación cruzada. Así, el modelo LADTree vuelve a superar al modelo del J4.8, ambos pertenecientes a la rama de los árboles de decisión, con lo cual se opta por seguir esta línea de los árboles, debido a que se desea entregar una propuesta de causas de la fuga de los clientes a la compañía, por ende, la red neuronal disminuye su prioridad en cuanto a ser el modelo más conveniente, por su característica de caja negra.

Cuadro 3.14: Tabla de resultados: Experimento 1

Criterios técnicos	Modelos				
	J48	LADTree	Random Tree	Random Forest	Red Neuronal
Accuracy [%]	89.54 %	89.76 %	68.29 %	88.76 %	90.03 %
Medida F [%]	88.94 %	89.31 %	62.64 %	87.70 %	89.23 %
AUC	0.928	0.948	0.672	0.943	0.946
Lift [%]	232.41	231.19	164.47	238.07	238.9
TN	1545	1536	1489	1581	1577
TP	852	867	339	795	833
FN	88	73	601	145	107
FP	192	201	248	156	160

3.2.2. Experimento 2: Refinamiento

Si bien en el primer experimento se obtuvieron buenos resultados, al momento de evaluar la predicción para meses posteriores (Mayo-Junio), se observa que la predicción simplemente no existe, es decir, ningún rut es considerado como fugado, lo cual al observar el gráfico 1.11, se comprueba su inexactitud. De esta forma, se empieza a investigar la procedencia del churn, cabe señalar que se sigue utilizando la base que se bajaba directamente del MySap, por ende, el problema solamente es visible en los meses de Abril y Mayo. Fue así, como se opta por usar la base del Product Manager del producto NGN como alternativa de base para predecir el churn respectivo. Sin embargo, en un inicio se trabaja con la primera base, en otras palabras, este experimento marca la primera transición de bases del proyecto.

A lo anterior, se suma el hecho de que dentro de los objetivos del proyecto se busca refinar el modelo propuesto, por lo tanto, la aplicación de nuevas transformaciones, así como también, de nuevo modelos, es necesaria.

Integración y Selección

Para la integración de bases de este experimento se mantienen la mayoría de las bases integradas en el experimento 1, además, se agregan nuevas bases para poder contrastar la información pertinente, de esta manera, se puede bosquejar de forma más precisa variables como el GIRO, la CLASIFICACIÓN, etc.

En esta etapa, se toma la base del product manager con 9111, esto se realiza, para seguir una línea de trabajo y evitar particularidades de las bases como la incrementalidad. Con ello, se continúa a escoger los clientes o llaves únicas, a partir de la base de datos **NGN_Instalados**. Esta base es el fundamento de apoyo para la gestión que elabora el área comercial entorno al producto, por consiguiente, usar dicha base para la realización del nuevo estudio, entregará resultados que puedan ser comparables de forma directa desde la perspectiva comercial. Dentro de los supuestos asignados a esta base se encuentra aquel que señala que los datos no eran incrementales, este supuesto se justifica en el hecho de que si los datos fuesen incrementales se posee una fecha que permite establecer la marginalidad correspondiente. Nótese que esta base también es usada en el experimento anterior, para rescatar variables como por ejemplo: cantidad de sucursales, tipo de conexión, accesos, planes (nombre contrato NGN), entre otras[7]. Sin embargo, al cabo de un tiempo se vuelve a evidenciar el tema de los datos incrementales, aún cuando es esta base con la que se trabaja operacionalmente, no obstante, el problema de datos incrementales, en esta ocasión, difiere al problema del experimento anterior, debido a que en esta ocasión se puede establecer la marginalidad de los datos, mientras que en el experimento anterior no. Ahora bien, este problema de marginalidad se refiere a que no se pueden obtener los clientes fugados en determinado período pues no se guardan todas las fechas de término de contrato, además, se añade el efecto de que la variable relacionada con el estado del contrato del cliente dentro del sistema no es inmediata, esto quiere decir, que si el cliente decide el término de contrato y lo expresa a través del call center, el sistema no cambia el estado de sus planes a renunciados hasta que se le retiran los equipos respectivos al cliente, acción que, acorde a entrevistas informales con el personal de la empresa, puede tardar semanas, meses e incluso años. Bajo esta temática se pierde la temporalidad de la fuga del cliente, o sea, no se puede determinar la fecha

[7]Todas las variables y sus sistemas respectivos pueden ser observados en la tabla 5.1

exacta en que el cliente se fuga, dada una base, por lo que no se puede entrenar modelos o algoritmos sin caer en un excesivo error. Sin embargo, en este nuevo escenario en entrevistas informales con el personal de la empresa se concluye que para el caso del mes posterior al de la entrega de la base, se puede establecer una marginalidad correspondiente con el objeto de obtener resultados que sean efectivos para la predicción del comportamiento de los clientes reales. Otro punto no menor es que la base pasa por un preprocesamiento anterior a su entrega al área comercial. Cabe enfatizar que la base de datos **NGN_Instalados** se obtiene de forma distinta, debido a que es una base que ya tiene un procesamiento previo antes de llegar al usuario final (en este caso el Product Manager de NGN) . Esta base resulta consultando a 3 plataformas distintas, tal y como se ilustra en la siguiente figura 3.2 que se puede contrastar con 1.10.

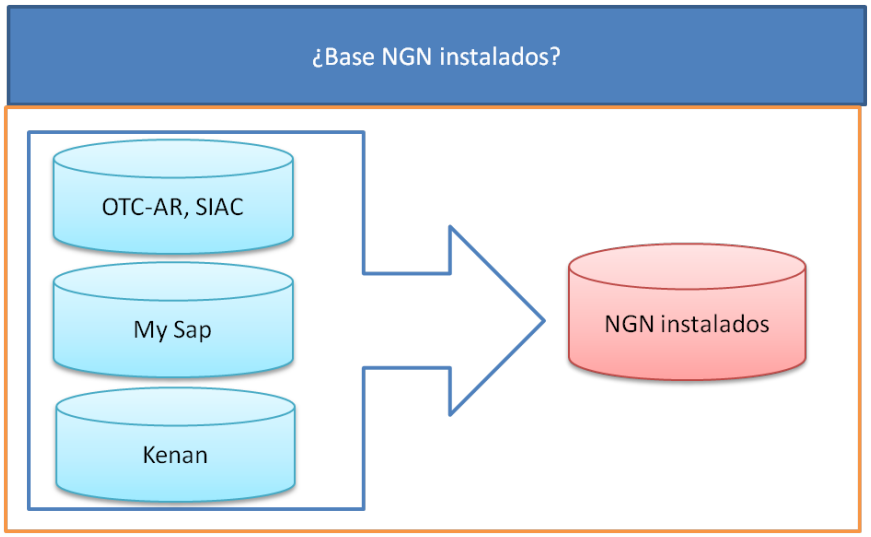

Figura 3.2: Origen de base de datos NGN Instalados

De forma transversal, las bases trabajadas en todo momento fueron las de **Proforma** y **Boletas Técnicas**. No obstante, la segunda base se integró de forma distinta al experimento 1, puesto que habían sido utilizados solamente como contador, es decir, contaba el número de reclamos técnicos efectuados por un cliente. En cambio, en este estudio, se toman más variables de esta base para enfatizar el tipo de reclamo dado que difieren en importancia y urgencia unos con otros. Sin embargo, para la extracción de las variables, se requiere de una base adicional que alude a los tipos de servicio asociados al NGN entregada directamente por el personal de la compañía. De esta forma se puede concretar una consulta que permitiese extraer los reclamos de NGN según responsables.

Una tercera base y, a su vez, la más relevante para el tema de la definición del churn en esta empresa, es la base de Reclamos comerciales del producto NGN **Oracle Workflow (OWF)** (cuya descripción puede observarse en la sección 5.1.1) . Sin embargo, la importancia de la base va por la perspectiva de las solicitudes de término, las cuales se registran posterior al ofrecimiento de un plan de retención al clientes, es decir, cuando el cliente llega a la empresa o llama a la misma, solicitando su término de servicio, se le ofrece de forma reactiva inmediata un plan de retención que cada ejecutivo de la empresa conoce, si el cliente rechaza esta retención entonces la empresa marca efectivamente esta solicitud como término de contrato; en caso de que el cliente acepte este

plan de retención, se le marca la solicitud como retención y esto se registra en el sistema. Por lo que no existe un margen de error a nivel de considerar una solicitud de término como retenido. Con esta nueva base se continúa a declarar los ruts fugados fundamentando esta decisión en los términos de su contrato, además, se aplica un estudio de los principales motivos por los cuales se retiraban dichos clientes. Posteriormente se extraen los reclamos comerciales de facturación por entrevistas informales con el personal comercial encargado del producto NGN. Para su extracción simplemente se toma de la base, las variables *RUT CLIENTE y TIPO*, y se realiza una tabla dinámica entre estas dos variables determinando la cantidad de reclamos de facturación por Rut, lo que se contrasta con la base que contiene los registros vigentes para obtener la variable completa.

Otra base que previamente no se había mencionado, y que ahora pretende jugar su rol respectivo es la que contiene los datos de los suscriptores, llamada base de datos **Suscriptores** (cuya descripción se encuentra en la sección 5.1.1) . El uso de ésta será para concebir una variable que describa la influencia de la competencia sobre los clientes de la empresa de telecomunicaciones. Dicha base consta de una gran cantidad de registros, la cual el software Microsoft Access no puede manejar por sus restricciones propias del programa relacionadas con el peso de la información en el sistema. Además, posee 3 campos, uno referente al Rut, otro al Ani[8] correspondiente y un tercer campo que declaraba la compañía a la cual pertenece ese Ani. Para tratar esta base se requiere de 2 bases adicionales, aquella consistente en los ruts vigentes del producto NGN y la que contiene la tipificación de las compañías presentes en la base de datos **Suscriptores**, estas base no son necesarias, sin embargo, acorta la demora de la extracción donde la última es facilitada por el personal de la compañá. Posteriormente se divide el archivo macro, usando el software Textpad, en 16 partes, cada una de 500 mil registros y la última de 160. Luego, para cada una de esas partes se asigna un archivo en Access con los ruts vigentes incorporados. De esta manera, mediante una consulta apropiada se extrae parte por parte la cantidad total de anis pertenecientes a los cliente NGN. Un punto relevante es que esta base sólo se usa para determinar con qué compañía competencia está asociado el cliente, no se estudia la cantidad de anis como fuerza de dicha relación con la competencia. Además, esta base de datos permitió descubrir tres variables más, referentes a la compañía: *COMPANY PHONE*, *COMPANY MOBILE*, *COMPANY CELL*, las tres variables binarias que expresan la fortaleza de la relación entre el cliente y la compañía.

Adicionalmente a las bases de datos anteriores, se utiliza una denominada **Órdenes terminadas** que se describe en la sección de 5.1.1 en los anexos. Ésta contiene 74 campos, de los cuales se contempla una porción la cual es mostrada en la tabla 5.30, es discriminada para este experimento por el plazo para preprocesarla (puesto que es una base que consta de muchos campos y no fue usada en la experiencia anterior).

La última base de datos a usar es la de **Seg_empresas** de forma directa (a diferencia del experimento anterior) ésta contiene la información de la categoría a la pertenece cada empresa ya sea por la clasificación que se le asigna, o bien, por el tamaño que tiene a nivel global. Dentro de la misma se puede apreciar que existe un desglose de las facturaciones de los otros productos que el cliente dispone aparte del NGN dentro de la empresa de telecomunicaciones, dicho desglose es convertido a variables concretas para luego determinar el producto que manda en el cliente, en otras palabras, aquel producto en el que el cliente gasta más y por lo tanto, es más relevante para él. Concretamente, de esta base se extraen variables como *CICLO VIDA, TAMAÑO, CAT CORP, VALOR, RETENCION*,

[8] Ani se refiere al número telefónico desde el punto de vista interno en una empresa

cada una de ellas descritas en la tabla 5.1.

A nivel de variables se incluyeron algunas de las bases anteriores (aquellas descargadas directamente de MySap), dentro de las cuales se encuentran las variables *GIRO y RUBRO* se refieren al nombre del negocio propiamente tal. Particularmente se escoge la variable *GIRO* del sistema KENAN presente en la base de datos de MySap denominada "Req292" . Esto se decide dado que este último sistema aplica para el tema de facturación, por lo que, acorde al personal de la empresa, se sugiere considerar el atributo tomado por este sistema, sin embargo, su formato es dudoso, es decir, dos categorías de esta variable pueden tener el mismo valor y estar representados de distinta forma, por ejemplo, "Ventas al por mayor" y "Vtas. X Mayor", disyuntiva que se soluciona al no contemplar un formato completo a priori. En el caso del atributo rubro, es un agrupación obtenida a partir de la base Giros Comerciales S.I.I._2008 (descrita en la sección de bases de datos en los anexos) que permite describir el ámbito específico del negocio cuando se agrupa con el atributo Giro. Se debe mencionar que los dos atributos descritos son variables ajenas a la empresa de telecomunicaciones.

Otras variables previamente abordadas, fueron cambiadas de origen, específicamente las variables *TAMAÑO y CLASIFICACIÓN* son extraídas de la base de datos **Seg_empresas**, puesto que en el experimento anterior se tomaron de la base directamente descargada de la plataforma MySap. La razón para esta elección radica en conversaciones con el personal cercano al servicio en cuestión. Las variables individuales contempladas a partir de esta base se muestran en la tabla 5.1.

Referente al tema de buscar una variable que permita dilucidar la relevancia del NGN, respecto a los otros productos en los clientes de NGN, se cuenta con una base de datos de **boletas de facturaciones**, en las cuales se posee la información completa del gasto de cada cliente. No obstante, por temas de tiempo; es decir, el hecho de que cada consulta en esta base tarda mucho tiempo respecto al tiempo que tarda la base Seg-empresas; y la cantidad de valores ausentes; pues la base de boleta de facturaciones posee 4629 instancias no válidas (valores perdidos) mientras que la base de datos **Seg_empresas** sólo presenta 400 valores perdidos se opta por usar la base de datos **Seg_empresas** para extraer la variable de importancia del producto.

Se mantienen las variables de reclamos técnicos del experimento 1, y las variables de facturación y consumo, así como también las variables que describen los planes del producto. El ICP también fue agregado, de manera independiente a la base **Seg_empresas**.

Preprocesamiento

Debido a que en este experimento la sección de preprocesamiento toma relevancia y es guía para el resto de los experimento, se muestra en el recuadro 3.4, una síntesis del mismo.

Registros La elección de registros implica explorar completamente la base de datos **NGN_Instalados**, la cual consta de un archivo excel, de dos pestañas, una contiene la información de todos los clientes que alguna vez han estado en el producto (independiente si está fugado o no) y la otra posee todos los planes que han sido instalados. Por lo tanto, la primera decisión de selección, es la de escoger los 5893 ruts a partir de la variable de *FECHA DE ÚLTIMA FACTURA* y *CANTIDAD DE FACTURAS*. Estas dos variables son claves para el diseño básico de la base general, la cual, posteriormente, servirá de base de entrenamiento y validación del modelo de minería de datos. Además, los ruts escogidos tienen el formato regular de ruts, es decir, 12345678-0 lo que se refiere a que están como

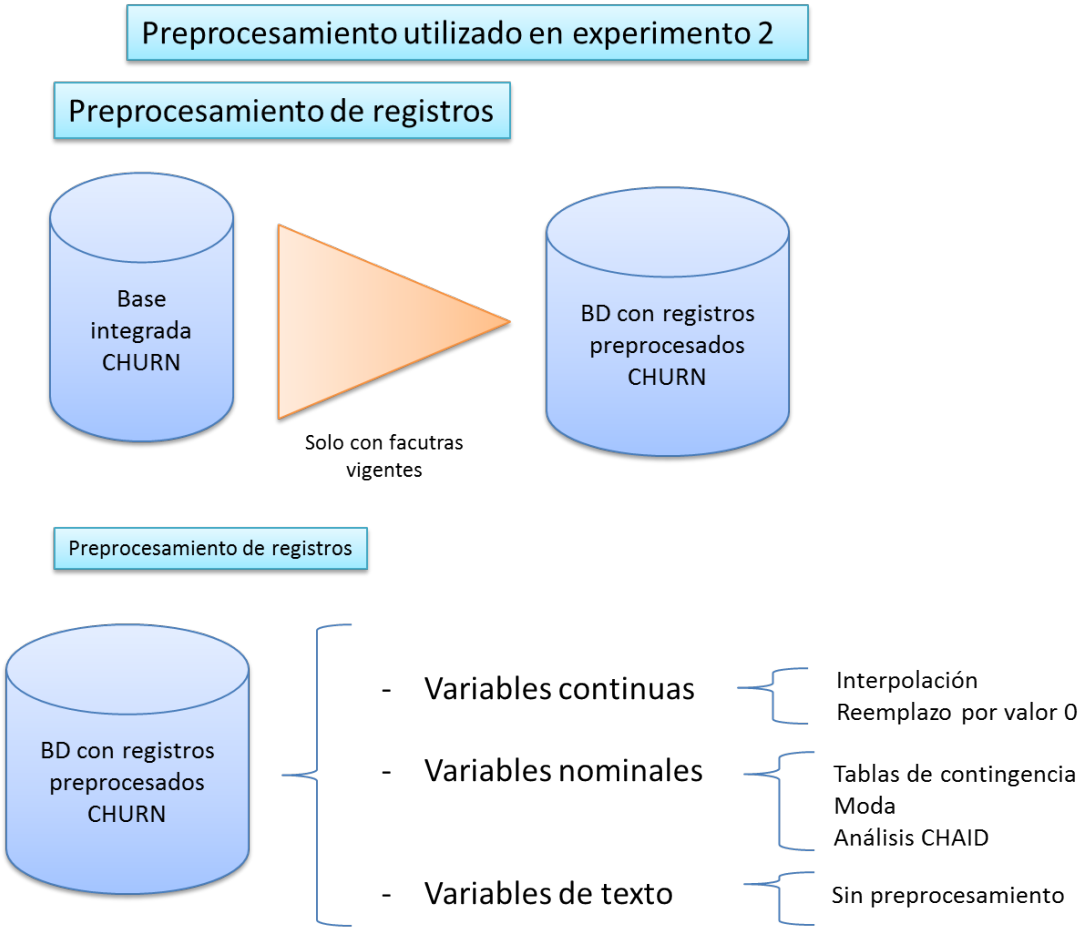

Figura 3.3: Síntesis de preprocesamiento aplicado en el experimento 2

datos en un único atributo. En cambio, en otras bases de datos como la de **boletas de reclamos técnicos** y en la pestaña planes de la base de datos **NGN_Instalados** la característica rut tiene otro formato, el cual, consiste en dos atributos referentes al cuerpo del rut (números previos al guión) y el dígito verificador, sin embargo, la solución que permite contrastar ambos casos ha sido expuesta en dos atributos separados ya solucionado previamente en la subsección 3.2.1.

Variables El tratamiento de los valores perdidos o ausentes en este experimento se basa en la tesis base, sin embargo, se refinaron determinados reemplazos, para obtener una base preprocesada más robusta y con mayor cantidad de información interpretable.

Un resumen con todas las variables, su cantidad de valores perdidos respectiva y la estrategia de reemplazo se describen en la tabla 5.13. Nótese que la cantidad de instancias es de 5893 en el conjunto de datos contemplado. Respecto a los valores fuera de rango, no se analizan en este experimento.

En esta ocasión se comienza a observar las posibles causas de dichos valores perdidos, de tal

manera de generar transformaciones adecuadas, además, en se detecta un problema de desbalanceo o rareza de clases explicado previamente en la sección 2.8, lo que significa que la mayoría de los clientes pertenecen a un clase, mientras que una minoría pequeña pertenece a la otra clase. Por lo tanto, el tratamiento de los valores ausentes debe ser mucho más riguroso, por ejemplo, en el caso de que una variable posea 98 % de los valores perdidos, no significa que tenga que ser descartada, puesto que ese 2 % de valores puede entregar un patrón relevante para la detección del comportamiento de fuga de los clientes. Por consiguiente, se utiliza una técnica de elaboración propia denominada "Tabla de presencia", que es una tabla análoga a la que el software SPSS entrega cuando se solicita un análisis de valores perdidos, mas esta "Tabla de presencia" muestra las relaciones o los patrones posibles de los valores perdidos entre las distintas variables y lo cuantifica bajo la misma forma que las tablas de contingencia, en otras palabras, es una tabla de contingencia de los valores perdidos cuyos valores son el porcentaje de coincidencia en la ausencia. La necesidad de elaborar dicha tabla es que no está implementada en ningún programa aún, tal como se requiere, pues es deseable calcular el porcentaje de perdida entre cada conjunto de variables. El resultado de dicha tabla aplicada se muestra en las tablas 5.14 a 5.19 y su código en lenguaje Visual Basic se puede apreciar en la sección 5.16.

Las ténicas de reemplazo utilizadas que se pueden apreciar en la tabla 5.13, están definidas pero no descritas respecto a la forma en que fueron ejecutadas. A continuación se procede a ahondar en mayor detalle sobre su significado:

- **Interpolación**: Esta técnica se usa sobre la misma variable, en pocas palabras, se ejecuta con los datos adyacentes al valor perdido, por lo tanto se utiliza en las variables continuas dado que se busca interpolar; que es que dado un conjunto de n+1 puntos (x_i, y_i) se encuentre una función de grado n f tal que: $f(x_i) = y_i \quad \forall i = 1, \ldots, n+1$ [40]; sin dejar valores fuera de rango ni alterando la distribución de la variable.

- **Tabla de contingencia, Moda**: Esta técnica conjunta es ejecutada en dos pasos, en el primeros se analizan las tablas de contingencia entre la variable, en la cual se desea reemplazar sus valores perdidos, y la variable especificada en la estrategia. Posteriormente, se reemplaza por la moda en una de las variables y luego, se ejecuta una simplificación del hot deck, en cuanto a reemplazar en base a las relaciones detectadas en las tablas de contingencia.

- **Tabla de contingencia, Moda, CHAID**: Esta técnica es una extensión de la anterior, puesto que el reemplazo por la moda originaba un gran sesgo teórico entre las variables nominales analizadas con la tabla de contingencia (*VALOR, CICLO_VIDA, RETENCIÓN, TAMAÑO, CAT_CORP*) . Por ello, se decide establecer una tabla de contingencia entre una variable y las otras, luego, se sustituye con la moda (los valores perdidos) en esta variable, y consecuentemente, se reemplaza por hot deck en el resto de las variables. Finalmente con las variables reemplazadas se ejecuta un árbol CHAID, para predecir, el campo previamente reemplazado. Donde este último algoritmo (CHAID) consiste en un árbol de decisión *"provisto de un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir qué registros darán un cierto resultado. Segmenta un conjunto de datos utilizando test de chi cuadrado para crear múltiples divisiones..., además, permite realizar tipologías de clientes en función de una variable criterio, así como realizar pronósticos con probabilidades conocidas"* [58].

Cabe destacar que las cinco variables: *VALOR, CICLO_VIDA, RETENCIÓN, TAMAÑO y CAT_CORP*, presentan una equivalencia en los valores perdidos, es decir, coinciden sus valores perdidos, por ende, no se puede reemplazar en base a alguna otra variable, además, no poseen relación alguna con el resto de las variables nominales.

- **Reemplazo por valor 0**: Esto equivale al mismo preprocesamiento efectuado en los experimentos anteriores, que es reemplazar los valores perdidos por el dato numérico 0.
- **Moda**: Equivale a reemplazar los valores perdidos por la moda de la variable.
- **Nada**: Significa dejar el valor perdido tal como está, en el caso de la variable *COMUNA*, al momento de probar modelos se deja en blanco, como una categoría extra.

Transformación

En la etapa de transformación, se empieza a visualizar nuevas formas de representar los datos transaccionales (referentes a la facturación y al consumo), así como también, aquellos datos fijos (como la descripción de planes), manteniendo el propósito de esta etapa que es sintetizar la información contenida en las variables que se desean incorporar.

De manera análoga, se inicia la nominalización de las variables que están en texto para poder ejecutar una mayor cantidad de análisis sin que el software que se usa obstaculice el proceso. Esto se efectúa para todas las variables exceptuando *COMUNA y GIRO*.

En la base de datos de la **Proforma**, se localizan las variables referentes a la facturación y al consumo, a las cuales se les aplica una transformación bajo una función, la cual es una variante de la transformada discreta de Fourier para un período de 6 meses. Sin embargo, para almacenar más información se crean dos variables que contemplan la facturación y el consumo. De esta manera, las 4 variables dedicadas a la facturación o al consumo (*ID_CURVA*), se aumentan a 4 para facturación y 4 variables para el consumo. Dos de esas variables son los promedios de los 6 meses (función que se mantuvo de la experiencia base, 1 para la facturación y otra para el consumo), mientras que las otras 6 fueron un promedio ponderado total(3 para la facturación y 3 para el consumo) . La fórmula respectiva que caracteriza estas 6 variables de promedio ponderado, que se denominan en esta memoria como "Trim", se expresan de la siguiente forma:

$$Trim_{facturacion} = \sum_{i=0}^{n-1} Facturacion_{mes_i} * exp\frac{-2 * \pi * i}{n} \tag{3.3}$$

$$Trim_{consumo} = \sum_{i=0}^{n-1} Consumo_{mes_i} * exp\frac{-2 * \pi * i}{n} \tag{3.4}$$

Por lo tanto, de la **Proforma** se extraen 8 variables, de las cuales 2 son los promedios de consumo y facturación, y las otras 6 son los promedios ponderados de los bimestres contemplados en el período de estudio, donde la variable *Trim_Facturacion(Consumo)_j* se refiere al promedio ponderado del bimestre j considerado. Un punto relevante es la tabla asociada a los valores de j la cual se expone en la tabla 3.15, el caso de las variables de facturación es análogo.

Cuadro 3.15: Transformación de variable Trim de consumo

Bimestre	Valores de j	Fórmula Trim
Mes 1 y 2	1	$Trim_{Consumo1} = \sum_{i=0}^{1} Consumo_{mesi+1} * exp\frac{-2*pi*i}{6}$
Mes 3 y 4	2	$Trim_{Consumo2} = \sum_{i=2}^{3} Consumo_{mesi+1} * exp\frac{-2*pi*i}{6}$
Mes 5 y 6	3	$Trim_{Consumo3} = \sum_{i=4}^{5} Consumo_{mesi+1} * exp\frac{-2*pi*i}{6}$

Para la base de las órdenes terminadas se busca clusterizar según alguna característica relevante que pudiese vislumbrarse al momento de hacer dicho análisis, o bien, categorizar de alguna forma parte de la información que está base contiene. En este experimento solamente se discretizan las órdenes terminadas acorde a la acción que significa. Esto último se puede apreciar en la tabla 3.16.

Cuadro 3.16: Transformación de variable referente a la base de órdenes terminadas

Valor	Categoría
0	Vacía o Capacitación
1	Instalación de algún servicio
2	Agregación de algún servicio
3	Eliminación de servicios
4	Retiro Masivo de servicios
5	Retiro singular de servicio
6	Ampliación de algún servicio o Aumento de velocidad
7	Cambio de algún servicio
8	Suspensión de algún servicio
9	Disminución de Velocidad
10	Habilitación o Regularización de servicios
11	Traslado (interno o singular)
12	Modificación a servicios

Ahora bien, dicha categorización aparece luego de una exploración a la base de órdenes terminadas, en donde existen variables como *FECHA DE EMISIÓN* y *FECHA DE EJECUCIÓN*, las cuales pueden ser utilizadas para manejar la marginalidad de la base de datos. De esta forma, la nominalización se crea mediante el siguiente procedimiento: se toma la variable *Trabajo* (definida en la tabla 5.1, luego se crea una función que categorice numéricamente los textos, la cual puede ser expresada como:

$$\text{Tipo trabajo} = \sum_{i=1}^{12} g_i \tag{3.5}$$

Donde los valores de dicha función se denotan en las tablas 5.31, 5.32, 5.33. Finalmente para describir por completo dicha transformación se hace mención al conjunto de variables seleccionadas con las que consta la base de *Ordenes terminadas*, las cuales se pueden apreciar en la tabla 5.30.

Cuadro 3.17: Clusterización de planes Experimento 2

Grupo	Anis	BA	Tec_Wiimax_Cobre	ADSL	Velocidad	NOMBRE
1	3.43	0	COBRE	0	0	Antiguos
2	6.84	0-1	WIIMAX	0	0-640-1024-1294-5000	Estándar Wiimax
3	2.65	1	COBRE	0	640	Velocidad 640 con Cobre
4	2.63	1	COBRE	0-1	2000-30000	ADSL 2+

Otro punto relevante respecto a esta base era saber si la instancia era una orden terminada de acción, o simplemente una cotización, para ello se creó una variable denominada *COTIZACION*, la cual es una dummización[9] de la variable *DESCRIPCION*, que señala los casos en que aparece la palabra cotización dentro de la variable *DESCRIPCION*. Su expresión respectiva es:

$$COTIZACION = \begin{cases} 1 & \text{si la variable } DESCRIPCION \text{ contiene la palabra COTIZACION} \\ 0 & \text{si no} \end{cases}$$

De esta manera, las variables que se pretenden usar para este experimento, de esta base, son: *Tipo trabajo* (variable creada), *Cotización* (variable creada), *Tipo serv1*, *Tipo serv2*, *Dentro del plazo* (la cual está definida a partir de las fechas de anulación de la provisión y la fecha pactada cliente), *Urgencia*.

Para la característica contractual de planes, se observa solamente una variable que se denomina *SAP_DESCRIPCION_ITEM*, la cual posee información de texto no filtrada, por lo tanto, se procede a dividir dicha variable en múltiples atributos que pudiesen describir de mejor manera un plan. De esta forma, se llega a un estimativo de la velocidad de cada plan, al cual, se agregan otras variables como la cantidad de Anis, si el plan tiene o no banda ancha, el tipo de plan, si posee o no ADSL superior y si utiliza tecnología de cable de cobre o Wiimax. Todas los atributos descritos previamente se encuentran en la base de datos **NGN_Instalados** en su segunda pestaña dedicada a los paquetes. Así, con estas 5 variables se clusterizan los planes usando la técnica de *Two step Cluster*, cuyos resultados finales se muestran en la tabla 3.17. Para describir los centroides de cada grupo respecto a las variables continuas se toma la media y en las variables nominales se observa la categoría con una frecuencia mayor al 50 %. En algunos casos donde las otras categorías son de frecuencia torno al 30 % de su total, o bien, eran escasas y sólo se daban en un grupo, se agregaron en forma conjunta. El detalle de los centroides se encuentra en las tablas 5.22 a 5.28. Para la decisión de la cantidad de clústers a usar, el programa ejecuta automáticamente la mejor opción según el indicador de información de Akaike (AIC)[10], el cual determina el número de grupo en relación al máximo cambio en el índice de información AIC, el resultado de éste puede ser observado en la tabla 5.21. Esto genera 4 variables denominadas *PLAN_TIPO_i*, las cuales consisten en contar el número de planes de estilo i, donde $i = 1, \ldots, 4$.

Nótese que la cantidad de planes es del orden de 174 mil registros, esto implica que esta clusterización se obtiente a partir de la base extraída directamente de la plataforma Mysap. Además, no

[9] Equivalente a una discretización binaria

[10] Descrito en la sección 2.7.3, cuya fórmula es 2.38

se etiquetan los planes.

De la base de datos **Seg_empresas** se extraen entre otras variables previamente mencionadas, variables del tipo "fact_(nombre del producto)", las cuales indican la facturación del cliente en cada producto. Cabe destacar, que la base de **Seg_empresas** contiene todos los registros de todas las empresas asociadas a la compañía, para lo cual, se debe contrastar los ruts de esta base, con la de los clientes de NGN vigentes. Posteriormente, se genera una variable llamada "fact_total" que expresa la suma de las variables *fact_(nombre del producto)*. Otra variable que se genera es la "Importancia de NGN" que se expresa como $\frac{fact\_NGN}{fact\_total}$. Una tercera variable se crea, denominada como "Producto principal", la cual se formula como la categorización del producto que presenta mayor facturación, es decir, aquel que cumple la condición $\frac{fact\_(nombredelproducto)}{fact\_total} > 0{,}5$ cuya categorización puede observarse en la tabla 5.29.

En lo que se refiere a los reclamos técnicos en un principio, se mantienen las variables que reflejan la cantidad total, por mes y divido por responsable, todo ello dentro de los 6 meses. Sin embargo, con estas 3 perspectivas se tiene un total de 18 variables, por lo que se procede a realizar una transformación que permita reducir la dimensionalidad de la base, sin restar información a la misma. Esta transformación tiene origen en el modelo RFM (descrito en la sección 2.7.4), que se refiere a Recency, Frequency y Mount. Un punto relevante es que en este experimento la frecuencia se tomó como la frecuencia de los reclamos, es decir, el promedio de los reclamos y el monto fue el total de los reclamos efectuados en el período de estudio, además, el Recency fue obtenido mediante el conteo de meses entre el mes más actual del estudio, y el mes más antiguo. Aún bajo esta arista, estas tres variables aportan mayor información interpretable que las 18 variables anteriores. Sin embargo, se pierde la información de la variable *Resp_falla*[11], es por ello, que se genera una variable adicional denominada *IMAGEN*, la cual tiene se define como valor -1 si la instancia posee más reclamos técnicos (fallas) de responsabilidad de la compañía que reclamos técnicos en los que el cliente de esa instancia fue el responsable y 1 en caso contrario. Cabe recordar que la variable *Resp_falla* es una variable interna, el cliente no tiene noción de ella, por ende, el nombre de la variable se refiere a un nombre teórico.

$$IMAGEN = \begin{cases} 1 & \text{si la variable } \textit{Resp\_falla} \text{ tiene el valor COMPANY}^{12} \\ 0 & \text{si no} \end{cases}$$

De las variables bases de datos de los **Suscriptores** (*Rut, Ani, Compañía*) se crean cuatro variables, las primeras tres, ya han sido mencionadas y descritas (*COMPANY PHONE, COMPANY MOBILE, COMPANY CELL*), que son variables binarias que para una instancia i adquieren el valor 1 en caso de que esta instancia (en particular, el rut) se encuentre en la base **Suscriptores** y, además, esté asociada a la compañía creadora de este piloto. La presencia de un rut en esta base respecto a la compañía interna puede presentarse en las tres aristas, es decir, en formato de teléfono, móviles del cliente y celulares del mismo, esto se da producto de la separación que existió previamente a la fusión dentro de la compañía. Por otro lado, la cuarta variable se denomina *COMPETENCIA*, esta variable se crea tomando todas las instancias que tengan NGN y que tengan una compañía distinta a las expresadas por las tres variables anteriores, posteriormente se crea una tabla dinámica que permite dilucidar los competidores más influyentes y en base a esto se extraen las 6 compañías etiquetadas como "más influyentes" y se crea una variable binaria por cada compañía, generando

[11] variable que indica de quien fue la responsabilidad de la falla técnica

6 variables parciales, que luego convergen a una sola variable mediante la selección priorizando desde la "más influyente a la menos influyente", de esta manera se llega a una variable nominal con 6 categorías según la frecuencia de ruts que posean al menos un ani en la competencia. Esta tabla de frecuencias se encuentra representada en 5.34, de esta manera se decide por considerar las categorías 1, 12, 27, 19, 3, 11. Un punto a enfatizar es que esta base se considera estática para todos los experimentos, debido a que se toma el supuesto de que sus registros no poseen alta variabilidad.

Modelamiento

En esta etapa se busca entrenar con la mayoría de los modelos adaptados en el programa de Rapidminer 5, no obstante, en un inicio se trabaja con los árboles de decisión puesto que en la experiencia anterior y en el experimento 1 estos clasificadores habían sido los más adecuados acorde a los requerimientos de la compañía. Sin embargo, aparte de los modelos tradicionales (sea naive bayes, SVM, KNN y árboles) se probaron algunos multiclasificadores, entre los cuales destacan el *Voting*. En particular, este modelo se ejecuta con la participación de dos modelos KNN, los cuales actuaban como votantes, el primer KNN se caracteriza por hacer uso de distancias numéricas y a la hora de estudiar las variables nominales usa una distancia euclídea mixta, mientras que el otro KNN se dejó para usar distancias nominales. En las tres iteraciones del modelo de Votación (donde solamente se cambiaron las medidas de distancia asociadas), fueron denominada Vote 1, Vote 2, y Vote 3.

En cuanto a qué modelo entrenar se privilegia a aquellos cuyo tiempo de respuesta fuese reducido, por ende, modelos como las redes neuronales fueron descartados en este experimento. Otro motivo de discriminación del entrenamiento del modelo era la efectividad de sus resultados, es decir, si el modelo implementado en Rapidminer 5 arroja un error que simplemente no deja ejecutar el modelo, éste se descartaba, la justificación de esto radica en que existen ciertos errores que en el software no tienen solución actual.

Finalmente en esta etapa, se pretende modelar el sistema para generar la validación inmediata, y así, entregar resultados potencialmente útiles en el corto plazo. Resta decir, que aún cuando este experimento se efectúa en el mes de Julio, busca la validación de la base de Marzo, con el nuevo refinamiento del KDD.

Un punto relevante es el de resumir la información de cada prueba de modelos, debido a que se estudian distintas combinaciones, para ver la robustez de cada modelo. Por ende, se utiliza una nomenclatura particular, la cual puede ser observada en las tablas 5.35 a 5.38 y, en base a dicha nomenclatura, se prueban las distintas configuraciones de cada modelo, las cuales pueden ser observadas en las tablas 5.39 a 5.42.

Un resumen de los modelos a aplicados en este experimento, se detalla en el esquema.

Evaluación

La evaluación en este experimento cambia de perspectiva, puesto que anteriormente se buscaba usar una base para entrenar y probar el modelo, mientras que en el experimento actual se pretende validar con una nueva base la efectividad del modelo, es decir, se entrena (y prueba) con una base y luego se valida con otra base. Además, para los clientes que se fugan en marzo, el entrenamiento de su predicción se realiza con la base de marzo.

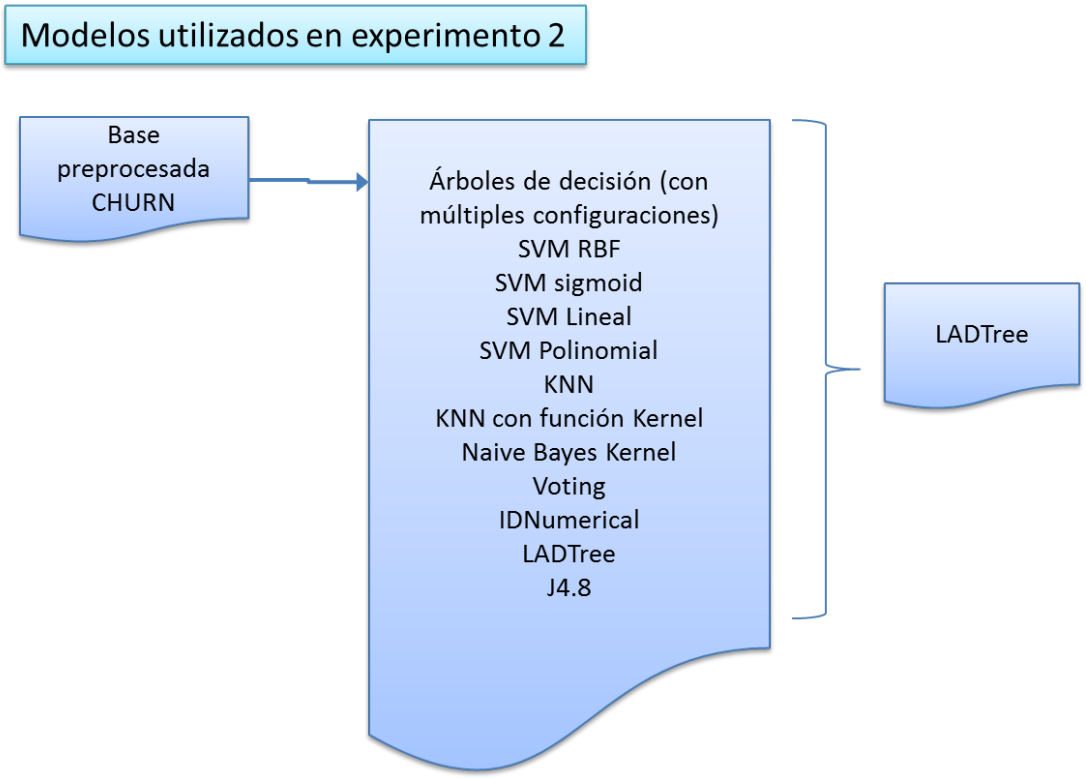

Figura 3.4: Modelos utilizados en el experimento 2

Las medición de los resultados se efectúa con las medidas típicas, es decir, *Accuracy*, Medida F y AUC, siendo esta calculada en el entrenamiento.

Finalmente los resultados son mostrados en las tablas 5.44 a 5.47, donde se puede apreciar que el modelo LADTree es superior al resto, con una F total de 75 % aproximadamente, lo cual es bastante bueno según los estándares generales. Esta medida es superior al *Accuracy* ya que realiza una medición más rigurosa que no depende excesivamente del criterio de corte que se use. Lamentablemente no se pudo rescatar el valor del criterio AUC, para validar aún más el modelo.

Sin embargo, otra parte de la validación depende de la parte comercial, la cual no se pudo efectuar debido a la relación costo-riesgo, que posee una alta incertidumbre, por lo tanto, se opta por continuar analizando los distintos prototipos hasta que se genere un piloto cuya precisión estadística sea aceptable para el personal que maneja el producto, por ende, este experimento solamente efectúa una validación técnica y para un mes.

3.2.3. Experimento 3: Predicción con clusterización

En este experimento se pretende descubrir una mayor explicación a las causas de la fuga del cliente, además, se busca establecer una temporalidad lo suficientemente válida para el modelo, debido a que al probar el KDD estipulado en el experimento 2, para el mes de Mayo, se constata una diferencia en los cálculos y error en los resultados los que se pueden apreciar en la tabla 3.19.

Cuadro 3.18: Resultados: Validación Marzo-Abril Experimento 2

Modelo: LADTREE			
Predicción	Vigencia real	Fuga real	Precisión
Vigente	5587	12	99.79 %
Fuga	89	42	32.06 %
Recall	98.43 %	77.78 %	
Med. F Class SI	45.41 %	F TOTAL	75.417 %
Med. F Class NO	99.10 %		
Accuracy	98.24 %		
Correctamente Clasificadas		5629	98.237 %
Incorrectamente Clasificadas		101	1.763 %

Además, se refinan algunas transformaciones que se encontraron sin el fundamento matemático suficiente para ser usadas de manera correcta. Por ende, en este nuevo experimento se propone la clusterización de los clientes como proyecto aparte, de manera tal que se pueda descubrir un perfil de cliente fugado, y así optimizar las acciones correctivas sobre los clientes de este estilo.

El resultado incondescendiente en la temporalidad de los meses, es aquel que se obtiene al probar el modelo para el mes de Mayo para predecir sus posibles fugados. De esta manera, se entrena tanto con el mes de Marzo, como en el mes de Abril, mas ninguno de los resultados finales fue satisfactorio. Para mayor detalle se muestran las tablas de confusión que representan esto, en donde los entrenamientos para los meses de Marzo y Abril se encuentran en 5.49, 5.50:

Cuadro 3.19: Tabla de confusión: Validación LADTree en experimento 2 para el mes de Mayo con entrenamiento en el mes de Marzo

Validación entrenando con marzo			
Categorías Predicción	Categorías Real		Precisión
	0	1	
0	5627	40	99.29 %
1	127	1	0.78 %
Recall	97.79 %	2.44 %	
Accuracy	97.12 %		
Medida F	50.08 %		
Medida F pos	1.18 %		

En donde se puede apreciar que los valores de la medida F (entendida como la medida total de la evaluación, contemplando ambas clases) son cercanos a un 50 % lo que implica una clase dominante sobre otra. Al observar la "medida F pos", la cual, considera solamente la medida F asociada a la clase de fuga, se denota un bajo porcentaje lo que indica una baja predicción desde el punto de vista de detección de fuga.

Cuadro 3.20: Tabla de confusión: Validación LADTree en experimento 2 para el mes de Mayo con entrenamiento en el mes de Abril

Validation entrenando con abril			
Categorías Predicción	Categorías Real		Precisión
	0	1	
0	5748	41	99.29 %
1	6	0	0.00 %
Recall	99.90 %	0.00 %	
Accuracy	99.19 %		
Medida F	49.80 %		
Medida F pos	Desconocida		

Integración

Bajo un aprendizaje incremental en el proyecto se mantiene el uso de las bases de datos anteriores discriminando **Órdenes terminadas** y agregando las variables *MOROSIDAD y CARTERA*, cuyas descripciones se muestran en la tabla 5.1. Cabe destacar que estas dos últimas provienen de bases individuales propias de un área determinada, por lo que se consideran como estáticas (cuando en realidad no lo son), debido a que solamente se prueba con ellas en este experimento.

Ahora bien, la eliminación de la base de datos **Órdenes terminadas**, se debe a que la información en ella, no se puede extraer de manera correcta, además, se agrega el hecho de que presenta variables multi-producto, por lo que hay ventas que dan origen a una orden de trabajo de instalaciones y postventas que dan origen a una orden de trabajo de retiros, modificaciones y traslados. Por ende, se determina que estas variables no están consolidadas como para efectos de predicción de fuga en el producto NGN.

Preprocesamiento

Registros El preprocesamiento de los registros se efectúa a partir de la pestaña "paquetes" de la base de datos **NGN_Instalados**, en la cual se caracterizan el estado de sus planes, en base a ella se consideran sólo aquellos planes cuyos estados actuales no son RENUNCIADA, ni ERROR. Posteriormente se realiza una cuenta en el software Microsoft Access que permite dilucidar la cantidad de ruts correspondientes que para la base de Agosto, resulta ser 6204. Sin embargo, uno de los ruts tiene gran cantidad de registros vacíos, por ende, se elimina, quedando así en una base con 6203 registros. No obstante, en la glosa que describe el estado de los planes en esta base, no se tiene información sobre si el cliente es nuevo o no. Por lo tanto, se procede a solucionar dicha problemática a través de las variables *FECHA PRIMERA FACTURACION* e *INGRESO CONTRATO SAP* una regla de un cliente nuevo. Cabe destacar, que las instancias asociadas a estos clientes nuevos no se borran, puesto que se busca usar la base para meses tanto anteriores como posteriores, mas la variable *Nuevo* no entra dentro del aprendizaje del modelo.

Variables Tras el extenso estudio de los valores perdidos en el experimento 2, se persiste con el trato a algunas variables y otras simplemente se eliminan (excepto la variable *GIRO*) debido a su alta varianza, la cual, puede ser observada en los gráficos 5.11 a 5.15. Estas variables se refieren principalmente a la ubicación espacial específica del cliente, por ende, se decide mantener variables más generales como COMUNA, en vez de la calle particular del cliente. También se opta por eliminar las variables de facturación de los otros productos, es decir, FACT_(Producto), debido a que las variables *IMPORTANCIA NGN, FACT_TOTAL y PRODUCTO PRINCIPAL*, las resumen. De esta manera, se puede apreciar la distribución de valores perdidos en la tabla 5.51, la cual tiene incorporada la estrategia respectiva para los valores perdidos. Muchas de esas estrategias fueron previamente explicadas en el experimento 2, no obstante, existen otras nuevas, las cuales se describen a continuación:

- **Ingreso Contrato - Delta interpolada**: Usando la variable *DELTA INTERPOLADA*, que significa el número de días transcurridos entre la primera instalación y el ingreso del contrato al sistema, se reemplazan los valores perdidos de la variable *PRIMERA_INSTALACION*.
- **Primera Instalación + Delta interpolada**: Es un reemplazo análogo al anteriormente explicado, donde en esta ocasión se busca usar la variable *PRIMERA_INSTALACION* y la variable *DELTA INTERPOLADA* para reemplazar el valor perdido de la variable *INGRESO CONTRATO SAP*.
- **Variable eliminada**: Equivale a la eliminación previa de la variable (*PLANES_RENUNCIADOS*) por cambio en la clusterización de los planes y el no aporte de información de esta variable en el modelo anterior.
- **Reemplazo por valor S/I**: En la variable *GIRO* se reemplazaron los valores perdidos por la categoría S/I preexistente en la variable, este valor indica la glosa ”Sin Infomación”, lo cual es equivalente a decir, sin información sobre el valor, que es un valor perdido del tipo MCAR.
- **Reemplazo por valor 2**: En la variable *COMUNA* (original), se realiza un reemplazo específico por las comunas que tenían un nombre similar, es decir, un error de escritura en la categoría. En la variable *COMUNA2* (Transformada), se sustituye por el valor 2 cambio que se explica en la sección 3.2.3.

Para el caso de los valores fuera de rango u *outliers* se detectan dos valores fuera de rango en las variables *TAMAÑO* y *RETENCION*. En la variable *TAMAÑO*, el valor fuera de rango es “Persona con Acceso”, lo cual se demuestra en la tabla de frecuencias de dicha variable para el mes de Mayo, es decir la tabla 5.53, su trato es la sustitución de dicho valor por otro correspondiente a Microempresa”, categoría perteneciente a la misma variable. En la variable *RETENCION*, se nota una categoría con baja frecuencia que se considera un *outlier*, por lo que se sustituye este valor por “Retención Mínima” en las instancias correspondientes. La justificación de dicho reemplazo se encuentra en la tabla de frecuencias de esta variable del mes de Mayo, es decir, la tabla 5.54.

Otros valores fuera de rango se encuentran en las variable GIRO, sus valores que no corresponden son “?” y “0” los cuales se encuentran en una frecuencia descrita en la tabla 5.55, cuya temporalidad se remonta al mes de Julio.

Cuadro 3.21: Transformación de variable Trim de facturación

Bimestre	Fórmula Trim
Mes 1, 2 y 3	$Trim_{Consumo1} = \sum_{i=0}^{2} Consumo_{mesi+1} * exp\frac{-2*\pi*i}{3}$
Mes 4, 5 y 6	$Trim_{Consumo2} = \sum_{i=0}^{2} Consumo_{mesi+4} * exp\frac{-2*\pi*i}{3}$

Transformación

En la etapa de transformación se consolidan las transformaciones previas TRIM y aquella asociada a los planes. Además, se crean dos nuevas variables: MOROSIDAD y CARTERA. Se cambian las variables nominales *COMUNA* y *GIRO*, utilizando bases externas y propias de algunas áreas comerciales adyacentes a NGN, en pos de reducir las categorías de cada variable. A esto se agrega la explicación de la etiquetación de determinados ruts como clientes nuevos. En particular, la variable *COMUNA* se cambia a través de una base externa (llamada **TIPIFICACION_CPO**, cuya descripción se encuentra en la sección de bases de datos en los anexos) en la cual se describen la glosa de los centros primarios, cuya descripción se encuentra en la tabla 5.1. Con esta base, se categorizan las comunas por lo centros primarios generando la variable *COMUNA2*, la cual, posee más valores perdidos que la COMUNA anterior. Al verificar la procedencia del aumento de los valores perdidos se descubre que se debe a la no existencia de un centro primario para las comunas de Santiago por separado, es decir, todas las comunas de la provincia de Santiago poseen el mismo centro primario expresado en el valor 2, no obstante, este valor solamente estaba asignado a la comuna de Santiago. Por ende, se opta por sustituir los valores perdidos por el valor 2.

La variable *GIRO* se nominaliza puesto que son cadenas de texto (naturaleza que presenta problemas al implementar la base en SPSS o Rapidminer), lo que conlleva a vislumbrar 408 categorías totales, dicho cambio genera la variable *GIRO_TRANS*.

Al analizar la fórmula de la variante de la transformada discreta de Fourier, expresada en las variables *TRIM* del experimento 2, se visualiza que esta transformación base no estaba bien aplicada en la base, debido a que se deseaba efectuar el promedio ponderado por trimestre, por ende, para implementar las fórmulas correspondientes 3.3, 3.4, se debe diferenciar a los trimestres como período (en otras palabras, el n disminuye a 3, en cada una de sus fórmulas), con lo cual, se obtienen distintos resultados, por lo que se vuelve a aplicar a toda la nueva base de Agosto. Con el formato mostrado en la tabla 3.21.

Dicha transformación, al igual que en el experimento anterior, es análoga para la facturación. De esta manera, las 12 variables de facturación y consumo, se convierten en 6 variables, 3 (PROMEDIO, TRIM1, TRIM2) para cada tema (consumo y facturación).

Para la variable de los planes, se persiste con el tema de la clusterización, que en esta ocasión arroja 5 grupos (1 grupo nace a partir de los clústers originados en el experimento 2) . Además, se arma un modelo de clasificación[13] de planes que permitiese etiquetar a los nuevos ingresados. Posteriormente se cuenta el número de planes de un determinado tipo i (donde esta variedad de categorías es entregada por la clusterización previamente descrita) y ese valor se almacena en la variable *PLAN_TIPO_i* con $i = 1, \ldots, 5$.

[13]Un árbol de decisión básico de Rapidminer, es decir, el modelo A1 de la tabla 5.39

Cuadro 3.22: Segmentación de planes Experimento 3

Características por conglomerados							
Conglomerado	N	Anis	Velocidad	BA	Tecnología	ADSL	Nombre
1	6422	(1 a 5)	Media (392 a 1714)	SI	COBRE	NO	Plan Estándar
2	1155	(1 a 5)	Alta (2000)	SI	MIXTO_COBRE	SI	Plan ADSL
3	1518	(1 a 6)	Media Baja (587 a 685)	SI	WIMAX	NO	Plan Wimax
4	3833	(1 a 5)	Nula (0 a 0)	NO	COBRE	NO	Plan Sin Banda Ancha
5	830	(1 a 25)	Baja (0 a 179)	NO	MIXTO_WIIMAX	BAJO	Plan Personalizado

Para la obtención de estos conglomerados se utiliza el algoritmo *Two-Step Clúster* con la opción de tomar al vecino más lejano. Un detalle de los resultados obtenidos se encuentra en las tablas 5.57 a 5.62, donde en esta última se usa el promedio para las variables continuas y las categoría con más de un 50 % de frecuencias a su favor.

La justificación de esta segmentación, difiere a la del experimentó 2 debido a que se requiere caracterizar concretamente los planes, caracterización que contempla un margen de error alto en caso de que se hubiese utilizado la clusterización del experimento 2, por ende, se agrega un clúster más, lo cual, no es un error sino que se divide un clúster anterior en dos y se pudo concretar una segmentación con un error menor. El resultado en cuanto a los centroides que se obtienen a partir de la segmentación se muestra en la tabla 3.22

En relación con las variables relacionadas con la morosidad y la cartera, su creación tiene como objetivo expresar los dos términos anteriormente mencionados, estas dos son binarias, y cada una expresa la presencia tanto de una instancia o cliente moroso como también de un cliente cartera, es decir:

$$Morosidad_i = \begin{cases} 1 & \text{si la instancia i es morosa} \\ 0 & \text{si no} \end{cases}$$

$$Cartera_i = \begin{cases} 0 & \text{si la instancia i no es cliente cartera} \\ 1 & \text{si la instancia i es cliente cartera} \end{cases}$$

Otra transformación es la categorización de todas las variables nominales, referentes a la base de datos **Seg_empresas**, las cuales se cambian por números tal y como se indica en la tabla 5.56.

La variable *DELTA INTERPOLADA* es creada en esta etapa para bosquejar la distancia entre las variables *INGRESO CONTRATO SAP* y *PRIMERA INSTALACION*. De esta forma, se apoya el reemplazo de los valores de dichas variables, no obstante, su expresión matemática viene dada por la resta *INGRESO CONTRATO SAP-PRIMERA INSTALACION*, además, se agrega el reemplazo por interpolación efectuado en la etapa anterior.

Los clientes nuevos se clasifican como tales al tomar en cuenta dos variables *INGRESO CONTRATO SAP* y *FECHA_PRIMERA_FACTURACION*. Bajo la regla: si una de estas dos variables tiene un valor de fecha dentro del mes anterior al mes actual de estudio (Julio en este caso, dado que se estaba trabajando con el mes de Agosto), entonces se toma como un cliente nuevo. Con esta regla 227 ruts son etiquetados como nuevos.

En este experimento se reformulan las variables RFM, para que puedan entregar información de manera directa, para ello, la frecuencia se cambia: mientras que en el experimento 2 se refería

Cuadro 3.23: Valores de TS1 y TS2 para el producto NGN

Producto	TS1	TS2
NGN	ACCE	PYME
NGN	FU	PYME
NGN	TFBA	VSAT

a la cantidad de reclamos promedio, en el experimento 3 alude al porcentaje que el cliente reclama en un período de 6 meses considerados, en otras palabras, si el cliente posee un valor de 50 % en esta nueva frecuencia, implica que de los 6 meses, el cliente reclamó $6 * 50\,\% = 3$ veces. Aparte de la frecuencia, se cambia la variable *Monto*, que en el experimento 2 se refiere al total de reclamos dentro de los 6 meses, mientras que en el experimento 3 toma la siguiente expresión $\frac{1}{n}\sum_i^n f(x_i)$ donde $f(x_i)$ es una función que adquiere el valor 1 cuando el cliente reclamó en el mes i y 0 en caso contrario y siendo n el número total de meses considerados para el estudio. Para la transformación del monto se utiliza la expresión $\frac{1}{\text{Veces que reclama}}\sum_i^n s(x_i)$ donde $s(x_i)$ es la cantidad de reclamos del mes i. Cabe señalar que está función en el caso de que las instancias que no reclaman adquiere el valor 0. La variable *Recency* se obtiene en base a la función $f(x_i)$, contando el mes más cercano en donde el cliente reclamó. Un percance detectado a partir del experimento 2, es el hecho de que la consulta para obtener los reclamos con sus responsables respectivos, hace referencia a una tercera base que contempla los valores de las variables *TS1 y TS2* que describen el servicio (mayor profundización en la tabla 5.1), que en el caso del NGN son 3 combinaciones posibles, observables en las tabla 3.23, las cuales en su forma original tienen multiplicados sus registros cuatro veces, por lo tanto, las cantidades de reclamos de cada instancias se descubren en un estado amplificado (por 4), por lo que se procede a eliminar los registros duplicados.

Modelamiento

Esta etapa difiere de las anteriores, debido a que se busca la clusterización de los clientes, es decir, se propone una agrupación de las instancias para determinar un modelo más específico y con menor gasto en cuanto al tiempo de cómputo. Es decir, hay un cambio en la estrategia del modelamiento más que en el algoritmo particular. Esta nueva perspectiva permite percibir las características de grupos de clientes que facilita la determinación de las causas posibles de la fuga y aún más, obtener de forma rápida de elaborar acciones correctivas.

En aspectos técnicos, se decide por presentar 3 clústers, esto es producto de un análisis jerárquico de conglomerados bajo un corte predeterminado, al cual se le agrega la utilización del modelo W-EM que es la implementación en WEKA[14] del algoritmo EM, que como capacidad adicional contiene la técnicas de validación cruzada para determinar el número probable de grupos en el conjunto de datos, sin embargo, debido a restricciones del equipo computacional con el que se cuenta, solamente se puede ejecutar un muestreo (estratificado para este caso), lo que convergió a un resul-

[14]Software usado principalmente para ejecutar la etapa de minería de datos del procedimiento KDD. Es ser gratuito y programable, debido a que es un software open source, además cuenta con una interfaz y se encuentra implementado en el software open source Rapidminer, ambos se encuentran definidos en la sección 1.5.2

tado de 3 grupos. No obstante, el tercer grupo no puede ser concretado con un solo nombre, por lo que al conversar con el experto en el negocio, se llegó a la conclusión de clusterizar este grupo solamente para obtener varios subgrupos. Esto también iba acompañado por la intuición de la naturaleza del tercer clúster, el cual se bosqueja en la tabla 5.86. La forma en que esta se ejecuta es equivalente a la primera clusterización. De esta manera, todo este proceder (que se detalla en la sección 5.4.5) se concreta en los siguientes grupos:

Cuadro 3.24: Etiquetas asignadas a los conglomerados: Experimento 3

Conglomerado	N	% del total	Nombre de Conglomerado
1	2136	36,66	Fieles
2	291	4,99	Rehén
3	3399	58,34	Mixto
Total	5826	100	

Cuadro 3.25: Etiquetas asignadas a los subconglomerados: Experimento 3

Conglomerado	N	% del total	Nombre de Conglomerado
1	1705	50,16	Los Indiferentes
2	859	25,27	Los Refugiados
3	835	24,57	Los Switchers
Total	3399	100,00	

Una vez ya generado los conglomerados por cliente e identificado los perfiles de los mismos, se procede a estudiar un modelo para cada grupo. Estos modelos incorporan una nueva característica al momento de aplicar el KDD, la nueva particularidad consiste en utilizar técnicas de submuestreo, es decir, se disminuye la cantidad de instancias de clase “no fugado” y se aumenta la cantidad de instancias de clase “fugado” (esto último se hace multiplicando la cantidad de instancias fugadas) . Finalmente el modelamiento de este experimento consta de varios modelos propuestos y sus resultados, uno para cada grupo, dichos modelos se pueden visualizar en las tablas 5.79 a 5.84. Cabe destacar que en el grupo 1 se aplicó una técnica de muestreo para manejar el desbalanceo de las clases presentes. Dicha técnica consistió en seleccionar aquellas instancias que no fuesen fugadas y eliminar el porcentaje seleccionado, es decir, tomando en cuenta la nomenclatura de las tablas, un “10” y “90” indica que un 10 % de las instancias fueron elegidas mientras que el 90 % fue desechada, por otro lado, si a esta glosa presente en la fila “Proporción entrenamiento” se le agrega las palabras “sin multiappend”, implica que no se duplicó las instancias fugadas, nótese que esta herramienta se aplicó solamente en el entrenamiento y prueba, dado que ne la validación se desconoce la variable *Fuga*.

Evaluación

Ahora bien, esta etapa cobra una mayor importancia, debido a que se **evalúa desde un punto de vista comercial**, perspectiva agregada a la evaluación técnica que se ha descrito anteriormente. La evaluación técnica en este experimento es equivalente a la del experimento 2, es decir, toma una subetapa de entrenamiento del modelo y se prueba (siendo ésta la primera evaluación del aprendizaje del modelo) . Posteriormente se usa el mismo modelo y se aplica sobre una nueva base, la cual en este caso es la base de Octubre y se mide la certeza para el mes en cuestión. Sin embargo, la evaluación técnica no es suficiente a la altura correspondiente del proyecto (siendo este el último experimento contemplado en la formación original del proyecto) . De esta forma, se procede a realizar la valorización de la predicción, para lo cual, se toman las instancias cuya clase es predicha exitosamente en la validación, éstas forman un total de 13 instancias, luego se suman los centroides de la variable *PROMEDIO_FACTURACION* de cada conglomerado a la cual pertenecen las 13 instancias predichas eficazmente. Por consiguiente, se entrega un total anual probable de pérdida ahorrada con el modelo actual. Los modelos "finales" (de este experimento) para cada grupo se muestran en la tabla 3.27, los universos de cada clúster en la tabla 3.26 y el principal resultado, con sus métricas técnicas, en las tablas 3.28, 3.29.

Cuadro 3.26: Universos involucrados en el experimento 3

Universos de la Predicción		
Número de K	Conglomerado	N
1	Leales	2105
2	Rehenes	288
3	Mixto	3438
Sub 1	Indiferentes	1633
Sub 2	Refugiados	913
Sub 3	Switchers	892
Total		5831

Cuadro 3.27: Modelos utilizados en el experimento 3

Grupo		Modelo
Clúster 1		Rule Induction (50 y 50)
Clúster 2		BayesNet
Clúster 3	Sub clúster 1	WFT
	Sub clúster 2	Naive Bayes
	Sub clúster 3	Logistic1

Ahora bien, respecto a la evaluación comercial, esta se considera de manera teórica, por lo tanto, solamente constata la valorización desde el punto de vista de los datos internos a la empresa, mas representa un punto de partida en el inicio de describir el valor de la herramienta ocupada, es así que

Cuadro 3.28: Resumen de Resultados finales: Experimento 3

Validación Septiembre con Agosto			
Categorías Predicción	Categorías Reales		Precisión
	No Fugados	Fugados	
No Fugados	5449	37	99,33 %
Fugados	332	13	3,77 %
Recall	94,26 %	26,00 %	

Cuadro 3.29: Métricas finales: Experimento 3

Medida	Valor
Accuracy(%)	93,7
AUC	0,62
Lift(%)	439,4
FMeasure (%)	6,6

los resultados entregados son:

Cuadro 3.30: Evaluación comercial: Experimento 3

Grupo	Nombre Grupo	Fugados	Monto promedio de facturacion	Costo Total
1	Leales	4	194,783.00	779,132.00
2	Rehenes	0	2,530,680.00	-
3.1	Indiferentes	1	75,849.00	75,849.00
3.2	Refugiados	5	188,884.00	944,420.00
3.3	Switchers	3	20,304.00	60,912.00
Costo Mensual Ahorrado				1,860,313.00
Costo Anual Ahorrado				22,323,756.00

Un punto relevante es el hecho de que esta valorización solamente considera el la fuga de ingresos al mes siguiente, y en el caso de la evaluación anual, ésta solamente es la multiplicación del costo mensual ahorrado por 12, esto quiere decir que no se valoriza la pérdida de ingreso por cliente, sino la fuga de ingreso, la diferencia radica en que el primer concepto (pérdida) alude al ciclo de vida del cliente, por ende, su fuga es una pérdida frente a lo esperado por la empresa, mientras que la fuga de ingreso expresa lo que se puede ahorrar independiente del ciclo de vida en el que se encuentre el cliente fugado.

3.3. Fase 3: Validación histórica del KDD

Una vez efectuado los tres experimentos, la funcionalidad del KDD queda concretada en la compañía, sin embargo, se decide estudiar la validez temporal del procedimiento en sí. Si bien, no se ha demostrado la existencia de un modelo de minería de datos absoluto para cada problema existente hasta los tiempos actuales, se da la generalidad de que como el mercado cambia, los datos cambian y, por lo tanto, las técnicas usadas en las etapas del procedimiento deben cambiar. Para establecer una seguridad mayor en los resultados obtenidos, se opta por crear una predicción para cada mes en la cual se tenga la información necesaria para aplicar el KDD.

El plazo de esta validación fue de dos semanas aproximadamente. Esta validación se efectúa fuera del proyecto general, cerca del 13 de Diciembre del año 2010, fecha en la que, teóricamente, el proyecto inicial estaba terminado. No obstante, en entrevistas informales con el personal de la compañía se llega a la extensión del proyecto en pos de establecer una validación comercial basada en un muestreo a las instancias que se predijesen como fugadas para un mes determinado, en el cual no se tuviese conocimiento alguno del valor de dicha clase. Ese particular mes es Enero del año 2011. No obstante, al igual que los otros experimentos, éste también está sujeto a la aplicación de todas las etapas del KDD, lo que en el caso de la validación histórica, se toma desde el mes Diciembre del año 2009, hasta el mes Noviembre del año 2010, formando un conjunto de 11 meses, dentro de los cuales se aplica el KDD en todo momento. Sin embargo, debido al escaso tiempo, se replican las etapas de Preprocesamiento y Transformación, manteniendo el formato de ejecución del experimento 3.

3.3.1. Consideraciones en Etapa Integración: Experimento 4

La etapa de integración es idéntica para los experimentos 4 y 5, sin embargo, para el experimento 4 su similitud es solamente a nivel de las fuentes requeridas para elaborar la base de prueba, puesto que al ser un experimento que alude al comportamiento histórico del modelo, se debe diferenciar que las bases de datos se integran 12 veces, en donde se debe separar el trato de las bases según la consideración temporal. Está última consiste en una división de los tipos de bases según el supuesto de cómo les afecta la variación en el tiempo:

- Bases No temporales: Estas bases son aquellas en las que se supuso una variación despreciable en sus valores según transcurre el tiempo. Las bases que entran en esta categoría fueron:
  - **Suscriptores**
  - **Segmentación empresas**
  - **NGN_Instalados**
- Bases Temporales: Son aquellas cuya variación entre meses resulta significativa debido a su naturaleza transaccional, generalmente, estas bases expresan de forma inmediata la influencia de externalidades en los datos. Dentro de las bases a las que se les asigna esta categoría están:
  - **Boletas Técnicas**
  - **OWF**

- **Proforma**

Agregado a lo anterior, se desechan las bases externas individuales que contienen las variables *MOROSIDAD y CARTERA*, puesto que su integración en el experimento 3 no fue significativa estadísticamente, además, la periodicidad de ambas era incierta, debido a que son bases que recién se habían contemplado. Por otro lado, se agrega la variable de la **OWF** referente al total de reclamos comerciales.

3.3.2. Experimento 4

Este experimento posee una particularidad correspondiente a la preparación de las bases de datos con las que se trabajará a modo de validación, puesto que lo que se pretende en este experimento es entregar un modelo consolidado en el tiempo. Por lo tanto, se debe preprocesar 11 bases de datos y preparar cada una de ellas para la elaboración de un experimento de validación. De esta forma, se toma la base de datos que contiene la historia de 6 meses previos a Diciembre 2009 (incluyendo este último mes) . Ésta se denomina "Base de Diciembre 2009", nomenclatura que permite establecer un orden en términos de las fases de evaluación por las que transcurre cada base de datos, por ejemplo, si se desea saber el rendimiento de un modelo determinado para el mes de Marzo, se puede entrenar y probar con las bases de Diciembre de 2009, Enero de 2010 o Febrero de 2010, y validar con el restante, es decir, si se escoge entrenar y probar con Diciembre de 2009, se puede validar el resultado en la base de Enero de 2010, Febrero de 2010 y predecir una base de posibles fugados para el mes de Marzo de 2010. Esta ventaja se utiliza para el experimento 6 y 7 con el objetivo de entregar resultados con un fundamento práctico, desde la perspectiva del negocio, en el tiempo. Lo anterior, puede ser apreciado en la figura 3.5.[15]

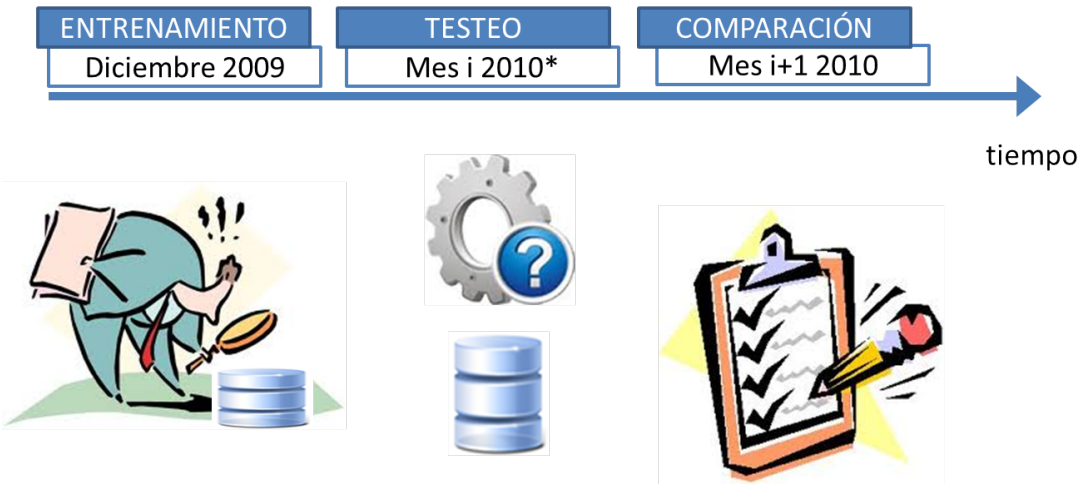

Figura 3.5: Esquema de pruebas de validación de modelos

En otras palabras, en la etapa de **Entrenamiento** se estudia bajo la supervisión de los valores

[15]Con i = (Enero, Febrero, Marzo, Abril, Mayo, Junio, Julio, Agosto, Septiembre, Octubre, Noviembre) e $i + 1$ siendo la referencia al mes siguiente.

conocidos de la variable *Fuga* los patrones posibles, dejando como base fija aquella preprocesada con las instancias vigentes para el mes de Diciembre 2009. Luego en la etapa de **Testeo** se corre el modelo deducido a partir de la etapa anterior para verificar la exactitud del modelo sobre un conjunto de meses en donde se conocen los valores de la variable *Fuga*. Posteriormente, en la **Comparación** se prueba sobre los meses con la perspectiva de que no se conocen los valores (es decir, no se añaden al proceso automático de validación del software, sino que se compara en forma manual). Cada una de ellas, en el experimento 4, son aplicadas para las bases preprocesadas de prueba y validación.

Preprocesamiento

Registros Una peculiaridad que se debe destacar en el experimento histórico 4, es que no se cuenta con las bases marginales suficientes como para establecer la vigencia de los clientes en cada mes, es decir, se cuenta con las bases de datos **NGN_Instalados** de Diciembre 2009, Marzo 2010, Julio 2010, Septiembre 2010, Octubre 2010 y Diciembre 2010. Por ende, para obtener los clientes Vigentes de las bases de los meses faltantes, tomando en cuenta el escaso tiempo, se procede a trabajar con la base de Diciembre 2010 y se comienzan a agregar registros acorde una heurística determinada. Ésta se elaboró de la siguiente manera: con el objeto de detectar a los clientes vigentes, se observa el atributo de FECHA_TERMINO presente en la pestaña paquetes de la base **NGN_Instalados**, usándolo se filtran las instancias de planes por mes. Por ejemplo, aquellas instancias con FECHA_TERMINO anterior a Julio, son consideradas para establecer los clientes vigentes de Agosto, pues un cliente que entra en Julio es nuevo, por lo tanto, si el cliente se fuga no presenta relevancia para el estudio del comportamiento de deserción de los cliente. Luego, se toma la variable *ESTADO_PLANES* y se consideran los planes que no fuesen ERROR, ni RENUNCIADA, finalmente se ejecuta una tabla dinámica que entrega la cantidad de clientes con identificador RUT. Sin embargo, esto soluciona sólo la mitad de la problemática puesto que los clientes fugados en cada mes siguen siendo un misterio. Por ende, se procede a elaborar una segunda estrategia que permita la detección de los clientes fugados y su marginalidad. En esta nueva estrategia se trabaja con la base del Oracle Workflow, la cual se encontraba disponible para todos los meses, por consiguiente, se posee la fuga para todos los meses. No obstante, se descubre un problema de marginalidad, pues se ha usado la base de Diciembre 2010 para obtener la vigencia de clientes en los meses anteriores. Por ejemplo, si un cliente está vigente en Abril pero se fuga en Mayo, su estado según la marginalidad de Diciembre 2010 es de Fugado para el mes de Abril, esto puede verse con más claridad en la figura 3.6, en donde para una instancia se tienen dos estados, uno marginal real (que es su estado en un mes determinado, Vigente) y su estado marginal causado por la incrementalidad de la base de datos **NGN_Instalados** (que para ese mismo mes indica que su estado es fugado).

La solución contemplada para manejar esta marginalidad errónea, es la de incluir la fuga en forma incremental, lo que quiere decir que si un cliente se presenta como fugado en Abril 2010, entonces su RUT es tomado como vigente para las bases marginales de los meses anteriores a Abril 2010, mas esta solución conlleva a una alteración en el comportamiento común de las bases de datos que posteriormente se muestra en la figura 3.7, en donde "Solución Marginal" se refiere a la descrita anteriormente para mantener las situaciones que afectan las instancias (es decir, que un cliente que estaba vigente en Junio de 2010 y renunciado en Diciembre 2010, se refleje en la base de datos como vigente en ese mes y no solamente como renunciado en el mes de Diciembre de 2010).

Esta alteración solamente afecta a los meses posteriores a Junio aunque se desconoce cuál fue

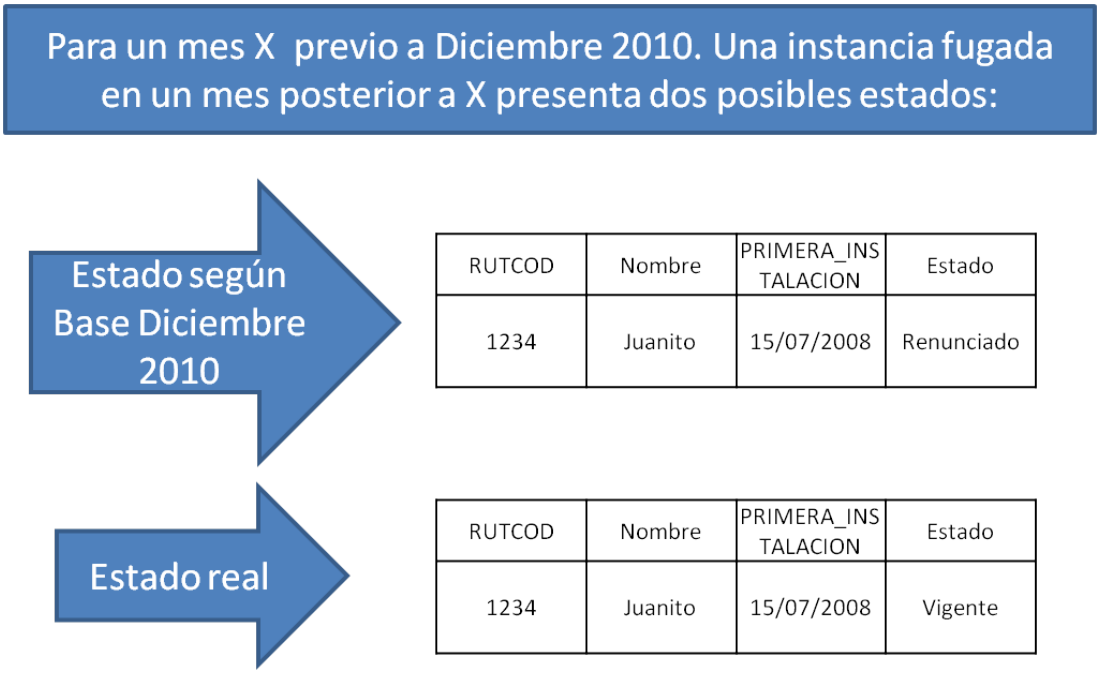

Figura 3.6: Esquema de pérdida de información con bases incrementales

la causa de una anomalía en el comportamiento común de la fuga.

Variables Respecto al preprocesamiento de variables, las variables *GIRO y COMUNA* vuelven a ser reemplazadas por la moda, para disminuir el tiempo de preprocesamiento. Sin embargo, en el experimento histórico (de aquí en adelante, denominado experimento 4) la variable *Sucursales* presenta valores contradictorios puesto que para que un cliente esté vigente debe tener al menos el valor 1 en la variable *Sucursales*, por lo tanto, el valor perdido se asume como un error de consulta y se reemplazan los *missings* correspondientes por dicho valor. Fuera de estas 3 variables, las etapas de preprocesamiento y transformación de variables son idénticas al experimento 3.

Modelamiento y Evaluación: Experimento 4

En la etapa de modelamiento, el experimento 4 sobresale por la maduración de la validación de resultados, separando según meses, las fases que existen al momento de implementar un modelo, como se puede apreciar en la figura 3.5, se debe tener un mes para entrenar y probar(dentro del mismo mes) el modelo, mes en el que se conoce completamente la categoría, posteriormente se aplica el modelo al mes siguiente para ver la efectividad en el tiempo del modelo. Esto se hace eliminando la variable objetivo e incorporándola, posterior a la aplicación del modelo, a la base de datos testeada. De esta manera, se hace lo que se denomina en esta memoria una validación primaria y técnica del modelo. Finalmente se procede a validar con la prueba del modelo sobre un tercer mes, todo ello, para comprobar que el modelo maneja el período de tiempo en el que se aplica.

En cuanto a la etapas de Modelamiento y evaluación en sí (es decir, los algoritmos y métricas), se

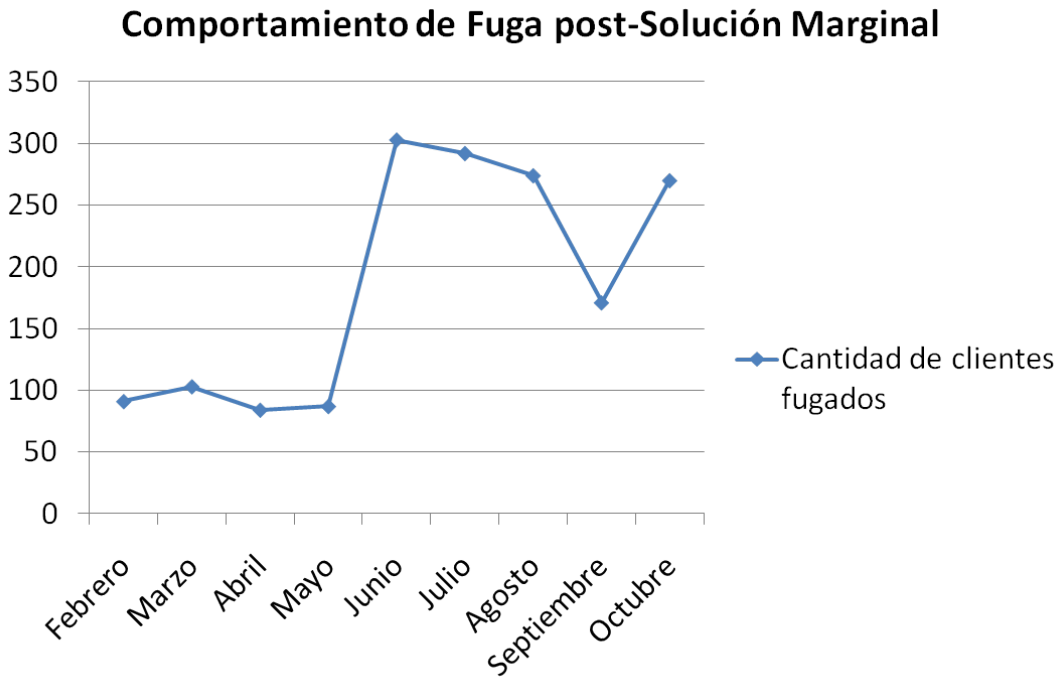

Figura 3.7: Comportamiento de clientes fugados posterior a la implementación de la solución marginal para las bases de datos

deja de lado la alternativa de clusterización surgida en el experimento 3, debido a su complejidad al momento de ser aplicada a los 11 meses y la dificultad que al aplicarla se debe considerarse el efecto de traslado inter-grupal, además, los resultados del experimento 3 fueron inferiores al experimento 2 en el período correspondiente. Por consiguiente, se utiliza el modelo del experimento 2 LADTree, dado que a nivel general se destaca en el experimento 2, lo que conlleva a asumir que posiblemente presentará una solidez en los primeros meses. Para aplicar el formato descrito previamente (figura 3.5 se opta por dejar el mes de Diciembre 2009 como mes de entrenamiento y prueba, luego para la primera validación se decide por un mes i y para la segunda validación se contempla un mes i+1, por ejemplo, si se desea ver la certeza en el mes de Marzo, se entrena y prueba con el mes de Diciembre 2009, luego se realiza la primera validación en el mes de Febrero y posteriormente se valida Marzo, en lo que es la segunda validación. De esta forma, se obtienen los resultados esquematizados en la tabla 3.31, y en los gráficos 3.8, 3.9.

El gráfico 3.8 expresa el error en cuanto a la cantidad de fugados que predice el modelo, respecto a los fugados que acierta, en sí, es un error relativo, mientras que el gráfico sucesor 3.9 bosqueja el acierto de los fugados acertados versus a la cantidad de clientes fugados reales.

Como se puede apreciar en el gráfico antecesor 3.9, el modelo resulta muy eficaz en los primeros meses de aplicación, no obstante, a medida que transcurre el tiempo, su certeza va disminuyendo hasta un punto mínimo distinto de cero, en otras palabras, el modelo no queda obsoleto solamente fuera de un rango de confianza. Obviamente mientras más largo el horizonte de predicción mayor

Cuadro 3.31: Validación Histórica: Experimento 4

Modelo considerado	LADTree				
Mes	Total fugados reales	Fugados acertados	Accuracy [%]	AUC	Medida F(clase 1) [%]
Febrero	91	61	97.91818314	0.908	46.04
Marzo	103	30	97.68760908	0.82	27.4
Abril	84	9	97.70914891	0.815	10.23
Mayo	87	9	97.77520948	0.817	10.47
Junio	303	7	94.66243508	0.68	3.65
Julio	292	11	94.95700335	0.694	5.98
Agosto	274	19	95.30041642	0.697	10.73
Septiembre	171	12	96.63432383	0.739	9.76
Octubre	270	14	95.41199939	0.699	8.56

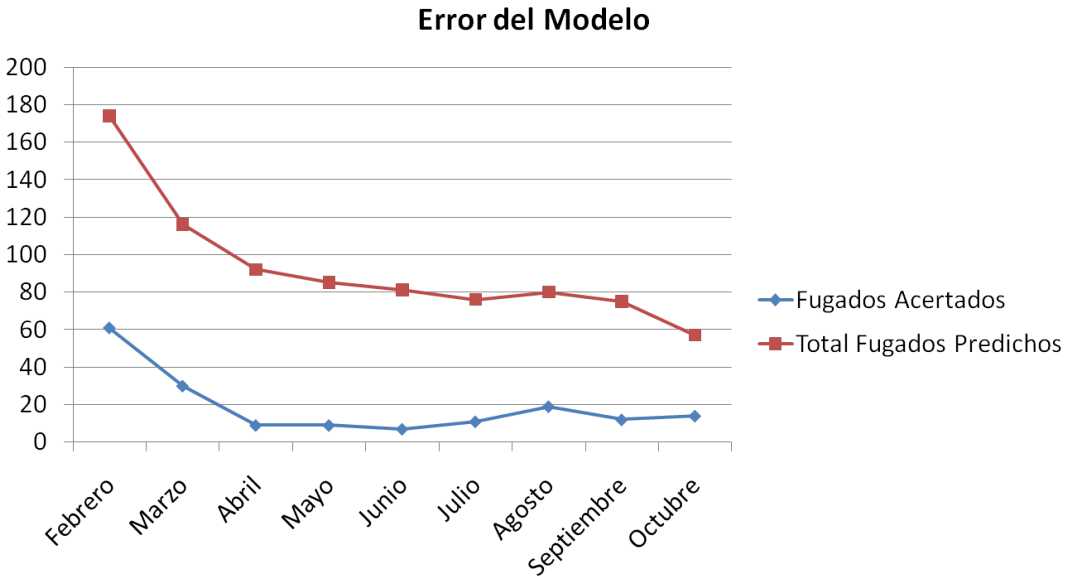

Figura 3.8: Gráfico de desacierto histórico del modelo LADTree

es el error, debido a que es mayor la incertidumbre comprendida a predecir.

Otra característica del experimento 4 que vale resaltar es la entrega de posibles causas que tiene dicha fuga, las cuales están expresadas en base a los pesos asignados por el algoritmo a las variables correspondientes. De esta manera, se bosqueja el resultado de los pesos en 3.10 y su interpretación como variables más influyentes dentro de las cuales se encuentran:

- Consumo en fechas anteriores
- Total de solicitudes comerciales que efectúa el cliente el mes anterior a su fuga
- Posesión previa de Entel phone
- Posesión de planes del tipo 1

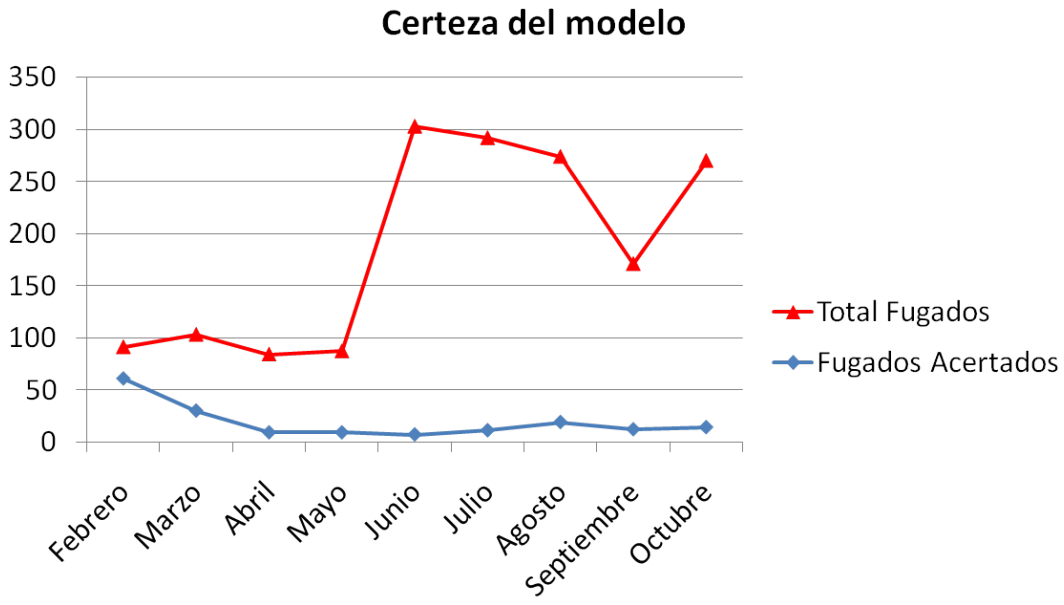

Figura 3.9: Gráfico de certeza histórica del modelo LADTree

- Posesión de planes del tipo 4
- La frecuencia de reclamos técnicos

Valorización de pérdida de ingresos en función del ciclo de vida del cliente

El objetivo del experimento 4; aparte de mostrar la validez de un modelo en el transcurso en el tiempo y que incluso entregue una interpretación posible de las causas de la fuga; es: la validación y valorización de un modelo para probar su veracidad con información extra. Por ende, el experimento 4 se contempla la evaluación monetaria del cliente como pérdida para la compañía. Para ello, se considera el ingreso que se deja de percibir del cliente tomando como referencia su etapa del ciclo de vida. Un cliente del producto NGN dura en promedio 17 meses, dato que fue facilitado en entrevistas informales con el personal del área comercial del producto y que además se valida con la base de datos **NGN_Instalados** accediendo a la pestaña paquetes y filtrando por la condición ESTADO_PLANES= RENUNCIADA, se copia y pega esas instancias en un libro nuevo, luego se busca la variable *INGRESO CONTRATO SAP* para cada contrato y se calcula el delta, posteriormente se promedian estos deltas y se obtiene un valor cercano a los 17 meses. Agregado a esto, en conversaciones con el área de Data Warehouse se señala que el contrato del producto NGN tiene una duración estándar de 24 meses en el sistema. En base a este valor del ciclo de vida se calcula el ingreso perdido según la figura 3.11. Nótese que se denomina ingreso perdido, porque el cliente teóricamente dura hasta su mes 17 o 24 como mínimo para que sea rentable a la empresa.

Donde el caso 1 corresponde a la valorización contemplando el hecho de que el cliente dura 17 meses. Mientras que el caso 2 se refiere a la valorización con una duración por cliente de 24 meses.

```
Text View  Weka Result
|   (1)Total Solicitudes comerciales < 1.5: 1.446,-1.446
|   (1)Total Solicitudes comerciales >= 1.5: 0.479,-0.479
|   |   (2)Entelphone = 0: -1.352,1.352
|   |   |   (4)Trim1_consumo < 6.198: -0.226,0.226
|   |   |   (4)Trim1_consumo >= 6.198: 2,-2
|   |   (2)Entelphone = 1: 0.396,-0.396
|   |   |   (3)Planes_tipo_1 < 5.5: 0.381,-0.381
|   |   |   (3)Planes_tipo_1 >= 5.5: -1.184,1.184
|   (5)Planes_tipo_3 < 2.5: 0.527,-0.527
|   |   (6)Trim2_fact < 30.242: -0.119,0.119
|   |   |   (7)Total Solicitudes comerciales < 0.5: 0.442,-0.442
|   |   |   (7)Total Solicitudes comerciales >= 0.5: -0.536,0.536
|   |   |   (9)Freq_rec < 0.083: 0.381,-0.381
|   |   |   (9)Freq_rec >= 0.083: -0.617,0.617
|   |   (6)Trim2_fact >= 30.242: 0.392,-0.392
|   |   |   (8)Planes_tipo_4 < 9.5: 0.326,-0.326
|   |   |   (8)Planes_tipo_4 >= 9.5: -1.124,1.124
|   (5)Planes_tipo_3 >= 2.5: -0.364,0.364
Legend: 0, 1
```

Figura 3.10: Resultado de ponderación de variables del modelo LADTree en el experimento 4

Caso 1: Valor pérdida cliente i=(17-3)*Promedio de facturación
Caso 2:Valor pérdida cliente i=(24-3)*Promedio de facturación

Figura 3.11: Procedimiento de valorización de fuga en base al ciclo de vida de un cliente

De esta manera, al evaluar monetariamente el período entre Diciembre 2009 y Noviembre 2010 se obtiene la tabla 3.12.

En esta tabla se contemplan todos los clientes etiquetados como fugados, por ende, no existe distinción entre los clientes que abandonaron la compañía voluntaria o involuntariamente, puesto que se desconoce una variable diferenciadora entre ambos tipos de fuga). Por lo tanto, se propone

Mes	No considera Traslado de servicio		Considera Traslado de servicio		Cantidad de fugados
	Ingreso Perdido Práctico (17 meses)	Ingreso Perdido Teórico (24 meses)	Ingreso Perdido Práctico (17 meses)	Ingreso Perdido Teórico (24 meses)	
dic-09	18.143.832	73.785.047	15.422.257	62.717.290	75
ene-10	55.302.718	89.961.038	47.007.311	76.466.882	70
feb-10	14.877.234	53.143.702	12.645.649	45.172.146	46
mar-10	14.390.782	54.652.766	12.232.165	46.454.851	67
abr-10	21.042.649	52.135.630	17.886.252	44.315.285	67
may-10	39.395.928	81.734.397	33.486.539	69.474.237	55
jun-10	5.934.240	45.564.353	5.044.104	38.729.700	54
jul-10	34.322.441	49.392.493	29.174.075	41.983.619	85
ago-10	25.958.534	45.572.319	22.064.754	38.736.472	71
sep-10	33.488.166	59.086.206	28.464.941	50.223.275	71
oct-10	20.310.728	111.120.330	17.264.118	94.452.280	44
nov-10	15.324.195	1.041.642.808	13.025.566	885.396.387	70
Total	**298.491.448**	**1.757.791.088**	**253.717.731**	**1.494.122.425**	**775**

Porcentaje que se traslada (15%)

Figura 3.12: Valorización histórica en base al ciclo de vida del cliente

un supuesto en base a entrevistas informales con el Product Manager del producto NGN: sólo un 15 % de los clientes que se fugan, se trasladan a un servicio superior, por ende, se determinan dos valorizaciones: una que contempla este supuesto y otra que no. Posteriormente, se contemplan los dos valores correspondientes a la vida promedio de un cliente en el producto NGN, o sea, los valores 17 meses y 24 meses. El escenario más realista de la tabla 3.12 es aquel correspondiente a la columna que considera el traslado de servicio y una vida del cliente de 17 meses.

3.3.3. Experimento 5

Como se menciona anteriormente, el experimento 5 no es más que la implementación del experimento 4 a la actualidad. Por consiguiente, la integración de bases es idéntica, así como también, el preprocesamiento de variables. Las diferencias radican en el establecimiento de la marginalidad de bases para el experimento 5 (mientras que en el experimento 4 se imputan en el mes i registros fugados del mes i+1, en el experimento 5 sólo se trabaja con las bases de datos que han sido extraídas mes a mes, es decir, Noviembre 2010, Diciembre 2010 y Enero 2011) , puesto que se pretende traspasar a una propuesta de base de posibles fugados para que el área comercial del producto efectúe una encuesta que permita evaluar desde un punto de vista comercial, la efectividad del modelo considerado.

Preprocesamiento

Esta etapa, en el experimento 5, se diferencia del experimento 4 en lo que se refiere a los registros, debido a que en el experimento 4 se añaden algunos con instancias de otras bases para poder generar la marginalidad propia de cada base, por lo tanto, en el experimento 5 se toman las bases para el mes de Noviembre y Diciembre, pues estos meses carecen del problema con la marginalidad,

dado que se le ha hecho un seguimiento a la base de datos **NGN_Instalados**, lo que se traduce en tener una marginalidad real.

Los registros etiquetados como vigentes vuelven a ser aquellos que ruts que no tienen sólo contratos cuyo ESTADO_PLANES es igual a Error o Renunciada.

Respecto a La etapa de transformación, ésta es idéntica al experimento 4, que a su vez, es similar al experimento 3.

Modelamiento y Evaluación: Experimento 5

En el modelamiento del experimento 5, se utiliza el algoritmo LADTree, usando el mes de Noviembre 2010 como entrenamiento y prueba primaria, Diciembre 2010 como mes de prueba de validación y se entrega una base con las posibles instancias pertenecientes a la clase "fuga" en el mes de Enero de 2011, pues el cliente puede dejar la empresa en cualquier día del mes de Enero.

La evaluación técnica se adquiere a finales de Enero, en la que obtiene el siguiente resultado expresado en la tabla:

Cuadro 3.32: Tabla de confusión: Experimento 5

Modelo	LADTree		
	Categorías Reales		
Categorías Predicción	0	1	Precisión
0	6248	44	99.30 %
1	111	3	2.63 %
Recall	98.25 %	6.38 %	
Accuracy	97.58 %		
Medida F (clase 1)	3.73 %		

La tabla 3.32, indica que de todas las instancias fugada realmente solamente se captura a un 1 %, lo que es un valor menor del esperado.

Por otro lado, la evaluación comercial se realiza a través de un MAS (muestreo aleatorio simple) sobre la base que contiene 114 instancias que se señalan como fugadas, la cual proviene de otra base de 6406 instancias. Por consiguiente, se ejecuta una encuesta telefónica en la que se generan 50 contactos con los clientes del producto NGN, de los cuales 14 expresan una "intención de fuga" (esta última variable fue creada por la persona encargada de hacer las entrevistas), correspondiente al 30 % de la muestra, como dato adicional, la persona responsable por las entrevistas posee experiencia previa relevante en este tipo de acciones. Ese 30 % de instancias que exponen una intención de fuga se evalúan monetariamente de la misma forma que en el experimento 3, es decir, se mide la fuga de ingresos mensual, que entrega como resultado la siguiente tabla:

Si bien al comparar los resultados técnicos con los comerciales se vislumbra una fuerte diferencia, se converge a un resultado real al valorizar la muestra, debido a que se detectan clientes que facturan un monto significativo respecto al producto. Sin embargo, esto plantea la duda sobre la veracidad de la variable *Fuga*, la cual, en conversaciones con el personal de sistemas solamente capta una parte de la fuga total, puesto que la variable *Fuga* es extraída de la plataforma Oracle Workflow

Cuadro 3.33: Valorización de la evaluación comercial del Experimento 5

Glosa	Cantidad de respuestas
Sin información	2
No desea fugarse	33
Sí desea cambiarse	15
Universo total	50
Efectividad del modelo	30.00 %
Valorización Evaluación Comercial	30,687,421
Valorización Evaluación Técnica	5,687,143

que sólo contiene los registros de los call center.

Un punto relevante en este experimento es el hecho de que la evaluación comercial crea una disyuntiva con la evaluación técnica. Por lo tanto, se presenta una problemática conjunta entre las validaciones técnicas previas, comercial y técnica en el mes donde se busca predecir.

3.4. Fase 4: Estrategia continua y concreta

En la fase final del proyecto se plantea otro refinamiento completo al KDD propuesto en los experimentos anteriores, con el objeto de aclarar la disyuntiva triple en las evaluaciones. Sin embargo, no se valida históricamente como el experimento anterior. Esta última fase comienza en Enero del año 2011 y dura hasta el mes de Abril del año 2011.

Para esta fase se busca tratar un problema que se ha dejado de lado en los experimentos anteriores, en la etapa de preprocesamiento, los valores de fuera de rango u *outliers*, los cuales, se encuentran directamente relacionados con el tema del desbalanceo o rarezas de clases.

Cabe destacar que esta fase consta de un experimento general que se puede dividir en experimento 6 y 7, los cuales poseen las mismas etapas de Integración y Transformación. Las diferencias en la etapa de Preprocesamiento se deben únicamente al tratamiento de los valores fuera de rango, mientras que en la etapa de Modelamiento resulta natural que difieran las configuraciones de los modelos usados (pues son dos experimentos ejecutados en dos períodos distintos), posteriormente en la etapa de Evaluación, la diferencia subyace en la evaluación comercial y monetaria presente en el experimento 7.

3.4.1. Integración

El experimento 6 tiene la misma integración que el experimento 5, ya sea en cuanto a las bases utilizadas y a las variables tomadas. Sin embargo, el experimento 7 posee una salvedad en esta etapa respecto a las bases finales[16], las cuales tienen un problema relacionado directamente con la marginalidad de aquellos meses que no se tienen, lo que impide la integración correcta, particularmente la base de datos **NGN_Instalados**, puesto que ocurre una ausencia de ésta en el mes Enero de 2011.

[16]Estas son las bases posteriores a las transformaciones y que ya contemplan los 6 meses de estudio

No obstante, el objetivo de la solución a este problema es ver qué clientes son vigentes en el mes j, de esta forma, se toman los clientes vigentes del mes j-1 y se sustraen aquellos que terminaron contrato en el mes j. El procedimiento para llevar esto a cabo fue extraer los ruts vigentes de Enero a partir de la combinación de la base de datos **NGN_Instalados** de Diciembre 2010 más los fugados en Febrero 2011.

3.4.2. Preprocesamiento

La cantidad de valores ausentes y estrategias para los experimentos 6 y 7 se muestran en las tablas 5.104 y 5.105 respectivamente.

En la base de reclamos comerciales y técnicos, se tiene una detalle previo a considerar el preprocesamiento de las bases, y esta es que en la glosa nombrada como RESOLUCION en las bases respectivas (**OWF** y **Boletas Técnicas**) se filtra por aquellos registros que son "Cursados" (automático o normal), "Rechazados" (Automático, por jefe o normal), dejando de lado los registros con la etiqueta "Anulados". Esto se debe a que los registros con el estado anulados, pueden deberse a errores del sistema por lo que no componen una parte del comportamiento del cliente.

Registros

Para ambos experimentos, se estudian en mayor detalle los estados de los contratos de cada cliente, por ende, se determina que para descubrir a los clientes o ruts vigentes, se deben usar aquellos contratos con estados FACTURAR y ACTIVADA, y el resto de los estados se descartan. Esto se justifica en que los estados INHABILITAR e INHABILITADO sólo tienen dos salidas para el cliente, en la primera el cliente vuelve a facturar, mientras que en la segunda, el cliente es retirado de la compañía por incumplimiento de contrato, lo cual, califica como una fuga involuntaria. El otro estado en cuestión es el de ACTIVADA NO INSTALADA, el cual, induce a identificar al cliente como nuevo, debido a que su servicio aún no se encuentra instalado, por lo que no se cuenta con la cantidad suficiente de información como para evaluar la predicción de fuga. Estos estados discriminados se agregan a la lista de los estados RENUNCIADA, ERROR, Friendly User.

Registros: Experimento 6 Ahora bien, el preprocesamiento de variables es equivalente para ambos experimentos desde un punto de vista de valores ausentes, sin embargo, en el caso de los valores fuera de rango, difiere. En el experimento 6 se determina, en un análisis exploratorio (de ensayo y prueba) que algunos registros añaden más ruido a la base de datos que información. Con este supuesto, se procede a establecer reglas sobre ciertas variables que permitieron encontrar aquellos registros que contienen dentro de los valores de sus variables, valores fuera de rango. Una descripción breve de las reglas se presenta en la tabla 3.34. Bajo la identificación de estos valores fuera de rango con sus respectivas instancias, se eliminan los registros(instancias) cuyo fundamento se encuentra en un breve estudio de 4 meses(de los cuales se muestran dos meses) sobre las variables *FAC1_Consumo, FAC2_Facturacion, Q_ANIS, Trim2_fact, Sucursales y Cat_Corp*. Los resultados de este estudio se expresan en las tablas 5.89 a 5.98; en ellas se puede apreciar la relación inherente entre la variable objetivo y las distintas variables seleccionadas para el establecimiento de reglas. Sintetizando lo anterior, se aplica una eliminación *listwise* a las instancias que tengan un valor *outlier* en algunas

de las variables sujetas a las reglas impuestas. La variación generada en cantidad de registros se representa en la tabla 3.35.

Cuadro 3.34: Reglas inducidas para la eliminación de valores fuera de rango

Variable	Condición
FAC1_Consumo	< 2
Q_ANIS	≤ 20
Trim2_fact	≤ 2000000
Cat_corp	$\neq 3$
FAC2_Facturacion	$\in [-0{,}2; 0{,}2]$
Sucursales	< 4

Cuadro 3.35: Variación en cantidad de registros en las bases de datos por mes

	Frecuencia de Variable Fuga_pos				
Categoría variable Fuga_pos	Agosto	Septiembre	Octubre	Noviembre	Diciembre
0	6242	6301	6249	6305	6359
1	142	85	67	45	47
Total	6384	6386	6316	6350	6406
Categoría variable Fuga_pos Post listwise	Agosto	Septiembre	Octubre	Noviembre	Diciembre
0	3693	3675	4379	4419	4421
1	75	60	58	30	38
Total	3768	3735	4437	4449	4459

Cabe destacar que la base considerada en Diciembre del 2010 es confeccionada de forma diferente a las otras, debido a que esta base se prepara en el mismo experimento 6, por ende, no presenta influencia alguna por el problema de la marginalidad al cual si están sometidas las otras bases provenientes del experimento 4. Aún así, posterior al *listwise* de los valores fuera de rango, todas las bases contempladas quedan en el mismo orden de magnitud.

Registros: Experimento 7 Para el experimento 7 se cambia la estrategia de los *outliers*, dado que eliminar vía listwise resulta en una alteración al comportamiento común de los clientes NGN, tal y como se pudo apreciar posteriormente en el mes de Enero 2011 (expresado en las tablas 5.99 a 5.103) . Dicho cambio se expresa en contemplar las instancias con este tipo de valores pues corresponden a los clientes con mayor facturación y determinan el comportamiento de un grupo reducido de clientes.

Variables

En la etapa de preprocesamiento se ocupan las estrategias del experimento 3 respecto a las todas las variables. Estos para el experimento 6, se denotan con sus estrategias correspondientes en la tabla

5.104. Por otro lado, para el experimento 7 se los valores perdidos y sus estrategias se representan en la tabla 5.105. Un punto relevante es que las variables extraídas de la base de datos **Seg_empresas** presentan una sincronización en sus valores perdidos donde no existe ningún fugado involucrado para todas las bases preprocesadas, por lo que se decide eliminar los casos que no contengan los cinco campos.

3.4.3. Transformación

Para ambos experimentos los temas de transformación fueron idénticos, las cuales, son descritas a continuación:

- La primera de las transformaciones realizadas tiene el propósito de refinar la transformación del promedio ponderado de los experimentos anteriores, para las variables de facturación y consumo de la base **Proforma**, puesto que las variables *TRIM* no están suficientemente fundamentadas y generalmente mantienen la correlación excesiva entre ellas como se refleja en los gráficos 5.1, 5.2, 5.3, 5.4. En base a este antecedente se opta por efectuar una transformación con el análisis de componentes principales (ACP), el cual se encuentra inserto en el análisis factorial. Este cambio permite reducir la dimensionalidad total de las variables de la base **Proforma**, desde 12 variables a 2, las que entregan una puntuación, por lo tanto, se mantienen las variables de *Promedios (comunes y ponderados)* para obtener una interpretación de la facturación y consumo del cliente, dichas variables promedios no se usan en el aprendizaje del modelo. Ahora bien, esta sustitución se encuentra fundamentada en la tabla 3.36 para las bases correspondientes a 3 de los 4 meses, donde para el caso de la base de Marzo de 2011, la matriz de factores generada no era definida positiva, por lo tanto, el software utilizado para crear dichos factores no facilita la métrica del KMO. Sin embargo, se puede argumentar que la adecuación es correcta bajo el criterio de la matriz de correlaciones reproducida, que señala que *"se estimen las diferencias entre correlaciones observadas y reproducidas"* [58] y si se tiene una gran cantidad de estas diferencias mayor a un parámetro de corte entonces el modelo factorial no bosqueja bien los datos. Bajo este criterio, se muestran las tablas de correlaciones observadas (tabla 5.111) y reproducidas (tabla 5.110) con sus diferencias respectivas (en la tabla 5.112), a las cuales si se les aplica un parámetro de corte de 10 % ninguna cantidad lo sobrepasa y si se aplica un parámetro del 5 % cerca del 10 % de los casos lo sobrepasan. Un análisis de los distintos valores para los parámetros se puede apreciar en 5.113.

Cuadro 3.36: Pruebas de calidad del análisis factorial en el Experimento 7

KMO y prueba de Bartlett				
Meses		dic-10	ene-11	feb-11
Medida de adecuación muestral de Kaiser-Meyer-Olkin.		0.902	0.901	0.919
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	178316.361	186998.918	202874.627
	gl	66	66	66
	Sig.	0	0	0

Los nombres para los dos factores que se generan son FAC1_Consumo y FAC2_Facturacion,

etiquetas que se crean contemplando el resultado de la matriz de factores rotados para todos los meses que se representa en la tabla 5.107. Los altos valores en las variables de *Consumo* para el factor 1 denotan la relación del mismo con las variables asociadas al consumo, análogamente se vislumbran altos valores para las variables de *Facturación* en el caso del factor 2. Las comunalidades presentadas en la tabla 5.106, describen que el procedimiento logra extraer de forma satisfactoria los factores a partir de las variables, así como también, la tabla 5.108 y 5.109 muestran que se pierde una cantidad reducida de información que se representa por la varianza de los datos.

- La segunda transformación que se realiza en estos experimentos es la de la EDAD, es decir, se origina una variable denominada EDAD a partir de la variable *PRIMERA_INSTALACION* y la supuesta fecha de aplicación de acciones correctivas o, en este caso, de la encuesta comercial (que para ambos experimentos fue el primer día del mes siguiente) . La razón para considerar la variable *PRIMERA_INSTALACION* en vez de *INGRESO CONTRATO SAP*, fue el hecho de que esta última deja de ser un buen referente, debido a que el cliente puede tener contratos hace más de un año pero contrata un contrato nuevo en este mes y vence el contrato anterior. Dicho caso aparecerá, entonces, como un cliente de vida igual a 1 mes, lo cual no es verídico. En cambio, la fecha de la primera instalación no se modifica, por lo tanto, sólo existe la fecha que prácticamente indica el instante en que el cliente ingresa a la empresa de telecomunicaciones.

- Una tercera transformación se compone de la variable *Competencia*, la cual presenta una distribución de frecuencias dispareja observable en los histogramas que se muestran en las figuras 5.5, 5.6, 5.7, 5.8. Respecto a lo anterior, se propone la creación de una variable binominal que indique si existe competencia o no, en vez de la compañía correspondiente.

- Cuando se transforman los reclamos comerciales, se toma en cuenta el promedio, el cual se aplica de la siguiente manera: Se suman las cantidades de reclamos por cada mes (o sea, se suman los reclamos hechos en los 6 meses), y se divide por el número de veces que el cliente "entra" en la base **OWF**, es decir, los reclamos comerciales promedio del cliente en las veces que interactúa con la compañía, lo que en forma matemática es $Reclamos_{comercial} = \frac{1}{\sum_i^6 IndRec_{\text{cliente j en mes i}}} \sum_{i=1}^{6} Reclamos_{\text{cliente i}}$, donde "Reclamos" es la cantidad de reclamos y la función IndRec es:

$$IndRec_{\text{cliente j en mes i}} = \begin{cases} 0 & \text{si el cliente j no reclama en el mes i} \\ 1 & \text{si el cliente j reclama en el mes i} \end{cases}$$

- La transformación de los reclamos técnicos es la misma que en el experimento anterior, es decir, aquella sustentada sobre el modelo RFM.

Modelamiento

Antes de proseguir a mencionar los principales modelos que se utilizan en estos experimentos se debe mencionar la presencia de una clusterización previa, presente en ambos experimentos. Por

ende, las diferencias y semejanzas de las clusterizaciones entre ambos experimento se describen a continuación:

- El experimento 6 considera una clusterización general notando a la variable *Fuga* como una variable cualquiera dentro del algoritmo (además, de ser la variable objetivo), en cambio, el experimento 7 contempla una clusterización por comportamiento de fuga, esto quiere decir que no toma parte en la clusterización como variable de aprendizaje.

- El número de clústers es definido por distintos criterios, en el primero se opta por un criterio basado en el diagrama jerárquico del algoritmo Clúster Jerárquico. En cambio, el experimento 7, además de este algoritmo, se utiliza la validación del criterio BIC.

- Respecto a las similitudes, la primera es que tienen ambos experimentos, es el tema de que la agrupación no se estudia a fondo, por lo tanto, la etiquetación de estos grupos solamente se basan en conclusiones de dos o tres variables principales y carece del análisis suficiente como para denominarse segmentación.

- Otra similitud es el hecho de que usan las mismas variables, las cuales se encuentran en la tabla 3.37:

Cuadro 3.37: Variables utilizadas para la generación de conglomerados en los experimentos 6 y 7

Variable	Tipo de variable
Edad	Continua
Plan_tipo_1	Continua
Plan_tipo_4	Continua
Plan_tipo_5	Continua
FAC1_Consumo	Continua
FAC2_Facturacion	Continua
Sucursales	Continua
Mount_rec	Continua
TotalSolicitudescomerciales	Continua
Icp	Continua
Imagen	Nominal

El descubrimiento de este set de variables se origina en el experimento 6, en base a ensayo y error de una segmentación óptima para las instancias correspondientes. La idea en estado bruto es la de identificar grupos de clientes fugados, otros que estuviesen a punto de irse y unos que no se fuguen en un período determinado. De esta manera, se comienza a probar distintos valores de K (grupos) tal que se vislumbre una descripción de los clientes con un grupo que abarque toda la fuga.

Modelamiento: Experimento 6

El modelamiento del experimento se basa en el ensayo y error de distintos valores de K y visualizando la descripción que generan respecto a las variables para cada uno de los grupos originado. Tras este análisis se deducen que las variables a usar son: *FUGA POS, IMAGEN, FAC2_Facturacion, FAC1_Consumo, EDAD, Sucursales, MOUNT REC, PLAN TIPO 1, PLAN TIPO 4, PLAN TIPO 5, TOTAL SOLICITUDES COMERCIALES, ICP*. Esta investigación permite almacenar la divergencia de la fuga inter-grupal, aplacando la variación de la misma. Un punto relevante es el hecho de que las variables sostienen la una distribución similar entre las bases finales de distintos meses, en otras palabras, los grupos se mantienen en el tiempo para la base de datos preprocesada.

Respecto al cálculo del número de conglomerados, se utiliza el criterio de *cross validation* o validación cruzada, ejecutando el modelo de clusterización de Rapidminer W-EM. El resultado de dicha acción, por base mensual preprocesada, se puede apreciar en la tabla 5.120. Cabe señalar que este criterio se usa para escalar el número de grupos puesto que lo que se busca en esta etapa de modelamiento es la segmentación de la base de datos. Sin embargo, este criterio genera un gasto computacional mayor lo que se traduce en realizar un muestreo estratificado de 1000 instancias para cada mes, en pos de obtener los resultados que se muestran en la tabla 5.120, 5.114. Por consiguiente, se puede decir que el número de clústers varía entre 4 y 5 para los últimos meses, mientras que se denota una posible externalidad en los meses de Agosto y Septiembre.

La distribución de las instancias en estos nuevos grupos fue la siguiente, para cada una de las bases mensuales preprocesadas:

Cuadro 3.38: Distribución de clústers por base mensual

Conglomerado	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Pasivos	1919	1729	1792	2254	1999	2263
Fugados	312	275	337	345	517	801
Insatisfechos	1110	1085	966	1128	1151	1063
Reactivos	424	679	640	710	782	332
Combinados	3765	3768	3735	4437	4449	4459

Cabe destacar que para llegar a la etiquetación o nombramiento de los grupos obtenidos en este experimento, se usan los centroides resultantes del algoritmo *Two Step Cluster*, cuyas tablas respectivas son 5.115 a 5.119. Estas últimas son parte del último mes segmentado, debido a que esta segmentación se realiza para cada mes contemplado en este experimento como base preprocesada de prueba y validación, las que, en este experimento, son: Agosto de 2010, Septiembre de 2010, Octubre de 2010, Noviembre de 2010 y Diciembre de 2010. No obstante, la caracterización de los grupos resultantes se efectúa particularmente en el mes de Diciembre para obtener una representación actualizada al momento de efectuar una prueba práctica en el negocio. Otra particularidad es el hecho de que cada base de prueba y validación (bases preprocesadas con datos históricos de 6 meses hasta el mes correspondiente) contiene en una variable correspondiente al grupo de pertenencia de la instancia en la base de prueba y validación anterior, es decir, una instancia en la base de Septiembre de 2010 tiene una variable llamada CLUSTER CON ICP, que indica el grupo de pertenencia al cual se llega mediante la segmentación sobre la base mencionada (Septiembre 2010), y además, cuenta

con otra variable (denominada "N-1") que apunta al grupo en donde se encontraba la instancia en la base de Agosto de 2010 bajo la segmentación que se lleva a cabo en esa base preprocesada.

Una vez elaborada las cinco bases (Agosto 2010, Septiembre 2010, Octubre 2010, Noviembre 2010 y Diciembre 2010), se divide la aplicación de modelos por grupos. Para el entrenamiento y prueba se considera la variable *CLUSTER CON ICP* en pos de dividir la base y estudiar los distintos modelos, luego con el objeto de validar el resultado anterior se toma la base sucesora preprocesada y se divide dicha base por la variable "N-1" para establecer veracidad en la aplicación. Por ejemplo, si se entrena y prueba con Agosto 2010, se valida con Septiembre 2010 haciendo un contraste entre la variable *CLUSTER CON ICP* de la base Agosto 2010 y la variable "N-1" de la base Septiembre 2010. Un punto importante a explicar es el hecho de que en caso de aplicar la misma variable (*CLUSTER CON ICP*) en la base sucesora (para el ejemplo Septiembre 2010) se entra en una discordancia, debido a que si se desea generar una predicción que contemple una campaña de acciones correctivas no se tendrá esta variable para un mes determinado (que en el ejemplo puede ser Octubre 2010), puesto que no se contará con las variables actualizadas para elaborar la segmentación correspondiente.

Contemplando la explicación anterior se presentan los resultados de los modelos prototipo para los grupos previamente descritos en la tabla 5.122, donde para cada uno se muestran los mejores 8 prototipos y sus configuraciones respectivas se muestran en la tabla 5.121.

Modelamiento: Experimento 7

El experimento 7 se origina en la necesidad de abordar los valores fuera de rango de una manera apropiada distinta de una técnica estilo *listwise*, debido a que ésta implica la eliminación de información de la base de datos. De esta forma, aludiendo a una de las técnicas sugeridas por Weiss en [107], se procede a agrupar a las instancias mediante una segmentación de dos etapas, en donde se utiliza la variable *Fuga* conocida (es decir, la variable "supervisora") para dividir la base preprocesada en dos, con ello se obtiene una subbase preprocesada conformada netamente por las instancias cuya clase es "No fugado" y una segunda consistente en instancias con una única clase "Fugada". Estas bases se denominan base NF y base F respectivamente. Un esquema que sintetiza los anterior puede ser apreciado en la figura 3.14:

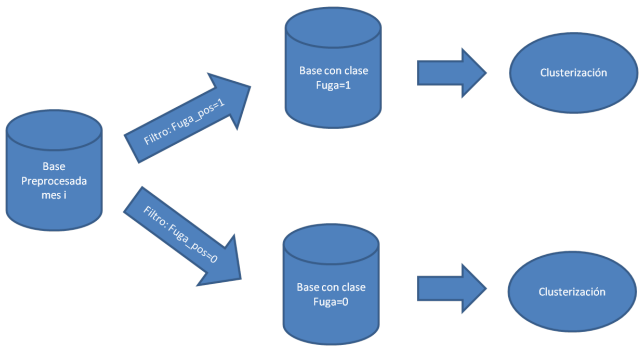

Figura 3.13: Procedimiento de clusterización del experimento 7

Ahora bien la segmentación realizada en este experimento pretende identificar los comporta-

mientos propios de cada grupo de cliente. Por ende, el procedimiento a seguir para la detección de los conglomerado fue partir por la división de la base de datos original. Para trabajar en este experimento, se consta de 3 bases de datos de entrenamiento, prueba y validación (Diciembre 2010, Enero 2011, Febrero 2011), y una base (del mes de Marzo 2011) que se usa en pos de entregar una predicción, bajo la cual, el área comercial puede ejecutar una encuesta y de esta forma evaluar comercialmente el modelo implementado. Por lo tanto, al dividir las bases de datos se tiene la tabla 3.39.

Cuadro 3.39: Cantidad de registros en las subbases generadas

Número de registros por subbases								
Mes	Base Diciembre 2010		Base Enero 2011		Base Febrero 2011		Base Marzo 2011	
Categoría Variable Fuga	1	0	1	0	1	0	1	0
Frecuencia	52	5457	57	5707	68	5653	0	5692

Posteriormente se ejecuta una segmentación en cada una de estas subbases, cuyo criterio de elección en cantidad de grupos fue el BIC. Los resultados de este criterio, que mide la variación en la información de la base de datos, son observados en la tabla 3.40, bajo la cual se puede apreciar que las bases se adaptan bien a un número de grupos igual a 3. Por consiguiente, se efectúa la clusterización correspondiente con el algoritmo *Two Step Cluster* que da como resultado las tablas 5.124 a 5.130, que permiten etiquetar los grupos pues para el primer grupo de la subbase NF se tienen cliente con pocas sucursales (pequeños) y con una buena imagen de la empresa, el segundo grupo tiene una cantidad similar de sucursales y un comportamiento parecido al grupo 1 en el resto de las variables, no obstante, tiene una mala imagen de la empresa, el tercer grupo posee varias sucursales y una cantidad de solicitudes mayor que los otros dos grupos. De esta forma, se pueden etiquetar los grupos como sigue: Clientes Pasivos (1), Clientes Reactivos (2), Clientes Grandes(3) . Análogamente en la subbase F de Febrero11, se pueden obtener las mismas etiquetas bajo el razonamiento señalado, sin embargo, los números son distintos, en esta nueva subbase los grupos se etiquetan de la siguiente manera: Clientes Grandes (1), Clientes Pasivos(2), Clientes Reactivos(3).

Una peculiaridad en relación a los segmentos resultantes, es el hecho de que tienen su análogo en su parte adversa, es decir, existen 3 grupos en la base NF, así como también existen 3 grupos en la base F, además, cada grupo de la base NF tiene un grupo que se comporta de manera similar en la base F, es por ello, que la etiquetación de los nombres es válida para ambas subbases.

Posterior a la clusterización que converge en segmentación, se busca modelar la situación para cada grupo, a través de distintos modelos y configuraciones, dentro de los cuales destacan la regresión logística y las Support Vector Machines, tal y como se contemplaron en el experimento 6. Los resultados de los prototipos de prueba, en el experimento 7, son mostrados en la tabla 5.131, cuyas configuraciones se representan en la tabla 5.132.

Un punto relevante es mencionar que estos resultados son válidos para la ejecución de una predicción posterior, esto se puede argumentar aludiendo al procedimiento de cómo se entrena y valida cada modelo, el cual queda expresado en la figura 3.14. En ella, se puede apreciar que se usa una base preprocesada de prueba y validación (Diciembre 2010 o Enero 2011), luego se efectúa la primera validación con una segunda base preprocesada (Enero 2011 o Febrero 2011) y una segunda

Cuadro 3.40: Resultados de métricas BIC para la detección del número de grupos para las subbases en el experimento 7

	Diciembre10 NF	Diciembre10 F	Enero11 NF	Enero11 F	Febrero11 NF	Febrero11 F	Marzo11 NF
Número de conglomerados	RMD(c)	RMD(c)	RMD(c)	RMD(c)	RMD(c)	RMD(c)	RMD(c)
1							
2	2.235	1.815	2.164	1.812	2.648	1.678	2.298
3	2.771	1.858	2.648	1.924	4.035	1.378	2.195
4	2.337	1.727	1.675	1.152	1.214	2.416	2.412
5	1.054	1.027	1.570	1.429	1.296	1.469	1.198
6	1.082	1.077	1.191	1.136	1.465	1.003	1.132
7	1.339	1.128	1.054	1.469	1.030	1.092	1.385
8	1.112	1.468	1.248	1.059	1.071	1.024	1.397
9	1.042	1.223	1.148	1.082	1.023	1.199	1.075
10	1.382	1.156	1.214	1.352	1.222	1.309	1.050
RMD: razones de las medidas de la distancia							
(c): Las razones de las medidas de la distancia se basan en el número actual de conglomerados frente al número de conglomerados anterior.							

validación con una tercera base (Febrero 2011 o Marzo 2011) . No obstante, en el caso de la base de Marzo 2011, tal y como señala la figura 3.14, no posee una variable de agrupación para las instancias fugadas. Esta ausencia se debe principalmente al hecho de que el objetivo de este último experimento es verificar a las instancias fugadas del mes de Marzo, cuya fuga puede ser obtenida una vez terminado el mes de Abril (de 2011), a través de una evaluación técnica y comercial. Por consiguiente, no se genera la primera división o primera etapa de la segmentación, por lo tanto, no existe una variable de agrupación. Sin embargo, con tal de conseguir una base preprocesada con las mismas variables, se opta por asumir que todas las instancias de Marzo 2011 fuesen no fugadas, y agrupar bajo la explicación de los grupos análogos previamente mencionada.

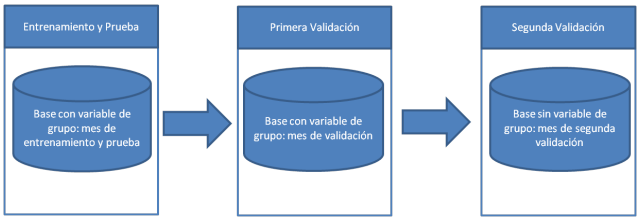

Figura 3.14: Esquema genérico de modelamiento en el experimento 7

Evaluación

La evaluación del experimento 6 solamente existe a nivel técnico, por lo cual, su evaluación final y modelo a usar es simplemente el mejor de los resultados de la tabla 5.122. Los cuales se muestran en las tablas 3.41, 3.42.

A diferencia de este experimento, el experimento 7 entabla una evaluación técnica sobre la incertidumbre, es decir, se mide la efectividad de un entregable sobre un grupo de clientes. Con la base de cliente predicho como posibles fugados, se tiene una variable que etiqueta a los clientes con el grupo al cual se han asignado. Esta variable en la base de posibles fugados, sólo tiene dos

Cuadro 3.41: Resultados finales experimento 6

Modelo	TN	FP	FN	TP	Grupo
RL 4	633	55	1	2	Reactivos
SVM 19	2184	33	15	2	Pasivos
SVM 25	1981	236	11	6	Pasivos
RL 13	1049	64	3	3	Insatisfechos
RL 14	261	74	1	3	Fugados
RL 15	281	54	2	2	Fugados

Cuadro 3.42: Resultados finales experimento 6 con métricas

Modelo	*Accuracy* [%]	*Recall* 0 [%]	*Recall* 1 [%]	Precisión 0 [%]	Precisión 1 [%]	Medida F general [%]	Medida F clase 0 [%]	Medida F clase 1 [%]
RL 4	91.896	92.006	66.667	99.842	3.509	62.586	95.764	6.667
SVM 19	97.851	98.512	11.765	99.318	5.714	53.795	98.913	7.692
SVM 25	88.944	89.355	35.294	99.448	2.479	56.074	94.132	4.633
RL 13	94.013	94.250	50.000	99.715	4.478	60.496	96.905	8.219
RL 14	77.876	77.910	75.000	99.618	3.896	61.727	87.437	7.407
RL 15	83.481	83.881	50.000	99.293	3.571	58.170	90.939	6.667

valores, Ausente o 99, el primero indica que el cliente perteneció a algún grupo de fuga en el mes anterior, en cambio, el segundo número representaba a los cliente denominados como GRANDES. Por consiguiente, se toma una cantidad distribuida de cada grupo. Cabe decir que el entregable de predicción consiste de una base de posibles fugados con 162 Ruts, de los cuales 70 pertenecen al grupo GRANDES o Gigantes y 92 eran de un grupo de fugados en el mes anterior. En las tablas 5.131 y 5.132 se presentan los modelos prototipo y sus configuraciones respectivas, de lo que se pueden destacar los siguientes:

Cuadro 3.43: Resultados finales experimento 7

Modelos	TN	FP	FN	TP	Observación
SVM 3	2165	97	0	67	No usa grupo pasivo en entrenamiento
LADTree	4933	31	0	67	Predice pocas instancias para Marzo
SVM 6	4949	15	0	67	No predice instancias para Marzo
SVM 10	4948	16	2	65	Predice pocas instancias para Marzo

Cuyas métricas pueden ser observadas en 5.133. Muchos modelos con buenas predicciones en la primera validación se descartan debido a su baja predicción de instancias declaradas como futuras fugas, de esta manera, la decisión sobre qué modelo escoger se sustenta en la observación propuesta en la tabla 3.43. Un punto no menor a destacar, es el hecho de que se escoge como parámetro de corte el valor 0.26818 (mediante el ensayo y el error), el cual se valida posteriormente al obtener la variable *Fuga* para el mes de Mayo, dicha validación se muestra en la tabla 5.134, así como también, en los gráficos 5.10, cabe destacar, que el criterio seleccionado es el número 4 debido

a que la cantidad de fugados totales que predice (atributo P en la tabla 5.134) es 162 y a que en conversaciones con el personal de la compañía contempla la capacidad de evaluación comercial. El modelo que es finalmente aplicado es el SVM 3. El resultado técnico sobre el contraste de la fuga real versus la predicción se expresa en la tabla. 3.44.

Cuadro 3.44: Resultados finales experimento 7 con métricas

	Categorías reales		
Categorías predichas	0	1	Precision
0	5492	38	0.993
1	157	5	0.031
Recall	0.972	0.116	
Accuracy	96.574 %		
Medida F Total	52.763 %		
Medida F clase 1	4.878 %		

Donde el gráfico de las curvas ROC para la primera validación puede ser observado en la figura 5.9.

El resultado de la tabla 3.44 muestra poca efectividad del modelo frente a la variable seleccionada o bien, frente al objetivo planteado, debido a que éste se formula en pos de preparar la predicción para el siguiente mes, estrategia que tiene poca utilidad si se desea entablar acciones correctivas. Sin embargo, para evitar la espera de los 6 meses posteriores para medir la completa efectividad, se procede a efectuar la encuesta telefónica que busca verificar los estados de los clientes. De esta forma, el esquema de predicción es similar a los experimento 4 y 5.

El procedimiento para entablar dicha encuesta se describe a continuación: para la elaboración de la evaluación comercial se divide la base de 162 ruts en 70 ruts grandes y 92 previamente fugados, luego dentro de cada uno se intenta mantener la proporción para lo cual se extrae el porcentaje respectivo que corresponde a un 43 % y un 57 % respectivamente. Luego se constata la cantidad de contactos que se pueden lograr a nivel de capacidad, que debido a procesos internos de la empresa se establece en 31 contactos. Con este valor se toma el porcentaje de cada grupo sobre 31, es decir, se escoge a 12 clientes grandes y 19 previamente fugados, sin embargo, en la práctica se logra contactar a los 31 ruts propuestos y a 12 más, llegando a un universo de 43 ruts. Finalmente se muestra el resultado de dicha encuesta:

Esta última tabla presenta otra perspectiva acerca de la efectividad del modelo, puesto que en la encuesta comercial sólo se contacta a un cliente, sin embargo, también se observa la presencia de ANIS que están fuera de servicio, dado que al intentar ser contactados una grabación da cuenta de la afirmación anterior. Agregado a este hecho, también se detectan ANIS que tienen como destino un estado de retiro según el sistema on-line SAP-CRM. Ahora bien, el hecho de que un ANI (teléfono que la empresa entrega) esté sin tono, no significa que el cliente esté fugado. No obstante, en un breve estudio con bases de datos anexas a las consideradas en el estudio se descubre que el ANI encontrado en la base de posibles fugados es el ANI Principal (es decir, es como el Teléfono representante del contrato), luego en conversaciones con el personal de la empresa se concluye que este hecho (el que el ANI principal estuviese fuera de servicio) es una consecuencia de la fuga del

Cuadro 3.45: Resultados encuesta comercial de llamados a teléfonos de contacto (ANIS Principales)

Descripción de ANIS	Cantidad	Porcentaje [%]
ANIS Vigentes	21	48.837
ANIS Sin tono	14	32.558
ANIS Contactado que señala mala atención	1	2.326
ANIS Con tono pero por sistema se retira	7	16.279
Total	43	100.000
ANIS en Vigencia	21	48.837
ANIS en Retiro	22	51.163

cliente.

A este resultado se añade su valorización respectiva teórica en caso de que este piloto se hubiese instalado como proyecto dentro de la empresa, esta valorización se efectúa de forma marginal y no en base al ciclo de vida. Dicha valorización se presenta a continuación:

Cuadro 3.46: Valorización monetaria de todos los experimentos

Experimento	Pérdida por período	
	Mensual ($/mes)	Anual ($/año)
3	1,408,173	16,898,076
5	30,687,421	368,249,052
7 (con supuestos)	48,587,700	583,052,400
7 (sin supuestos)	96,089,900	1,153,078,800

Cabe destacar que esta valorización está sujeta a un supuesto en general, el cual es que las grandes empresas (clientes pertenecientes al grupo Grandes) en caso de estar fugadas solamente han renunciado al 50 % de sus contratos, porcentaje que se justifica al realizar un estudio sobre la cantidad de contratos promedio que tienen los clientes en el grupo Grandes. A esta valorización se puede agregar el posible impacto real, en base a supuestos, esto último tiene por objeto analizar el beneficio monetario que significa la implementación de un proyecto de minería de datos, en donde se entablaron dos supuestos, el primero es la mantención del porcentaje de efectividad del modelo, lo cual se justifica con el resultado de la encuesta comercial, el segundo supone dos casos, el peor con una tasa de un 15 % de retención ejecutando acciones correctivas y el mejor con una tasa de 50 % de retención sobre la base de posibles fugados. Evaluación monetaria que se muestra en la tabla 3.47.

Cuadro 3.47: Valorización monetaria de aplicación teórica de acciones correctivas

	Fuga de ingresos prevenida ($/mes)		Fuga de ingresos prevenida ($/año)	
	Peor caso	Mejor caso	Peor caso	Mejor caso
Ingreso	7,288,155	24,293,850	87,457,860	291,526,200
Peor caso	15 % retenido por acciones correctivas de la predicción			
Mejor caso	50 % retenido por acciones correctivas de la predicción			

Capítulo 4

Conclusiones

Al momento de partir la experimentación de la memoria, se cuenta con una tesis anterior, que contempla un estudio de la aplicación del KDD, dejando de lado la evaluación monetaria del impacto que puede dejar el mismo en la empresa. Sin embargo, el tener un documento inicial resulta de bastante ayuda debido a que permite un inicio acelerado en un procedimiento KDD sobre el producto NGN, siendo una de las etapas que a nivel nacional resulta en un gran gasto de tiempo, puesto que las empresas nacionales por miedo a que se generen fuga de datos plantean una política de restricción frente a personas ajenas.

En el entorno experto del KDD se verifica el hecho de que la etapa de preprocesamiento es la más extensa. Además, en este documento, se contextualiza dicha creencia a nivel nacional, pues si bien el preprocesamiento genera un gran impacto sobre los resultados de los modelos (debido a que el KDD es un proceso cíclico pero secuencial, como se puede apreciar en 5.17), la realidad nacional constata que la integración domina al preprocesamiento en tiempo y relevancia, porque si la integración de datos no es adecuada, independiente de las técnicas de preprocesamiento que se usen, el resultado ya está sujeto a error. Por lo tanto, el primer requerimiento que debe contemplarse bajo esta perspectiva al intentar aplicar KDD a una empresa o mercado, es el origen y mantenimiento de las fuentes de información.

La conclusión respecto a cuál es el mejor segundo requerimiento mínimo es el bosquejo de la situación actual iniciando por cómo se calcula la variable que se desea predecir. En esta memoria, se descubre que el churn en la compañía tiene dos características relevantes: primero no es el churn de clientes y segundo, no se sabe qué cliente se van, es decir, el cálculo es global. Este esquema sobre cómo se obtiene la variable objetivo permite establecer diferencias y ventajas del piloto sobre la situación actual, además, de aumentar el conocimiento sobre las bases. En concreto, se concluye que el bosquejo del cálculo de fuga en el producto o servicio determinado permite generar ventajas sobre la situación actual y disminuye el error en la etapa de integración de bases. Si no se bosqueja no se detectan anomalías posteriores tanto en la implementación del piloto como en la operación del mismo. Por ejemplo, en la tesis base no existe este problema dado que la variable estado fue incorporada a la base entregada a la persona analista, posteriormente se nota que esta variable se encuentra desfasada de la realidad y la fuente de dicha variable no se puede averiguar concretamente debido a que es traspasada directamente al analista presente en ese instante, por ende, no se puede analizar el camino que toma dicha variable en tiempos posteriores dado que los atributos del

producto cambian en el período respectivo (2 años).

En un tono más específico relacionado con la selección de los registros, el establecimiento de las instancias vigentes se dificulta si no existe una estandarización, ni tampoco una o más variables que declaren con certeza dicha característica. Esto converge en añadir instancias inadecuadas o ya retiradas, lo que incrementa el número de valores fuera de rango que, en grandes cantidades o en un problema de rareza de clases, conlleva a un resultado erróneo. Agregado al tema de identificación de registros vigentes, se tiene la etiquetación de las instancias fugadas, las cuales tampoco tienen una estandarización pero si una base externa señalada como indicatriz. Por ende, puede darse el caso de tener 67 ruts fugados en esta base externa, mas al ejecutar el contraste con la base final preprocesada se obtengan 58 ruts están etiquetados como tales. Por lo anterior, se concluye trabajar con las bases de datos que el personal de la compañía utiliza, pues si bien, puede tener más ruido que las bases de sistemas, muestran el presente del producto o servicio, por lo tanto, es la última *"foto"* del mercado respectivo. Además, representa el trabajo actual y real del producto, por lo que si se descubre un error en estas bases significa que se debe trabajar sobre la reparación de las mismas antes de poder aplicar el KDD.

Por otro lado, la selección de variables es un tema a tratar directamente con el personal de la empresa mediante entrevistas informales, donde se debe vislumbrar dos argumentos para decidir la incorporación o eliminación nuevas variables y registros, estos son: La periodicidad de las fuentes de información que contienen dichas variables y la información que puede extraerse a partir de ellas. Al aplicar KDD en una empresa, la etapa de integración cobra relevancia, la cual viene determinada por las variables usadas y su utilidad, por lo tanto, se concluye que la selección de variables se debe efectuar a través del análisis de posibles transformaciones o imputaciones que puedan insertarse de tal forma que quede un resultado de fácil interpretación. Además, puede darse el caso de que el personal de la empresa señale una o más variables como importantes, no obstante, previo a insertarlas a la base integrada, se debe visualizar su histograma o bien, estudiar alternativas de transformación para su mejor interpretación.

Dado que la interpretación es una etapa de alta relevancia en este tipo de proyectos, originar una descripción de las instancias que permita establecer causales de fuga, así como también, propuestas de abordamiento para generar una acción de retención, resulta vital. Por lo que se concluye que una buena herramienta principalmente para analizar los datos de un conjunto determinado es la segmentación. Esto se debe a que bosqueja una exploración rápida del mercado y familiarización con el producto o servicio. Además, en caso de que la base de datos presente una estructura ideal para clusterizar se puede agregar un modelamiento específico para cada uno de los grupos detectados y caracterizados. No obstante, en caso de que la base de datos carezca de una estructura de clusterización, no se recomienda crear un modelamiento específico, mas sí se sugiere una segmentación que ayude a conocer las instancias. En un problema de rareza de clases, como el detectado en esta memoria, el uso de la segmentación ayuda tanto a su descubrimiento como a su tratamiento. En síntesis, se la segmentación es una herramienta que debe usarse principalmente a modo de exploración de una base de datos para ahorrar el tiempo que toma conocer el producto o servicio y el mercado en el que están insertos. Sin embargo, el tiempo dedicado a la realización de una segmentación válida técnicamente (clusterización) resulta extenso, puesto que las métricas de calidad de la clusterización están sujetas al número de registros de la base trabajada.

La transición del experimento 2 y 3 sugiere una problemática relacionada con el período de

fuga considerado, debido a que se plantea un entrenamiento de 6 meses, por lo que la evaluación de la predicción entregada sólo se obtiene 6 meses después de la aplicación correspondiente. En esta memoria, siempre se busca un predictor que permita el descubrimiento del comportamiento de los cliente mes a mes, lo que probablemente es un enfoque incorrecto a menos que se trate de una mejora operacional como un predictor de fallas. Esto deja la siguiente conclusión: la observación de los procesos a los cuales están sujetos las mejoras dentro de una empresa deben ser contemplados y analizados tanto por la parte técnica del proyecto como por la parte estratégica, puesto que en caso de que no se haga los resultados son precisos pero no prácticos, lo que converge en un gasto mayor de horas-hombre para establecer la periodicidad dedicada a la elaboración de las acciones correctivas. Sin embargo, puede relajarse esta condición en el caso de que el objetivo de la aplicación del KDD sea aumentar la confianza y conocimiento sobre esta tecnología.

En el experimento 4, las principales conclusiones radican en la marginalidad de las bases, lo que contempla una consideración de las bases operacionales. Este ámbito del KDD dentro de los textos examinados, no es tratado de ningún modo, ni siquiera presentado como tal. Por lo que para llevar a cabo un proyecto de minería de datos se debe contar con dos tipos de fuentes de datos como mínimo y estas son un Data Warehouse y las bases Marginales, tal y como puede observarse en la figura 4.1, en donde, generalmente para tener un *data warehouse* se asume la existencia de bases operacionales, lo que es un error debido a que no siempre es así.

Figura 4.1: Pirámide de información

También este experimento permite concluir que al momento de establecer marginalidad a partir de un *data warehouse* incremental se añade un comportamiento erróneo y distinto del común, lo que en concreto dificulta establecer un modelo no influenciable por la temporalidad de los datos.

En cierto sentido, un *data warehouse* puede definirse como una fuente de datos consolidada, no obstante, no siempre guarda el efecto teórico en la realidad, donde al referirse a esta fuente como única válida se pierde la posibilidad de análisis dentro de él en busca de errores, en otras palabras, si un Data Warehouse se implementa bajo una cultura que prescinde de otras bases para darle prioridad, no tendrá la misma confiabilidad que presenta en la teoría.

El experimento 5, plantea la disyuntiva de resultados entre la evaluación comercial y la evaluación técnica, puesto que para que un modelo sea robusto y confiable, ambas evaluaciones deben ser consistentes. Frente a esto se concluye que si existe una discrepancia en ambos valores, dicha diferencia indica que el KDD ha sido ejecutado consecuentemente (que es distinto de correctamente) con las variables del sistema. Esto quiere decir que las transformaciones están fundamentadas y lo

que posiblemente está generando dicha disyuntiva se asumen como dos causas: El algoritmo que se usa no es el adecuado y la base de datos tiende a tener gran parte de las instancias en un estado previo a la deserción definitiva. La segunda causa detectada es que la variable fuga escogida no contemple el total de la información, con lo cual, en la evaluación comercial se detecta está fuga escondida, la que no está presente en los datos. La solución que se propone es analizar el algoritmo primero, pues es más directo, mientras que el análisis de la variable objetivo conlleva a plantear el proceso KDD completo, desde la integración de bases hasta la evaluación a efectuar. Otra causa probable de este efecto puede deberse al hecho de que en ningún momento del piloto se genera un estudio acerca del criterio de corte, la razón de ello es que no se desea caer en el error de sobreajuste ni de establecer una tendencia errónea, sin embargo, al momento de entregar una predicción la utilización del mismo es necesaria debido a que en el experimento 5 es probable que un estudio del criterio de corte hubiese entregado una mejor prueba y validación para la evaluación técnica y comercial o por lo menos puede señalar una guía para no eliminar modelos con un bajo número de posibles fugados en la predicción entregable.

En lo que respecta a las transformaciones, se concluye que, dentro del refinamiento de ellas, se deben contemplar dos ámbitos: la interpretación y la fundamentación estadística que contiene. Un ejemplo de no contemplar el primer aspecto es realizar un análisis ACP sin mantener variables que permitan una interpretación sobre las variables originales, lo que en este documento se muestra en el experimento 6. En cambio, si no se añade el segundo ámbito, tal y como sucede con la función TRIM, lo que ocurre es la aparición de inconsistencias en los análisis de datos sin un motivo aparente, además, puede degenerar en entregar nuevas variables que se correlacionan a un alto nivel con otras variables perteneciente a la base de datos preprocesada, lo que sesgará el algoritmo aplicado. Otro ejemplo que se adhiere al primer ámbito es mantener una base conceptual sobre cada transformación aplicada, debido a que en el experimento 3 se denota un error (ocurrido en el experimento 2) de implementación inadecuada del modelo RFM a la base de datos preprocesada, constatando que cada transformación debe realizarse bajo un objetivo de obtener mayor información a partir de una o más variables originales.

La evaluación técnica debe ejecutarse de manera rigurosa y conciente de cada paso efectuado, de tal manera que se tenga una explicación probable a cada resultado inesperado, por ejemplo, en el experimento 5 se captura un 1 % de las instancias fugadas reales, sin embargo, dado que se sigue un proceso riguroso y metódico, se puede deducir que el flujo del comportamiento de los clientes para el producto NGN puede tener efectos de estacionalidad, o bien, puede ser un caso en que el comportamiento sea completamente aleatorio. Otro ejemplo, es el caso de la disyuntiva entre la evaluación comercial y la evaluación técnica, que se traduce en una problemática entre las validaciones técnica y comercial que posee distintas causas entre las que se destacan: variable fuga errónea, validación comercial sesgada, validación técnica errónea y modelo inestable.

Otra conclusión es que estudiar la cadena de valor del producto puede entregar perspectivas útiles a la ejecución del procedimiento KDD. Un ejemplo de ello se encuentra en el experimento 4, donde se descubre que la fuga de los clientes puede ocurrir en cualquier momento del mes lo que resta el efecto de estacionalidad de la variable objetivo, es decir, los clientes no son retirados en un día predeterminado o no se fugan el mismo día. De esta manera, el procedimiento KDD cambia, de su manera usual de entrenamiento; aprendizaje con variables de un mes y predección en el mismo; un nuevo formato de aprendizaje con variables de un mes y predicción para la fuga en

el mes siguiente. La razón que justifica este proceder es el hecho de que el cliente tiene la libertad de declarar su fuga voluntaria en cualquier instante, por lo tanto, el primero de cada mes ya puede existir uno o varios fugado, lo que conlleva a adquirir un sesgo inicial. Cabe señalar que esto es útil en los problemas en que una o más bases pertenecientes a la integración no están disponibles a partir del primero de cada mes, si no que se entreguen a mediados del mismo.

Un punto importante cuando se utiliza la segmentación como parte del modelamiento de la aplicación de la minería de datos es la aparición de traslado de instancias entre segmentos. Usar esta técnica en esta etapa resulta en un alto riesgo debido a que el analista busca resolver dos problemas de clasificación, el primero es averiguar en qué grupo se encuentra al cliente y el segundo es saber si se fuga o no. Para lo cual se propone verificar su validez a través del análisis de la matriz de traslados inter-grupal y añadir dichos pesos (o probabilidades de transición las que se pueden obtener vía probabilidades estacionarias de Markov) al modelo, o bien, validar que los grupos se mantienen en el tiempo con sus mismas características. Estas propuestas permiten almacenar la divergencia de la fuga inter-grupal, aplacan la varianza del conjunto de datos.

Cabe agregar que en la etapa de integración se concluye que se debe investigar aparte del personal relacionado con el producto, puesto que pueden existir otras variables que describan procesos inherentes en el producto o prácticas no documentadas en el mismo. Un ejemplo, es la inexistencia de la variable de renegociación en el producto NGN, la cual no se encuentra documentada, ni tampoco su proceso dentro de la cadena de valor del producto.

4.1. Propuestas

Las propuestas que nacen en base a las conclusiones previamente mencionadas son:

- Estudiar dos perspectivas, por una parte la técnica desde la observación del parámetro de corte que en esta memoria no pudo ser abordado en un tiempo adecuado, lo que conlleva a que al momento de entregar una predicción a la empresa para efectuar la encuesta comercial se cambiasen los modelos correspondientes y se busca aquel modelo que entrega un buen resultado en la primera validación y a su vez, un número considerable (entre 50 y 200) posibles fugados para la validación final. La segunda perspectiva es la de inspección de la variable FUGA, puesto que ésta, según conversaciones informales, solamente captura el 40 % de la fuga real, además, dentro de esta fuga expresada en la variable *Estado contrato* como la categoría de "Término de contrato", no se tiene el seguimiento de la misma, por lo que la indicación de esta categoría no necesariamente acaba en fuga. Acorde al estudio comercial del experimento 5 existen clientes que terminan un contrato para renegociar uno nuevo, lo cual, no se puede tomar como fuga.

- Utilizar nuevas variables, por ejemplo, la variable DESCUENTO de la Proforma, que representa si en el mes i-ésimo se le dio un descuento a un cliente, o bien, generar nuevas transformaciones que ajusten de manera adecuada las variables de la base Ordenes Terminadas. Además, se pueden recuperar variables eliminadas como aquellas asociadas a la importancia de los otros productos para el cliente.

- Usar las bases de Suscriptores y Morosidad como bases temporales, solicitándolas mensualmente y efectuando un seguimiento a las mismas, además, de a la base NGN Instalados.
- Realizar la evaluación técnica de las predicciones entregadas en un período de 6 meses, para observar la efectividad de los modelos usados.
- Ajustar el modelamiento en Rapidminer al software adquirido por la compañía SAS.
- Refinar el procedimiento de esta memoria en base a la actual discontinuación de la base Boletas Técnicas consecuencia de la fusión interna a la compañía.
- Actualizar la evaluación monetaria de la fuga en base al ciclo de vida del cliente y estudiar este efecto a través de análisis de supervivencia y evaluar a los cliente en base a ambas predicciones (de fuga y de tiempo de vida) a través de la matriz de riesgo sugerida en la tabla 5.135 basada en [101].
- Crear una base que sea combinación de aquellas usadas en este documento para facilitar la labor de integración de datos.
- Automatizar la predicción de cliente del producto NGN para un usuario final.

Ahora bien, se deben plantear sugerencias de acciones correctivas, estas son:

- Estudiar el efecto del cambio de precio en la retención, para ello se debe aplicar el procedimiento KDD correspondiente al experimento 7 y establecer una metodología de evaluación de resultados equivalente a la expresada por la competencia expresada en el gráfico 1.6. Posteriormente evaluar monetariamente esta acción.
- Establecer una conexión entre el departamento operacional encargado de la reparación de fallas, en caso de que esto sea muy difícil, se propone generar un mecanismo de control sobre los ejecutivos de cada cliente. De esta manera, usando el modelo RFM sobre una base similar a Boletas Técnicas se puede tener un indicador sobre qué clientes demandan mayor atención, en particular aquellos que están siendo predichos como fugados.
- Rediseñar algunos planes tomando como referencia la clusterización de los mismos efectuada en el experimento 3. Para lo cual, se pueden entablar encuestas con los usuarios u observar que necesidades no se están cubriendo tanto por el producto NGN como su mejora Trunk IP.
- Evaluar la posibilidad de obtener la variable de renegociación para no caer en fugas erróneas.

4.2. Trabajo Futuro

Las propuestas previamente mencionadas permiten la mejora del proceso implementado en este trabajo particular referente a la predicción de fuga de un producto en particular. Como se ha mencionad previamente que el cliente se fugue del producto no es una acción equivalente a que el cliente se retire voluntariamente de la compañía, dado que en una empresa de telecomunicaciones, el cliente posee múltiples productos. Por lo anterior, se presenta un bosquejo de trabajo futuro pensando en

la visión general del churn, es decir, el churn de la compañía dado que este tema es lo que falta por hacer, estudio que únicamente se puede efectuar contemplando las plataformas principales de utilización. En el esquema 1.10 se visualizan las principales plataformas y bases de datos que se manejan internamente en la empresa. Con ese esquema el trabajo a futuro propuesto es el siguiente:

- **Detección de fuga voluntaria de clientes:** Para poder establecer una fuga voluntaria de un cliente de la empresa, se debe verificar dos puntos relevantes, el primero es que no tenga contrato alguno de ningún producto, lo cual se puede verificar mediante dos vías: verificando las plataformas de ventas, es decir, USV y Visión MAD, o bien, tomar la base única de clientes (BUN, que es una base de datos transversal a la compañía que contiene todos los clientes de la red fija) y la base de clientes de la red móvil. El segundo punto relevante es comprobar que todos sus servicios (comerciales y técnicos) se encuentren en estado de retiro en la base de datos de servicios (expresada en el diagrama 1.10 por el nombre de "VUS"). Finalmente para distinguir entre fuga voluntaria de involuntaria se debe constatar si el cliente fue desafiliado de la empresa por el área de cobranzas, para ello, en la plataforma "SAP RMCA" se encuentra la historia de bloqueos y morosidad de los clientes en su totalidad.

- **Selección de fuentes y variables relevantes:** En este trabajo se utilizaron bases de datos locales que en su mayoría estaban enfocadas netamente en el producto. Para estudiar la situación de un cliente particular en cada producto se debe recopilar dicha información, la cual se encuentra principalmente en las bases de datos que respaldan las plataformas USV, Visión MAD, Móvil, solamente referente a los contratos. Un punto muy relevante es que la factura se efectúa por contrato no por servicio particular. Lo que conlleva a estudiar al cliente por el contrato y no por los servicios que posee. Una guía para iniciar el trabajo futuro es hablar con el DWH para que se genere la extracción de estas bases de datos. Complementando adicionalmente la información anterior, se puede reutilizar la base de datos de **Boletas Técnicas** debido a que contiene los reclamos de los centros de asistencia, por lo que no discrimina por producto. Agregado a lo anterior, la reutilización de la BD **Ordenes Terminadas** también puede ser de gran utilidad dado que son bases de datos transversales a la compañía aunque esta última se aplica a la red fija solamente. Posterior a la recopilación de instalaciones y reclamos, la información de los descuentos y el consumo para cada factura se puede verificar en los prefacturadores principalmente (para el tráfico sobre fuera de su zona estos descuentos son de díficil extracción dado que son bases de tipo marginal), existen 2 en su mayoría, uno para tráfico, o paquetes para empresas (por ejemplo NGN) y un segundo que solamente ve los servicios privados (es decir, internet, TI, etc.). Finalmente con la información extraída de la plataforma "SAP RMCA" se puede verificar la facturación real y los estados de cuenta para cada cliente. Cabe destacar que, si bien esta plataforma (SAP RMCA) permite muchas acciones, esta posee una gran cantidad de registros que con las herramientas actuales no se pueden procesar de manera individual, lo que hace necesario un compromiso con el área de sistemas para generar una extracción válida y suficiente de la información requerida

- **Ejecución del resto de las etapas del KDD:** Gran parte de las transformaciones implementadas en este trabajo son aplicables al resto de las bases de datos, así como también el preprocesamiento. Ahora bien, en el modelamiento se requiere de modelos de rápida ejecución, por

ello, la herramienta Rapidminer no puede ser usada en este trabajo por ahora (se está trabajando en una versión de mayor capacidad), además, la empresa cuenta con licencias del software SAS, mas está no ha sido renovada en años y no comprende el módulo de inteligencia de negocios. Esto implica que las etapas de preprocesamiento y transformación son manejables incluso sobre el volumen de datos a contemplar, no obstante, dificulta la tarea del modelamiento del problema general, el cual se debería llevar a cabo mediante la programación del mismo en JAVA usando la librería de Rapidminer como ayuda.

- **Ejecución de análisis de fuga de ingresos en base al ciclo de vida:** Para establecer el costo real que significa para la compañía la fuga voluntaria de clientes se propone realizar un estudio sobre la predicción del ciclo de vida para poder determinar la fuga de ingresos real al momento de que un cliente pide término de contrato total. Esto podría utilizarse, además, como valor agregado para la elaboración de campañas comerciales para cada uno de los cliente al momento de realizar la evaluación comercial final y ejecutar las acciones correctivas. Análogo a esto puede ser usado un algoritmo de segmentación para los mismo propósitos.

Aparte del trabajo futuro propuesto, resulta interesante analizar la posibilidad de implementar las siguientes ideas:

- **Análisis de la red social de los dispositivos móviles:** El año 2012 se implementará la portabilidad numérica y los registros de todos los teléfonos móviles quedarán a libre uso por las compañías de telecomunicaciones, lo que permitirá efectuar un mapa que permita establecer a los usuarios que mantienen una mayor red y que influyen más en su entorno, de esta manera, se podrán implementar campañas más enfocadas a estos usuarios para que recomienden los servicios de la empresa. Además, permitirá la detección efectiva de llamada que generen fraude tanto para la compañía como para los usuarios.

- **Análisis de procesos internos en la compañía:** Actualmente se crean muchos reporte referentes a las actividades de cada área en el DWH, además, no se puede bosquejar la importancia de la consistencia de datos puesto que no se puede evaluar cuantitativamente los valores asociados a dichas inconsistencias. La segunda idea interesante a realizar es una red similar a la anteriormente explicada, pero inter-áreas, que permita evaluarlas en base a los resultados, a lo que se agrega la valorización de tareas críticas e identificar los punto críticos reales de pérdida de ingreso en el proceso. Un punto relevante para lograr lo anterior es el hecho de la compañía actualmente cuenta con un sistema de costeo por actividad, es decir, para cada momento del ciclo de vida del producto existe un costo asociado.

5

Anexos

5.1. Anexo 1: Bases de datos y variables utilizadas

5.1.1. Bases de datos

- **Oracle Workflow (OWF)**: Esta base contiene los registros con fecha de las solicitudes comerciales de los clientes ya sea de término de contrato como también de solicitud de servicios, aunque éste último tema no se tiene en su totalidad debido a que en esta base sólo se registran las postventas que se realizan por el call center de la compañía y algunas presenciales, en otras palabras, es aquella que comprende las transacciones que el cliente efectúa vía telefónica con la empresa, es decir, la postventas, el cambio del servicio, término de contrato entre otros. Proviene de la plataforma denominada WEBI y su periodicidad es mensual.

- **Proforma**: Es la base generada por el área del *data warehouse* destinada a temas de facturación y consumo. Proviene directamente del sistema Kenan. Su periodicidad es mensual, no obstante, se debe tener cuidado dado que esta base se almacena cada tres meses, es decir, si se desea averiguar la facturación del año 2009, se debe consultar directamente al *data warehouse* comercial, sin embargo, dicha base está sin procesar y el rut no posee el dígito verificador, por ende, se debe mantener un seguimiento a esta base. Esta base se borra cada tres meses, es decir, se tiene las bases de marzo-abril-mayo y al llegar el 15 de Junio la base de marzo se borra y quedan las bases de abril-mayo-junio, pero no es que se borren completamente, esta aclaración se hace para decir que una vez pasado tres meses, la obtención de estos datos resulta más complicada.

- **Boletas Técnicas**: Base destinada a ser reportes acerca de los reclamos desde una perspectiva técnica. En otras palabras, describe todo el comportamiento de mantenimiento que el cliente recibe, en otras palabras, en ella se encuentran todas las reparaciones a fallas en que el cliente ha avisado a la compañía. Por ende, su origen es el área de operaciones. Proviene directamente de MySap. Su periodicidad es mensual. Esta base fue de fácil acceso y se contiene los datos actuales más los históricos de 2 años atrás.

- **Segmentación de empresas (seg_empresas)**: Esta base contiene la totalidad de clientes y sus características descriptivas, es decir, su tamaño, clasificación, categoría, plan de retención,

entre otros. Su origen es el área del *data warehouse*. Proviene directamente de MySap. su periodicidad es mensual. Esta base, por razones no descubiertas tarda más del tiempo normal en cargarse en el software Microsoft Excel 2007.

- **NGN Instalados**: Esta base contiene solamente los detalles de los clientes del producto NGN, sin embargo, se agrega una base secundaria referente a los planes que poseen dichos clientes así como su estado de vigencia del plan correspondiente (un contrato son varios planes). Su periodicidad es mensual en la teoría, no obstante en la práctica es trimestral.
- **Suscriptores**: Es una base que consta en sus registros de todos los teléfonos fijos existentes visibles en el mercado de las telecomunicaciones, por lo tanto, es una base de gran tamaño, cercano a los 7,5 millones de registros. Debido a esto, se toma como una base cuya variación es despreciable, es decir, estática y se obtiene por petición al área del *data warehouse*.
- **Órdenes terminadas**: Ésta contiene todas las acciones de instalación, activación, retiro y suspensión, entre otros, que realiza el área de operaciones. Por ende, consta de una cantidad superior a 50 atributos. Esta base es de generación diaria (ordenes terminadas en el día), más un consolidado del mes (ordenes terminadas del mes). Por tanto cada base sólo contiene las órdenes de trabajo terminadas en el lapso de tiempo correspondiente. Además, es multiproducto y son ordenes de trabajo entre las que se pueden encontrar instalaciones y postventas. Agregado a esto, su origen se encuentra en la gerencia de redes y luego se preprocesa en la gerencia de operaciones, siendo esta última la que se utiliza.
- **Ruts Friendly User(FU), corporaciones y mayoristas**: Es una base que contiene ruts pertenecientes a grandes holdings, corporaciones y mayoristas, su origen es informal dentro de la compañía, es decir, es una planilla que se transfiere entre las distintas áreas de la empresa. Aparte de las gigantescas empresas que alberga, también contiene ruts Friendly User, que equivalen a identificadores falsos que se usan para probar el comportamiento del sistema frente a ingreso de nuevos productos o promociones.
- **Giros Comerciales S.I.I._2008**: Esta base es interna en la compañía y se confecciona a partir de la clasificación que utiliza el servicio de impuestos internos desde el 2008.
- **TIPIFICACION_CPO**: Es una base individual y estática, sin un área responsable y de carácter público. En ella se describe la relación entre los centros primarios y las provincias, así como también, aquella que existe entre las provincias y las comunas de cada región del país, por lo tanto, permite establecer la interacción entre los centros primarios y las comunas. Un punto relevante es que un centro primario puede tener más de una comuna asociada, mientras que una comuna solo tiene un centro primario asociado. La explicación de los centros primarios se expone en la tabla 5.1. Esta nueva base consta de tres campos, el nombre de la comuna, el número asociado y la región a la cual pertenece dicha comuna.

5.1.2. Descripción de variables

Las variables son descritas mediante la siguiente tabla:

Cuadro 5.1: Descripción de todas las variables

Variable Final	Tipo de variable	Base	Observación
TEC RUT	Nominal	NGN Instalados paquetes	Es el rut del cliente. También denominada RUTCOD.
TS1	Nominal	Boletas técnicas	Es el prefijo que ayuda a identificar los productos en la empresa
TS2	Nominal	Boletas técnicas	Es el sufijo que ayuda a distinguir los productos más específicos en la empresa. En conjunto con TS1 permiten la identificación de todos los productos en la empresa.
COD PROD	Nominal	Base externa individual	Es la variable que contiene los TS1 y TS2 unidos del producto NGN.
GIRO	Cadena	NGN Instalados rut	Es el giro asociado tanto a la facturación como al declarado por el cliente.
GIRO Trans	Nominal	NGN Instalados rut	Es una categorización de la variable GIRO donde las cadenas de texto son reemplazadas por números para el entendimiento del software usado.
PRIMERA INSTALACION	Fecha	NGN Instalados rut	Es la fecha en la que el cliente entró a la empresa.
INGCPI	Binominal	NGN Instalados rut	También llamada INGRESO CONT POST INST (ingreso contrato posterior a la instalación), representa si el cliente tuvo una instalación del servicio con anterioridad al ingreso del contrato (en cuyo caso tiene asignada la categoría SI) o no (cuya categoría es NO).
INGRESO CONTRATO SAP	Fecha	NGN Instalados rut	Es la variable asociada a la fecha en que se ingresa una contrato del producto NGN al sistema MySap.
Continúa en la página siguiente			

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Dias de antig¨edad reales	Continua	NGN Instalados rut	Es la cantidad de días transcurridos entre una fecha actual (primer día del mes que se busca predecir) y la fecha expresada en la variable INGRESO CONTRATO SAP.
DELTA INTERPOLADA	Continua	NGN Instalados rut	Es la cantidad de días transcurridos entre una fecha actual (primer día del mes que se busca predecir) y la fecha expresada en la variable PRIMERA INSTALACION.
ComunaFinal	Cadena de texto	NGN Instalados rut	Es la comuna en la que reside el cliente.
CÓDIGO DE ÁREA	Nominal	TIPIFICACION CPO	Es el código de área asociado a la comuna que resume a las comunas por números asociados a las provincias.
SUCURSALES	Entero	NGN Instalados rut	Se refiere al número de sucursales que posee el cliente.
Sucursales nominalizadas (SUC NOM)	Nominal	NGN Instalados rut	Es una discretización de la variable SUCURSALES acorde a intervalos descritos en este documento.
CANT CPO	Entero	NGN Instalados rut	Es la cantidad de centros primarios; es decir, los números que previamente se deben ingresar al teléfono para que la compañía pueda identificar la provincia del teléfono receptor; que posee el cliente.
Q SUCURSALES CON BA	Entero	NGN Instalados rut	Es la cantidad de sucursales con banda ancha que tiene el cliente.
Q SUCURSALES SIN BA	Entero	NGN Instalados rut	Es la cantidad de sucursales sin banda ancha que tiene el cliente.
Q ACCESOS	Entero	NGN Instalados rut	Cantidad de Accesos con Banda Ancha contratados.
Continúa en la página siguiente			

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Q ANIS	Entero	NGN Instalados rut	Es la cantidad de teléfonos o números telefónicos del clientes.
Q FACTURACION OTROS PROD	Entero	Segmentación empresas	Se refiere a la cantidad facturada en otros productos que no sean NGN, que el cliente posee.
Q OTROS PRODUCTOS	Entero	Segmentación empresas	Es la cantidad de otros productos aparte del NGN que el cliente posee.
Q SUCURSALES CENTRO	Entero	NGN Instalados rut	Es la cantidad de sucursales ubicadas en la Zona Centro del país.
Q SUCURSALES NORTE	Entero	NGN Instalados rut	Es la cantidad de sucursales ubicadas en la Zona Norte del país.
Q SUCURSALES SUR	Entero	NGN Instalados rut	Es la cantidad de sucursales ubicadas en la Zona Sur del país.
PRIMERA FACTURA	Fecha	NGN Instalados rut	Indica la fecha de la primera factura en caso de que se hubiese llevado a cabo en los últimos 3 meses previos a la consulta.
FECHA DE ÚLTIMA FACTURA	Fecha	NGN Instalados rut	Indica la fecha de la última factura en caso de que se hubiese llevado a cabo en los últimos 3 meses previos a la consulta.
CANTIDAD DE FACTURAS	Entero	NGN Instalados rut	Indica la cantidad de facturas realizadas en caso de que se hubiese llevado a cabo en los últimos 3 meses previos a la consulta (por lo que sus valores son 0, 1, 2, 3.
Promedio fact	Continua	Proformas Pyme	Es el promedio de facturación del cliente en los últimos 6 meses.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Trim1 fact	Continua	Proformas Pyme	Es el promedio ponderado de facturación en el primer trimestre del período de estudio.
Trim2 fact	Continua	Proformas Pyme	Es el promedio ponderado de facturación en el segundo trimestre del período de estudio.
ID CURVA FACTURACION	Nominal	Proformas Pyme	Transformación ejecutada en la tesis base que discretiza las facturaciones de 8 meses en una sola variable.
FAC2 Facturacion	Continua	Proformas Pyme	Es el factor obtenido efectuando un análisis de componentes principales sobre las variables de facturación de cada mes, durante el período de 6 meses.
MONTO FACT MES i	Entero	Proforma Pyme Mes i	Es el monto facturado, donde i hace referencia al mes, es decir, son 6 variables equivalentes cuya única diferencia es el mes en el que son extraídas, por lo que i puede tomar los valores 1, 2, 3, 4, 5, 6. Cabe agregar que a estas variables se les llama también, variables de facturación y vienen expresadas en pesos.
Promedio consumo	Continua	Proformas Pyme	Es el promedio de consumo del cliente en los últimos 6 meses.
Trim1 consumo	Continua	Proformas Pyme	Es el promedio ponderado de consumo en el primer trimestre del período de estudio.
Trim2 consumo	Continua	Proformas Pyme	Es el promedio ponderado de consumo en el segundo trimestre del período de estudio.
ID CURVA CONSUMO	Nominal	Proformas Pyme	Transformación ejecutada en la tesis base que discretiza los consumo de 8 meses en una sola variable.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
FAC1 Consumo	Continua	Proformas Pyme	Es el factor obtenido efectuando un análisis de componentes principales sobre las variables de consumo de cada mes, durante el período de 6 meses.
MONTO CONSUMO MES i	Continua	Proforma Pyme Mes i	Es el monto de consumo expresado en minutos mediante la división: (SEGUNDOS MES i) /60, es decir, la variable original viene en SEGUNDOS. El índice i hace referencia al mes, es decir, son 6 variables equivalentes cuya única diferencia es el mes en el que son extraídas, por lo que i puede tomar los valores 1, 2, 3, 4, 5, 6. Cabe agregar que a estas variables se les llama también, variables de consumo.
RESP FALLA	Binominal	Boletas técnicas	Indica si el cliente o la compañía es responsable de la falla, sus categorías son COMPANY o Cliente
Recencyreal	Nominal	Boletas técnicas	Señala hace cuantos meses que el cliente no reclama, tomando como referencia el mes actual, sus categorías son 0,1,2,3,4,5,6,7, donde 7 es una categoría como otros (aquellos que no reclaman hace mucho tiempo).

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Freq rec	Continua	Boletas técnicas	Señala la frecuencia porcentual de ocasiones en las que el cliente reclamó en base a una falla del producto [Experimento 3]. Así como también, se refiere a la cantidad promedio de reclamos que hace un cliente dentro del período de estudio [Experimento 2].
Imagen	Binominal	Boletas técnicas	Señala si la mayoría de las fallas eran de responsabilidad del cliente (1) o de la empresa (-1)
Mount rec	Continua	Boletas técnicas	Es la cantidad promedio de reclamos técnicos que el cliente realiza en un período de 6 meses [Experimento 3]. A su vez, también puede ser usada como el total de reclamos efectuados por el cliente en el período de 6 meses [Experimento 2].
COMPANY i	Entero	Boletas técnicas	Es la cantidad de reclamos técnicos que son catalogados como responsabilidad de la empresa, donde i hace referencia al mes, es decir, son 6 variables equivalentes cuya única diferencia es el mes en el que son extraídas, por lo que i puede tomar los valores 1, 2, 3, 4, 5, 6. Un nombre alternativo para esta variable es Q RECL TEC RESP EMPRESA.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Cliente i	Entero	Boletas técnicas	Es la cantidad de reclamos técnicos que son catalogados como responsabilidad del cliente, donde i hace referencia al mes, es decir, son 6 variables equivalentes cuya única diferencia es el mes en el que son extraídas, por lo que i puede tomar los valores 1, 2, 3, 4, 5, 6. Un nombre alternativo para esta variable es Q RECL TEC RESP CLIENTE.
Totales i	Entero	Boletas técnicas	Es la cantidad de reclamos técnicos totales del cliente, donde i hace referencia al mes, es decir, son 6 variables equivalentes cuya única diferencia es el mes en el que son extraídas, por lo que i puede tomar los valores 1, 2, 3, 4, 5, 6.
Q Rec com cliente	Entero	Oracle Workflow mes i	Es la cantidad de reclamos comerciales con responsabilidad del cliente.
Q Rec com emp	Entero	Oracle Workflow mes i	Es la cantidad de reclamos comerciales con responsabilidad de la empresa.
fact ngn	Continua	Segmentación empresas	Es la facturación del cliente en el producto NGN.
fact bdl	Continua	Segmentación empresas	Es la facturación del cliente en el producto Bundling.
fact dtc	Continua	Segmentación empresas	Es la facturación del cliente en el producto Datacom.
fact ded	Continua	Segmentación empresas	Es la facturación del cliente en el producto Dedicado.
fact ba	Continua	Segmentación empresas	Es la facturación del cliente en el producto Banda Ancha.
fact tl	Continua	Segmentación empresas	Es la facturación del cliente en el producto Telefonía.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
fact dc	Continua	Segmentación empresas	Es la facturación del cliente en el producto Datacenter.
fact pu	Continua	Segmentación empresas	Es la facturación del cliente en el producto Público.
fact ic	Continua	Segmentación empresas	Es la facturación del cliente en el producto Conmutado.
fact mc	Continua	Segmentación empresas	Es la facturación del cliente en el producto Merconet.
fact va	Continua	Segmentación empresas	Es la facturación del cliente en el producto Va y portal.
fact total	Continua	Segmentación empresas	Es la facturación total del cliente.
Importancia NGN	Continua	Segmentación empresas	Es la relevancia que tiene el producto NGN por sobre los otros productos que el cliente posea.
ProductoPrincipal	Nominal	Segmentación empresas	Es el producto más improtante del cliente que representa más del 50 % de su facturación.
Tamaño numero	Nominal	Segmentación empresas	Es el tamaño de la empresa, donde 1 es microempresa, 2 es pequeña, 3 es mediana y 4 es Grande. Se destaca que esta variable es equivalente a la variable Nom Tam.
Ciclo vida	Nominal	Segmentación empresas	Se refiere al ciclo de vida que la empresa cree que el cliente tiene, 0 es no Clasificado, 1 eezs Nuevo, 2 es Crecimiento y 3 es Madurez. Se destaca que esta variable es equivalente a la variable Nom Clas.
Valor R Num	Nominal	Segmentación empresas	Es el valor del cliente asignado por la empresa, 1 es bajo, 2 es medio bajo, 3 es medio, 4 es medio alto y 5 es alto.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Reten R	Nominal	Segmentación empresas	Es el programa de retención que tiene el cliente asociado 1 es retención mínima, 2 es media y 3 es full.
Cat corp RN	Nominal	Segmentación empresas	Es la categoría de corporaciones que tiene el cliente, 1 es normal, 2 es superior, 3 es premium y 4 es especial.
Rec fac	Entero	Oracle Workflow	Es la cantidad de reclamos comerciales del cliente que tiene el cliente. También se denomina Q REC COM FACT en algunos experimentos.
ICP reload	Continua	Segmentación empresas	Es el índice de comportamiento de pago del cliente (variable continua interna a la compañía), también es asociada como variable ICP.
Riesgo Alto	Binominal	Segmentación empresas	Es la discretización del índice de comportamiento de pago del cliente, cuyas categorías son 1 para riesgo alto y 0 para riesgo bajo. Esta variable es equivalente a Tipo ICP descrita en la tesis base.
BA	Binominal	NGN Instalados paquetes	Indica si el plan tiene banda ancha incluída.
Anis	Entero	NGN Instalados paquetes	Son los teléfonos o accesos del plan.
Tec WIIMAX COBRE	Cadena de texto	NGN Instalados paquetes	Proporciona la información acerca de la tecnología, que puede ser mediante cobre o usando el dispositivo Wiimax que permite virtualidad en el servicio (es decir, no necesita un medio físico para conectar los accesos);.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
ADSL	Binominal	NGN Instalados paquetes	Indica si el plan contiene ADSL 2+ que es una característica que entrega mayor velocidad al plan.
ESTADO FINAL	Cadena de texto	NGN Instalados paquetes	Indica el estado del plan el cual puede ser: ACTIVADA, Error, Facturar, FRIENDLY USER, Inhabilitar, Inhabilitado, INSTALADA NO ACTIVADA y RENUNCIADA.
USV CANAL	Cadena de texto	NGN Instalados paquetes	Indica el canal mediante el cual se vendió el plan, sus categorías pueden ser: CACE (Centros de la empresa), DEALER (Vendedor individual), Ejecutivo COMPANY, EVI, No definido, NO IDENTIFICADO.
Nom CAN	Nominal	NGN Instalados paquetes	Es la nominalización (cambio de cadenas de texto por número) de la variable USV CANAL.
SAP DESCRIPCION ITEM	Cadena de texto	NGN Instalados paquetes	Es una variable de texto que describe el plan en términos generales donde se indican características como las de las variables ADSL, BA, Anis y Velocidad. No existe una estandarización pertinente, consiste de 61 categorías distintas. También se denomina DESCRIPCION PLANES o variable planes.
VELOCIDAD	Nominal	NGN Instalados paquetes	Es una variable extraída a partir de la variable SAP DESCRIPCION ITEM que indica la velocidad asociada al plan respectivo.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Planes tipo 1	Entero	NGN Instalados paquetes	Es la cantidad de planes tipo 1 que posee el cliente, tanto para el experimento 2 como para los posteriores.
Planes tipo 2	Entero	NGN Instalados paquetes	Es la cantidad de planes tipo 2 que posee el cliente, tanto para el experimento 2 como para los posteriores.
Planes tipo 3	Entero	NGN Instalados paquetes	Es la cantidad de planes tipo 3 que posee el cliente, tanto para el experimento 2 como para los posteriores.
Planes tipo 4	Entero	NGN Instalados paquetes	Es la cantidad de planes tipo 4 que posee el cliente, tanto para el experimento 2 como para los posteriores.
Planes tipo 5	Entero	NGN Instalados paquetes	Es la cantidad de planes tipo 5 que posee el cliente, desde el experimento 3 en adelante.
Nom plan real	Nominal	NGN Instalados paquetes	Es la categorización de los planes que se efectúa en el experimento 2, en base a los tipos de planes ofrecidos en el producto NGN.
N Planes Renunciados	Entero	NGN Instalados paquetes	Cantidad de planes renunciados que el cliente tiene en su período de vida.
Competencia	Nominal	Suscriptores	Es la competencia asociada al cliente con valores 1, 3, 11, 12, 19, 27.
Competencia 2	Binominal	Suscriptores	Esta variable indica si el cliente está afecto a la competencia o no. Sus categorías son 1 para señalar que el cliente está afecto a la competencia y 0 en caso contrario.
COMPANY MOBILE	Binominal	Suscriptores	Indica si el cliente pertenece a COMPANY MOBILE (1) o no (0)

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
COMPANY PHONE	Binominal	Suscriptores	Indica si el cliente pertenece a COMPANY PHONE (1) o no (0)
COMPANY CELL	Binominal	Suscriptores	Indica si el cliente pertenece a COMPANY CELL (1) o no (0)
Cartera	Binominal	Base externa CARTERA	Indica si el cliente es del tipo cartera(1) o no (0)
MOROSIDAD	Nominal	Base externa MOROSIDAD (Consulta SAP-R3)	Indica si el cliente es moroso (1 o 2) o no lo es (0). También es denominada Moroso VALORES.
Reclamos comerciales generales	Continua	Oracle Workflow mes i	Promedio de reclamos comerciales generales del cliente dentro de los 6 meses
Elimina sucursal Acumulado	Continua	Oracle Workflow mes i	Es un tipo de solicitud comercial que se refiere a la eliminación de un plan del contrato. En particular, esta variable señala la eliminación de sucursales de forma acumulada
Retención NGN Acumulado	Continua	Oracle Workflow mes i	Tipo de solicitud comercial referente a la cantidad de veces que se le ha asignado retención al cliente, dentro de los seis meses.
Total Solicitudes comerciales	Continua	Oracle Workflow mes i	Es el total de solicitudes comerciales que el cliente ha realizado en el período de los 6 meses.
Solicitudes técnicas	Continua	Oracle Workflow mes i	Es el total de solicitudes técnicas que el cliente ha realizado en el período de los 6 meses.
Edad	Continua	NGN Instalados rut	Es la cantidad de días que tiene el cliente en la empresa desde que se realizó su primera instalación hasta el último día del mes (en este caso 31 de Diciembre)

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Corporaciones	Binominal	Ruts Friendly User(FU), corporaciones y mayoristas	Indica 1 si el rut pertenece a corporaciones, mayoristas o friendly users y 0 si no.
Clasificacion por comportamiento en ordenes terminadas	Nominal	Ordenes Terminadas	Es una variable que categoriza a los clientes en base a los valores de la variable Tipo trabajo que presentan las ordenes terminadas asociadas a ellos.
Trabajo	Cadena de texto	Ordenes Terminadas	Es la variable que describe todos los movimientos de cada transacción de la base de ordenes terminadas, en términos generales.
Tipo trabajo	Cadena de texto	Ordenes Terminadas	Es una categorización de la variable Trabajo que destaca los movimientos de cada transacción, que señala el comportamiento de los clientes según si agrega cosas, retira, cambia, instala, suspende, habilita, etc.
DESCRIPCION	Cadena de texto	Ordenes Terminadas	Es una variable que describe los detalles principales de una orden terminada, en particular, es una cadena de texto con 242 categorías.
COTIZACION	Binominal	Ordenes Terminadas	Indica si la glosa de la variable DESCRIPCION presenta la palabra cotización o no.
Dentro del plazo	Binominal	Ordenes Terminadas	Definida a partir de las fechas de anulación de la provisión y la fecha pactada cliente, representa si la orden se encuentra en el plazo pactado con el cliente o no.
Urgencia	Binominal	Ordenes Terminadas	Expresa si la orden del cliente requiere de ser atendida en forma urgente.

Continúa en la página siguiente

Cuadro 5.1 – Continuación de la página anterior

Variable Final	Tipo de variable	Base	Observación
Estado contrato sap	Cadena de texto	Base de My sap (Req292)	Se refiere al estado en el que se encuentra el contrato en la empresa, el cual puede ser: vigentes, anulados, renunciados y nuevos.
K3New	Nominal	-	Es la clusterización primaria siendo 1, los leales, el 2 son lo rehenes y el 3 es el grupo mixto
K33New	Nominal	-	Es la clusterización secundaria siendo 1, los refugiados, el 2 son lo indiferentes y el 3 es el grupo de los switchers
Cluster real	Nominal	-	Esta variable expresa un tipo de agrupación obtenido por el método *Two Step cluster*, con las variables mencionadas. La descripción de dichos clúster se encuentra en la viñeta "Grupos" y en las viñetas "Centroides 1", "Centroides 2z "Frecuencias", de este archivo
Fuga	Binominal	Oracle Workflow mes i	Esta variable expresa la fuga del mes posterior a la base preprocesada.

5.1.3. Variables por experimento

Cuadro 5.2: Variables que se usan por experimentos

Variables	Experimentos						
	1	2	3	4	5	6	7
TEC RUT	X	X	X	X	X	X	X
TS1		X	X	X	X	X	X
TS2		X	X	X	X	X	X
COD PROD		X	X	X	X	X	X
Continúa en la página siguiente							

Cuadro 5.2 – Continuación de la página anterior

Variables	Experimentos						
GIRO	X			X	X	X	X
GIRO Trans			X				
PRIMERA INSTALACION	X	X	X	X	X	X	X
INGCPI	X						
INGRESO CONTRATO SAP	X	X					
Dias de antigedad reales		X	X				
DELTA INTERPOLADA						X	X
ComunaFinal		X	X	X	X	X	X
CÓDIGO DE ÁREA				X	X	X	X
SUCURSALES	X	X	X	X	X	X	X
Sucursales nominalizadas (SUC NOM)	X						
CANT CPO	X	X	X	X	X	X	X
Q SUCURSALES CON BA	X	X	X	X	X		
Q SUCURSALES SIN BA	X	X	X	X	X		
Q ACCESOS	X			X	X		
Q ANIS	X			X	X		
Q FACTURACION OTROS PROD	X	X					
Q OTROS PRODUCTOS	X	X					
Q SUCURSALES CENTRO	X	X					
Q SUCURSALES NORTE	X	X					
Q SUCURSALES SUR	X	X					
PRIMERA FACTURA		X					
FECHA DE ÚLTIMA FACTURA		X	X				
CANTIDAD DE FACTURAS		X					
Promedio fact	X	X	X	X	X		
Trim1 fact		X	X	X	X		
Trim2 fact		X	X	X	X		
ID CURVA FACTURACION	X						
FAC2 Facturacion						X	X
MONTO FACT MES i	X	X	X	X	X	X	X
Promedio consumo	X	X	X	X	X		
Trim1 consumo		X	X	X	X		
Trim2 consumo		X	X	X	X		
ID CURVA CONSUMO	X						
FAC1 Consumo						X	X
MONTO CONSUMO MES i	X	X	X	X	X	X	X
RESP FALLA	X	X	X	X	X	X	X
Recencyreal		X	X	X	X	X	X
Freq rec		X	X	X	X	X	X

Continúa en la página siguiente

Cuadro 5.2 – Continuación de la página anterior

Variables	Experimentos						
Imagen		X	X	X	X	X	X
Mount rec		X	X	X	X	X	X
COMPANY i	X	X	X	X	X	X	X
Cliente i	X	X	X	X	X	X	X
Totales i	X	X					
Q Rec com cliente	X	X					
Q Rec com emp	X	X					
fact ngn	X	X	X				
fact bdl	X	X					
fact dtc	X	X					
fact ded	X	X					
fact ba	X	X					
fact tl	X	X					
fact dc	X	X					
fact pu	X	X					
fact ic	X	X					
fact mc	X	X					
fact va	X	X					
fact total	X	X					
Importancia NGN		X	X				
ProductoPrincipal		X	X				
Tamaño numero		X	X	X	X	X	X
Ciclo vida		X	X	X	X	X	X
Valor R Num			X	X	X	X	X
Reten R			X	X	X	X	X
Cat corp RN			X	X	X	X	X
Rec fac	X	X	X	X	X	X	X
ICP reload		X		X	X		
Riesgo Alto	X					X	X
BA		X	X	X	X		
Anis		X	X	X	X		
Tec WIIMAX COBRE		X	X	X	X		
ADSL		X	X	X	X		
ESTADO FINAL		X	X	X	X		
USV CANAL		X					
Nom CAN	X	X					
SAP DESCRIPCION ITEM		X	X	X	X		
VELOCIDAD		X	X	X	X		
Planes tipo 1		X	X	X	X		

Continúa en la página siguiente

Cuadro 5.2 – Continuación de la página anterior

Variables	Experimentos						
Planes tipo 2		X	X	X	X		
Planes tipo 3		X	X	X	X		
Planes tipo 4		X	X	X	X		
Planes tipo 5			X	X	X		
Nom plan real	X						
N Planes Renunciados	X	X					
Competencia		X	X	X	X	X	X
Competencia 2						X	X
COMPANY MOBILE		X	X	X	X	X	X
COMPANY PHONE		X	X	X	X	X	X
COMPANY CELL		X	X	X	X	X	X
Cartera				X	X		
MOROSIDAD				X	X		
Reclamos comerciales generales						X	X
Elimina sucursal Acumulado						X	X
Retención NGN Acumulado						X	X
Total Solicitudes comerciales						X	X
Solicitudes técnicas						X	X
Edad						X	X
Corporaciones					X		X
Clasificacion por comportamiento en ordenes terminadas		X					
Trabajo							
Tipo trabajo		X					
DESCRIPCION							
COTIZACION		x					
Dentro del plazo							
Urgencia							
Estado contrato sap	X						
K3New				X	X		
K33New				X	X		
Cluster real							X
Fuga	X	X	X	X	X	X	X

5.2. Anexo 2: Experimento 1

5.2.1. Correlaciones entre variables

Cuadro 5.3: Tabla de correlaciones experimento 1

Correlaciones de Pearson	Variables													
Variables	Nom Clas	Nom Tam	ID CURVA FACTURA-CION	ID CURVA CONSUMO	INGCPI	Nom plan real	Nom CAN	Tipo ICP	Sucursales no-minalizadas	PRIMERA INSTALA-CION	INGRESO CONTRATO SAP	PROMEDIO FACTURA-CION	PROMEDIO CONSUMO	SUCURSALES
Nom Clas	1.000	0.195	-0.007	-0.014	0.005	0.017	0.040	-0.002	0.349	0.000	-0.001	0.174	0.296	0.312
Nom Tam	0.195	1.000	0.063	0.080	-0.068	0.030	-0.080	-0.057	0.155	-0.116	-0.135	0.031	0.157	0.173
ID CURVA FACTURA-CION	-0.007	0.063	1.000	0.692	0.013	-0.087	-0.281	-0.022	0.114	0.138	0.173	0.109	0.272	0.111
ID CURVA CONSUMO	-0.014	0.080	0.692	1.000	0.007	-0.083	-0.280	-0.025	0.101	0.092	0.122	0.070	0.294	0.093
INGCPI	0.005	-0.068	0.013	0.007	1.000	-0.012	0.001	0.025	0.037	-0.049	0.207	0.017	0.023	0.030
Nom plan real	0.017	0.030	-0.087	-0.083	-0.012	1.000	0.039	0.011	0.124	0.034	0.018	0.019	0.058	0.166
Nom CAN	0.040	-0.080	-0.281	-0.280	0.001	0.039	1.000	0.068	-0.033	-0.132	-0.113	-0.024	-0.108	-0.025
Tipo ICP	-0.002	-0.057	-0.022	-0.025	0.025	0.011	0.068	1.000	0.012	0.007	0.017	0.008	0.023	0.015
Sucursales no-minalizadas	0.349	0.155	0.114	0.101	0.037	0.124	-0.033	0.012	1.000	-0.046	-0.037	0.280	0.575	0.846
PRIMERA INSTALA-CION	0.000	-0.116	0.138	0.092	-0.049	0.034	-0.132	0.007	-0.046	1.000	0.953	0.013	0.028	-0.054
INGRESO CONTRATO SAP	-0.001	-0.135	0.173	0.122	0.207	0.018	-0.113	0.017	-0.037	0.953	1.000	0.016	0.039	-0.044
PROMEDIO FACTURA-CION	0.174	0.031	0.109	0.070	0.017	0.019	-0.024	0.008	0.280	0.013	0.016	1.000	0.333	0.310
PROMEDIO CONSUMO	0.296	0.157	0.272	0.294	0.023	0.058	-0.108	0.023	0.575	0.028	0.039	0.333	1.000	0.619
SUCURSALES	0.312	0.173	0.111	0.093	0.030	0.166	-0.025	0.015	0.846	-0.054	-0.044	0.310	0.619	1.000
CANT CPO	0.266	0.202	0.105	0.096	0.021	0.149	-0.059	-0.003	0.662	-0.053	-0.042	0.118	0.403	0.708
Q SUCURSA-LES CON BA	0.202	0.111	0.084	0.069	0.025	0.119	-0.050	0.013	0.618	-0.056	-0.051	0.259	0.401	0.807
Q SUCURSA-LES SIN BA	0.278	0.149	0.101	0.087	0.021	0.130	-0.019	0.009	0.665	-0.018	-0.009	0.208	0.560	0.694
Q ACCESOS BA	0.206	0.113	0.076	0.061	0.025	0.120	-0.032	0.013	0.618	-0.056	-0.051	0.259	0.398	0.807
Q ANIS	0.333	0.149	0.128	0.103	0.047	0.106	-0.025	0.017	0.734	-0.007	0.000	0.559	0.667	0.804
Q SUCURSA-LES CENTRO	0.285	0.188	0.092	0.106	0.049	0.164	0.000	0.022	0.703	-0.039	-0.032	0.377	0.576	0.789
Q SUCURSA-LES NORTE	0.255	0.011	0.113	-0.002	0.028	0.178	0.014	0.003	0.560	0.014	0.025	0.458	0.405	0.603
Q SUCURSA-LES SUR	0.471	-0.001	0.167	0.087	0.012	0.128	-0.083	0.069	0.699	0.037	0.034	0.729	0.513	0.771
Q RECL TEC RESP EMPRESA	0.343	0.207	0.233	0.235	0.008	0.091	-0.094	-0.003	0.583	0.067	0.067	0.268	0.650	0.639
Q RECL TEC RESP CLIEN-TE	0.317	0.184	0.164	0.161	0.012	0.069	-0.058	-0.020	0.444	0.075	0.075	0.214	0.484	0.481
Q Rec com fact	0.094	0.018	0.052	0.045	0.036	0.037	0.007	0.005	0.157	-0.043	-0.043	0.188	0.168	0.145
Q Rec com emp	0.175	0.080	0.073	0.092	0.030	0.092	-0.037	0.013	0.360	-0.011	-0.009	0.196	0.369	0.373
Q Rec com cliente	0.091	0.077	0.017	0.037	0.027	0.068	0.012	0.016	0.230	-0.071	-0.071	0.077	0.162	0.217
Dias de antige-dad reales	-0.001	0.119	-0.138	-0.092	0.043	-0.036	0.132	-0.008	0.045	-0.998	-0.957	-0.013	-0.027	0.053
ICP Preproce-sado	-0.008	-0.018	0.015	0.000	0.027	-0.020	0.010	-0.071	-0.010	-0.002	0.000	-0.003	-0.011	-0.009
Q otros pro-ductos	0.254	0.306	0.320	0.287	0.031	0.081	-0.187	-0.025	0.269	0.148	0.157	0.124	0.356	0.267
Q fact otros productos	0.398	0.205	0.093	0.081	0.018	0.032	-0.020	-0.007	0.273	0.043	0.047	0.315	0.368	0.284
Fuga nomina-lizada	0.012	-0.068	-0.538	-0.530	-0.063	0.153	0.502	0.039	-0.113	-0.201	-0.200	-0.074	-0.250	-0.103

Cuadro 5.4: Tabla de correlaciones experimento 1 parte 2

Correlaciones de Pearson	Variables													
Variables	CANT CPO	Q SUCURSALES CON BA	Q SUCURSALES SIN BA	Q ACCESOS BA	Q ANIS	Q SUCURSALES CENTRO	Q SUCURSALES NORTE	Q SUCURSALES SUR	Q RECL TEC RESP EMPRESA	Q RECL TEC RESP CLIENTE	Q Rec com fact	Q Rec com emp	Q Rec com cliente	Dias de antigedad reales
Nom Clas	0.266	0.202	0.278	0.206	0.333	0.285	0.255	0.471	0.343	0.317	0.094	0.175	0.091	-0.001
Nom Tam	0.202	0.111	0.149	0.113	0.149	0.188	0.011	-0.001	0.207	0.184	0.018	0.080	0.077	0.119
ID CURVA FACTURACION	0.105	0.084	0.101	0.076	0.128	0.092	0.113	0.167	0.233	0.164	0.052	0.073	0.017	-0.138
ID CURVA CONSUMO	0.096	0.069	0.087	0.061	0.103	0.106	-0.002	0.087	0.235	0.161	0.045	0.092	0.037	-0.092
INGCPI	0.021	0.025	0.021	0.025	0.047	0.049	0.028	0.012	0.008	0.012	0.036	0.030	0.027	0.043
Nom plan real	0.149	0.119	0.130	0.120	0.106	0.164	0.178	0.128	0.091	0.069	0.037	0.092	0.068	-0.036
Nom CAN	-0.059	-0.050	-0.019	-0.032	-0.025	0.000	0.014	-0.083	-0.094	-0.058	0.007	-0.037	0.012	0.132
Tipo ICP	-0.003	0.013	0.009	0.013	0.017	0.022	0.003	0.069	-0.003	-0.020	0.005	0.013	0.016	-0.008
Sucursales nominalizadas	0.662	0.618	0.665	0.618	0.734	0.703	0.560	0.699	0.583	0.444	0.157	0.360	0.230	0.045
PRIMERA INSTALACION	-0.053	-0.056	-0.018	-0.056	-0.007	-0.039	0.014	0.037	0.067	0.075	-0.043	-0.011	-0.071	-0.998
INGRESO CONTRATO SAP	-0.042	-0.051	-0.009	-0.051	0.000	-0.032	0.025	0.034	0.067	0.075	-0.043	-0.009	-0.071	-0.957
PROMEDIO FACTURACION	0.118	0.259	0.208	0.259	0.559	0.377	0.458	0.729	0.268	0.214	0.188	0.196	0.077	-0.013
PROMEDIO CONSUMO	0.403	0.401	0.560	0.398	0.667	0.576	0.405	0.513	0.650	0.484	0.168	0.369	0.162	-0.027
Sucursales	0.708	0.807	0.694	0.807	0.804	0.789	0.603	0.771	0.639	0.481	0.145	0.373	0.217	0.053
CANT CPO	1.000	0.586	0.482	0.583	0.475	0.467	0.350	0.426	0.469	0.393	0.061	0.234	0.167	0.050
Q SUCURSALES CON BA	0.586	1.000	0.142	1.000	0.571	0.644	0.445	0.600	0.511	0.399	0.083	0.272	0.163	0.055
Q SUCURSALES SIN BA	0.482	0.142	1.000	0.138	0.650	0.514	0.469	0.552	0.452	0.325	0.131	0.291	0.155	0.019
Q ACCESOS BA	0.583	1.000	0.138	1.000	0.571	0.649	0.445	0.599	0.512	0.399	0.089	0.272	0.165	0.055
Q ANIS	0.475	0.571	0.650	0.571	1.000	0.731	0.631	0.780	0.588	0.433	0.217	0.404	0.208	0.009
Q SUCURSALES CENTRO	0.467	0.644	0.514	0.649	0.731	1.000	0.435	0.471	0.643	0.550	0.200	0.474	0.227	0.037
Q SUCURSALES NORTE	0.350	0.445	0.469	0.445	0.631	0.435	1.000	0.714	0.413	0.332	0.281	0.269	0.097	-0.014
Q SUCURSALES SUR	0.426	0.600	0.552	0.599	0.780	0.471	0.714	1.000	0.584	0.402	0.129	0.357	0.210	-0.033
Q RECL TEC RESP EMPRESA	0.469	0.511	0.452	0.512	0.588	0.643	0.413	0.584	1.000	0.764	0.134	0.314	0.161	-0.066
Q RECL TEC RESP CLIENTE	0.393	0.399	0.325	0.399	0.433	0.550	0.332	0.402	0.764	1.000	0.087	0.254	0.106	-0.074
Q Rec com fact	0.061	0.083	0.131	0.089	0.217	0.200	0.281	0.129	0.134	0.087	1.000	0.276	0.194	0.043
Q Rec com emp	0.234	0.272	0.291	0.272	0.404	0.474	0.269	0.357	0.314	0.254	0.276	1.000	0.299	0.011
Q Rec com cliente	0.167	0.163	0.155	0.165	0.208	0.227	0.097	0.210	0.161	0.106	0.194	0.299	1.000	0.071
Dias de antigedad reales	0.050	0.055	0.019	0.055	0.009	0.037	-0.014	-0.033	-0.066	-0.074	0.043	0.011	0.071	1.000
ICP Preprocesado	-0.010	-0.008	-0.008	-0.007	-0.013	-0.008	-0.027	-0.024	-0.009	-0.004	-0.015	-0.009	-0.009	0.001
Q otros productos	0.238	0.158	0.257	0.156	0.299	0.233	0.202	0.340	0.460	0.411	0.074	0.193	0.072	-0.148
Q fact otros productos	0.247	0.194	0.238	0.193	0.376	0.301	0.066	0.233	0.490	0.580	0.104	0.170	0.076	-0.043
Fuga nominalizada	-0.114	-0.095	-0.086	-0.083	-0.103	-0.106	-0.095	-0.128	-0.232	-0.167	-0.057	-0.101	-0.022	0.203

Cuadro 5.5: Tabla de correlaciones experimento 1 parte 3

Correlaciones de Pearson	Variables			
Variables	ICP Preprocesado	Q otros productos	Q fact otros productos	Fuga nominalizada
Nom Clas	-0.008	0.254	0.398	0.012
Nom Tam	-0.018	0.306	0.205	-0.068
ID CURVA FACTURACION	0.015	0.320	0.093	-0.538
ID CURVA CONSUMO	0.000	0.287	0.081	-0.530
INGCPI	0.027	0.031	0.018	-0.063
Nom plan real	-0.020	0.081	0.032	0.153
Nom CAN	0.010	-0.187	-0.020	0.502
Tipo ICP	-0.071	-0.025	-0.007	0.039
Sucursales nominalizadas	-0.010	0.269	0.273	-0.113
PRIMERA INSTALACION	-0.002	0.148	0.043	-0.201
INGRESO CONTRATO SAP	0.000	0.157	0.047	-0.200
PROMEDIO FACTURACION	-0.003	0.124	0.315	-0.074
PROMEDIO CONSUMO	-0.011	0.356	0.368	-0.250
Sucursales	-0.009	0.267	0.284	-0.103
CANT CPO	-0.010	0.238	0.247	-0.114
Q SUCURSALES CON BA	-0.008	0.158	0.194	-0.095
Q SUCURSALES SIN BA	-0.008	0.257	0.238	-0.086
Q ACCESOS BA	-0.007	0.156	0.193	-0.083
Q ANIS	-0.013	0.299	0.376	-0.103
Q SUCURSALES CENTRO	-0.008	0.233	0.301	-0.106
Q SUCURSALES NORTE	-0.027	0.202	0.066	-0.095
Q SUCURSALES SUR	-0.024	0.340	0.233	-0.128
Q RECL TEC RESP EMPRESA	-0.009	0.460	0.490	-0.232
Q RECL TEC RESP CLIENTE	-0.004	0.411	0.580	-0.167
Q Rec com fact	-0.015	0.074	0.104	-0.057
Q Rec com emp	-0.009	0.193	0.170	-0.101
Q Rec com cliente	-0.009	0.072	0.076	-0.022
Dias de antigedad reales	0.001	-0.148	-0.043	0.203
ICP Preprocesado	1.000	-0.022	-0.008	0.002
Q otros productos	-0.022	1.000	0.461	-0.332
Q fact otros productos	-0.008	0.461	1.000	-0.077
Fuga nominalizada	0.002	-0.332	-0.077	1.000

5.2.2. Tabla de frecuencias y valores perdidos

Cuadro 5.6: Tabla de Frecuencias de la variable Nom CAN

Variable	Nom_CAN	
Categorías	Frecuencia	Porcentaje
1	4	0,04
2	1122	12,57
4	5729	64,20
5	2	0,02
6	2067	23,16
Total	8924	100,00

Cuadro 5.7: Tabla de frecuencias de la variable Nom plan

Variable	Nom_plan	
Categorías	Frecuencia	Porcentaje
1	905	10,14
2	433	4,85
3	3217	36,05
4	1	0,01
5	120	1,34
6	2525	28,29
8	254	2,85
10	457	5,12
Total válido	7912	88,66
Perdidos por el Sistema	1012	11,34
Total de observaciones	8924	100,00

Cuadro 5.8: Tabla de frecuencias de la variable TIPO ICP

Variable	Tipo_ICP		
Categoría Original	Nominalización	Frecuencia	Porcentaje
Riesgo Bajo	0	5174	57,98
Riesgo Alto	1	2327	26,08
Total Válidos		7501	84,05
Perdidos por el Sistema		1423	15,95
Total		8924	100,00

Cuadro 5.9: Tabla de valores perdidos de la variable INGRESO CONTRATO SAP

Variables	Valores completos	Valores perdidos
INGRESO_CONTRATO_SAP_COMERCIAL	8053	421
INGRESO_CONTRATO_SAP_SISTEMA	8605	319
INGRESO_CONTRATO_SAP_COMBINACIóN	8658	151

Cuadro 5.10: Tabla de Frecuencias de la variable INGRESO CONT POST INST

Variable	INGRESO_CONT_POST_INST
Categorías	Frecuencia
SI	200
NO	8314
Total frecuencias	8514
Total Muestra	8924
Missings	410

Cuadro 5.11: Tabla de frecuencias de la variable SUCURSALES nominalizada

Sucursales categorizadas	Frecuencia	Porcentaje	Porcentaje acumulado
1	5883	65,9	65,9
2	2587	29,0	94,9
3	316	3,5	98,5
4	89	1,0	99,5
5	38	0,4	99,9
6	11	0,1	100
Total	8924	100	

Cuadro 5.12: Tabla de frecuencias de la variable FUGA nominalizada

Variable	Fuga_nominalizada	Porcentaje válido
Categoria	Frecuencia	
-1	5794	64,93
1	3130	35,07
Total	8924	100

5.3. Anexo 3: Experimento 2

5.3.1. Estrategias para tratamiento de valores perdidos

Cuadro 5.13: Tabla de valores perdidos y estrategias: Experimento 2

Variables	Cantidad de valores perdidos	Estrategia de reemplazo
fact ngn	100	Interpolación
fact bdl	100	Interpolación
fact dtc	100	Interpolación
fact ded	100	Interpolación
fact ba	100	Interpolación
fact tl	100	Interpolación
fact dc	100	Interpolación
fact pu	100	Interpolación
fact ic	100	Interpolación
fact mc	100	Interpolación
fact va	100	Interpolación
fact total	100	Interpolación
tamano	100	Tabla de contingencia con ciclo vida y cat corp, Moda
Ciclo Vida	100	Hot deck
Valor	100	CHAID
Retención	100	CHAID
cat corp	100	Tabla de contingencia con ciclo vida, Moda y luego reemplazo con CHAID
GIRO	2262	Reemplazo por datos de otra base de datos (plataforma Kenan)
TELEFONO CONTACTO	93	Eliminada por histograma
CALLE FACT	18	Eliminada por histograma
NRO CASA	75	Eliminada por histograma
REFERENCIA	4424	Eliminada por histograma
ORIENTACION	5034	Eliminada por histograma
NRO PISO	4523	Eliminada por histograma
COMUNA	18	Nada
INGRESO CONTRATO SAP	153	Moda
PRIMERA FACTURA	570	Moda
ULTIMA FACTURA	570	Moda
Continúa en la página siguiente		

Cuadro 5.13 – Continuación de la página anterior

Variables	Cantidad de valores perdidos	Estrategia de reemplazo
FACTURACION Octubre09	974	Reemplazo por valor 0
FACTURACION Noviembre09	934	Reemplazo por valor 0
FACTURACION Diciembre09	898	Reemplazo por valor 0
FACTURACION Enero10	829	Reemplazo por valor 0
FACTURACION Febrero10	792	Reemplazo por valor 0
FACTURACION Marzo10	809	Reemplazo por valor 0
FACTURACION Abril10	766	Reemplazo por valor 0
FACTURACION Mayo10	809	Reemplazo por valor 0
FACTURACION Junio10	879	Reemplazo por valor 0
CONSUMO Octubre09	974	Reemplazo por valor 0
CONSUMO Noviembre09	934	Reemplazo por valor 0
CONSUMO Diciembre09	898	Reemplazo por valor 0
CONSUMO Enero10	829	Reemplazo por valor 0
CONSUMO Febrero10	792	Reemplazo por valor 0
CONSUMO Marzo10	809	Reemplazo por valor 0
Consumo Abril	766	Reemplazo por valor 0
Consumo Mayo	809	Reemplazo por valor 0
Consumo Junio	879	Reemplazo por valor 0
Reclamos facturacion	4753	Reemplazo por valor 0
Plan Tipo 1	14	Interpolación
Plan Tipo 2	14	Interpolación
Plan Tipo 3	14	Interpolación
Plan Tipo 4	14	Interpolación
N Planes Renunciados	14	Interpolación
ICP	214	Reemplazo por valor 0
Cliente Octubre	5358	Reemplazo por valor 0
COMPANY Octubre	4769	Reemplazo por valor 0
Cliente Noviembre	5354	Reemplazo por valor 0
COMPANY Noviembre	4853	Reemplazo por valor 0
Cliente Diciembre	5311	Reemplazo por valor 0
COMPANY Diciembre	4812	Reemplazo por valor 0
Cliente Enero	5297	Reemplazo por valor 0
COMPANY Enero	4872	Reemplazo por valor 0
Cliente Febrero	5369	Reemplazo por valor 0
COMPANY Febrero	4951	Reemplazo por valor 0

Continúa en la página siguiente

Cuadro 5.13 – Continuación de la página anterior

Variables	Cantidad de valores perdidos	Estrategia de reemplazo
Cliente Marzo	5284	Reemplazo por valor 0
COMPANY Marzo	4404	Reemplazo por valor 0

Cuadro 5.14: Tabla de presencia variables parte 1: Experimento 2

Variables	fact ngn	fact bdl	fact dtc	fact ded	fact ba	fact tl	fact dc	fact pu	fact ic	fact mc	fact va	fact total	tamano	Ciclo Vida	valor	retencion	cat corp
fact ngn	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact bdl	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact dtc	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact ded	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact ba	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact tl	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact dc	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact pu	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact ic	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact mc	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact va	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
fact total	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
tamaño	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Ciclo Vida	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
valor	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
retencion	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
cat corp	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
GIRO	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %
TELEFONO CONTACTO	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %	99 %
CALLE FACT	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %
NRO CASA	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %
REFERENCIA	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
ORIENTACION	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
NRO PISO	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMUNA	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %	89 %
INGRESO CONTRATO SAP	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %	61 %
PRIMERA FACTURA	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %
ULTIMA FACTURA	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %	87 %
FACTURACION Octubre09	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %
FACTURACION Noviembre09	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %
FACTURACION Diciembre09	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %
FACTURACION Enero10	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %
FACTURACION Febrero10	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %
FACTURACION Marzo10	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %
FACTURACION Abril10	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %
FACTURACION Mayo10	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %
FACTURACION Junio10	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %
CONSUMO Octubre09	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %
CONSUMO Noviembre09	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %	92 %
CONSUMO Diciembre09	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %	91 %
CONSUMO Enero10	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %
CONSUMO Febrero10	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %
CONSUMO Marzo10	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %	90 %
Consumo Abril	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %	95 %
Consumo Mayo	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %
Consumo Junio	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %	97 %
Reclamos facturacion	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
Plan Tipo 1	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %
Plan Tipo 2	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %
Plan Tipo 3	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %
Plan Tipo 4	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %
N Planes Renunciados	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %	57 %
ICP	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %	59 %
Cliente Octubre	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMPANY Octubre	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
Cliente Noviembre	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMPANY Noviembre	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
Cliente Diciembre	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMPANY Diciembre	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
Cliente Enero	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMPANY Enero	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
Cliente Febrero	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMPANY Febrero	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
Cliente Marzo	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %
COMPANY Marzo	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %	98 %

Cuadro 5.15: Tabla de presencia variables parte 2: Experimento 2

	GIRO	TELEFONO CONTACTO	CALLE FACT	NRO CASA	REFERENCIA	ORIENTACION	NRO PISO	COMUNA	INGRESO CONTRATO SAP	PRIMERA FACTURA	ULTIMA FACTURA
fact ngn	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact bdl	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact dtc	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact ded	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact ba	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact tl	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact dc	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact pu	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact ic	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact mc	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact va	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
fact total	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
tamano	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
Ciclo Vida	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
valor	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
retencion	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
cat corp	75 %	99 %	98 %	98 %	24 %	15 %	23 %	98 %	41 %	24 %	24 %
GIRO	0 %	98 %	100 %	97 %	28 %	8 %	18 %	100 %	99 %	90 %	90 %
TELEFONO CONTACTO	56 %	0 %	100 %	99 %	24 %	13 %	23 %	100 %	99 %	89 %	89 %
CALLE FACT	44 %	100 %	0 %	0 %	0 %	0 %	0 %	0 %	56 %	50 %	50 %
NRO CASA	24 %	99 %	76 %	0 %	17 %	1 %	1 %	76 %	89 %	85 %	85 %
REFERENCIA	63 %	98 %	100 %	99 %	0 %	12 %	20 %	100 %	97 %	89 %	89 %
ORIENTACION	59 %	98 %	100 %	99 %	23 %	0 %	12 %	100 %	97 %	90 %	90 %
NRO PISO	59 %	98 %	100 %	98 %	21 %	2 %	0 %	100 %	97 %	90 %	90 %
COMUNA	44 %	100 %	0 %	0 %	0 %	0 %	0 %	0 %	56 %	50 %	50 %
INGRESO CONTRATO SAP	86 %	99 %	95 %	95 %	24 %	14 %	21 %	95 %	0 %	39 %	39 %
PRIMERA FACTURA	62 %	98 %	98 %	98 %	18 %	11 %	24 %	98 %	84 %	0 %	0 %
ULTIMA FACTURA	62 %	98 %	98 %	98 %	18 %	11 %	24 %	98 %	84 %	0 %	0 %
FACTURACION Octubre09	65 %	98 %	99 %	98 %	20 %	12 %	23 %	99 %	85 %	41 %	41 %
FACTURACION Noviembre09	64 %	98 %	99 %	98 %	20 %	12 %	23 %	99 %	84 %	39 %	39 %
FACTURACION Diciembre09	63 %	99 %	99 %	98 %	20 %	12 %	24 %	99 %	84 %	37 %	37 %
FACTURACION Enero10	60 %	98 %	99 %	98 %	19 %	11 %	23 %	99 %	85 %	31 %	31 %
FACTURACION Febrero10	58 %	98 %	99 %	98 %	18 %	11 %	22 %	99 %	88 %	28 %	28 %
FACTURACION Marzo10	57 %	98 %	99 %	98 %	19 %	11 %	22 %	99 %	88 %	30 %	30 %
FACTURACION Abril10	53 %	98 %	99 %	98 %	19 %	11 %	22 %	99 %	95 %	35 %	35 %
FACTURACION Mayo10	53 %	98 %	99 %	98 %	20 %	11 %	23 %	99 %	98 %	42 %	42 %
FACTURACION Junio10	55 %	98 %	99 %	98 %	20 %	11 %	23 %	99 %	98 %	47 %	47 %
CONSUMO Octubre09	65 %	98 %	99 %	98 %	20 %	12 %	23 %	99 %	85 %	41 %	41 %
CONSUMO Noviembre09	64 %	98 %	99 %	98 %	20 %	12 %	23 %	99 %	84 %	39 %	39 %
CONSUMO Diciembre09	63 %	99 %	99 %	98 %	20 %	12 %	24 %	99 %	84 %	37 %	37 %
CONSUMO Enero10	60 %	98 %	99 %	98 %	19 %	11 %	23 %	99 %	85 %	31 %	31 %
CONSUMO Febrero10	58 %	98 %	99 %	98 %	18 %	11 %	22 %	99 %	88 %	28 %	28 %
CONSUMO Marzo10	57 %	98 %	99 %	98 %	19 %	11 %	22 %	99 %	88 %	30 %	30 %
Consumo Abril	53 %	98 %	99 %	98 %	19 %	11 %	22 %	99 %	95 %	35 %	35 %
Consumo Mayo	53 %	98 %	99 %	98 %	20 %	11 %	23 %	99 %	98 %	42 %	42 %
Consumo Junio	55 %	98 %	99 %	98 %	20 %	11 %	23 %	99 %	98 %	47 %	47 %
Reclamos facturacion	60 %	98 %	100 %	99 %	23 %	12 %	22 %	100 %	97 %	88 %	88 %
Plan Tipo 1	43 %	93 %	50 %	50 %	14 %	7 %	7 %	50 %	57 %	50 %	50 %
Plan Tipo 2	43 %	93 %	50 %	50 %	14 %	7 %	7 %	50 %	57 %	50 %	50 %
Plan Tipo 3	43 %	93 %	50 %	50 %	14 %	7 %	7 %	50 %	57 %	50 %	50 %
Plan Tipo 4	43 %	93 %	50 %	50 %	14 %	7 %	7 %	50 %	57 %	50 %	50 %
N Planes Renunciados	43 %	93 %	50 %	50 %	14 %	7 %	7 %	50 %	57 %	50 %	50 %
ICP	66 %	99 %	99 %	98 %	26 %	10 %	25 %	99 %	64 %	42 %	42 %
Cliente Octubre	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	89 %	89 %
COMPANY Octubre	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	88 %	88 %
Cliente Noviembre	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	89 %	89 %
COMPANY Noviembre	60 %	98 %	100 %	99 %	23 %	12 %	22 %	100 %	97 %	88 %	88 %
Cliente Diciembre	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	89 %	89 %
COMPANY Diciembre	60 %	98 %	100 %	99 %	23 %	12 %	22 %	100 %	97 %	88 %	88 %
Cliente Enero	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	89 %	89 %
COMPANY Enero	60 %	98 %	100 %	99 %	23 %	12 %	22 %	100 %	97 %	88 %	88 %
Cliente Febrero	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	89 %	89 %
COMPANY Febrero	60 %	98 %	100 %	99 %	23 %	12 %	22 %	100 %	97 %	89 %	89 %
Cliente Marzo	60 %	98 %	100 %	99 %	23 %	12 %	21 %	100 %	97 %	89 %	89 %
COMPANY Marzo	60 %	99 %	100 %	99 %	22 %	12 %	22 %	100 %	97 %	87 %	87 %

Cuadro 5.16: Tabla de presencia variables parte 3: Experimento 2

	FACTURACION Octubre09	FACTURACION Noviembre09	FACTURACION Diciembre09	FACTURACION Enero10	FACTURACION Febrero10	FACTURACION Marzo10	FACTURACION Abril10
fact ngn	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact bdl	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact dtc	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact ded	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact ba	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact tl	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact dc	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact pu	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact ic	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact mc	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact va	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact total	20 %	21 %	20 %	21 %	22 %	22 %	60 %
tamano	20 %	21 %	20 %	21 %	22 %	22 %	60 %
Ciclo Vida	20 %	21 %	20 %	21 %	22 %	22 %	60 %
valor	20 %	21 %	20 %	21 %	22 %	22 %	60 %
retencion	20 %	21 %	20 %	21 %	22 %	22 %	60 %
cat corp	20 %	21 %	20 %	21 %	22 %	22 %	60 %
GIRO	85 %	85 %	85 %	85 %	85 %	85 %	84 %
TELEFONO CONTACTO	84 %	84 %	86 %	86 %	86 %	86 %	86 %
CALLE FACT	50 %	44 %	44 %	44 %	44 %	44 %	44 %
NRO CASA	80 %	77 %	80 %	80 %	81 %	80 %	80 %
REFERENCIA	82 %	83 %	84 %	85 %	85 %	85 %	86 %
ORIENTACION	83 %	84 %	84 %	85 %	86 %	86 %	86 %
NRO PISO	83 %	84 %	85 %	86 %	86 %	86 %	87 %
COMUNA	50 %	44 %	44 %	44 %	44 %	44 %	44 %
INGRESO CONTRATO SAP	5 %	5 %	5 %	17 %	39 %	39 %	75 %
PRIMERA FACTURA	0 %	0 %	0 %	0 %	0 %	0 %	12 %
ULTIMA FACTURA	0 %	0 %	0 %	0 %	0 %	0 %	12 %
FACTURACION Octubre09	0 %	11 %	18 %	28 %	34 %	34 %	41 %
FACTURACION Noviembre09	7 %	0 %	7 %	19 %	26 %	26 %	33 %
FACTURACION Diciembre09	11 %	4 %	0 %	13 %	20 %	20 %	28 %
FACTURACION Enero10	15 %	9 %	5 %	0 %	9 %	10 %	18 %
FACTURACION Febrero10	19 %	13 %	9 %	5 %	0 %	2 %	11 %
FACTURACION Marzo10	21 %	15 %	11 %	8 %	4 %	0 %	9 %
FACTURACION Abril10	25 %	19 %	15 %	12 %	8 %	4 %	0 %
FACTURACION Mayo10	30 %	25 %	22 %	20 %	18 %	15 %	12 %
FACTURACION Junio10	35 %	30 %	27 %	26 %	24 %	22 %	20 %
CONSUMO Octubre09	0 %	11 %	18 %	28 %	34 %	34 %	41 %
CONSUMO Noviembre09	7 %	0 %	7 %	19 %	26 %	26 %	33 %
CONSUMO Diciembre09	11 %	4 %	0 %	13 %	20 %	20 %	28 %
CONSUMO Enero10	15 %	9 %	5 %	0 %	9 %	10 %	18 %
CONSUMO Febrero10	19 %	13 %	9 %	5 %	0 %	2 %	11 %
CONSUMO Marzo10	21 %	15 %	11 %	8 %	4 %	0 %	9 %
Consumo Abril	25 %	19 %	15 %	12 %	8 %	4 %	0 %
Consumo Mayo	30 %	25 %	22 %	20 %	18 %	15 %	12 %
Consumo Junio	35 %	30 %	27 %	26 %	24 %	22 %	20 %
Reclamos facturacion	81 %	82 %	82 %	83 %	84 %	84 %	85 %
Plan Tipo 1	29 %	29 %	43 %	50 %	50 %	50 %	21 %
Plan Tipo 2	29 %	29 %	43 %	50 %	50 %	50 %	21 %
Plan Tipo 3	29 %	29 %	43 %	50 %	50 %	50 %	21 %
Plan Tipo 4	29 %	29 %	43 %	50 %	50 %	50 %	21 %
N Planes Renunciados	29 %	29 %	43 %	50 %	50 %	50 %	21 %
ICP	36 %	37 %	39 %	39 %	40 %	40 %	63 %
Cliente Octubre	82 %	83 %	83 %	85 %	85 %	85 %	86 %
COMPANY Octubre	80 %	80 %	81 %	83 %	84 %	83 %	84 %
Cliente Noviembre	82 %	83 %	83 %	85 %	85 %	85 %	86 %
COMPANY Noviembre	81 %	81 %	82 %	83 %	84 %	84 %	85 %
Cliente Diciembre	82 %	83 %	83 %	84 %	85 %	85 %	86 %
COMPANY Diciembre	81 %	81 %	82 %	83 %	84 %	83 %	84 %
Cliente Enero	83 %	83 %	84 %	85 %	85 %	85 %	86 %
COMPANY Enero	81 %	82 %	82 %	83 %	84 %	83 %	84 %
Cliente Febrero	83 %	83 %	84 %	85 %	85 %	85 %	86 %
COMPANY Febrero	82 %	82 %	83 %	84 %	84 %	84 %	85 %
Cliente Marzo	82 %	83 %	84 %	85 %	85 %	85 %	86 %
COMPANY Marzo	80 %	81 %	81 %	82 %	83 %	82 %	83 %

Cuadro 5.17: Tabla de presencia variables parte 4: Experimento 2

	FACTURACION Mayo10	FACTURACION Junio10	CONSUMO Octubre09	CONSUMO Noviembre09	CONSUMO Diciembre09	CONSUMO Enero10	CONSUMO Febrero10	CONSUMO Marzo10	Consumo Abril
fact ngn	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact bdl	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact dtc	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact ded	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact ba	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact tl	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact dc	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact pu	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact fc	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact mc	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact va	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
fact total	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
tamano	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
Ciclo Vida	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
valor	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
retencion	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
cat corp	77 %	77 %	20 %	21 %	20 %	21 %	22 %	22 %	60 %
GIRO	83 %	82 %	85 %	85 %	85 %	85 %	85 %	85 %	84 %
TELEFONO CONTACTO	85 %	85 %	84 %	84 %	86 %	86 %	86 %	86 %	86 %
CALLE FACT	44 %	44 %	50 %	44 %	44 %	44 %	44 %	44 %	44 %
NRO CASA	77 %	76 %	80 %	77 %	80 %	80 %	81 %	80 %	80 %
REFERENCIA	85 %	84 %	82 %	83 %	84 %	85 %	85 %	85 %	86 %
ORIENTACION	86 %	85 %	83 %	84 %	84 %	85 %	86 %	86 %	86 %
NRO PISO	86 %	85 %	83 %	84 %	85 %	86 %	86 %	86 %	87 %
COMUNA	44 %	44 %	50 %	44 %	44 %	44 %	44 %	44 %	44 %
INGRESO CONTRATO SAP	91 %	90 %	5 %	5 %	5 %	17 %	39 %	39 %	75 %
PRIMERA FACTURA	17 %	18 %	0 %	0 %	0 %	0 %	0 %	0 %	12 %
ULTIMA FACTURA	17 %	18 %	0 %	0 %	0 %	0 %	0 %	0 %	12 %
FACTURACION Octubre09	42 %	41 %	0 %	11 %	18 %	28 %	34 %	34 %	41 %
FACTURACION Noviembre09	35 %	34 %	7 %	0 %	7 %	19 %	26 %	26 %	33 %
FACTURACION Diciembre09	29 %	29 %	11 %	4 %	0 %	13 %	20 %	20 %	28 %
FACTURACION Enero10	22 %	22 %	15 %	9 %	5 %	0 %	9 %	10 %	18 %
FACTURACION Febrero10	16 %	16 %	19 %	13 %	9 %	5 %	0 %	2 %	11 %
FACTURACION Marzo10	15 %	15 %	21 %	15 %	11 %	8 %	4 %	0 %	9 %
FACTURACION Abril10	7 %	8 %	25 %	19 %	15 %	12 %	8 %	4 %	0 %
FACTURACION Mayo10	0 %	2 %	30 %	25 %	22 %	20 %	18 %	15 %	12 %
FACTURACION Junio10	10 %	0 %	35 %	30 %	27 %	26 %	24 %	22 %	20 %
CONSUMO Octubre09	42 %	41 %	0 %	11 %	18 %	28 %	34 %	34 %	41 %
CONSUMO Noviembre09	35 %	34 %	7 %	0 %	7 %	19 %	26 %	26 %	33 %
CONSUMO Diciembre09	29 %	29 %	11 %	4 %	0 %	13 %	20 %	20 %	28 %
CONSUMO Enero10	22 %	22 %	15 %	9 %	5 %	0 %	9 %	10 %	18 %
CONSUMO Febrero10	16 %	16 %	19 %	13 %	9 %	5 %	0 %	2 %	11 %
CONSUMO Marzo10	15 %	15 %	21 %	15 %	11 %	8 %	4 %	0 %	9 %
Consumo Abril	7 %	8 %	25 %	19 %	15 %	12 %	8 %	4 %	0 %
Consumo Mayo	0 %	2 %	30 %	25 %	22 %	20 %	18 %	15 %	12 %
Consumo Junio	10 %	0 %	35 %	30 %	27 %	26 %	24 %	22 %	20 %
Reclamos facturacion	84 %	83 %	81 %	82 %	82 %	83 %	84 %	84 %	85 %
Plan Tipo 1	14 %	7 %	29 %	29 %	43 %	50 %	50 %	50 %	21 %
Plan Tipo 2	14 %	7 %	29 %	29 %	43 %	50 %	50 %	50 %	21 %
Plan Tipo 3	14 %	7 %	29 %	29 %	43 %	50 %	50 %	50 %	21 %
Plan Tipo 4	14 %	7 %	29 %	29 %	43 %	50 %	50 %	50 %	21 %
N Planes Renunciados	14 %	7 %	29 %	29 %	43 %	50 %	50 %	50 %	21 %
ICP	74 %	74 %	36 %	37 %	39 %	39 %	40 %	40 %	63 %
Cliente Octubre	85 %	84 %	82 %	83 %	83 %	85 %	85 %	85 %	86 %
COMPANY Octubre	84 %	82 %	80 %	80 %	81 %	83 %	84 %	83 %	84 %
Cliente Noviembre	85 %	84 %	82 %	83 %	83 %	85 %	85 %	85 %	86 %
COMPANY Noviembre	84 %	83 %	81 %	81 %	82 %	83 %	84 %	84 %	85 %
Cliente Diciembre	85 %	84 %	82 %	83 %	83 %	84 %	85 %	85 %	86 %
COMPANY Diciembre	84 %	82 %	81 %	81 %	82 %	83 %	84 %	83 %	84 %
Cliente Enero	85 %	84 %	83 %	83 %	84 %	85 %	85 %	85 %	86 %
COMPANY Enero	84 %	83 %	81 %	82 %	82 %	83 %	84 %	83 %	84 %
Cliente Febrero	85 %	84 %	83 %	82 %	83 %	84 %	85 %	85 %	86 %
COMPANY Febrero	84 %	83 %	82 %	82 %	83 %	84 %	84 %	84 %	85 %
Cliente Marzo	85 %	84 %	82 %	83 %	84 %	85 %	85 %	85 %	86 %
COMPANY Marzo	82 %	81 %	80 %	81 %	81 %	82 %	83 %	82 %	83 %

Cuadro 5.18: Tabla de presencia variables parte 5: Experimento 2

	Consumo Mayo	Consumo Junio	Reclamos facturacion	Plan Tipo 1	Plan Tipo 2	Plan Tipo 3	Plan Tipo 4	N Planes Renunciados	ICP	Cliente Octubre	COMPANY Octubre
fact ngn	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact bdl	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact dtc	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact ded	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact ba	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact tl	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact dc	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact pu	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact ic	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact mc	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact va	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
fact total	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
tamano	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
Ciclo Vida	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
valor	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
retencion	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
cat corp	77 %	77 %	2 %	94 %	94 %	94 %	94 %	94 %	12 %	2 %	5 %
GIRO	83 %	82 %	17 %	100 %	100 %	100 %	100 %	100 %	97 %	6 %	16 %
TELEFONO CONTACTO	85 %	85 %	23 %	99 %	99 %	99 %	99 %	99 %	98 %	6 %	17 %
CALLE FACT	44 %	44 %	6 %	61 %	61 %	61 %	61 %	61 %	83 %	0 %	6 %
NRO CASA	77 %	76 %	15 %	91 %	91 %	91 %	91 %	91 %	93 %	7 %	17 %
REFERENCIA	85 %	84 %	17 %	100 %	100 %	100 %	100 %	100 %	96 %	6 %	17 %
ORIENTACION	86 %	85 %	17 %	100 %	100 %	100 %	100 %	100 %	96 %	7 %	17 %
NRO PISO	86 %	85 %	18 %	100 %	100 %	100 %	100 %	100 %	96 %	7 %	17 %
COMUNA	44 %	44 %	6 %	61 %	61 %	61 %	61 %	61 %	83 %	0 %	6 %
INGRESO CONTRATO SAP	91 %	90 %	5 %	96 %	96 %	96 %	96 %	96 %	50 %	0 %	2 %
PRIMERA FACTURA	17 %	18 %	1 %	99 %	99 %	99 %	99 %	99 %	78 %	0 %	0 %
ULTIMA FACTURA	17 %	18 %	1 %	99 %	99 %	99 %	99 %	99 %	78 %	0 %	0 %
FACTURACION Octubre09	42 %	41 %	8 %	99 %	99 %	99 %	99 %	99 %	86 %	1 %	2 %
FACTURACION Noviembre09	35 %	34 %	6 %	99 %	99 %	99 %	99 %	99 %	86 %	0 %	0 %
FACTURACION Diciembre09	29 %	29 %	6 %	99 %	99 %	99 %	99 %	99 %	85 %	0 %	0 %
FACTURACION Enero10	22 %	22 %	5 %	99 %	99 %	99 %	99 %	99 %	84 %	0 %	1 %
FACTURACION Febrero10	16 %	16 %	5 %	99 %	99 %	99 %	99 %	99 %	84 %	0 %	2 %
FACTURACION Marzo10	15 %	15 %	6 %	99 %	99 %	99 %	99 %	99 %	84 %	1 %	2 %
FACTURACION Abril10	7 %	8 %	7 %	99 %	99 %	99 %	99 %	99 %	90 %	1 %	3 %
FACTURACION Mayo10	0 %	2 %	9 %	99 %	99 %	99 %	99 %	99 %	93 %	2 %	4 %
FACTURACION Junio10	10 %	0 %	10 %	99 %	99 %	99 %	99 %	99 %	94 %	3 %	5 %
CONSUMO Octubre09	42 %	41 %	8 %	99 %	99 %	99 %	99 %	99 %	86 %	1 %	2 %
CONSUMO Noviembre09	35 %	34 %	6 %	99 %	99 %	99 %	99 %	99 %	86 %	0 %	0 %
CONSUMO Diciembre09	29 %	29 %	6 %	99 %	99 %	99 %	99 %	99 %	85 %	0 %	0 %
CONSUMO Enero10	22 %	22 %	5 %	99 %	99 %	99 %	99 %	99 %	84 %	0 %	1 %
CONSUMO Febrero10	16 %	16 %	5 %	99 %	99 %	99 %	99 %	99 %	84 %	0 %	2 %
CONSUMO Marzo10	15 %	15 %	6 %	99 %	99 %	99 %	99 %	99 %	84 %	1 %	2 %
Consumo Abril	7 %	8 %	7 %	99 %	99 %	99 %	99 %	99 %	90 %	1 %	3 %
Consumo Mayo	0 %	2 %	9 %	99 %	99 %	99 %	99 %	99 %	93 %	2 %	4 %
Consumo Junio	10 %	0 %	10 %	99 %	99 %	99 %	99 %	99 %	94 %	3 %	5 %
Reclamos facturacion	84 %	83 %	0 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	15 %
Plan Tipo 1	14 %	7 %	14 %	0 %	0 %	0 %	0 %	0 %	79 %	0 %	0 %
Plan Tipo 2	14 %	7 %	14 %	0 %	0 %	0 %	0 %	0 %	79 %	0 %	0 %
Plan Tipo 3	14 %	7 %	14 %	0 %	0 %	0 %	0 %	0 %	79 %	0 %	0 %
Plan Tipo 4	14 %	7 %	14 %	0 %	0 %	0 %	0 %	0 %	79 %	0 %	0 %
N Planes Renunciados	14 %	7 %	14 %	0 %	0 %	0 %	0 %	0 %	79 %	0 %	0 %
ICP	74 %	74 %	4 %	99 %	99 %	99 %	99 %	99 %	0 %	3 %	4 %
Cliente Octubre	85 %	84 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	0 %	14 %
COMPANY Octubre	84 %	82 %	15 %	100 %	100 %	100 %	100 %	100 %	96 %	4 %	0 %
Cliente Noviembre	85 %	84 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	6 %	15 %
COMPANY Noviembre	84 %	83 %	15 %	100 %	100 %	100 %	100 %	100 %	96 %	4 %	12 %
Cliente Diciembre	85 %	84 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	15 %
COMPANY Diciembre	84 %	82 %	15 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	13 %
Cliente Enero	85 %	84 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	6 %	15 %
COMPANY Enero	84 %	83 %	15 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	13 %
Cliente Febrero	85 %	84 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	6 %	15 %
COMPANY Febrero	84 %	83 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	14 %
Cliente Marzo	85 %	84 %	16 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	15 %
COMPANY Marzo	82 %	81 %	15 %	100 %	100 %	100 %	100 %	100 %	96 %	5 %	13 %

Cuadro 5.19: Tabla de presencia variables parte 6: Experimento 2

	Cliente Noviembre	COMPANY Noviembre	Cliente Diciembre	COMPANY Diciembre	Cliente Enero	COMPANY Enero	Cliente Febrero	COMPANY Febrero	Cliente Marzo	COMPANY Marzo
fact ngn	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact bdl	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact dtc	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact ded	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact ba	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact tl	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact dc	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact pu	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact ic	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact mc	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact va	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
fact total	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
tamano	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
Ciclo Vida	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
valor	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
retencion	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
cat corp	0 %	3 %	2 %	8 %	3 %	2 %	7 %	4 %	9 %	14 %
GIRO	5 %	14 %	7 %	15 %	7 %	14 %	6 %	14 %	7 %	22 %
TELEFONO CONTACTO	3 %	15 %	5 %	14 %	5 %	14 %	6 %	13 %	6 %	31 %
CALLE FACT	6 %	11 %	0 %	22 %	0 %	6 %	0 %	6 %	6 %	17 %
NRO CASA	4 %	19 %	5 %	17 %	8 %	12 %	3 %	13 %	11 %	25 %
REFERENCIA	7 %	15 %	8 %	16 %	7 %	15 %	6 %	14 %	8 %	23 %
ORIENTACION	7 %	15 %	7 %	16 %	8 %	15 %	6 %	14 %	8 %	23 %
NRO PISO	7 %	16 %	8 %	16 %	8 %	16 %	7 %	14 %	8 %	24 %
COMUNA	6 %	11 %	0 %	22 %	0 %	6 %	0 %	6 %	6 %	17 %
INGRESO CONTRATO SAP	1 %	2 %	2 %	3 %	8 %	10 %	7 %	14 %	8 %	18 %
PRIMERA FACTURA	0 %	0 %	0 %	0 %	0 %	0 %	1 %	0 %	2 %	3 %
ULTIMA FACTURA	0 %	0 %	0 %	0 %	0 %	0 %	1 %	0 %	2 %	3 %
FACTURACION Octubre09	3 %	4 %	4 %	7 %	6 %	7 %	4 %	7 %	5 %	11 %
FACTURACION Noviembre09	1 %	2 %	3 %	4 %	5 %	6 %	4 %	6 %	4 %	9 %
FACTURACION Diciembre09	1 %	1 %	2 %	3 %	4 %	5 %	3 %	5 %	3 %	8 %
FACTURACION Enero10	1 %	1 %	0 %	1 %	2 %	2 %	2 %	3 %	3 %	5 %
FACTURACION Febrero10	1 %	1 %	1 %	1 %	1 %	1 %	1 %	1 %	2 %	3 %
FACTURACION Marzo10	0 %	1 %	1 %	1 %	0 %	0 %	1 %	1 %	2 %	3 %
FACTURACION Abril10	1 %	2 %	1 %	1 %	1 %	1 %	1 %	1 %	1 %	2 %
FACTURACION Mayo10	1 %	3 %	1 %	2 %	2 %	2 %	1 %	2 %	1 %	3 %
FACTURACION Junio10	2 %	4 %	2 %	3 %	2 %	3 %	1 %	2 %	2 %	4 %
CONSUMO Octubre09	3 %	4 %	4 %	7 %	6 %	7 %	4 %	7 %	5 %	11 %
CONSUMO Noviembre09	1 %	2 %	3 %	4 %	5 %	6 %	4 %	6 %	4 %	9 %
CONSUMO Diciembre09	1 %	1 %	2 %	3 %	4 %	5 %	3 %	5 %	3 %	8 %
CONSUMO Enero10	1 %	1 %	0 %	1 %	2 %	2 %	2 %	3 %	3 %	5 %
CONSUMO Febrero10	1 %	1 %	1 %	1 %	1 %	1 %	1 %	1 %	2 %	3 %
CONSUMO Marzo10	0 %	1 %	1 %	1 %	0 %	0 %	1 %	1 %	2 %	3 %
Consumo Abril	1 %	2 %	1 %	1 %	1 %	1 %	1 %	1 %	1 %	2 %
Consumo Mayo	1 %	3 %	1 %	2 %	2 %	2 %	1 %	2 %	1 %	3 %
Consumo Junio	2 %	4 %	2 %	3 %	2 %	3 %	1 %	2 %	2 %	4 %
Reclamos facturacion	5 %	13 %	6 %	14 %	7 %	13 %	5 %	12 %	7 %	21 %
Plan Tipo 1	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Plan Tipo 2	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Plan Tipo 3	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Plan Tipo 4	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
N Planes Renunciados	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
ICP	1 %	6 %	3 %	7 %	2 %	5 %	4 %	5 %	7 %	14 %
Cliente Octubre	6 %	13 %	6 %	14 %	7 %	14 %	5 %	12 %	7 %	22 %
COMPANY Octubre	5 %	11 %	5 %	12 %	6 %	11 %	5 %	11 %	6 %	20 %
Cliente Noviembre	0 %	13 %	6 %	14 %	6 %	14 %	5 %	12 %	7 %	22 %
COMPANY Noviembre	4 %	0 %	5 %	12 %	6 %	12 %	5 %	10 %	6 %	19 %
Cliente Diciembre	5 %	13 %	0 %	13 %	6 %	13 %	5 %	12 %	7 %	21 %
COMPANY Diciembre	4 %	11 %	4 %	0 %	6 %	12 %	5 %	10 %	6 %	19 %
Cliente Enero	5 %	13 %	6 %	14 %	0 %	12 %	5 %	12 %	7 %	21 %
COMPANY Enero	5 %	12 %	5 %	13 %	5 %	0 %	4 %	10 %	6 %	19 %
Cliente Febrero	6 %	14 %	6 %	15 %	6 %	13 %	0 %	11 %	7 %	21 %
COMPANY Febrero	5 %	12 %	6 %	13 %	5 %	12 %	4 %	0 %	6 %	19 %
Cliente Marzo	5 %	14 %	6 %	14 %	6 %	13 %	5 %	12 %	0 %	20 %
COMPANY Marzo	5 %	11 %	5 %	11 %	5 %	11 %	4 %	9 %	4 %	0 %

5.3.2. Descripción de variables: Base de datos boletas técnicas

Cuadro 5.20: Variables contempladas de la base Boletas Técnicas

Variables Finales	Observaciones
DIAS	Se refiere a los días que toma resolver el reclamo .
Categoria	Es la categoría de cliente a nivel macro.
Categoria Detalle	Es el detalle de la categoría anterior.
Tipo Problema	Es el la clase de problema que se evalúa en el reclamo.
Causa	Es la causa probable del problema.
Nivel 1	Es la descripción de un punto de vista general del problema.
Nivel 2	Es un nivel más detallado del problema.
Segmento	Es el segmento al cual pertenece al cliente según el tipo de empresa que es.
Subsegmento	Es el segmento más detallado al que pertenece el cliente
Categoría Cliente	Es la categoría a priori del cliente, lo que se puede asociar a la clasificación del cliente en NGN instalados
Rut Cliente	Es el cuerpo del rut del cliente
Dv	Es el dígito verificador
Sigla Cliente	La sigla asociada al cliente por parte de la empresa.
RespFalla	Señala el responsable de la falla.
Ts1	Es el tipo de servicio, en codificación general.
Ts2	Es el tipo de servicio, en codificación más específica.
servicio crit	Se refiere a que si el servicio es crítico o no.
Tipo BR	Sin conocimiento.
Tipo Soluc	Notifica si la solución entregada al cliente es otra a la pedida.
TpoTotal	El tiempo total dedicado al reclamo.
Decis ef	Sin conocimiento
Conclus	Señala si el reclamo fue existoso o no

5.3.3. Análisis de conglomerados para los planes

Cuadro 5.21: Criterio AIC: Experimento 2

Agrupación automática				
Número de conglomerados	Criterio de información de Akaike (AIC)	Cambio en AIC(a)	Razón de cambios en AIC(b)	Razón de medidas de distancia(c)
1	893255.505			
2	509833.470	-383422.035	1.000	2.004
3	318470.439	-191363.031	0.499	1.644
4	202091.535	-116378.904	0.304	1.908
5	141093.736	-60997.799	0.159	1.370
6	96562.505	-44531.231	0.116	1.728
7	70802.720	-25759.785	0.067	1.844
8	56840.930	-13961.790	0.036	1.266
9	45820.480	-11020.450	0.029	1.804
10	39723.069	-6097.411	0.016	1.053
11	33932.248	-5790.821	0.015	1.355
12	29666.209	-4266.038	0.011	1.457
13	26746.443	-2919.767	0.008	1.091
14	24073.248	-2673.195	0.007	1.103
15	21652.732	-2420.515	0.006	1.301

Cuadro 5.22: Distribución de conglomerados para los planes: Experimento 2

Distribución de conglomerados			
Conglomerado	N	% de combinados	% del total
1	62,999	36.3 %	36.1 %
2	21,267	12.2 %	12.2 %
3	64,702	37.3 %	37.0 %
4	24,670	14.2 %	14.1 %
Combinados	173,638	100.0 %	99.4 %
Casos excluidos	999		0.6 %
Total	174,637		100.0 %

Cuadro 5.23: Centroides de conglomerados para los planes 1: Experimento 2

Centroides	Anis	
Conglomerado	Media	Desv. típica
1	3.43	2.425
2	6.84	9.928
3	2.65	1.544
4	2.63	1.650
Combinados	3.44	4.150

Cuadro 5.24: Centroides de conglomerados para los planes 2: Experimento 2

Variable: BA	Valores			
	0		1	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	62,999	88.6 %	0	0.0 %
2	8,084	11.4 %	13,183	12.9 %
3	0	0.0 %	64,702	63.1 %
4	0	0.0 %	24,670	24.1 %
Combinados	71,083	100.0 %	102,555	100.0 %

Cuadro 5.25: Centroides de conglomerados para los planes 3: Experimento 2

Variable: TEC WIMAX COBRE	Valores			
	COBRE		WIMAX	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	62,999	40.6 %	0	0.0 %
2	2,708	1.7 %	18,559	99.6 %
3	64,702	41.7 %	0	0.0 %
4	24,600	15.9 %	70	0.4 %
Combinados	155,009	100.0 %	18,629	100.0 %

Cuadro 5.26: Centroides de conglomerados para los planes 4: Experimento 2

Variable: ADSL	Valores			
	0		1	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	62,999	37.3 %	0	0.0 %
2	20,584	12.2 %	683	14.3 %
3	64,702	38.3 %	0	0.0 %
4	20,579	12.2 %	4,091	85.7 %
Combinados	168,864	100.0 %	4,774	100.0 %

Cuadro 5.27: Centroides de conglomerados para los planes 5.1: Experimento 2

Variable: Velocidad	Valores							
	0		640		1,000		1,024	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	62,999	88.6 %	0	0.0 %	0	0.0 %	0	0.0 %
2	8,088	11.4 %	12,751	16.5 %	257	100.0 %	41	100.0 %
3	0	0.0 %	64,702	83.5 %	0	0.0 %	0	0.0 %
4	0	0.0 %	0	0.0 %	0	0.0 %	0	0.0 %
Combinados	71,087	100.0 %	77,453	100.0 %	257	100.0 %	41	100.0 %

Cuadro 5.28: Centroides de conglomerados para los planes 5.2: Experimento 2

Variable: Velocidad	Valores							
	1,294		2,000		5,000		30,000	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	0	0.0 %	0	0.0 %	0	0.0 %	0	0.0 %
2	23	100.0 %	60	0.3 %	46	100.0 %	1	0.0 %
3	0	0.0 %	0	0.0 %	0	0.0 %	0	0.0 %
4	0	0.0 %	20,579	99.7 %	0	0.0 %	4,091	100.0 %
Combinados	23	100.0 %	20,639	100.0 %	46	100.0 %	4,092	100.0 %

5.3.4. Valores perdidos, tabla de frecuencia y explicación de variables creadas

Cuadro 5.29: Glosa de variable Producto principal

Glosa	Categoría	Producto asociado a nomenclatura
Ngn	1	NGN
bdl	2	Bundling
dtc	3	Datacom
ded	4	Dedicado
ba	5	Banda Ancha
tl	6	Telefonía
dc	8	Datacenter
pu	11	Público
ic	7	Conmutado
mc	9	Merconet
va	10	VA y portal

Cuadro 5.30: Variables contempladas de la base Órdenes terminadas

Variable	Cantidad de Valores presentes	Porcentaje de Valores presentes
Rut Cliente	2911	100 %
Cotizacion	2911	100 %
Tipo Trabajo	2911	100 %
Descripcion	494	17 %
Estado Proy	374	13 %
Urgencia Proy	2889	99 %
Trabajo	2907	100 %
Tipo Termino	2911	100 %
Dentro Plazo	2911	100 %
Dias Ejecucion Prov	2266	78 %
Dias Atraso Prov	2230	77 %
Sigla Cliente	2910	100 %
Segmento Cliente	2885	99 %
Fecha Emision Prov	2869	99 %
Fecha Ejecucion Prov	2265	78 %
Fecha Fin Malla Prov	2910	100 %
Fecha Pactada Clie	2826	97 %
Fec Carta Cliente	44	2 %
F Anulacion Prov	2844	98 %
Producto	1739	60 %
Tipo Serv1	2911	100 %
Tipo Serv2	2911	100 %
Velocidad	1135	39 %

Cuadro 5.31: Valores de la función g

x	g(x)	Observación
0	0	La función g(x) vale 1 para toda orden que contenga la palabra CAPACITACION o su valor sea nulo
1	1(x=1)	La función g(x) vale 1 para toda orden que contenga la palabra INSTALACION
2	1(x=2)	La función g(x) vale 1 para toda orden que contenga la palabra AGREGA
3	1(x=3)	La función g(x) vale 1 para toda orden que contenga la palabra ELIMINA
4	1(x=4)	La función g(x) vale 1 para toda orden que contenga la palabra RETIRO MASIVO
5	1(x=5)	La función g(x) vale 1 para toda orden que contenga la palabra RETIRO y que no contenga la palabra MASIVO
6	1(x=6)	La función g(x) vale 1 para toda orden que contenga la palabra AMPLIACION o AUMENTO
7	1(x=7)	La función g(x) vale 1 para toda orden que contenga la palabra CAMBIO
8	1(x=8)	La función g(x) vale 1 para toda orden que contenga la palabra SUSPENSION
9	1(x=9)	La función g(x) vale 1 para toda orden que contenga la palabra DISMINUCION
10	1(x=10)	La función g(x) vale 1 para toda orden que contenga la palabra HABILITACION o REGULARIZACION
11	1(x=11)	La función g(x) vale 1 para toda orden que contenga la palabra TRASLADO
12	1(x=12)	La función g(x) vale 1 para toda orden que contenga la palabra MODIFICACION

Cuadro 5.32: Descripción de valores de la función g

x	g(x)	Descripción de fórmula en Excel	Valor de categoría que entrega
0	0	Valor perdido o fuera de rango	0
1	1(x=1)	SI(ESNUMERO(ENCONTRAR(İNSTALACION";F2));1;0)	1
2	1(x=2)	SI(ESNUMERO(ENCONTRAR(.ᴬGREGA";F2));2;0)	2
3	1(x=3)	SI(ESNUMERO(ENCONTRAR(.ᴱᴵIMINA";F2));3;0)	3
4	1(x=4)	SI(ESNUMERO(ENCONTRAR(RETIRO MASIVO";F2));4;0)	4
5	1(x=5)	SI(ESNUMERO(ENCONTRAR(RETIRO";F2));5;0)-((SI(ESNUMERO(ENCONTRAR("MASIVO";F2));5;0)))	5
6	1(x=6)	SI(O(ESNUMERO(ENCONTRAR(.ᴬMPLIACION";F2));ESNUMERO(ENCONTRAR(.ᴬUMENTO";F2)));6;0)	6
7	1(x=7)	SI(ESNUMERO(ENCONTRAR(ÇAMBIO";F2));7;0)	7
8	1(x=8)	SI(ESNUMERO(ENCONTRAR("SUSPENSION";F2));8;0)	8
9	1(x=9)	SI(ESNUMERO(ENCONTRAR("DISMINUCION";F2));9;0)	9
10	1(x=10)	SI(O(ESNUMERO(ENCONTRAR("HABILITACION";F2));ESNUMERO(ENCONTRAR(REGULARIZACION";F2)));10;0)	10
11	1(x=11)	SI(ESNUMERO(ENCONTRAR("TRASLADO";F2));11;0)	11
12	1(x=12)	SI(ESNUMERO(ENCONTRAR("MODIFICACION";F2));12;0)	12

Cuadro 5.33: Descripción de valores de la función g

x	g(x)	Valor de categoría que entrega	Significado de Categoría
0	0	0	Vacía o Capacitación
1	1(x=1)	1	Instalación de algún servicio
2	1(x=2)	2	Agregación de algún servicio
3	1(x=3)	3	Eliminación de servicios
4	1(x=4)	4	Retiro Masivo de servicios
5	1(x=5)	5	Retiro singular de servicio
6	1(x=6)	6	Ampliación de algún servicio o Aumento de velocidad
7	1(x=7)	7	Cambio de algún servicio
8	1(x=8)	8	Suspensión de algún servicio
9	1(x=9)	9	Disminución de Velocidad
10	1(x=10)	10	Habilitación o Regularización de servicios
11	1(x=11)	11	Traslado(interno o singular)
12	1(x=12)	12	Modificación a servicios

Cuadro 5.34: Cantidad de ruts que poseen al menos un ani con otro suscriptor distinto de COMPANY

Compañía	Ruts con al menos un Ani en la competencia
No poseen anis en los suscriptores de la competencia	5689
1	2089
12	360
27	270
19	203
3	105
11	97
2	28
6	22
26	16
30	14
7	13
16	11
32	7
Total de ruts considerados	8924

5.3.5. Configuraciones de modelos

Cuadro 5.35: Glosa Explicativa de columnas de configuración del modelo árboles de decisión

Árboles	Significado
A	Criterio
B	Minimal Size
C	Minimal Leaf
D	Minimal gain
E	Maximal depth
F	Confidence
G	Number of preprunning

Cuadro 5.36: Glosa Explicativa de columnas de configuración del modelo SVM

SVM	Significado
A	SVM Type
B	Kernel Type
C	Gamma
D	C
E	Cache Size
F	Epsilon
G	Class Weights

Cuadro 5.37: Glosa Explicativa de columnas de configuración del modelo *Naive Bayes*

NBK	Significado
A	Laplace correction
B	Estimation mode
C	Minimum bandwith
D	Number of Kernels
E	Use application grid
F	Band selection
G	Bandwith

Cuadro 5.38: Glosa Explicativa de columnas de configuración del modelo KNN

KNN	Significado
A	K (número de vecinos cercanos)
B	Weighted Vote
C	Measure Types
D	Mixed Measure

Cuadro 5.39: Configuraciones usadas del modelo árboles de decisión

Árbol ID	A	B	C	D	E	F	G
A1	Gain ratio	4	2	0.1	20	0.25	3
A2	Gain ratio	20	2	0.1	20	0.2	3
A3	Gain ratio	50	2	0.1	20	0.25	3
A5	Gain ratio	2	2	0.1	20	0.25	3
A6	Gain ratio	10	2	0.1	20	0.25	3
A7	Gain ratio	20	2	0.1	20	0.25	3
A8	Gain ratio	1	2	0.1	20	0.25	3
A9	Gain ratio	5	2	0.1	20	0.25	3
A10	Gain ratio	4	2	1	20	0.25	3
A11	Gain ratio	4	2	2	20	0.25	3
A12	Gain ratio	4	2	0.5	20	0.25	3
A13	Gain ratio	4	2	0.01	20	0.25	3
A14	Gain ratio	4	2	0.3	20	0.25	3
A15	Gain ratio	4	2	0.1	1	0.25	3
A16	Gain ratio	4	2	0.1	5	0.25	3
A17	Gain ratio	4	2	0.1	10	0.25	3
A18	Gain ratio	4	2	0.1	50	0.25	3
A19	Gain ratio	4	2	0.1	20	0.1	3
A20	Gain ratio	4	2	0.1	20	0.5	3
A21	Gain ratio	4	2	0.1	20	1	3
A22	Gain ratio	4	2	0.1	20	0.01	3
A23	Gain ratio	4	2	0.1	20	0.25	0
A24	Gain ratio	4	2	0.1	20	0.25	5
A25	Gain ratio	4	2	0.1	20	0.25	10
A26	Gain ratio	4	2	0.1	20	0.25	20
A27	Info Gain	4	2	0.1	20	0.25	3
A28	Gini	4	2	0.1	20	0.25	3

Cuadro 5.40: Configuraciones usadas del modelo SVM

SVM id	A	B	C	D	E	F	Parámetro extra
S1	CSVC	RBF	0	0	80	0.001	
S2	CSVC	RBF	0	100	80	0.001	
S3	CSVC	RBF	0	10	80	0.001	
S4	CSVC	RBF	0	50	80	0.001	
S5	CSVC	RBF	0	-1	80	0.001	
S6	CSVC	SIGMOID	0	0	80	0.001	
S7	CSVC	SIGMOID	0	100	80	0.001	
S8	CSVC	SIGMOID	0	1000	80	0.001	
S9	CSVC	LINEAR	0	0	80	0.001	
S10	CSVC	LINEAR	0	10	80	0.001	
S11	CSVC	POLYNOMINAL	0	0	80	0.001	
S12	CSVC	PRECOMPUTED	0	0	80	0.001	
S13	CSVC	PRECOMPUTED	0	100	80	0.001	
S14	CSVC	PRECOMPUTED	0	0	80	0.001	Coef0:10
S15	CSVC	RBF	100	0	80	0.001	
S16	CSVC	RBF	10	0	80	0.001	
S17	CSVC	RBF	1	0	80	0.001	
S18	NUCLASS	RBF	0	0	80	0.001	Nu:0.1
S19	NUCLASS	SIGMOID	0	0	80	0.001	Nu:0.1
S20	CSVC	RBF	0	0	80	0.00000001	

Cuadro 5.41: Configuraciones usadas del modelo KNN

KNN id	A	B	C	D
KN1	1	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN2	1	SI	MIXED MEASURE	MIXED EUCLIDEAN
KN3	2	NO	NOMINAL	NOMINAL DISTANCE
KN4	2	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN5	2	SI	MIXED MEASURE	MIXED EUCLIDEAN
KN6	3	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN7	3	SI	MIXED MEASURE	MIXED EUCLIDEAN
KN8	4	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN9	4	SI	MIXED MEASURE	MIXED EUCLIDEAN
KN10	5	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN11	5	SI	MIXED MEASURE	MIXED EUCLIDEAN
KN12	6	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN13	6	SI	MIXED MEASURE	MIXED EUCLIDEAN
KN14	15	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN15	30	NO	MIXED MEASURE	MIXED EUCLIDEAN
KN16	30	SI	MIXED MEASURE	MIXED EUCLIDEAN

Cuadro 5.42: Configuraciones usadas del modelo Bayes

NBK ID	A	B	C	D	E	F	G
NB1	SI	GREEDY	1	10	NO		
NB2	SI	FULL	1	10	NO	FIX	0.4
NB3	SI	FULL	1	10	NO	FIX	0.1
NB4	SI	FULL	1	10	NO	FIX	0.7
NB5	SI	FULL	1	10	NO	FIX	1
NB6	SI	GREEDY	1	5	NO		
NB7	SI	GREEDY	1	1	NO		
NB8	SI	GREEDY	1	15	NO		
NB9	SI	GREEDY	1	30	NO		
NB10	SI	GREEDY	1	40	NO		
NB11	SI	GREEDY	1	50	NO		
NB12	SI	GREEDY	1	100	NO		
NB13	SI	GREEDY	0.01	10	NO		
NB14	SI	GREEDY	1	500	NO		
NB16	SI	GREEDY	0.05	10	NO		

Cuadro 5.43: Otros Modelos usados en el experimento 2

Otros modelos	Observacion
Vote 1	Modelo de votación con 2 KNN
Vote 2	Modelo de votación con 2 KNN
Vote 3	Modelo de votación con 2 KNN
IDNumerical 1	Parámetro minimal gain: 0.1
IDNumerical 2	Parámetro minimal gain: 0.01
Ladtree	Iteraciones 5
J4.8	Modelo de árboles estándar

5.3.6. Resultados y evaluación de los modelos probados

Cuadro 5.44: Resultados de modelo árboles de decisión, Experimento 2

Árbol ID	TN	TP	FP	FN	ACCURACY	F	AUC(Entrenamiento)
A1	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A2	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A3	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A5	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A6	5610	23	66	31	98.31 %	66.44 %	AUC: 0.880 +/- 0.032
A7	5585	23	91	31	97.87 %	64.72 %	AUC: 0.877 +/- 0.031
A8	5639	5	37	49	98.50 %	54.91 %	AUC: 0.962 +/- 0.013
A9	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A10	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
A11	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
A12	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
A13	5581	25	95	29	97.84 %	65.68 %	AUC: 0.992 +/- 0.003
A14	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
A15	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A16	5616	23	60	31	98.41 %	66.98 %	AUC: 0.880 +/- 0.032
A17	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A18	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A19	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A20	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A21	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A22	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A23	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A24	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A25	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A26	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.032
A27	5620	23	56	31	98.48 %	67.39 %	AUC: 0.880 +/- 0.033
A28	5620	11	56	43	98.27 %	58.75 %	AUC: 0.982 +/- 0.006

Cuadro 5.45: Resultados de modelo SVM, Experimento 2

SVM id	TN	TP	FP	FN	ACCURACY	F	AUC(Entrenamiento)
S1	5557	2	119	52	97.02 %	50.58 %	AUC: 0.561 +/- 0.064
S2	5473	3	203	51	95.57 %	50.63 %	AUC: 0.188 +/- 0.026
S3	5512	3	164	51	96.25 %	50.88 %	AUC: 0.241 +/- 0.036
S4	5482	3	194	51	95.72 %	50.68 %	AUC: 0.194 +/- 0.027
S5	5557	2	119	52	97.02 %	50.58 %	AUC: 0.561 +/- 0.064
S6	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
S7	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
S8	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
S9	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.510 +/- 0.004
S10	5676	0	0	54	99.06 %	No se puede calcular	AUC: 0.500 +/- 0.000
S11	0	54	5676	0	0.94 %	No se puede calcular	AUC: 0.000 +/- 0.000
S12	5478	2	198	52	95.64 %	50.07 %	AUC: 0.495 +/- 0.042
S13	5478	2	198	52	95.64 %	50.07 %	AUC: 0.503 +/- 0.022
S14	5478	2	198	52	95.64 %	50.07 %	AUC: 0.495 +/- 0.042
S15	5676	0	0	54	99.06 %	No se puede calcular	AUC: 1.000 +/- 0.000
S16	5676	0	0	54	99.06 %	No se puede calcular	AUC: 1.000 +/- 0.000
S17	5676	0	0	54	99.06 %	No se puede calcular	AUC: 1.000 +/- 0.000
S18	0	54	5676	0	0.94 %	No se puede calcular	AUC: 0.000 +/- 0.000
S19	0	54	5676	0	0.94 %	No se puede calcular	AUC: 0.000 +/- 0.000
S20	5557	2	119	52	97.02 %	50.58 %	AUC: 0.561 +/- 0.064

Cuadro 5.46: Resultados de modelo Naive Bayes, Experimento 2

NBK ID	TN	TP	FP	FN	ACCURACY	F	AUC(Entrenamiento)
NB1	4129	34	1540	20	72.74 %	58.14 %	AUC: 0.988 +/- 0.010
NB2	5563	12	31	40	98.74 %	62.41 %	AUC: 0.978 +/- 0.005
NB3	5630	11	27	43	98.77 %	61.95 %	AUC: 0.986 +/- 0.007
NB4	5626	8	37	46	98.55 %	57.77 %	AUC: 0.981 +/- 0.011
NB5	5637	4	31	50	98.58 %	54.34 %	AUC: 0.978 +/- 0.009
NB6	2873	39	2801	15	50.84 %	55.39 %	AUC: 0.968 +/- 0.023
NB7	2025	50	3646	4	36.24 %	56.56 %	AUC: 0.926 +/- 0.044
NB8	5388	16	277	38	94.49 %	56.94 %	AUC: 0.986 +/- 0.007
NB9	4483	29	1180	25	78.92 %	57.65 %	AUC: 0.970 +/- 0.020
NB10	4858	24	805	30	85.39 %	57.29 %	AUC: 0.979 +/- 0.011
NB11	5102	20	561	34	89.59 %	56.83 %	AUC: 0.977 +/- 0.012
NB12	5070	23	593	31	89.09 %	57.92 %	AUC: 0.982 +/- 0.011
NB13	4569	25	1094	29	80.36 %	56.44 %	AUC: 0.991 +/- 0.005
NB14	5664	2	12	52	98.88 %	54.10 %	AUC: 0.985 +/- 0.005
NB16	3163	40	2503	14	56.00 %	56.86 %	

Cuadro 5.47: Resultados de modelo KNN, Experimento 2

KNN id	TN	TP	FP	FN	ACCURACY	F	AUC(Entrenamiento)
KN1	5656	9	20	45	98.87 %	61.44 %	AUC: 0.500 +/- 0.000
KN2	5656	9	20	45	98.87 %	61.44 %	AUC: 0.500 +/- 0.000
KN3	5673	0	3	54	99.01 %	49.75 %	AUC: 0.935 +/- 0.146
KN4	5656	9	20	45	98.87 %	61.44 %	AUC: 0.847 +/- 0.227
KN5	5656	9	20	45	98.87 %	61.44 %	AUC: 0.847 +/- 0.227
KN6	5616	20	60	34	98.36 %	64.97 %	AUC: 0.996 +/- 0.004
KN7	5616	20	60	34	98.36 %	64.97 %	AUC: 0.996 +/- 0.004
KN8	5619	20	57	34	98.41 %	65.24 %	AUC: 0.996 +/- 0.003
KN9	5619	20	57	34	98.41 %	65.24 %	
KN10	5595	25	81	29	98.08 %	66.54 %	AUC: 0.996 +/- 0.003
KN11	5595	25	81	29	98.08 %	66.54 %	AUC: 0.997 +/- 0.003
KN12	5595	25	81	29	98.08 %	66.54 %	AUC: 0.994 +/- 0.003
KN13	5595	25	81	29	98.08 %	66.54 %	AUC: 0.996 +/- 0.003
KN14	5599	17	77	37	98.01 %	61.73 %	AUC: 0.989 +/- 0.004
KN15	5629	13	47	41	98.46 %	61.04 %	AUC: 0.979 +/- 0.006
KN16	5628	13	48	41	98.45 %	60.95 %	AUC: 0.982 +/- 0.005

Cuadro 5.48: Resultados de otros modelos, Experimento 2

Modelos	TN	TP	FP	FN	ACCURACY	F	AUC(Entrenamiento)
Vote 1	5658	14	18	40	98.99 %	66.88 %	AUC: 1.000 +/- 0.000
Vote 2	5604	25	72	29	98.24 %	67.21 %	
Vote 3	5630	22	46	32	98.64 %	67.87 %	AUC: 0.986 +/- 0.024
IDNumerical 1	5642	19	34	35	98.80 %	67.45 %	AUC: 0.955 +/- 0.013
IDNumerical 2	5611	21	65	33	98.29 %	65.21 %	AUC: 0.963 +/- 0.012
Ladtree	5587	42	89	12	98.24 %	75.42 %	
J4.8	5603	11	73	43	97.98 %	57.81 %	

Cuadro 5.49: Tabla de confusión: Entrenamiento LADTree en experimento 2 para el mes de Marzo

Training abril			
Categorías Predicción	Categorías Real		Precisión
	0	1	
0	3981	94	97.69 %
1	39	166	80.98 %
Recall	99.03 %	63.85 %	
Accuracy	96.89 %		
AUC	0.942		
Medida F	85.20 %		
Medida F (clase=1)	71.09 %		

Cuadro 5.50: Tabla de confusión: Entrenamiento LADTree en experimento 2 para el mes de Abril

Training abril			
Categorías Predicción	Categorías Real		Precisión
	0	1	
0	5683	47	99.18 %
1	0	0	0.00 %
Recall	100.00 %	0.00 %	
Accuracy	99.18 %		
AUC	0.942		
Medida F	49.79 %		
Medida F (clase=1)	Desconocida		

5.4. Anexo 4: Experimento 3

5.4.1. Estrategias para el tratamiento de valores perdidos y estudio de valores fuera de rango

Cuadro 5.51: Tabla de valores perdidos y estrategias: Experimento 3

Variable	Valores Válidos	Valores Perdidos	Estrategia
fact total 1	5472	731	Interpolación
Importancia Ngn	6203	0	Ninguna
Productoprincipal	6203	0	Ninguna
PRIMERA INSTALACION	5755	448	Ingreso Contrato + Delta interpolada
INGRESO CONTRATO SAP mejorado	5728	475	Primera Instalación + Delta interpolada
INGRESO CONTRATO SAP 2	6115	88	Ninguna
Delta interpolada	5728	475	Interpolación
Sucursales	6203	0	Ninguna
CANT CPO	6203	0	Ninguna
Q SUCURSALES CON BA	6203	0	Ninguna
Q SUCURSALES SIN BA	6203	0	Ninguna
Q ACCESOS BA	6203	0	Ninguna
Q ANIS	6203	0	Ninguna
cantidad facturas	6180	23	Reemplazo por valor 0
Promedio fact	6203	0	Ninguna
Trim1 fact	6203	0	Ninguna
Trim2 fact	6203	0	Ninguna
Promedio consumo	6203	0	Ninguna
Trim1 consumo	6203	0	Ninguna
Trim2 consumo	6203	0	Ninguna
Rec fac	6203	0	Ninguna
Competencia	6203	0	Ninguna
COMPANY MOBILE	6203	0	Ninguna
COMPANY PHONE	6203	0	Ninguna
COMPANY CELL	6203	0	Ninguna
Planes tipo 1	6203	0	Ninguna
Planes tipo 2	6203	0	Ninguna
Planes tipo 3	6203	0	Ninguna
Planes tipo 4	6203	0	Ninguna
Planes tipo 5	6203	0	Ninguna
Planes Renunciados	0	6203	Variable eliminada

Continúa en la página siguiente

Cuadro 5.51 – Continuación de la página anterior

Variable	Valores Válidos	Valores Perdidos	Estrategia
ICP reload	6203	0	Ninguna
Recencyreal	6203	0	Ninguna
Freq rec	6203	0	Ninguna
Mount rec	6203	0	Ninguna
Imagen	6203	0	Ninguna
CorpOMay	6203	0	Ninguna
Delta interpolada 1	6200	3	Ninguna
fact total 1 1	6197	6	Ninguna
Importancia Ngn 1	6203	0	Ninguna
Tamaño numero	5472	731	Ninguna
Ciclo vida	5472	731	Ninguna
Valor R Num	5472	731	Ninguna
Reten R	5472	731	Ninguna
Cat corp RN	5472	731	Ninguna
Comuna	6180	23	Transformación particular
Comuna 2	3939	2264	Reemplazo por valor 2
GIRO	4011	2192	Reemplazo por valor S/I
Facturacion Marzo	4982	1221	Reemplazo por valor 0
Facturacion Abril	5030	1173	Reemplazo por valor 0
Facturacion Mayo	5046	1157	Reemplazo por valor 0
Facturacion Junio	5050	1153	Reemplazo por valor 0
Facturacion Julio	5126	1077	Reemplazo por valor 0
Facturacion Agosto	5138	1065	Reemplazo por valor 0
Consumo Marzo	4983	1220	Reemplazo por valor 0
Consumo Abril	5031	1172	Reemplazo por valor 0
Consumo Mayo	5047	1156	Reemplazo por valor 0
Consumo Junio	5051	1152	Reemplazo por valor 0
Consumo Julio	5127	1076	Reemplazo por valor 0
Consumo Agosto	5139	1064	Reemplazo por valor 0
Rec Mar Cl	786	5417	Reemplazo por valor 0
Rec Mar Em	448	5755	Reemplazo por valor 0
Rec Abr Cl	1343	4860	Reemplazo por valor 0
Rec Abr Em	143	6060	Reemplazo por valor 0
Rec May Cl	376	5827	Reemplazo por valor 0
Rec May Em	412	5791	Reemplazo por valor 0
Rec Jun Cl	992	5211	Reemplazo por valor 0
Rec Jun Em	393	5810	Reemplazo por valor 0

Continúa en la página siguiente

Cuadro 5.51 – Continuación de la página anterior

Variable	Valores Válidos	Valores Perdidos	Estrategia
Rec Jul Cl	1052	5151	Reemplazo por valor 0
Rec Jul Em	574	5629	Reemplazo por valor 0
Rec Ago Cl	1008	5195	Reemplazo por valor 0
Rec Ago Em	828	5375	Reemplazo por valor 0

Cuadro 5.52: Tabla de valores perdidos en los meses considerados: Experimento 3

Variables	Valores Perdidos		
	Julio	Agosto	Septiembre
RUTCOD	0	0	0
fact total 1	731	731	731
Importancia Ngn	6203	6203	6203
Productoprincipal	6203	6203	6203
Tamaño post outlier	731	731	731
Ciclo vid miss	731	731	731
ValorReal	731	731	731
Retencionreal	731	731	731
Cat corp real	731	731	731
GIRO	2192	2192	2192
GIRO2			622
TELEFONO CONTACTO	29	23	23
PRIMERA INSTALACION	45	23	23
CALLE FACT	23	45	23
ComunaFinal			0
COMUNA	45	22	23
INGRESO CONTRATO SAP mejorado	475	475	23
Delta interpolada	475	452	475
Sucursales	23	0	23
CANT CPO	23	0	23
Q SUCURSALES CON BA	23	0	23
Q SUCURSALES SIN BA	23	0	23
Q ACCESOS BA	23	0	23
Q ANIS	23	0	23
cantidad facturas	23	0	23
Promedio fact	848	789	689
Trim1 fact	1117	1042	979
Trim2 fact	970	928	847
Promedio consumo	1031	974	948
Trim1 consumo	1300	1233	1217
Trim2 consumo	1217	1184	1172
Rec fac	5054	4845	5712
Competencia post outliers	3030	3030	3030
COMPANY MOBILE	4584	4584	4584
COMPANY PHONE	983	983	983
COMPANY CELL	4248	4248	4248
Planes tipo 1	2645	2618	2597
Planes tipo 2	5506	5457	5440
Planes tipo 3	5049	5046	5042
Planes tipo 4	4739	4735	4732
Planes tipo 5	5705	5701	5698
Planes Renunciados	6203	6203	6203
ICP reload	731	731	731
Recencyreal	3090	3172	7
Freq rec	2886	2748	2860
Mount rec	2886	2748	2860
Imagen	0	0	0

Cuadro 5.53: Tabla de valores fuera de rango en variable tamaño

Variable Tamaño	Frecuencia	Porcentaje [%]
Categorías	134	2.31
Grande	605	10.44
Mediana	936	16.15
Microempresa	947	16.34
Pequeña	3159	54.51
Personas con Acceso	14	0.24
Total	5795	100

Cuadro 5.54: Tabla de valores fuera de rango en variable retención

Variable Retencion	Frecuencia	Porcentaje [%]
Categorías	134	2.31
Full Retención	2044	35.27
Retención	803	13.86
Retención Media	2553	44.06
Retención Mínima	261	4.50
Total	5795	100

Cuadro 5.55: Tabla de valores fuera de rango en variable Giro

GIRO		
Categoría	Cantidad	Porcentaje [%]
?	2169	34.97
Vacío	23	0.37
S/I	608	9.80
Otros	3403	54.86
Total	6203	

Cuadro 5.56: Tabla de valores fuera de rango en variables de la base Seg empresas

Variable	Nominalización	Categoría
Cat_Corp	1	NORMAL
	2	SUPERIOR
	3	PREMIUM
	4	ESPECIAL
Ciclo_vida	0	0 No Clasificado
	1	1 Nuevo
	2	2 Crecimiento
	3	3 Madurez
Retención	1	Retención Mínima
	2	Retención Media
	3	Full Retención
Valor	1	Bajo
	2	Medio Bajo
	3	Medio
	4	Medio Alto
	5	Alto
Tamaño	1	Microempresa
	2	Pequeña
	3	Mediana
	4	Grande

5.4.2. Refinamiento en el análisis de conglomerados de planes

Cuadro 5.57: Distribución de conglomerados para los planes: Experimento 3

Distribución de conglomerados

Conglomerado	N	% de combinados	% del total
1	6422	46.68	46.68
2	1155	8.40	8.40
3	1518	11.03	11.03
4	3833	27.86	27.86
5	830	6.03	6.03
Combinados	13758	100	100
Total	13758		100

Cuadro 5.58: Centroides de conglomerados para los planes 1: Experimento 3

Centroides	Variables			
	ANIS		Velocidad Real	
Conglomerado	Media	Desv. típica	Media	Desv. típica
1	3.35	2.37	1052.80	661.24
2	2.98	2.21	2000	0
3	3.67	3.04	636.21	49.15
4	3.98	2.82	4.15E-14	0
5	11.32	13.94	23.42	155.89
Combinados	4.01	4.64	730.94	754.61

Cuadro 5.59: Centroides de conglomerados para los planes 2: Experimento 3

Variable:Ba	Valores			
	0		1	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	0	0	6422	70.42
2	0	0	1155	12.67
3	0	0	1518	16.65
4	3833	82.63	0	0
5	806	17.37	24	0.26
Combinados	4639	100	9119	100

Cuadro 5.60: Centroides de conglomerados para los planes 3: Experimento 3

Variable:ADSL	Valores			
	0		1	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	6422	51.34	0	0
2	0	0	1155	92.55
3	1518	12.13	0	0
4	3833	30.64	0	0
5	737	5.89	93	7.45
Combinados	12510	100	1248	100

Cuadro 5.61: Centroides de conglomerados para los planes 4: Experimento 3

Variable: TEC WIMAX COBRE	Valores			
	COBRE		WIMAX	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	6422	55.16	0	0.00
2	1127	9.68	28	1.32
3	0	0.00	1518	71.74
4	3833	32.92	0	0.00
5	260	2.23	570	26.94
Combinados	11642	100	2116	100

Cuadro 5.62: Centroides de conglomerados para los planes en resumen: Experimento 3

Planes Centroides	Velocidad	Anis	BA	Adsl	Tec_wiimax_cobre
1	1052	3	1	0	COBRE
2	2000	3	1	1	COBRE
3	636	4	1	0	WIIMAX
4	0	4	0	0	COBRE
5	23	11	0	0	WIIMAX

5.4.3. Análisis de conglomerados de clientes

Cuadro 5.63: Centroides de los conglomerados 1: Experimento 3

Conglomerado	Sucursales [cantidad]	CPO [cantidad]	Q_SUCURSALES_CON_BA [cantidad]	Q_SUCURSALES_SIN_BA [cantidad]
1	0 a 4	0 a 2	0 a 3	0 a 2
2	2 a 33	1 a 9	0 a 24	0 a 17
3	0 a 3	1 a 2	0 a 2	0 a 1

Cuadro 5.64: Centroides de los conglomerados 2: Experimento 3

Conglomerado	Q_ACCESOS_BA [cantidad]	Q_ANIS [cantidad]	Promedio_fact[pesos]	Trim1_fact [pesos]	Trim2_fact [pesos]
1	0 a 3	0 a 19	194.783	204.232	202.86
2	0 a 24	8 a 139	2.530.680	2.694.454	2.469.269
3	0 a 2	0 a 12	90.77	86.841	93.134

Cuadro 5.65: Centroides de los conglomerados 3: Experimento 3

Conglomerado	Trim1_consumo [minutos]	Trim2_consumo [minutos]	Promedio_consumo [minutos]	Reclamos_fact [minutos]
1	1.734	1.573	1.583	1
2	10.355	10.455	9.845	5
3	669	665	654	1

Cuadro 5.66: Centroides de los conglomerados 4: Experimento 3

Conglomerado	Planes_tipo_1 [cantidad]	Planes_tipo_2 [cantidad]	Planes_tipo_3 [cantidad]	Planes_tipo_4 [cantidad]	Planes_tipo_5 [cantidad]
1	1	0	0	0	0
2	6	1	1	4	1
3	1	0	0	0	0

Cuadro 5.67: Centroides de los conglomerados 5: Experimento 3

Conglomerado	ICP_reload [%]	Recency real [meses]	Freq_rec [%]	Monto_rec [cantidad]
1	93,24 %	3	23 %	4
2	94,86 %	1	66 %	12
3	75,97 %	4	14 %	3

Cuadro 5.68: Centroides de los conglomerados 6: Experimento 3

Conglomerado	Tamaño	Ciclo Vida	Valor	Competencia
1	Pequeña_Micro	Madurez_Crecimiento	Alto	MIX_NO
2	Pequeña_Grande	Madurez_No	Alto_Medio	SI
3	Pequeña_Mediana	Madurez_No	Medio_Bajo	MIX_NO

Cuadro 5.69: Centroides de los conglomerados 7: Experimento 3

Conglomerado	Retencion real	Cat_corp_real	COMPANY MOBILE	COMPANY PHONE	COMPANY CELL
1	FULL	Normal_Especial	NO	SI	MIX_NO
2	FULL_MEDIA	Premium_Superior	MIX_SI	SI	MIX_SI
3	MEDIA_BAJA	Normal_Especial	NO	SI	MIX_NO

Cuadro 5.70: Centroides de los conglomerados 8: Experimento 3

Conglomerado	Imagen	Cartera	Morosidad [%]
1	MIX_BUENA	MIX_NO	9
2	MALA	MIX_SI	26
3	MIX_BUENA	NO	12

Cuadro 5.71: Centroides de los subconglomerados 1: Experimento 3

Conglomerado	Sucursales [cantidad]	CPO [cantidad]	Q_SUCURSALES_CON_BA [cantidad]	Q_SUCURSALES_SIN_BA [cantidad]
1	1 a 1	1 a 1	1 a 1	0 a 0
2	0 a 4	1 a 2	0 a 3	0 a 2
3	0 a 4	0 a 2	0 a 3	0 a 2

Cuadro 5.72: Centroides de los subconglomerados 2: Experimento 3

Conglomerado	Q_ACCESOS_BA [cantidad]	Q_ANIS [cantidad]	Promedio_fact [pesos]	Trim1_fact [pesos]	Trim2_fact [pesos]
1	1 a 1	1 a 5	75.849	74.84	80.57
2	0 a 3	1 a 19	188.884	189.888	191.88
3	0 a 3	0 a 14	20.304	5.339	17.205

Cuadro 5.73: Centroides de los subconglomerados 3: Experimento 3

Conglomerado	Trim1_consumo [minutos]	Trim2_consumo [minutos]	Promedio_consumo [minutos]	Rec_fac [cantidad]
1	492	464	469	1
2	1.648	1.604	1.564	1
3	24	108	96	0

Cuadro 5.74: Centroides de los subconglomerados 4: Experimento 3

Conglomerado	Planes_tipo_1 [cantidad]	Planes_tipo_2 [cantidad]	Planes_tipo_3 [cantidad]	Planes_tipo_4 [cantidad]	Planes_tipo_5 [cantidad]
1	1	0	0	0	0
2	1	0	0	1	0
3	1	0	0	0	0

Cuadro 5.75: Centroides de los subconglomerados 5: Experimento 3

Conglomerado	ICP_reload [%]	Recency real [meses]	Freq_rec [%]	Mount_rec [cantidad]
1	92,88 %	4	14 %	3
2	96,66 %	3	25 %	5
3	20,16 %	5	3 %	1

Cuadro 5.76: Centroides de los subconglomerados 6: Experimento 3

Conglomerado	Tamaño	Ciclo Vida	Valor	Competencia
1	Pequeña	Madurez_No	Medio	MIX_NO
2	Mediana_Grande	Madurez_No	Medio	MIX_SI
3	Pequeña	Madurez	Bajo	MIX_NO

Cuadro 5.77: Centroides de los subconglomerados 7: Experimento 3

Conglomerado	Retencion real	Cat_corp_real	COMPANY MOBILE	COMPANY PHONE	COMI
1	MEDIA	NORMAL	NO	SI	
2	MEDIA	ESPECIAL_SUPERIOR	MIX_NO	SI	M
3	MíNIMA	NORMAL	NO	MIX_NO	

Cuadro 5.78: Centroides de los subconglomerados 8: Experimento 3

Conglomerado	Imagen	Cartera	Morosidad [%]
1	MIX_BUENA	NO	8
2	MIX_MALA	MIX_SI	12
3	BUENA	NO	20

Cuadro 5.79: Experimento 3: Resultados para clúster 1

Clúster K=1										
Modelo	LADTree	BFTree	WFT	Ridor	Jrip	J48	NBTree	Logistic	Naive Bayes	Vote (LADTree, Logistic, Naive Bayes
Accuracy [%]	60.19 %	95.72 %	93.73 %	92.07 %	96.48 %	95.39 %	98.62 %	78.91 %	77.15 %	91.92 %
AUC (validación)	0.571	0.541	0.606	0.5	0.573	0.511	0.555	0.507	0.605	0.501
Lift [%]	132.83	321.62	307	260.81	589.64	294.82	1547.79	141.67	130.62	157.74
Medida F positiva [%]	2.10 %	4.26 %	4.35 %	2.34 %	7.50 %	3.96 %	12.12 %	2.20 %	2.04 %	2.30 %
TN	1258	2013	1970	1936	2028	2006	2074	1656	1619	1933
FP	830	75	118	152	60	82	14	432	469	155
FN	8	15	14	15	14	15	15	12	12	15
TP	9	2	3	2	3	2	2	5	5	2
P	17	17	17	17	17	17	17	17	17	17
N	2088	2088	2088	2088	2088	2088	2088	2088	2088	2088
Total	2105	2105	2105	2105	2105	2105	2105	2105	2105	2105

5.4.4. Resultados de modelos probados por cada conglomerado o clúster

Cuadro 5.80: Experimento 3: Resultados para clúster 1 con técnica de muestreo en el entrenamiento

Clúster K=1											
Modelos	Logistic						J48				Rule Induction
Proporción entrenamiento	10 y 90	20 y 80	30 y 70	30 y 70 sin multiappend	40 y 60	40 y 60 sin multiappend	10 y 90	20 y 80	30 y 70	40 y 60	50 y 50
Accuracy [%]	68.08 %	80.19 %	81.09 %	85.75 %	84.80 %	87.98 %	54.82 %	73.82 %	91.02 %	96.06 %	93.63 %
AUC (validación)	0.593	0.574	0.497	0.676	0.499	0.609	0.582	0.519	0.572	0.513	0.563
Medida F positiva [%]	2.89 %	2.80 %	1.00 %	3.23 %	0.62 %	3.80 %	2.26 %	1.78 %	4.06 %	2.35 %	5.63 %
TN	1423	1682	1705	1800	1784	1847	1143	1549	1912	2021	1967
FP	665	406	383	288	304	241	945	539	176	67	121
FN	7	11	15	12	16	12	6	12	13	16	13
TP	10	6	2	5	1	5	11	5	4	1	4
P	17	17	17	17	17	17	17	17	17	17	17
N	2088	2088	2088	2088	2088	2088	2088	2088	2088	2088	2088
Total	2105	2105	2105	2105	2105	2105	2105	2105	2105	2105	2105

Cuadro 5.81: Experimento 3: Resultados para clúster 2

Clúster K=2

Modelos	Árbol 1	Árbol 2	BayesNet	LADTree	J48	SVM 1	SVM 2	SVM 3
Accuracy [%]	97.25	97.25	98.97	96.91	95.88	98.28	1.03	98.28
AUC (validación)	0.488	0.488	0.575	0.557	0.484	0.497	0.497	0.497
Lift [%]	0	0	Desconocida	0	0	0	0	0
Medida F positiva [%]	Desconocida	Desconocida	Desconocida	Desconocida	Desconocida	Desconocida	Desconocida	Desconocida
TP	0	0	0	0	0	0	3	0
TN	283	283	288	282	279	286	0	286
FP	5	5	0	6	9	2	288	2
FN	3	3	3	3	3	3	0	3
P	3	3	3	3	3	3	3	3
N	288	288	288	288	288	288	288	288
Total	291	291	291	291	291	291	291	291

Cuadro 5.82: Experimento 3: Resultados para clúster 3, subclúster 1

SubClúster K3=1																
Modelos	LADTree	Naive Bayes	Logistic	Rule Induction	BayesNet	Conjunctive Rule	WDNTB	Jrip	Oner	PART	Ridor	ZeroR	BFTree	NBTree	Random Forest	WFT
Accuracy [%]	92.04	46.36	85.49	93.75	99.14	99.51	99.51	95.65	95.65	97.61	96.26	99.51	97.67	99.33	99.33	97.73
AUC (validación)	0.58	0.43	0.531	0.389	0.609	0.5	0.629	0.543	0.5	0.49	0.5	0.5	0.553	0.583	0.481	0.57
Lift [%]	0	93.21	88.37	212.63	0	Desconocida	Desconocida	314.04	0	0	371.14	Desconocida	637.89	0	0	658.47
Medida F positiva [%]	Desconocida	0.9	0.84	1.92	Desconocida	Desconocida	Desconocida	2.74	Desconocida	Desconocida	3.17	Desconocida	5	Desconocida	Desconocida	5.13
TN	1503	753	1395	1530	1619	1625	1625	1561	1562	1594	1571	1625	1594	1622	1622	1595
FP	122	872	230	95	6	0	0	64	63	31	54	0	31	3	3	30
FN	8	4	7	7	8	8	8	7	8	8	7	8	7	8	8	7
TP	0	4	1	1	0	0	0	1	0	0	1	0	1	0	0	1
P	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
N	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625	1625
Total	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633	1633

Cuadro 5.83: Experimento 3: Resultados para clúster 3, subclúster 2

SubClúster K3=2																	
Modelos	J48	Rule Induction	SVM	Naive Bayes	Logistic	LADTree	BayesNet	W-DNTB	Jrip	OneR	PART	Ridor	ZeroR	BFTree	NBTree	WFT	Naive Bayes (kernel)
Accuracy [%]	96.28	92	98.47	82.37	92.55	79.52	98.88	98.58	95.51	93.32	95.29	95.95	98.69	94.19	98.58	97.26	98.57
AUC (validación)	0.541	0.551	0.503	0.606	0.583	0.533	0.486	0.399	0.524	0.5	0.482	0.5	0.5	0.517	0.494	0.52	0.446
Lift [%]	0	234.1	0	250.27	131.18	85.01	0	0	245.43	0	0	524.71	Desconocida	176.94	0	0	0
Medida F positiva [%]	Desconocida	5.19	Desconocida	5.85	2.86	2.09	Desconocida	Desconocida	4.65	Desconocida	Desconocida	9.76	Desconocida	3.64	Desconocida	Desconocida	Desconocida
TN	879	838	899	754	844	724	900	900	871	852	870	874	901	859	900	888	894
FP	22	63	2	147	57	177	1	1	30	49	31	27	0	42	1	13	1
FN	12	10	12	7	11	10	12	12	11	12	12	10	12	11	12	12	12
TP	0	2	0	5	1	2	0	0	1	0	0	2	0	1	0	0	0
P	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
N	901	901	901	901	901	901	901	901	901	901	901	901	901	901	901	901	895
Total	913	913	913	913	913	913	913	913	913	913	913	913	913	913	913	913	907

Cuadro 5.84: Experimento 3: Resultados para clúster 3, subclúster 3

SubClúster K3=3																	
Modelos	Logistic	Bagging(Logistic)	BayesNet	LADTree	Naive Bayes	Rule Induction	SVM	J48	Conjunctive	DNTB	Jrip	OneR	Ridor	BFTree	NBTree	WFT	Logistic1
Accuracy [%]	98.77	98.88	98.88	98.77	98.88	98.88	98.88	98.88	98.77	98.65	98.77	98.65	98.77	¡	98.88	98.77	98.54
AUC (validación)	0.636	0.701	0.46	0.549	0.5	0.5	0.54	0.5	0.499	0.622	0.5	0.5	0.5	0.5	0.479	0.678	0.629
Lift [%]	356.8	446	Desconocida	2973.33	Desconocida	Desconocida	Desconocida	Desconocida	0	0	2973.33	0	0	2973.33	Desconocida	0	2973.33
Medida F positiva [%]	26.67	28.57	Desconocida	15.38	Desconocida	Desconocida	Desconocida	Desconocida	Desconocida	Desconocida	15.38	Desconocida	Desconocida	15.38	Desconocida	Desconocida	31.58
TN	879	880	882	880	882	882	882	882	881	880	880	880	881	880	882	881	876
FP	3	2	0	2	0	0	0	0	1	2	2	2	1	2	0	1	6
FN	8	8	10	9	10	10	10	10	10	10	9	10	10	9	10	10	7
TP	2	2	0	1	0	0	0	0	0	0	1	0	0	1	0	0	3
P	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
N	882	882	882	882	882	882	882	882	882	882	882	882	882	882	882	882	882
Total	892	892	892	892	892	892	892	892	892	892	892	892	892	892	892	892	892

5.4.5. Análisis de conglomerados de clientes: Explicación y conclusiones

Los centroides de la segmentación que se ejecuta en el experimento 3 se concretan en las siguientes tablas mostradas a continuación. Posterior a su exposición se analizan en detalle cada una de ella, en pos de determinar la característica representativa de cada uno de los grupos resultantes.

Centroides Variables continuas

Las variables continuas entregan un aproximado, es decir, los centroides tienen una desviación estándar atribuida y con un sentido de distancia. Para el bosquejo de estos centroides se agrega el promedio en ciertas variables, mientras que en otras se coloca un rango de opciones dentro de un mismo intervalos que no es más que el promedio menos y más la desviación correspondiente.

Según la tabla 5.63 se puede concluir que el conglomerado 2 es aquel que posee más servicios asociados a la empresa de Telecomunicaciones, debido a que su valor en cuanto a sucursales es mucho mayor que el resto de los conglomerados.

Por otro lado, acorde a la tabla 5.64 se puede observar que el conglomerado 2 factura una cantidad muy superior al resto de los *clústers*, esto es concordante con la deducción anterior de que el conglomerado 2 presenta gran enlace con la empresa (dada su cantidad de sucursales). No obstante, en esta segunda tabla se puede diferenciar el resto de los clústers. Mientras que el *clúster* 1 factura del orden de 200 mil pesos, el clúster 3 factura la mitad, además, la cantidad de sucursales del grupo 3 es menor que la del grupo 1. Sin embargo, la variable Trim2 fact que corresponde al promedio ponderado del segundo trimestre (en este caso: Junio, Julio y Agosto), es mayor que el promedio general y que Trim1 Fact(primer trimestre, es decir, Marzo, Abril, Mayo). Esto indica un crecimiento en la facturación del *clúster 3*, caso que no es igual en el *clúster* 1, debido a que en este *clúster* la facturación ponderada promedio del segundo trimestre es levemente menor a la del primer trimestre, por lo que, se puede decir que el *clúster* 3 es "nuevo en la compañía" y está probando el producto, mientras que el *clúster* 1 es más "maduro", dado que su facturación es mayor pero su variación marginal es menor que la del *clúster* 3.

En la tabla 5.65 se aprecia el consumo, en una vista general, se observa que el único conglomerado que aumenta su consumo entre trimestres es el 2, sin embargo, todos los conglomerados se encuentran sobre el promedio general. Cabe decir, que las variables consumo originadas son subestimadas, debido a que la facturación en la empresa se carga por "minuto gordo"[1], mientras que en este caso se están valorizando en minutos reales con decimales. Además de las variables de consumo, se encuentran las variables de reclamos de facturación comerciales, los cuales dejan ver que el conglomerado 2 tiene más problemas en los pagos que el resto, lo cual, es de esperarse debido a que el conglomerado 2 posee mayor cantidad de sucursales que el resto de lo *clústers*.

En la tabla 5.66 se encuentra la cantidad de los planes según su tipo, cuya segmentación puede ser visualizada en la tabla 3.22. Como se puede apreciar, los planes que más se tienen son el 1 (que es clásico) y el 4(que está relacionado con la conectividad dentro de la empresa), el resto se presenta en forma escasa en el caso del NGN.

En la tabla 5.67 se vislumbra la presencia de variables continuas, además, se denota el pago a través de la variable ICP reload, el cual, mide la forma histórica en que el cliente se ha comportado

[1]Minuto gordo se refiere a que si el cliente habla más de 30 segundos, se le añade un minuto de facturación a su llamada, por lo que si el cliente habla 1,6 minutos realmente, se le facturan 2 minutos

con la empresa. En el análisis, esta variable indica que el conglomerado que peor se comporta es el 3, lo que fomenta aún más el hecho de que posiblemente los clientes de este conglomerado, recién vengan ingresando a la empresa de telecomunicaciones. Las otras tres variables se refieren al modelo RFM el cual, fue usado para procesar la variable Respfalla, de esta manera, se descubren las respuestas a: ¿Hace cuanto tiempo reclamó? (Recency), ¿Cada cuántos meses reclama?(Frecuencia) y ¿Cuánto reclama cuando lo hace?(Monto). Al observar estas tres variables, se deduce que el conglomerado 2 tiende a reclamar bastante seguido, el valor 1 en la variable *Recency* indica que el cliente reclamó hace 2 meses, el valor 3 es 4 meses, y 4 corresponde a los 5 meses. A esto se agrega que la frecuencia de reclamos (Freq rec) delata que el conglomerado número 2 presenta muchas más fallas técnicas respecto al resto, notificando a un perfil de cliente que reclama constantemente. Finalmente el monto de reclamos (Mount rec), revela que cuando al cliente del conglomerado 2 le falla el producto, lo hace de forma múltiple (12 reclamos en promedio en un mes, equivale a 2 reclamos aproximadamente por día laboral). El resto de los conglomerados no presenta un alto valor de reclamo, aunque el conglomerado 1 tiene un valor mayor que el conglomerado 3.

Variables Nominales

Un alcance a la interpretación de los conglomerados obtenidos es que presentan en forma de porcentajes de dominancia de categorías, es decir, la colocación de un nombre en las siguientes tablas va asociado a un porcentaje de presencia de la categoría. Por ejemplo, en la tabla 5.68, se tiene en tamaño la categoría "Pequeña Micro", dicha categoría no existe en la variable "Tamaño", sin embargo, esa categoría significa que el valor "Pequeña" presenta un alto porcentaje y "Microempresa" presenta el segundo más alto porcentaje dentro del conglomerado. El corte para detener la agregación de más categoría es el hecho de que se abarque más del 75 %, en otras palabras, "Pequeña" y "Micro" están presentes en esa categoría "híbrida" porque entre las dos se encuentra más del 75 % de los clientes pertenecientes al conglomerado 1. Es el mismo caso para la agregación del prefijo "Mix" a las categorías que solamente contienen 2 valores. Bajo este tipo de variables binominales, "MIX_NO" significa que el "NO" tiene un porcentaje superior al 50 % pero no superior al 75 %. El caso es análogo para la categoría "SI".

La tabla 5.68 bosqueja 4 variables, Tamaño, Ciclo de vida, Valor y Competencia. En ellas se puede apreciar que el conglomerado 2 contiene las empresas de mayor tamaño, además, este mismo tiene empresas que, en el ciclo de vida, se encuentran en una etapa madura dentro de la empresa y son de medio y alto valor. Un punto relevante a destacar en este conglomerado es el hecho de que está afecto a la competencia, es decir, no es nuevo en el mercado de las telecomunicaciones como cliente. A diferencia del grupo 2, los otros dos tienen menor relación con la competencia, por ejemplo, el conglomerado 1 es de gran valor, definido como tal por la empresa, además, sus miembros están en una etapa de crecimiento y madurez en su ciclo de vida. En cambio, el grupo 3 tiene un valor bajo debido a que posiblemente son aquellos clientes nuevos o con una relación degradada con la empresa.

En la tabla 5.69 se denotan las variables de Retención, la categoría de la empresa (Cat Corp) y el conocimiento acerca de la compañía por parte del cliente. En base a ella, se puede establecer que el conglomerado 1 es de alto valor, entonces tiene asociado un plan de retención Full, además, es de categoría especial y normal, al igual que el conglomerado 3. Sin embargo, el conglomerado 3 tiene asociado un plan de retención media-baja debido a que es de poco valor. A diferencia de

los 2 conglomerados descritos anteriormente, el número 2 conoce a la empresa completamente y la relación con ella es bastante estrecha debido a que tiene COMPANY MOBILE, COMPANY PHONE y COMPANY CELL, a lo que se añade el hecho de que las categorías dominantes en este conglomerado, en cuanto a la clasificación hecha por la empresa, son Premium y Superior.

Finalmente en la tabla 5.70, se presentan las variables Imagen, Cartera y Morosidad. Con estas variables, se deduce que el conglomerado 2 tiene características de ser un "Rehén" de la empresa, lo que implica que el cliente está tan ligado a la empresa, que a pesar de que conoce el mercado de las telecomunicaciones, tiene muchos planes con la empresa, en otras palabras, le resulta bastante costoso cambiarse a otra compañía. Además, presenta mayor morosidad que los otros conglomerados, con un 26 % de los clientes pertenecientes a ese conglomerado con estado moroso. En lo que se refiere a la variable cartera, resulta familiar que existan más de este tipo de clientes en el conglomerado 2, debido a que el "ser cartera" tiene asociado el tema de la retención. De esta forma, tanto el conglomerado 1 como el 2 entran en esta categoría debido a que en la tabla de la retención estos dos conglomerados tenían los valores Full. Por lo tanto, el conglomerado 1 tiene un buen servicio por parte de la empresa, por ende, tiende a ser más fiel a la misma, además, no conoce tanto el mercado como el conglomerado 2 y su tamaño es de Pequeña o Microempresa, por lo que este conglomerado puede denominarse como "Los Clientes Fieles", que si bien pueden no ser completamente leales, tienen potencial para convertirse en tales. El conglomerado 3 tiene tres vías detectadas: la primera es que pueden ser clientes que tienen planes con la empresa en temas no críticos, es decir, tienen plan es contratados como servicios adicionales a lo que es su competencia central propia en su rubro, la segunda vía, corresponde a que sean clientes nuevos en el tema de las telecomunicaciones y han escogido a la empresa para comenzar a familiarizarse y la tercera vía detectada es que estos clientes sean mercenarios en progreso, lo cual es poco probable debido a que la imagen no es mala.

En resumen las conclusiones llevan a la siguiente tabla:

Cuadro 5.85: Etiquetas asignadas a los conglomerados 9: Experimento 3

Conglomerado	N	% del total	Nombre de Conglomerado
1	2136	36,66	Fieles
2	291	4,99	Rehén
3	3399	58,34	Mixto
Total	5826	100	

Para escoger el número de subclústers existentes, se vuelve a utilizar el dendrograma presente en la clusterización jerárquica. Constrastando las ramas del dendrograma se puede apreciar la exis-

Cuadro 5.86: Nomenclatura sugerida para subconglomerados: Experimento 3

Conglomerado	Posibles integrantes
Mixto	Nuevos
	Servicios no críticos
	Mercenarios

Cuadro 5.87: Distribucion de los subconglomerados: Experimento 3

Conglomerado	N [Cantidad de clientes]	% del total
1	1705	501,618,123
2	859	252,721,389
3	835	245,660,488
Total	3399	100

tencia de tres clústers dentro del grupo 3, dicha comparación se hace respecto a la división de ramas. De esta manera, se entrega la siguiente distribución según subclústers del grupo 3:

Para k=2 no se muestran resultados debido a que la clústerización de k=3 entrega información más completa acerca de los nombres de los subgrupos, y para K=4 tampoco, debido a que se considera un outlier y eso convergerá a considerar un subgrupo redundante, lo cual es apreciable en la tabla de contingencia presente en los anexos, además, se agrega la misma tabla pero para comparar entre K=2 y K=3, nótese que K se refiere al número de clúster deseables.

Para la conglomeración se utiliza el mismo algoritmo que al inicio, es decir, el *Two-Step Clúster*, de esta manera se describen las distintas características de los segmentos.

5.4.6. Análisis de Subconglomerados: Explicación y conclusiones

A continuación se muestra un análisis de cada subconglomerado según los valores del "cliente representativo" determinado por el algoritmo utilizado (Two-Step Clúster).

Para la tabla 5.71 se descubre que el subconglomerado 1 resulta tener pocas sucursales, por lo que probablemente se trata de un cliente nuevo que recién está probando el servicio. En cambio el resto de los subconglomerados presentan cantidades superiores a 1 sucursal (o plan), y tienen más variedad en las mismas, mientras que el subconglomerado 1 solamente posee sucursales con banda ancha, los otros dos poseen, además de lo anterior, planes sin banda ancha.

En la tabla 5.72 se puede apreciar el comportamiento del cliente en su facturación, es así, como se puede apreciar que el subconglomerado 3 se diferencia del subconglomerado 2 por la cantidad facturada, la cual se diferencia en que una es 10 veces la otra, es decir, estos subconglomerados (2 y 3) tienen las misma características técnicas pero facturan de forma distinta.

En la tabla 5.73 se denota el comportamiento de los subconglomerados respecto a su consumo y a los reclamos de facturación efectuado en el mes. Además, se vuelve a recalcar una fuerte diferencia entre el subconglomerado 3 versus el 2, debido a que el consumo de este último supera por mucho más al primero. No obstante, esto ya era deducible por la tabla anterior en la que se apreciaba que la facturación del subconglomerado 2 era mayor que la del 3. Sin embargo, se agrega el hecho de que el subconglomerado 2 tiene problemas con la facturación dado que reclama, al igual que el subconglomerado 1, mientras que el subconglomerado 3 no reclama por facturación.

En la tabla 5.74 se puede agregar una característica al subconglomerado 2 y es el hecho de que posee más de un tipo de plan a diferencia de los otros dos subconglomerados. El tipo de plan extra que posee dicho subconglomerado es el 4, el cual hace referencia a los planes sin banda ancha, agregado a que los planes tipo 1 que hacen referencia a los planes estándar.

La variable ICP, expresada en la tabla 5.75, conjuntamente con otras variables, es bastante baja en el subcconglomerado 3, lo que implica que los clientes de este grupo son malos al momento de pagar. Esto último entrega, la percepción de poco encanto con el producto y poco compromiso con el mismo, lo que se valida al observar la frecuencia de reclamos, así como también, la cantidad de planes que tienen. Bajo esta percepción, se puede representar a este grupo de clientes como aquellos con alto riesgo de deserción. Dentro de esta índole están los terroristas que "se caracterizan por sus bajos niveles de Encanto y de Compromiso futuro"[**?**]. Para el subconglomerado 1 y 2, en cambio, su ICP es mucho mayor, por lo que muestra mayores niveles de compromiso hacia la compañía, sin embargo, puede que su encanto sea menor, debido a que tienen mayor frecuencia, no obstante, la afirmación anterior aún no puede ser concretada.

En la tabla 5.76, se puede apreciar que el subconglomerado 3 no contiene clientes nuevos en su mayoría, dado que el valor de la variable ciclo de vida que la empresa le asigna es el de "Madurez", con ello, se descarta la posibilidad de que sean nuevos en la empresa. A diferencia de este subconglomerado, el 1 posee empresas pequeñas y su ciclo de vida varía entre madurez y no estar clasificado, que indica que aún no se ha estudiado el caso. El subconglomerado 2, por su parte, contiene empresas de mediano tamaño (y grande), además, tiende a tener empresas que poseen conocimiento de la competencia (sin embargo, esto último no se puede concluir a ciencia cierta, debido a que la frecuencia de empresas dentro del subconglomerado 2 no supera el 75 % para la variable competencia.

Las variables presentadas en la tabla 5.77, señalan que el subconglomerado 3, no está ligado o comprometido a la empresa de telecomunicaciones, debido a que no posee COMPANY MOBIL, ni COMPANY CELL, además, presenta una tendencia a no poseer COMPANY PHONE adicional al servicio prestado por este producto. Esto sugiere que antes del servicio NGN, no poseía ninguna relación con la compañía. En cambio, los subconglomerado 1 y 2 sí tenían dicha relación con la compañía, es más, en el caso del subconglomerado 2 esta relación llega a la telefonía móvil.

La variable Imagen, expresada en la tabla 5.78,muestra valores positivos en el primer clúster, por lo que la idea de que sean terroristas queda desechada, es más, entra el tema de que poseen un alto nivel de encanto, no obstante, su compromiso es bajo por las deducciones realizadas anteriormente, con lo que se concluye que el clúster uno alberga a "los Mercenarios". Ahora bien, dentro de los mercenarios, como se ha mencionado anteriormente existen tres subcategorías adicionales. Es por ello, que se debe clarificar específicamente a que subcategoría de mercenarios pertenece el subconglomerado 3. Debido a que no se poseen variables que indiquen nociones del precio del paquete o los precios a los cuales el cliente fue expuesto por la competencia, no se puede deducir si este conglomerado alberga mercenarios que sean "Sensibles al precio". Sin embargo, se descarta el hecho de que sean "Negociadores" dado que la variable competencia señala que este conglomerado tiene una tendencia a no tener tantos clientes que conozcan una marca adicional a la empresa de telecomunicaciones. Es así, como se puede concluir que el subconglomerado 3 es denominado "Los Switchers". Sin embargo, el subconglomerado 1, tiene casi tantas empresas que tengan una buena como mala imagen de la compañía, no obtante, su porcentaje de morosidad resulta considerablemente menor que el resto de los subconglomerados, es por ello, que se tiene en ese grupo a empresas comprometidas con la empresa, pero que no necesariamente están encantadas con la compañía. El nombre que se le asigna a este grupo entonces, es de "Los Indiferentes", pues, no presentan reacciones que se puedan ver visualizadas en las variables de manera relevante.

Por último, el subconglomerado 2, presenta características similares al conglomerado de "Los Rehenes", no obstante, no se puede deducir que la imagen que tiene este subconglomerado 2 sea mala, por ende, se les denominará como "Los Refugiados".

Cuadro 5.88: Etiquetas asignadas a los subconglomerados: Experimento 3

Conglomerado	N	% del total	Nombre de Conglomerado
1	1705	50,16	Los Indiferentes
2	859	25,27	Los Refugiados
3	835	24,57	Los Switchers
Total	3399	100,00	

5.5. Anexo 5: Experimentos 6 y 7

5.5.1. Análisis de relación entre atributos y variable objetivo

Cuadro 5.89: Tabla de contingencia Sucursales*Fuga, Experimento 6 para el mes de Noviembre del año 2010

	Categorías Fuga_pos		
Categorías Sucursales	0	1	Frecuencias Sucursales
1	4421	33	4454
2	850	4	854
3	334	2	336
Mayor a 3	700	6	706
Total	6305	45	6350

Cuadro 5.90: Tabla de contingencia Q_ANIS*Fuga, Experimento 6 para el mes de Noviembre del año 2010

	Categorías Fuga_pos		
Categorías Q_ANIS	0	1	Frecuencias Q_ANIS
0	9	0	9
2	2132	18	2150
3	96	0	96
4	1569	7	1576
5	158	1	159
6	483	3	486
7	83	0	83
8	472	5	477
9	42	1	43
10	236	1	237
11	42	0	42
12	138	1	139
13	28	0	28
14	86	1	87
15	33	0	33
16	70	1	71
17	12	0	12
18	61	0	61
19	13	0	13
20	59	1	60
Mayor a 20	483	5	488
Total	6305	45	6350

Cuadro 5.91: Tabla de contingencia Filtro_Consumo*Fuga, Experimento 6 para el mes de Noviembre del año 2010

	Categorías Fuga_pos		
Categorías Filtro_Consumo	0	1	Frecuencias Filtro_Consumo
0	142	2	144
1	6163	43	6206
Total	6305	45	6350

Cuadro 5.92: Tabla de contingencia Filtro_Facturacion*Fuga, Experimento 6 para el mes de Noviembre del año 2010

	Categorías Fuga_pos		
Categorías Filtro_Facturacion	0	1	Frecuencias Filtro_Facturacion
0	356	1	357
1	5949	44	5993
Total	6305	45	6350

Cuadro 5.93: Tabla de contingencia Cat_Corp*Fuga, Experimento 6 para el mes de Noviembre del año 2010

	Categorías Fuga_pos		
Categorías Cat_Corp	0	1	Frecuencias Cat_Corp
1	3809	25	3834
2	601	5	606
3	390	3	393
4	1505	12	1517
Total	6305	45	6350

Cuadro 5.94: Tabla de contingencia Sucursales*Fuga, Experimento 6 para el mes de Diciembre del año 2010

	Categorías Fuga_pos		
Categorías Sucursales	0	1	Frecuencias Sucursales
1	4457	33	4490
2	862	3	865
3	334	2	336
Mayor a 3	706	9	715
Total	6359	47	6406

Cuadro 5.95: Tabla de contingencia Q_ANIS*Fuga, Experimento 6 para el mes de Diciembre del año 2010

	Categorías Fuga_pos		
Categorías Q_ANIS	0	1	Frecuencias Q_ANIS
0	9	0	9
2	2157	16	2173
3	99	1	100
4	1586	10	1596
5	155	1	156
6	487	2	489
7	82	2	84
8	475	1	476
9	45	0	45
10	235	0	235
11	43	1	44
12	138	2	140
13	26	0	26
14	84	1	85
15	35	0	35
16	72	0	72
17	12	1	13
18	60	2	62
19	12	0	12
20	60	0	60
Mayor a 20	487	7	494
Total	6359	47	6406

Cuadro 5.96: Tabla de contingencia Filtro_Consumo*Fuga, Experimento 6 para el mes de Diciembre del año 2010

	Categorías Fuga_pos		
Categorías Filtro_Consumo	0	1	Frecuencias Filtro_Consumo
0	145	4	149
1	6214	43	6257
Total	6359	47	6406

Cuadro 5.97: Tabla de contingencia Filtro_Facturacion*Fuga, Experimento 6 para el mes de Diciembre del año 2010

	Categorías Fuga_pos		
Categorías Filtro_Facturacion	0	1	Frecuencias Filtro_Facturacion
0	377	5	382
1	5982	42	6024
Total	6359	47	6406

Cuadro 5.98: Tabla de contingencia Cat_Corp*Fuga, Experimento 6 para el mes de Diciembre del año 2010

	Categorías Fuga_pos		
Categorías Cat_Corp	0	1	Frecuencias Cat_Corp
1	3863	21	3884
2	605	2	607
3	386	9	395
4	1505	15	1520
Total	6359	47	6406

Cuadro 5.99: Tabla de contingencia Sucursales*Fuga, Experimento 6 para el mes de Enero del año 2011

	Categorías Fuga_pos		
Categorías Sucursales	0	1	Frecuencias Sucursales
1	3631	33	3664
2	869	9	878
3	372	7	379
Mayor a 3	835	8	843
Total	5707	57	5764

Cuadro 5.100: Tabla de contingencia Q_ANIS*Fuga, Experimento 6 para el mes de Enero del año 2011

	Categorías Fuga_pos		
Categorías Q_ANIS	0	1	Frecuencias Q_ANIS
0	2	0	2
2	1899	21	1920
3	84	1	85
4	1205	9	1214
5	126	0	126
6	487	6	493
7	72	2	74
8	419	8	427
9	43	1	44
10	241	1	242
11	38	0	38
12	148	1	149
13	24	0	24
14	99	0	99
15	32	0	32
16	73	0	73
17	20	0	20
18	55	1	56
19	14	0	14
20	61	0	61
Mayor a 20	565	6	571
Total	5707	57	5764

Cuadro 5.101: Tabla de contingencia Filtro_Consumo*Fuga, Experimento 6 para el mes de Enero del año 2011

	Categorías Fuga_pos		
Categorías Filtro_Consumo	0	1	Frecuencias Filtro_Consumo
0	1283	13	1296
1	4424	44	4468
Total	5707	57	5764

Cuadro 5.102: Tabla de contingencia Filtro_Facturacion*Fuga, Experimento 6 para el mes de Enero del año 2011

	Categorías Fuga_pos		
Categorías Filtro_Facturacion	0	1	Frecuencias Filtro_Facturacion
0	1517	16	1533
1	4190	41	4231
Total	5707	57	5764

Cuadro 5.103: Tabla de contingencia Cat_Corp*Fuga, Experimento 6 para el mes de Enero del año 2011

	Categorías Fuga_pos		
Categorías Cat_Corp	0	1	Frecuencias Cat_Corp
1	3250	29	3279
2	573	5	578
3	390	3	393
4	1494	20	1514
Total	5707	57	5764

5.5.2. Estrategias para el tratamiento de valores perdidos

Cuadro 5.104: Tabla de valores perdidos y estrategias para el Experimento 6

Variables	ago10	sep10	oct10	nov10	dic10	Estrategia
PRIMERA INSTALACION	87	1	1	0	0	Reemplazo por Moda
Sucursales	2	1	1	0	0	Reemplazo por valor 1
Q ANIS	2	1	1	0	0	Reemplazo por valor 2
Q SUCURSALES CON BA	2	1	1	0	0	Reemplazo por valor 2
Q SUCURSALES SIN BA	2	1	1	0	0	Reemplazo por valor 2
CANT CPO	2	1	1	0	0	Reemplazo por valor 2
Recency	3000	2985	2943	2947	2953	Reemplazo por valor 7
Frecuencia cada cuanto reclama	2899	2894	2674	2686	2702	Reemplazo por valor 0
Mount prom rec	2899	2894	2674	2686	2702	Reemplazo por valor 0
Imagen	0	0	0	0	0	
Plan tipo 1	2599	2657	2658	2699	2682	Reemplazo por valor 0
Plan tipo 2	5540	5502	5414	5407	5380	Reemplazo por valor 0
Plan tipo 3	5160	5151	5093	5115	5088	Reemplazo por valor 0
Plan tipo 4	4799	4820	4769	4800	4779	Reemplazo por valor 0
Plan tipo 5	5838	5846	5786	5819	5791	Reemplazo por valor 0
Fact6	1359	1383	1372	1387	1525	Reemplazo por valor 0
Fact5	1317	1337	1319	1499	1277	Reemplazo por valor 0
Fact4	1269	1280	1431	1251	1329	Reemplazo por valor 0
Fact3	1209	1398	1182	1303	1337	Reemplazo por valor 0
Fact2	1329	1142	1232	1312	1367	Reemplazo por valor 0
Fact1	1074	1186	1240	1342	1340	Reemplazo por valor 0
Cons6	1359	1616	1599	1613	1777	Reemplazo por valor 0
Cons5	1317	1571	1548	1754	1578	Reemplazo por valor 0
Cons4	1269	1520	1692	1554	1582	Reemplazo por valor 0
Cons3	1209	1672	1492	1558	1584	Reemplazo por valor 0
Cons2	1329	1463	1496	1562	1566	Reemplazo por valor 0
Cons1	1074	1466	1498	1546	1564	Reemplazo por valor 0

Continúa en la página siguiente

Cuadro 5.104 – Continuación de la página anterior

Variables	ago10	sep10	oct10	nov10	dic10	Estrategia
Reclamos comerciales de facturación Monto promedio	5255	5308	5240	5235	4993	Reemplazo por valor 0
Reclamos comerciales generales	6353	6360	6292	6326	6299	Reemplazo por valor 0
Elimina sucursal Acumulado	6146	6172	6123	6158	6139	Reemplazo por valor 0
Retención NGN Acumulado	6307	6304	6222	6252	6233	Reemplazo por valor 0
Término de contrato Acumulado	5936	5973	6009	6029	5975	Reemplazo por valor 0
Total Solicitudes comerciales	4181	4188	4226	4240	4230	Reemplazo por valor 0
Solicitudes técnicas	4956	5025	4997	5038	5073	Reemplazo por valor 0
GIRO	2376	2376	2346	2341	2294	Reemplazo por valor S/I
COMUNA	254	253	253	252	242	Reemplazo por valor Santiago
Cat.Corp	853	853	847	892	910	
Tamaño	853	853	847	892	910	
Clase Valor	853	853	847	892	910	
Ciclo Vida	853	853	847	892	910	
Icp	853	853	903	892	966	Reemplazo por media
Retencion	853	853	847	892	910	
Competencia	2858	2865	2843	2854	2851	Reemplazo por valor 0
COMPANY MOBILE	4612	4620	4576	4607	4594	Reemplazo por valor 0
COMPANY PHONE	790	861	906	965	994	Reemplazo por valor 0
COMPANY CELL	4213	4228	4198	4226	4223	Reemplazo por valor 0
Extras	6246	6252	6229	6279	6316	Eliminación de variable
Fuga	0	6252	0	0	0	
Trim1 fact	1178	1186	1270	1193	1116	Reemplazo por valor 0
Trim2 fact	1010	965	1006	1159	1152	Reemplazo por valor 0
Promedio facturacion	920	891	938	964	915	Reemplazo por valor 0

Continúa en la página siguiente

Cuadro 5.104 – Continuación de la página anterior

Variables	ago10	sep10	oct10	nov10	dic10	Estrategia
Trim1 cons	1360	1389	1466	1405	1391	Reemplazo por valor 0
Trim2 cons	1248	1268	1288	1364	1379	Reemplazo por valor 0
Promedio consumo	1103	1124	1137	1168	1156	Reemplazo por valor 0
Tamaño numero	853	853	847	892	910	Reemplazo por valor 1
Ciclo vida	853	853	847	892	910	Reemplazo por valor 0
Valor R Num	853	853	847	892	910	Reemplazo por valor 1
Reten R	853	853	847	892	910	Reemplazo por valor 1
Cat corp RN	853	853	847	892	910	Reemplazo por valor 1
Fuga pos A	0	0	0	0	0	
Competencia 2	2858	2865	2843	2854	2997	Reemplazo por valor 0
Riesgo Alto	1896	1902	4156	4166	0	Reemplazo por valor 0
Missing	5533	5534	No existe	No existe	No existe	Eliminación de variable
FAC2 Facturacion	1550	1702	1744	1878	1974	Reemplazo por valor 0
FAC1 Consumo	1550	1702	1744	1878	1974	Reemplazo por valor 0
Edad	5	5	1	1	0	Reemplazo por valor 0
Grupo Fuga	No existe	No existe	0	No existe	No existe	No se le da tratamiento
predictionK3New	No existe	No existe	0	No existe	6316	No se le da tratamiento
Sub1	No existe	No existe	5829	No existe	6316	No se le da tratamiento
Sub K3New	No existe	No existe	5829	No existe	6316	No se le da tratamiento
Añadidos	No existe	No existe	6268	6346	No existe	No se le da tratamiento
Corporaciones	No existe	No existe	No existe	6313	6282	Reemplazo por valor 0
K3New	No existe	No existe	No existe	No existe	0	
ICP Reload	No existe	No existe	No existe	No existe	966	Reemplazo por valor 0
Sub2	No existe	No existe	No existe	No existe	6316	No se le da tratamiento

Cuadro 5.105: Tabla de valores perdidos y estrategias para el Experimento 7

Variables	dic10	ene11	feb11	mar11	Estrategia
RUTCOD	0	0	0	0	Ninguna
PRIMERA INSTALACION	175	250	184	188	Eliminación Listwise
Sucursales	0	250	184	188	Eliminación Listwise
Q ANIS	5	252	186	190	Eliminación Listwise
Q SUCURSALES CON BA	853	1004	952	960	Reemplazo por valor 0
Q SUCURSALES SIN BA	3926	4022	3955	3929	Reemplazo por valor 0
CANT CPO	0	250	184	188	Eliminación Listwise
Recencyreal	2258	2455	2609	2536	Reemplazo por valor 7
Freq rec	2019	2455	2307	2236	Reemplazo por valor 0
Mount rec	2019	2455	2307	2236	Reemplazo por valor 0
Imagen	0	0	0	0	Ninguna
Plan tipo 1	2404	2624	2538	2559	Reemplazo por valor 0
Plan tipo 2	4665	4870	4753	4710	Reemplazo por valor 0
Plan tipo 3	4427	4683	4583	4563	Reemplazo por valor 0
Plan tipo 4	4186	4387	4298	4270	Reemplazo por valor 0
Plan tipo 5	5045	5274	5218	5197	Reemplazo por valor 0
Rec fac	4378	4490	4552	4616	Reemplazo por valor 0
Reclamos comerciales generales	5493	5747	5707	5211	Reemplazo por valor 0
Elimina sucursal Acumulado	5356	5595	5558	5523	Reemplazo por valor 0
Continúa en la página siguiente					

Cuadro 5.105 – Continuación de la página anterior

Variables	dic10	ene11	feb11	mar11	Estrategia
Retención NGN Acumulado	5429	5683	5639	5613	Reemplazo por valor 0
Total Solicitudes comerciales	3657	3769	3806	3840	Reemplazo por valor 0
Solicitudes técnicas	4396	4557	4535	4510	Reemplazo por valor 0
GIRO2	2014	2126	2055	5684	Reemplazo por valor S/I
ComunaFinal	187	267	195	199	Reemplazo por valor Santiago
Icp	407	542	573	591	Reemplazo por valor 0
Competencia	2468	2532	2497	2490	Reemplazo por valor 0
COMPANY MOBILE	3968	4126	4085	4071	Reemplazo por valor 0
COMPANY PHONE	763	806	851	868	Reemplazo por valor 0
COMPANY CELL	3624	3768	3747	3732	Reemplazo por valor 0
Fuga anterior	5473	5764	5721	5692	Reemplazo por valor 0
Trim1 fact	621	675	693	533	Reemplazo por valor 0
Trim2 fact	512	619	594	618	Reemplazo por valor 0
Promedio fact	426	503	475	525	Reemplazo por valor 0
Trim1 consumo	808	854	859	847	Reemplazo por valor 0
Trim2 consumo	700	822	811	832	Reemplazo por valor 0
Promedio consumo	579	681	666	686	Reemplazo por valor 0
Tamaño numero	407	433	427	449	Eliminación Listwise
Ciclo vida	407	433	427	449	Eliminación Listwise

Continúa en la página siguiente

Cuadro 5.105 – Continuación de la página anterior

Variables	dic10	ene11	feb11	mar11	Estrategia
Valor R Num	407	433	427	449	Eliminación Listwise
Reten R	407	433	427	449	Eliminación Listwise
Cat corp RN	407	433	427	449	Eliminación Listwise
FAC1 Consumo	1229	0	0	0	Reemplazo por valor 0
FAC2 Facturacion	1230	0	0	0	Reemplazo por valor 0
Edad	0	0	1	0	Reemplazo por valor 0
Competencia 2	2468	2532	2496	2490	Reemplazo por valor 0
Riesgo Alto	4020	1356	1349	1358	Reemplazo por valor 0
Corporacion	5489	5741	5698	5669	Reemplazo por valor 0
Fuga pos	0	0	0	0	Ninguna
filter	0	0	0	0	Ninguna
Cluster New	459	652	655	616	Ninguna por tipo NMAR
Cluster real	459	652	655	616	Ninguna por tipo NMAR
TSC 9924	5457	5711	5657	No existe	Ninguna por tipo NMAR
Cluster real fugados	5457	5711	5657	No existe	Ninguna por tipo NMAR
TSC 5039	407	599	591	616	Ninguna por tipo NMAR
N1	No existe	5764	5721	5692	Ninguna por tipo NMAR
Missings	No existe	5331	5294	5243	Ninguna
VAR00002	No existe	851	No existe	No existe	Ninguna
ICP Reload	No existe	542	573	591	Reemplazo por valor 0

Continúa en la página siguiente

Cuadro 5.105 – Continuación de la página anterior

Variables	dic10	ene11	feb11	mar11	Estrategia
INGRESO CONTRATO SAP	No existe	250	No existe	No existe	Eliminación Listwise
Q ACCESOS BA	No existe	1004	No existe	No existe	Reemplazo por valor 0
Fact6	976	851	821	700	Reemplazo por valor 0
Fact5	726	796	823	826	Reemplazo por valor 0
Fact4	668	784	803	775	Reemplazo por valor 0
Fact3	657	761	743	763	Reemplazo por valor 0
Fact2	637	699	723	748	Reemplazo por valor 0
Fact1	577	766	701	700	Reemplazo por valor 0
Cons6	976	1054	1039	1041	Reemplazo por valor 0
Cons5	726	1020	1028	1001	Reemplazo por valor 0
Cons4	668	1001	985	977	Reemplazo por valor 0
Cons3	657	958	952	975	Reemplazo por valor 0
Cons2	637	929	948	981	Reemplazo por valor 0
Cons1	577	999	952	939	Reemplazo por valor 0
Término de contrato Acumulado	No existe	5469	5455	5467	Reemplazo por valor 0

5.5.3. Correlaciones entre variables de facturación y consumo

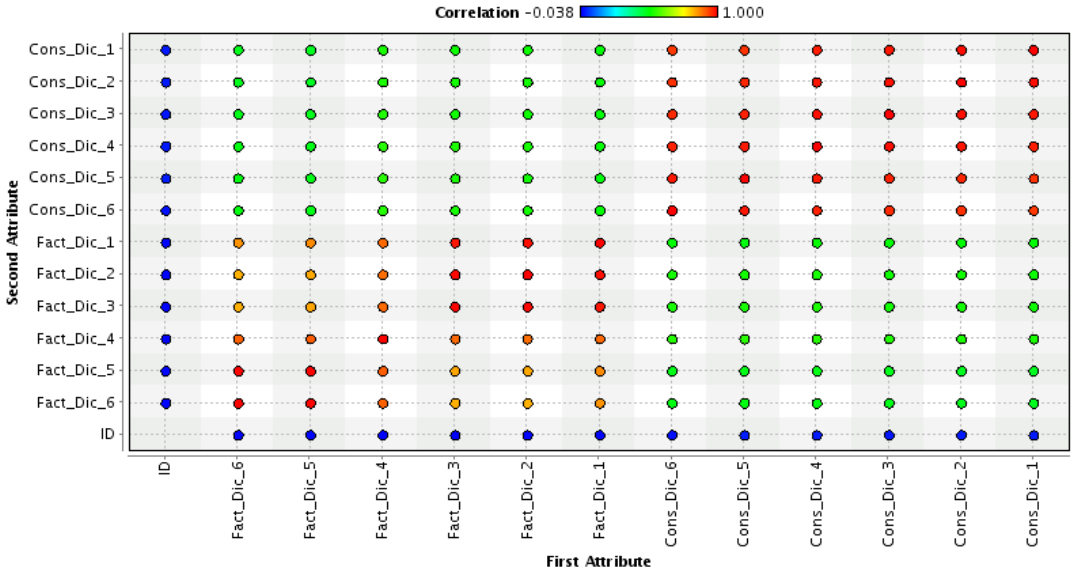

Figura 5.1: Gráfico de Correlaciones entre variables de Facturación y Consumo de la base Diciembre de 2010

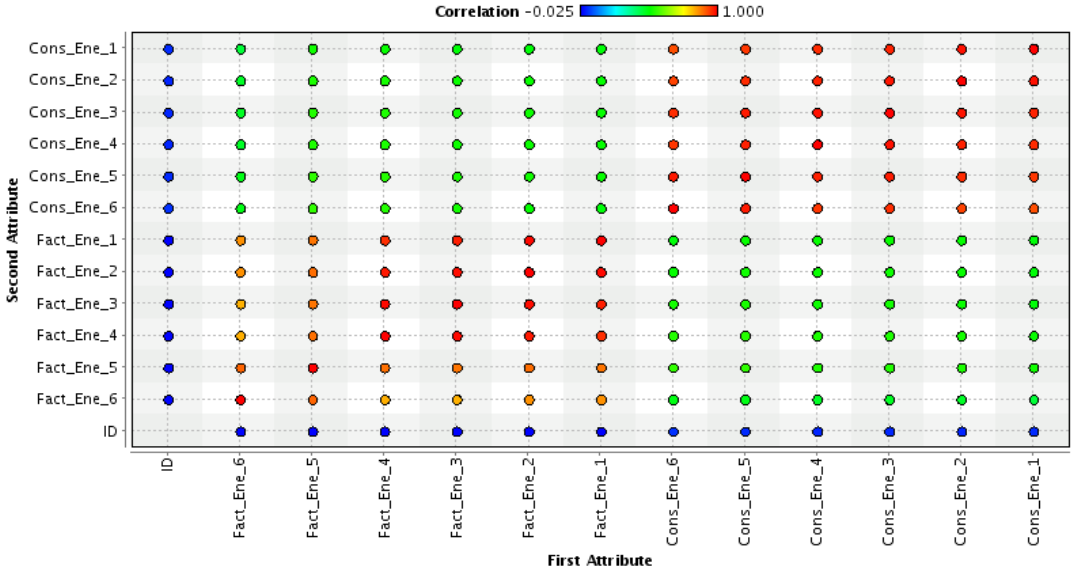

Figura 5.2: Gráfico de Correlaciones entre variables de Facturación y Consumo de la base Enero de 2011

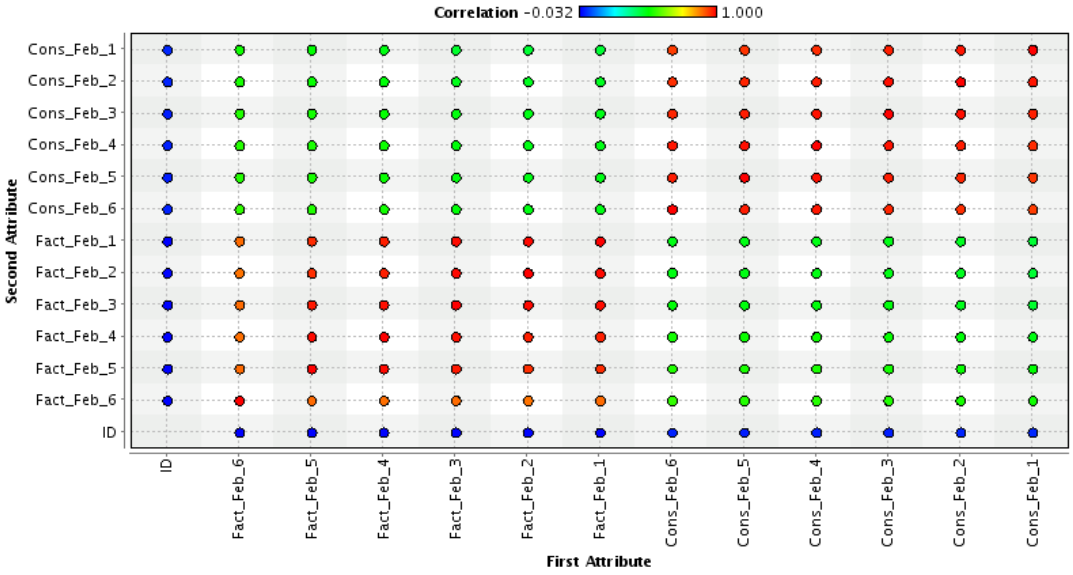

Figura 5.3: Gráfico de Correlaciones entre variables de Facturación y Consumo de la base Febrero de 2011

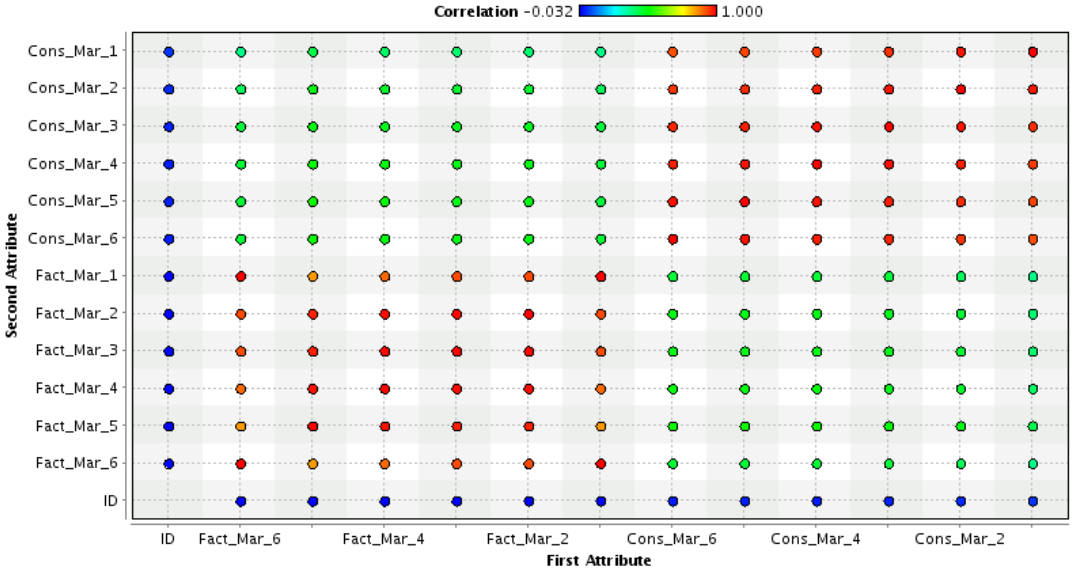

Figura 5.4: Gráfico de Correlaciones entre variables de Facturación y Consumo de la base Marzo de 2011

5.5.4. Análisis factorial

Cuadro 5.106: Tabla de comunalidades del análisis factorial en el Experimento 7

		Extracción de factores para mes			
		dic-10	ene-11	feb-11	mar-11
Variables	Inicial	Extracción	Extracción	Extracción	Extracción
Fact6	1	0.883	0.824	0.863	0.916
Fact5	1	0.894	0.887	0.967	0.924
Fact4	1	0.914	0.954	0.978	0.969
Fact3	1	0.926	0.961	0.990	0.986
Fact2	1	0.926	0.977	0.976	0.986
Fact1	1	0.942	0.960	0.978	0.916
Cons6	1	0.936	0.943	0.963	0.961
Cons5	1	0.962	0.972	0.973	0.974
Cons4	1	0.981	0.973	0.984	0.982
Cons3	1	0.970	0.986	0.984	0.983
Cons2	1	0.981	0.973	0.978	0.977
Cons1	1	0.957	0.958	0.958	0.946
Método de extracción: Análisis de Componentes principales.					

Cuadro 5.107: Tabla de rotación de factores del análisis factorial en el Experimento 7

	Meses							
	dic-10		ene-11		feb-11		mar-11	
	Componente		Componente		Componente		Componente	
Variables	1	2	1	2	1	2	1	2
Fact6	0.227	0.912	0.231	0.878	0.311	0.875	0.199	0.936
Fact5	0.223	0.919	0.300	0.893	0.275	0.944	0.272	0.922
Fact4	0.294	0.910	0.278	0.937	0.257	0.955	0.240	0.955
Fact3	0.277	0.922	0.261	0.945	0.230	0.968	0.230	0.966
Fact2	0.273	0.923	0.252	0.956	0.223	0.962	0.229	0.966
Fact1	0.241	0.940	0.248	0.948	0.221	0.964	0.199	0.936
Cons6	0.929	0.271	0.930	0.278	0.943	0.269	0.946	0.257
Cons5	0.944	0.265	0.946	0.279	0.952	0.260	0.953	0.257
Cons4	0.954	0.266	0.950	0.266	0.957	0.261	0.958	0.254
Cons3	0.951	0.255	0.956	0.268	0.959	0.256	0.961	0.242
Cons2	0.957	0.256	0.951	0.262	0.959	0.243	0.964	0.218
Cons1	0.944	0.256	0.947	0.249	0.951	0.230	0.959	0.161
”Método de extracción: Análisis de componentes principales.								
Método de rotación: Normalización Varimax con Kaiser.”								
La rotación ha convergido en 3 iteraciones.								

Cuadro 5.108: Tabla de Varianza explicada por los componentes del análisis factorial en el Experimento 7

Varianza total explicada mes Diciembre 2010				Varianza total explicada mes Enero 2011				Varianza total explicada mes Febrero 2011				Varianza total explicada mes Marzo 2011			
Componente	Autovalores	% de la varianza	% acumulado	Componente	Autovalores	% de la varianza	% acumulado	Componente	Autovalores	% de la varianza	% acumulado	Componente	Autovalores	% de la varianza	% acumulado
1	8.537	71.140	71.140	1.000	8.655	72.128	72.128	1.000	8.674	72.282	72.282	1.000	8.386	69.886	69.886
2	2.735	22.788	93.927	2.000	2.713	22.608	94.736	2.000	2.919	24.327	96.609	2.000	3.134	26.119	96.005
3	0.390	3.247	97.174	3.000	0.286	2.387	97.123	3.000	0.161	1.338	97.946	3.000	0.275	2.290	98.295
4	0.116	0.963	98.137	4.000	0.115	0.959	98.082	4.000	0.085	0.707	98.653	4.000	0.107	0.889	99.184
5	0.095	0.793	98.930	5.000	0.090	0.754	98.836	5.000	0.070	0.585	99.238	5.000	0.037	0.304	99.488
6	0.041	0.346	99.275	6.000	0.045	0.379	99.214	6.000	0.028	0.234	99.472	6.000	0.018	0.153	99.641
7	0.030	0.253	99.528	7.000	0.038	0.320	99.535	7.000	0.023	0.191	99.664	7.000	0.014	0.115	99.756
8	0.022	0.181	99.709	8.000	0.019	0.158	99.693	8.000	0.013	0.105	99.769	8.000	0.010	0.080	99.836
9	0.015	0.122	99.831	9.000	0.018	0.147	99.840	9.000	0.009	0.075	99.844	9.000	0.009	0.074	99.910
10	0.011	0.096	99.927	10.000	0.010	0.084	99.924	10.000	0.008	0.069	99.913	10.000	0.007	0.062	99.973
11	0.005	0.043	99.971	11.000	0.006	0.050	99.973	11.000	0.006	0.053	99.966	11.000	0.003	0.027	100.000
12	0.004	0.029	100.000	12.000	0.003	0.027	100.000	12.000	0.004	0.034	100.000	12.000	0.000	0.000	100.000
Método de extracción: Análisis de Componentes principales.															

Cuadro 5.109: Tabla de saturaciones del análisis factorial en el Experimento 7

Mes	Componente	Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
		Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
dic-10	1	8.537	71.140	71.140	5.774	48.113	48.113
	2	2.735	22.788	93.927	5.498	45.814	93.927
ene-11	1	8.655	72.128	72.128	5.791	48.257	48.257
	2	2.713	22.608	94.736	5.578	46.479	94.736
feb-11	1	8.674	72.282	72.282	5.845	48.708	48.708
	2	2.919	24.327	96.609	5.748	47.900	96.609
mar-11	1	8.386	69.886	69.886	5.811	48.421	48.421
	2	3.134	26.119	96.005	5.710	47.584	96.005

Cuadro 5.110: Matriz de correlaciones reproducidas para las variables de facturación y consumo para el mes de Marzo de 2011

	Fact6	Fact5	Fact4	Fact3	Fact2	Fact1	Cons6	Cons5	Cons4	Cons3	Cons2	Cons1
Fact6	0.916	0.917	0.942	0.950	0.950	0.916	0.430	0.430	0.429	0.419	0.397	0.342
Fact5	0.917	0.924	0.946	0.953	0.953	0.917	0.495	0.496	0.495	0.485	0.464	0.409
Fact4	0.942	0.946	0.969	0.978	0.978	0.942	0.473	0.474	0.472	0.462	0.440	0.384
Fact3	0.950	0.953	0.978	0.986	0.986	0.950	0.466	0.467	0.465	0.455	0.432	0.375
Fact2	0.950	0.953	0.978	0.986	0.986	0.950	0.465	0.466	0.464	0.454	0.432	0.375
Fact1	0.916	0.917	0.942	0.950	0.950	0.916	0.430	0.430	0.429	0.419	0.397	0.342
Cons6	0.430	0.495	0.473	0.466	0.465	0.430	0.961	0.967	0.972	0.972	0.968	0.949
Cons5	0.430	0.496	0.474	0.467	0.466	0.430	0.967	0.974	0.978	0.978	0.974	0.955
Cons4	0.429	0.495	0.472	0.465	0.464	0.429	0.972	0.978	0.982	0.983	0.979	0.960
Cons3	0.419	0.485	0.462	0.455	0.454	0.419	0.972	0.978	0.983	0.983	0.979	0.961
Cons2	0.397	0.464	0.440	0.432	0.432	0.397	0.968	0.974	0.979	0.979	0.977	0.960
Cons1	0.342	0.409	0.384	0.375	0.375	0.342	0.949	0.955	0.960	0.961	0.960	0.946

Cuadro 5.111: Matriz de correlaciones para las variables de facturación y consumo para el mes de Marzo de 2011

	Fact6	Fact5	Fact4	Fact3	Fact2	Fact1	Cons6	Cons5	Cons4	Cons3	Cons2	Cons1
Fact6	1.000	0.845	0.893	0.925	0.927	1.000	0.429	0.429	0.430	0.425	0.394	0.346
Fact5	0.845	1.000	0.985	0.969	0.969	0.845	0.493	0.497	0.490	0.476	0.467	0.411
Fact4	0.893	0.985	1.000	0.992	0.990	0.893	0.472	0.475	0.472	0.459	0.441	0.381
Fact3	0.925	0.969	0.992	1.000	0.992	0.925	0.465	0.465	0.464	0.454	0.434	0.375
Fact2	0.927	0.969	0.990	0.992	1.000	0.927	0.466	0.465	0.464	0.453	0.433	0.373
Fact1	1.000	0.845	0.893	0.925	0.927	1.000	0.429	0.429	0.430	0.425	0.394	0.346
Cons6	0.429	0.493	0.472	0.465	0.466	0.429	1.000	0.986	0.972	0.963	0.948	0.921
Cons5	0.429	0.497	0.475	0.465	0.465	0.429	0.986	1.000	0.983	0.971	0.956	0.932
Cons4	0.430	0.490	0.472	0.464	0.464	0.430	0.972	0.983	1.000	0.987	0.970	0.943
Cons3	0.425	0.476	0.459	0.454	0.453	0.425	0.963	0.971	0.987	1.000	0.981	0.954
Cons2	0.394	0.467	0.441	0.434	0.433	0.394	0.948	0.956	0.970	0.981	1.000	0.980
Cons1	0.346	0.411	0.381	0.375	0.373	0.346	0.921	0.932	0.943	0.954	0.980	1.000

Cuadro 5.112: Matriz de variaciones entre matrices de correlaciones común y reproducida para las variables de facturación y consumo para el mes de Marzo de 2011

	Fact6	Fact5	Fact4	Fact3	Fact2	Fact1	Cons6	Cons5	Cons4	Cons3	Cons2	Cons1
Fact6	8.42 %	8.58 %	5.39 %	2.67 %	2.55 %	8.42 %	0.22 %	0.23 %	0.42 %	1.59 %	0.54 %	1.34 %
Fact5	8.58 %	7.58 %	4.01 %	1.61 %	1.66 %	8.58 %	0.34 %	0.18 %	0.94 %	1.85 %	0.68 %	0.53 %
Fact4	5.39 %	4.01 %	3.05 %	1.46 %	1.22 %	5.39 %	0.13 %	0.08 %	0.00 %	0.75 %	0.25 %	0.69 %
Fact3	2.67 %	1.61 %	1.46 %	1.38 %	0.56 %	2.67 %	0.15 %	0.34 %	0.18 %	0.14 %	0.34 %	0.05 %
Fact2	2.55 %	1.66 %	1.22 %	0.56 %	1.36 %	2.55 %	0.11 %	0.28 %	0.10 %	0.28 %	0.44 %	0.45 %
Fact1	8.42 %	8.58 %	5.39 %	2.67 %	2.55 %	8.42 %	0.22 %	0.23 %	0.42 %	1.59 %	0.54 %	1.34 %
Cons6	0.22 %	0.34 %	0.13 %	0.15 %	0.11 %	0.22 %	3.89 %	1.91 %	0.00 %	0.87 %	2.09 %	3.01 %
Cons5	0.23 %	0.18 %	0.08 %	0.34 %	0.28 %	0.23 %	1.91 %	2.61 %	0.49 %	0.69 %	1.90 %	2.54 %
Cons4	0.42 %	0.94 %	0.00 %	0.18 %	0.10 %	0.42 %	0.00 %	0.49 %	1.76 %	0.43 %	0.88 %	1.83 %
Cons3	1.59 %	1.85 %	0.75 %	0.14 %	0.28 %	1.59 %	0.87 %	0.69 %	0.43 %	1.72 %	0.18 %	0.76 %
Cons2	0.54 %	0.68 %	0.25 %	0.34 %	0.44 %	0.54 %	2.09 %	1.90 %	0.88 %	0.18 %	2.33 %	2.03 %
Cons1	1.34 %	0.53 %	0.69 %	0.05 %	0.45 %	1.34 %	3.01 %	2.54 %	1.83 %	0.76 %	2.03 %	5.39 %

Cuadro 5.113: Escenarios considerados para la comparación de diferencias de las matrices de correlaciones observadas y reproducidas

Escenario	1	2	3	4	5	6	7	8	9	10	11
Porcentaje bajo el parámetro	100 %	100 %	94 %	94 %	94 %	90 %	89 %	86 %	75 %	56 %	0 %
Porcentaje sobre el parámetro	0	0	5.56 %	6.25 %	6.25 %	9.72 %	11.11 %	13.89 %	25.00 %	44.44 %	100.00 %
Valores que cumplen condición	0	0	8	9	9	14	16	20	36	64	144
Condición	>10 %	>9 %	>8 %	>7 %	¿6 %	>5 %	>4 %	>3 %	>2 %	>1 %	>0 %
Total valores	144.000	144.000	144.000	144.000	144.000	144.000	144.000	144.000	144.000	144.000	144.000

5.5.5. Histogramas variable competencia para distintos meses

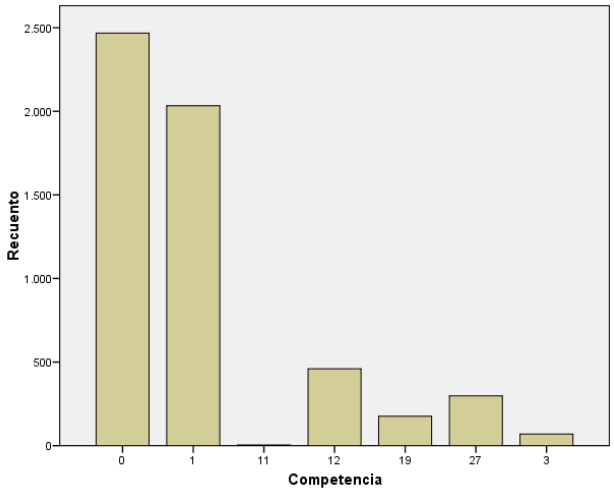

Figura 5.5: Histograma de variable Competencia de la base de Diciembre de 2010

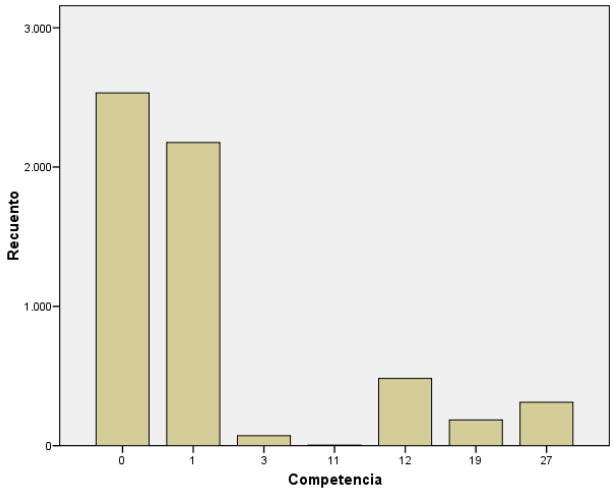

Figura 5.6: Histograma de variable Competencia de la base de Enero de 2011

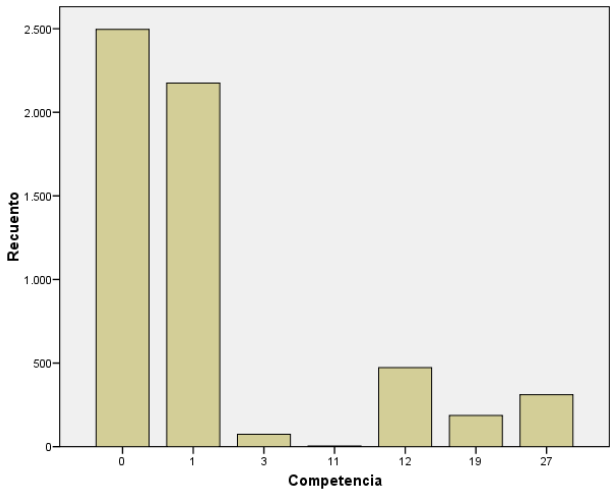

Figura 5.7: Histograma de variable Competencia de la base de Febrero de 2011

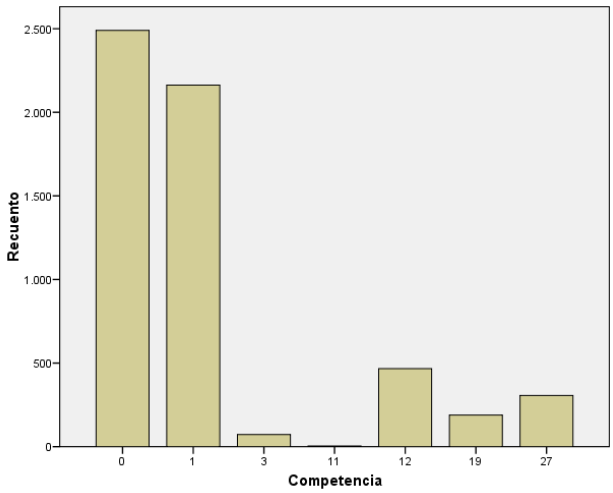

Figura 5.8: Histograma de variable Competencia de la base de Marzo de 2011

5.5.6. Experimento 6: Análisis de conglomerados de clientes

Cuadro 5.114: Resultados clusterización de muestra para el experimento 6

Cluster ID	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
0	687	54	280	646	98	180
1	313	40	129	69	680	491
2	No tiene	143	99	140	70	164
3	No tiene	420	114	55	152	106
4	No tiene	49	137	90	No tiene	59
5	No tiene	69	29	No tiene	No tiene	No tiene
6	No tiene	14	10	No tiene	No tiene	No tiene
7	No tiene	62	34	No tiene	No tiene	No tiene
8	No tiene	29	87	No tiene	No tiene	No tiene
9	No tiene	51	20	No tiene	No tiene	No tiene
10	No tiene	31	61	No tiene	No tiene	No tiene
11	No tiene	38	No tiene	No tiene	No tiene	No tiene

Cuadro 5.115: Resultados clusterización del experimento 6

	Categorías Fuga_pos Julio		Categorías Fuga_pos Agosto		Categorías Fuga_pos Octubre		Categorías Fuga_pos Noviembre		Categorías Fuga_pos Diciembre	
Conglomerado	0	1	0	1	0	1	0	1	0	1
Pasivos	1919	0	1729	0	2229	25	1999	0	2263	0
Fugados	222	90	214	61	320	25	487	30	766	35
Insatisfechos	1110	0	1081	4	1126	2	1151	0	1063	0
Reactivos	415	9	669	10	704	6	782	0	329	3
Total	3666	99	3693	75	4379	58	4419	30	4421	38

Cuadro 5.116: Centroides de variables continuas en el experimento 6 de la base Diciembre parte 1

	Sucursales		Mount_prom_rec		Plan_tipo_1		Plan_tipo_4		Plan_tipo_5	
Conglomerado	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica
1	1.045	0.208	0.563	0.850	0.575	0.515	0.095	0.294	0.000	0.000
2	2.055	0.676	1.411	1.189	1.099	0.912	0.644	0.768	0.000	0.000
3	1.386	0.652	0.929	1.050	0.217	0.468	0.114	0.363	0.563	0.576
4	1.050	0.218	1.539	0.889	0.555	0.499	0.069	0.253	0.000	0.000
Combinados	1.253	0.543	0.975	1.043	0.638	0.644	0.189	0.469	0.042	0.216

Cuadro 5.117: Centroides de variables continuas en el experimento 6 de la base Diciembre parte 2

	Total Solicitudes comerciales		Icp		FAC1_Consumo		FAC2_Facturacion		Edad	
Conglomerado	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica
1	0.576	0.958	0.983	0.048	-0.251	0.195	-0.011	0.033	1026.923	485.354
2	0.526	0.963	0.991	0.035	-0.042	0.292	-0.042	0.074	1150.388	439.461
3	0.502	0.880	0.593	0.501	-0.099	0.236	-0.012	0.055	597.666	460.671
4	0.479	0.771	0.985	0.049	-0.244	0.203	-0.018	0.041	987.984	494.651
Combinados	0.539	0.913	0.956	0.177	-0.201	0.236	-0.018	0.048	1007.858	494.661

Cuadro 5.118: Frecuencia variable Fuga en el experimento 6 de la base Diciembre

	Categorías Fuga_pos			
	0		1	
Conglomerado	Frecuencia	Porcentaje [%]	Frecuencia	Porcentaje [%]
1	2263	51.188	0	0.000
2	766	17.326	35	92.105
3	329	7.442	3	7.895
4	1063	24.044	0	0.000
Combinados	4421	100.000	38	100.000

Cuadro 5.119: Frecuencia variable Imagen en el experimento 6 de la base Diciembre

	Categorías Imagen			
	-1		1	
Conglomerado	Frecuencia	Porcentaje [%]	Frecuencia	Porcentaje [%]
1	0	0.000	2263	77.262
2	372	24.314	429	14.647
3	95	6.209	237	8.091
4	1063	69.477	0	0.000
Combinados	1530	100.000	2929	100.000

Cuadro 5.120: Detección de número de clústers con algoritmo W-EM

Cluster ID	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
0	687	54	280	646	98	180
1	313	40	129	69	680	491
2	No tiene	143	99	140	70	164
3	No tiene	420	114	55	152	106
4	No tiene	49	137	90	No tiene	59
5	No tiene	69	29	No tiene	No tiene	No tiene
6	No tiene	14	10	No tiene	No tiene	No tiene
7	No tiene	62	34	No tiene	No tiene	No tiene
8	No tiene	29	87	No tiene	No tiene	No tiene
9	No tiene	51	20	No tiene	No tiene	No tiene
10	No tiene	31	61	No tiene	No tiene	No tiene
11	No tiene	38	No tiene	No tiene	No tiene	No tiene

5.5.7. Experimento 6: Configuraciones y evaluación de modelos probados

Cuadro 5.121: Configuraciones de modelos para el experimento 6

Modelo	Tipo	C	Gamma	Parámetro adicional
SVM 1	Rbf	0	0	Ninguno
SVM 2	Sigmoid	0	0.9	Ninguno
RL 1	anova	10	2	degree: 2
RL 2	anova	10	25	degree: 2
RL 3	anova	10	20	degree: 2
RL 4	anova	0.5	20	degree: 2
SVM 4	rbf	0	1	Ninguno
SVM 6	rbf	0	0.5	Ninguno
SVM 10	rbf	10	0.95	Ninguno
SVM 16	sigmoid	0	0	Ninguno
SVM 19	sigmoid	0	0.8	Ninguno
SVM 20	sigmoid	5	0	Ninguno
SVM 23	rbf	3	0.37	Ninguno
RL 7	anova	5	4	degree: 2
SVM 25	rbf	3	0.37	Ninguno
SVM 26	rbf	3	0.37	Ninguno
SVM 27	rbf	3	0.37	Ninguno
RL 8	anova	1	5	degree: 2
RL 9	anova	1	2	degree: 2
RL 10	anova	10	2	degree: 5
RL 11	anova	1	1	degree: 2
RL 12	anova	10	2	degree: 4
RL 13	anova	100	2	degree: 4
SVM 28	rbf	10	2	Ninguno
SVM 29	rbf	0.5	0.00005	Ninguno
SVM 30	rbf	0.5	0.00005	Ninguno
SVM 35	Sigmoid	1	0	Ninguno
SVM 36	Sigmoid	0	0.1	Ninguno
SVM 37	rbf	30	0	Ninguno
Rl 14	anova	50	2	degree: 2
RL 15	anova	0.5	20	degree: 2
SVM 41	Rbf	5	0	Ninguno

Cuadro 5.122: Resultados de modelos prototipo para el experimento 6

Modelo	TN	FP	FN	TP	Grupo
SVM 1	507	181	1	2	Reactivos
SVM 2	568	120	1	2	Reactivos
RL 1	594	94	1	2	Reactivos
RL 2	630	58	1	2	Reactivos
RL 3	631	57	1	2	Reactivos
RL 4	633	55	1	2	Reactivos
RL 4	618	70	1	2	Reactivos
RL 4	631	57	1	2	Reactivos
SVM 16	1448	769	7	10	Pasivos
SVM 19	2184	33	15	2	Pasivos
SVM 20	1960	257	14	3	Pasivos
SVM 10	1383	834	6	10	Pasivos
SVM 1	1101	1116	6	11	Pasivos
SVM 6	1531	686	9	8	Pasivos
SVM 23	1897	338	10	7	Pasivos
RL 7	1762	455	9	8	Pasivos
SVM 25	1981	236	11	6	Pasivos
SVM 26	1947	270	11	6	Pasivos
SVM 27	1898	319	10	7	Pasivos
SVM 1	1055	58	3	3	Insatisfechos
RL 8	969	144	3	3	Insatisfechos
RL 9	1026	87	3	3	Insatisfechos
RL 1	1038	75	3	3	Insatisfechos
RL 11	1012	101	3	3	Insatisfechos
RL 1	1045	68	3	3	Insatisfechos
SVM 28	112	1001	0	6	Insatisfechos
SVM 1	398	715	1	5	Insatisfechos
SVM 1	668	445	2	4	Insatisfechos
RL 13	1049	64	3	3	Insatisfechos
SVM 29	1067	46	3	3	Insatisfechos
SVM 30	1062	51	3	3	Insatisfechos
SVM 1	248	87	1	3	Fugados
RL 14	261	74	1	3	Fugados
SVM 16	233	102	1	3	Fugados
SVM 35	225	110	1	3	Fugados
SVM 36	125	210	0	4	Fugados
SVM 37	255	80	1	3	Fugados
SVM 16	185	150	0	4	Fugados
RL 15	281	54	2	2	Fugados
SVM 1	210	125	0	4	Fugados
SVM 4	209	126	0	4	Fugados
SVM 41	113	222	0	4	Fugados

5.5.8. Experimento 7: Análisis de conglomerados de clientes

Cuadro 5.123: Distribución de conglomerados en base Febrero11 NF del Experimento 7

Conglomerado	Numero de instancias	% de combinados	% del total
1	2922	57.679	51.689
2	1950	38.492	34.495
3	194	3.829	3.432
Combinados	5066	100.000	89.616
Casos excluidos	587	Casos excluidos	10.384
Total	5653	Total	100.000

Cuadro 5.124: Descripción de centroides de conglomerados de la base base Febrero11 NF del Experimento 7

	Variables									
	Sucursales		Mount_rec		TotalSolicitudescomerciales		Icp		Edad	
Conglomerado	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica
1	1.808	1.920	0.618	0.936	0.515	0.918	0.962	0.155	1110.334	496.812
2	2.371	2.584	1.697	0.978	0.540	0.892	0.973	0.122	1144.522	487.584
3	22.082	18.618	3.986	3.270	1.345	2.733	0.835	1.265	1284.887	384.461
Combinados	2.801	5.726	1.162	1.367	0.557	1.050	0.961	0.285	1130.178	490.591

Cuadro 5.125: Descripción de centroides de conglomerados de la base base Febrero11 NF del Experimento 7

	Variables									
	Plan_tipo_1		Plan_tipo_4		Plan_tipo_5		FAC1_Consumo		FAC2_Facturacion	
Conglomerado	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica
1	0.692	0.916	0.338	0.867	0.071	0.288	-0.155	0.380	-0.027	0.103
2	0.895	1.183	0.492	1.130	0.099	0.343	0.016	0.523	-0.040	0.102
3	6.670	12.940	7.232	9.619	1.680	3.032	3.178	3.471	0.862	5.329
Combinados	0.999	2.948	0.662	2.485	0.144	0.733	0.039	1.023	0.002	1.059

Cuadro 5.126: Descripción de centroides de conglomerados de la base base Febrero11 NF del Experimento 7

	Categorías Variable Imagen			
	-1		1	
Conglomerado	Frecuencia	Porcentaje [%]	Frecuencia	Porcentaje [%]
1	0	0.000	2922	98.783
2	1950	92.505	0	0.000
3	158	7.495	36	1.217
Combinados	2108	100.000	2958	100.000

Cuadro 5.127: Distribución de conglomerados en base Febrero11 F del Experimento 7

Conglomerado	Numero de instancias	% de combinados	% del total
1	7	10.938	10.294
2	32	50.000	47.059
3	25	39.063	36.765
Combinados	64	100.000	94.118
Casos excluidos	4	Casos excluidos	5.882
Total	68	Total	100.000

Cuadro 5.128: Descripción de centroides de conglomerados de la base base Febrero11 F del Experimento 7

	Variables									
	Sucursales		Mount_rec		TotalSolicitudescomerciales		Icp		Edad	
Conglomerado	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica
1	14.429	20.663	3.238	2.780	1.900	2.835	0.763	0.350	886.429	496.613
2	1.781	1.313	0.469	0.706	0.411	0.590	0.979	0.076	946.938	497.711
3	3.840	4.180	1.521	0.838	0.779	1.026	0.979	0.038	1084.320	485.464
Combinados	3.969	7.922	1.183	1.422	0.718	1.242	0.955	0.140	993.984	490.723

Cuadro 5.129: Descripción de centroides de conglomerados de la base base Febrero11 F del Experimento 7

	Variables									
	Plan_tipo_1		Plan_tipo_4		Plan_tipo_5		FAC1_Consumo		FAC2_Facturacion	
Conglomerado	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica	Media	Desv. típica
1	2.000	4.472	1.143	2.193	0.714	0.951	1.541	3.439	1.011	1.929
2	0.563	0.878	0.250	0.622	0.000	0.000	-0.290	0.191	0.046	0.261
3	1.080	1.412	0.320	0.627	0.160	0.374	-0.057	0.355	0.011	0.102
Combinados	0.922	1.802	0.375	0.934	0.141	0.432	0.001	1.225	0.138	0.698

Cuadro 5.130: Descripción de centroides de conglomerados de la base base Febrero11 F del Experimento 7

	Categorías Variable Imagen			
	-1		1	
Conglomerado	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1	3	10.714	4	11.111
2	0	0.000	32	88.889
3	25	89.286	0	0.000
Combinados	28	100.000	36	100.000

5.5.9. Experimento 7: Resultados y configuraciones de modelos probados

Cuadro 5.131: Resultados de modelos prototipo para el experimento 7

Modelos	TN	FP	FN	TP	Observacion	*Accuracy*	*Recall* 1	Precisión 1	Medida F clase 1
SVM 1	4558	634	10	57		87.754 %	85.07 %	8.25 %	15.04 %
SVM 2	4554	638	12	55		87.640 %	82.09 %	7.94 %	14.47 %
SVM 3	4355	837	13	54		83.837 %	80.60 %	6.06 %	11.27 %
SVM 3	4776	416	8	59	Usa secuencia Dic-Feb	91.938 %	88.06 %	12.42 %	21.77 %
SVM 4	3704	1488	0	67		71.706 %	100.00 %	4.31 %	8.26 %
SVM 5	3713	1479	0	67		71.877 %	100.00 %	4.33 %	8.31 %
SVM 3	2170	92	0	67	No usa grupo pasivo en entrenamiento ni en validación	96.050 %	100.00 %	42.14 %	59.29 %
SVM 3	4838	126	1	66	No usa grupo grandes en entrenamiento	97.476 %	98.51 %	34.38 %	50.97 %
SVM 6	4949	15	0	67		99.702 %	100.00 %	81.71 %	89.93 %
SVM 7	4838	126	1	66		97.476 %	98.51 %	34.38 %	50.97 %
SVM 8	4717	247	2	65		95.051 %	97.01 %	20.83 %	34.30 %
SVM 9	4804	160	1	66		96.800 %	98.51 %	29.20 %	45.05 %
SVM 10	4948	16	2	65		99.642 %	97.01 %	80.25 %	87.84 %
Naive Bayes	4666	298	1	66		94.057 %	98.51 %	18.13 %	30.63 %
LADTree	4933	31	0	67		99.384 %	100.00 %	68.37 %	81.21 %
SVM 3	4979	172	0	54	Usa secuencia Dic-Ene y no usa grupo pasivo en entrenamiento	96.695 %	100.00 %	23.89 %	38.57 %
SVM 3	4162	989	4	50	Usa secuencia Dic-Ene y no usa grupo reactivo en entrenamiento	80.922 %	92.59 %	4.81 %	9.15 %
SVM 3	4333	818	4	50	Usa secuencia Dic-Ene y no usa grupo grandes en entrenamiento	84.207 %	92.59 %	5.76 %	10.85 %
SVM 3	5124	68	51	16	Usa secuencia Ene-Feb y no usa grupo grandes en entrenamiento	97.737 %	23.88 %	19.05 %	21.19 %
SVM 3	5047	145	47	20	Usa secuencia Ene-Feb y no usa grupo pasivo en entrenamiento	96.349 %	29.85 %	12.12 %	17.24 %
SVM 3	5054	138	51	26	Usa secuencia Ene-Feb y no usa grupo reactivo en entrenamiento	96.413 %	33.77 %	15.85 %	21.58 %
SVM 11	2262	2930	10	57	Usa secuencia Ene-Feb y no usa grupo pasivo en entrenamiento	44.096 %	85.07 %	1.91 %	3.73 %

Cuadro 5.132: Configuraciones de modelos para el experimento 7

Modelo	Tipo	c	gamma	w class 0	w class 1
SVM 1	rbf	0.3	0	0.6	1
SVM 2	rbf	0.5	0	0.6	1
SVM 3	rbf	0	0	1	1
SVM 4	rbf	0.5	1	1	1
SVM 5	rbf	0.5	0.5	1	1
SVM 6	Sigmoid	0	0	1	1
SVM 7	Sigmoid	0	1	1	1
SVM 8	Sigmoid	0	0.5	1	1
SVM 9	Sigmoid	0	1.5	1	1
SVM 10	rbf	10	2	1	1
SVM 11	rbf	0	1	1	1

Cuadro 5.133: Métricas de modelos finales contemplados para el experimento 7

Modelos	*Recall* clase 0	*Recall* clase 1	Precisión clase 0	Precisión clase 1	*Accuracy*	Medida F Total	Medida F clase 1
SVM 3	95.712 %	100.000 %	100.00 %	40.854 %	95.835 %	81.906 %	58.009 %
LADTree	99.376 %	100.000 %	100.00 %	68.367 %	99.384 %	91.282 %	81.212 %
SVM 6	99.698 %	100.000 %	100.00 %	81.707 %	99.702 %	95.139 %	89.933 %
SVM 10	99.678 %	97.015 %	99.960 %	80.247 %	99.642 %	94.044 %	87.838 %

ROC exp7.png

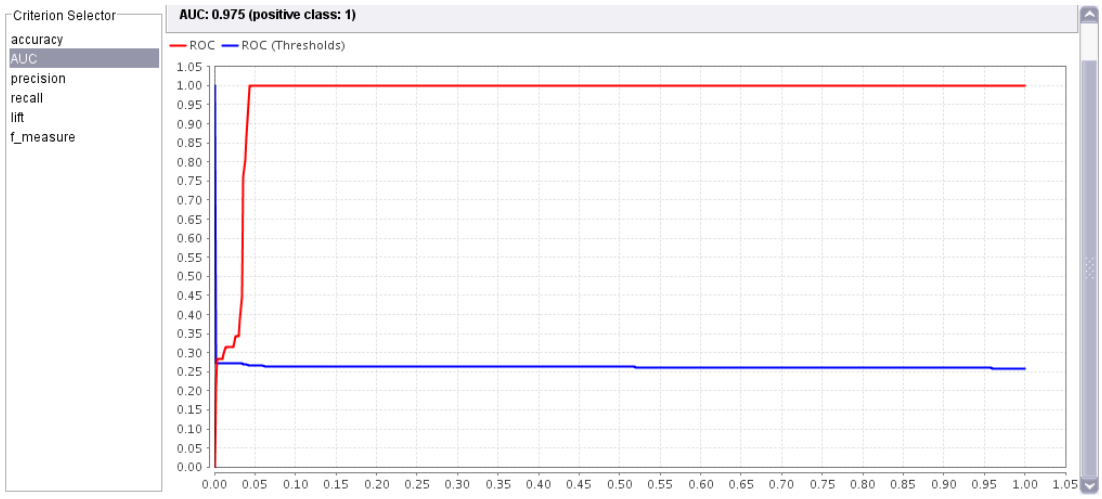

Figura 5.9: Gráfico de las curvas ROC en el experimento 7 del modelo SVM 3

5.5.10. Experimento 7: Criterios de corte

Cuadro 5.134: Criterios de corte en el Experimento 7

Corte	Valor	TN	FN	FP	TP	P	*Accuracy* [%]	Precisión clase 1 [%]	*Recall* clase 1 [%]	Medida F clase 1 [%]
1	0.26495	158	3	2233	16	2249	7.220	0.711	84.211	1.411
2	0.26665	2116	15	275	4	279	87.967	1.434	21.053	2.685
3	0.26765	2116	15	275	4	279	87.967	1.434	21.053	2.685
4	0.26818	2232	16	159	3	162	92.739	1.852	15.789	3.315
5	0.26865	2302	16	89	3	92	95.643	3.261	15.789	5.405
6	0.26875	2302	16	89	3	92	95.643	3.261	15.789	5.405
7	0.26905	2370	19	21	0	21	98.340	0	0	-
8	0.26915	2370	19	21	0	21	98.340	0	0	-

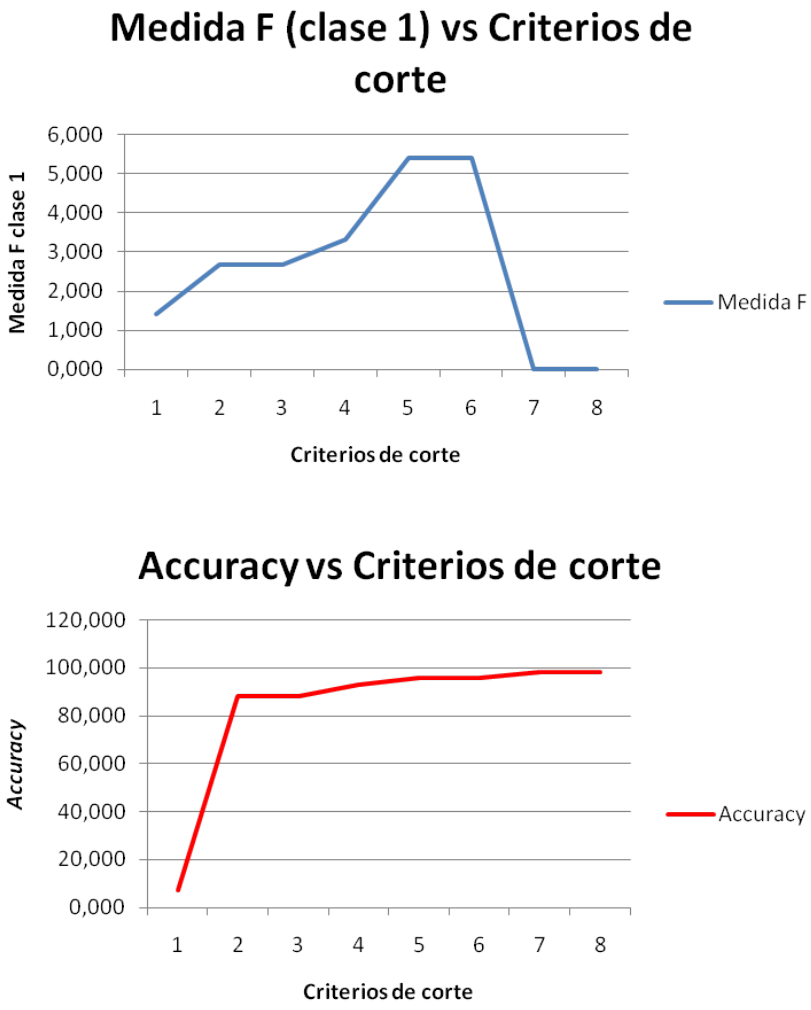

Figura 5.10: Gráficos de validación de criterios de corte

5.6. Anexo 7: Histogramas de variables eliminadas

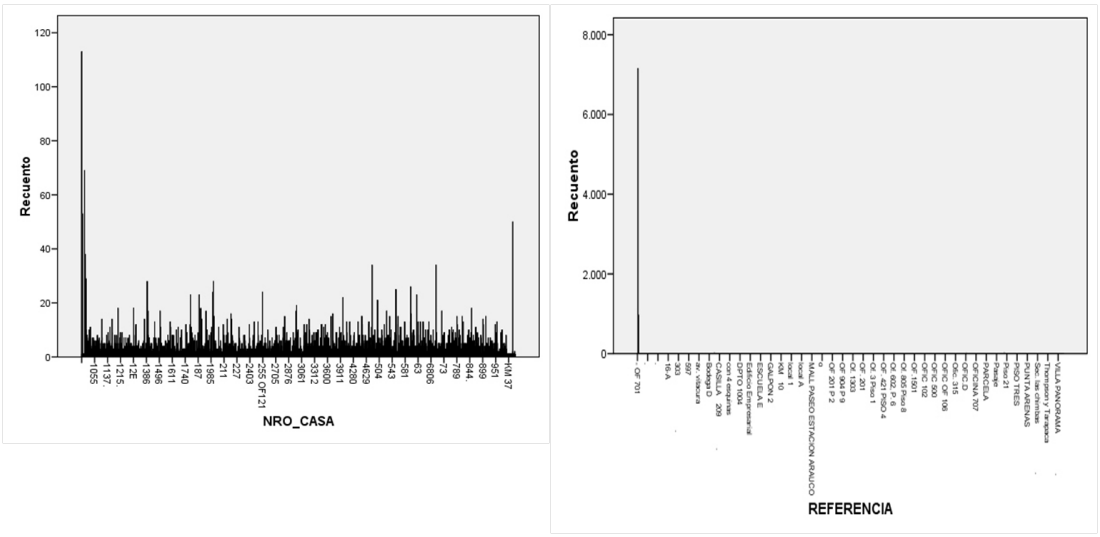

Figura 5.11: Histogramas de variables eliminadas: NRO Casa y Referencia

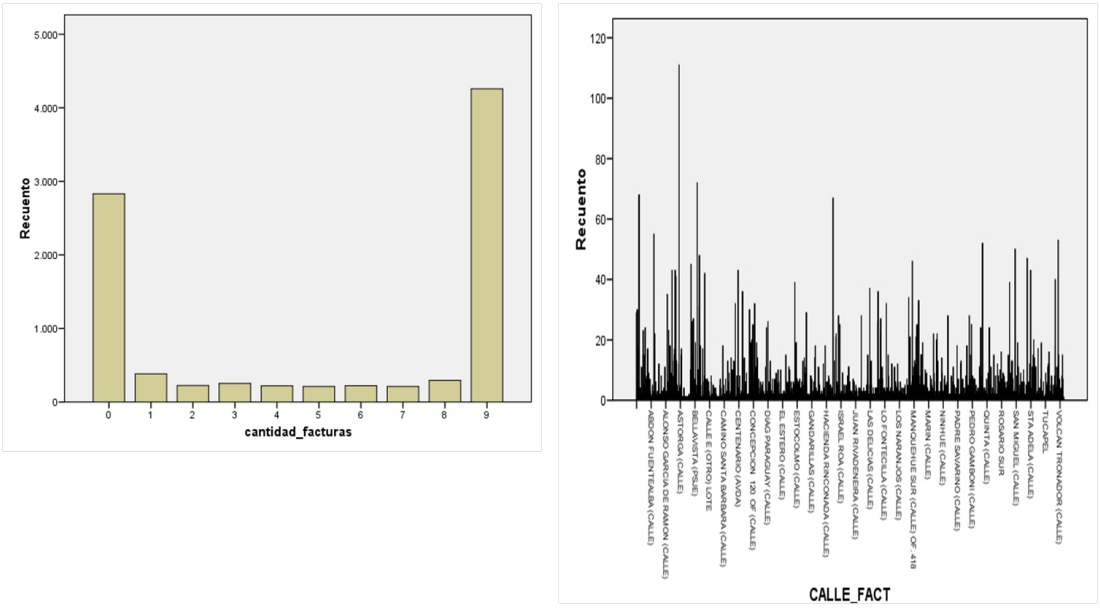

Figura 5.12: Histogramas de variables eliminadas: cantidad facturas y Calle Fact

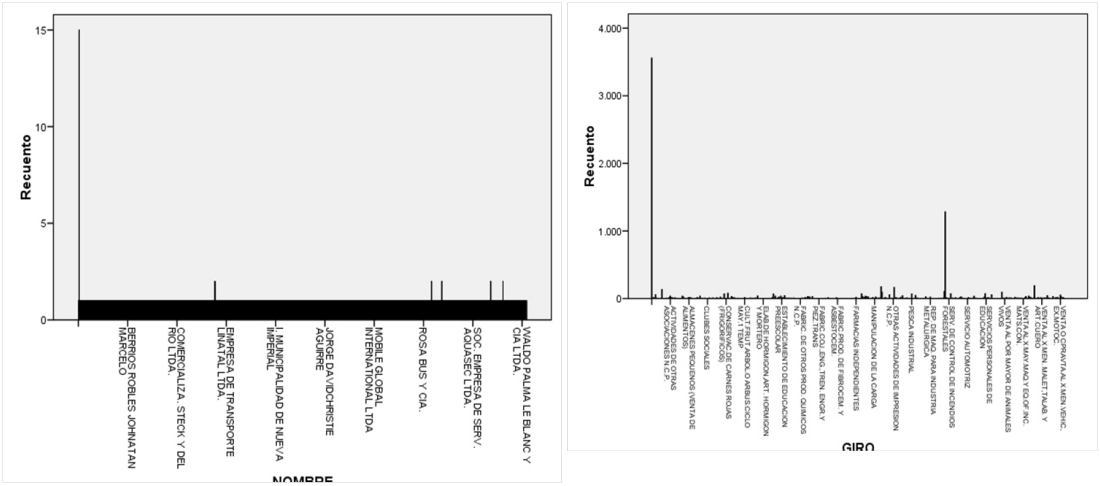

Figura 5.13: Histogramas de variables eliminadas: Nombre y Giro

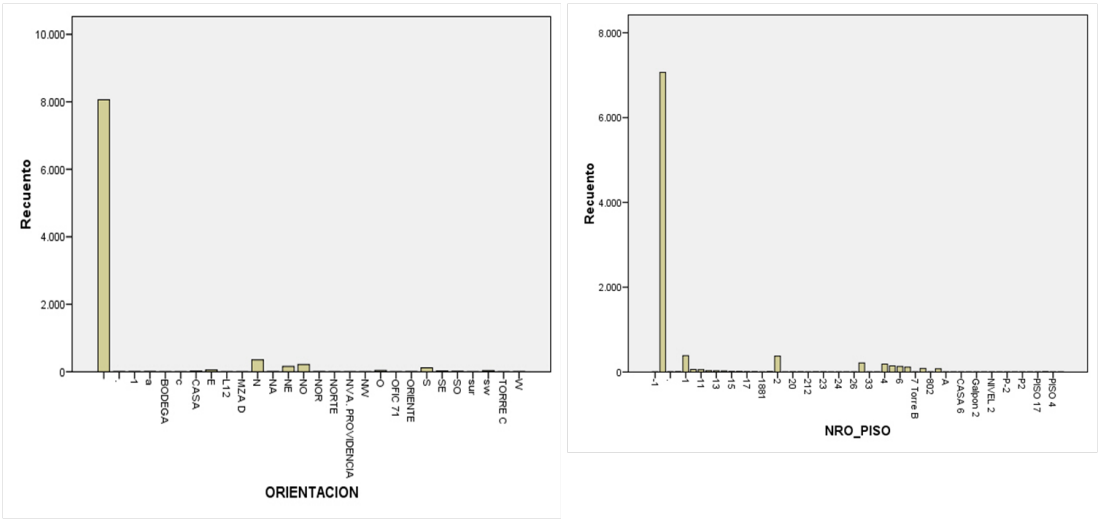

Figura 5.14: Histogramas de variables eliminadas: Orientación y NRO Piso

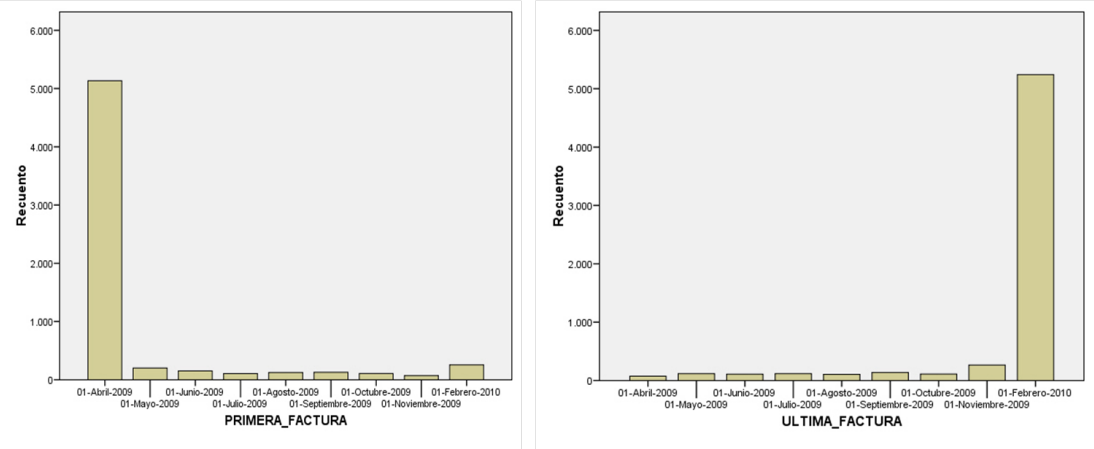

Figura 5.15: Histogramas de variables eliminadas: Primera factura y Última factura

5.7. Anexo 8: Tablas y figuras misceláneas

Cuadro 5.135: Matriz de riesgo e ingresos

		Probabilidad de abandono		
		Alto	Medio	Bajo
Life-time Value	Alto	prioridad A	prioridad B	prioridad C
	Medio	prioridad B	prioridad B	prioridad C
	Bajo	prioridad C	prioridad C	prioridad C

```
Dim var(78, 1) As String                                          \*Definición de vector que contiene las variables y
Dim w As String                                                   aquellos que se usan como auxiliares.
Dim w1 As String
Dim w2 As Integer

\*Se rellena el vector vars con los nombres de las variables

Sheets(``Hoja3'').Select                                          \*Selección de la hoja  donde se mostrarán los
w1 = ``B''                                                        resultados.
For j = 15 To 78                                                  \*Aplicación de ciclo sobre la tabla dinámica.
  Sheets(``Hoja4'').Select                                        \*Selección de la hoja de la tabla dinámica.
  ' ActiveSheet.PivotTables(``Tabla dinámica1'').PivotSelect ``Valores[All]'', \_
  '     xlLabelOnly + xlFirstRow, True
    With ActiveSheet.PivotTables(``Tabla dinámica1'').PivotFields( \_
      var(j, 1))
      .Orientation = xlPageField
      .Position = 1                                               \*Limpieza de los filtros de la tabla dinámica y
    End With                                                      aplicación de nuevo filtro en base a la variable (j,1)
    ActiveSheet.PivotTables(``Tabla dinámica1'').PivotFields(var(j, 1)) \_   que sólo cuente los valores vacíos.
      .ClearAllFilters
    ActiveSheet.PivotTables(``Tabla dinámica1'').PivotFields(var(j, 1)) \_
      .CurrentPage = ``(blank)''

    Range(``B4:b81'').Select                                      \*Copiado de los resultados originados por la tabla
    Selection.Copy                                                dinámica.

Sheets(``Hoja2'').Select                                          \*Selección de la hoja auxiliar.
     w2 = (j - 13)
    w = LTrim(Str(w2))
     Cells(3, 5) = w1
    w1 = w1 + w
    Cells(2, 5) = w1                                              \*Pegado del resultado, en forma traspuesta y sin
     Sheets(``Hoja3'').Select                                     espacios extras.
    Range(w1).Select
    Selection.PasteSpecial Paste:=xlPasteAll, Operation:=xlNone, SkipBlanks:=\_
      False, Transpose:=True

Next j                                                            \*Ampliación de índice del ciclo.

End Sub                                                           \*Fin de la macros.
```

Figura 5.16: Código visual basic de la tabla de presencia

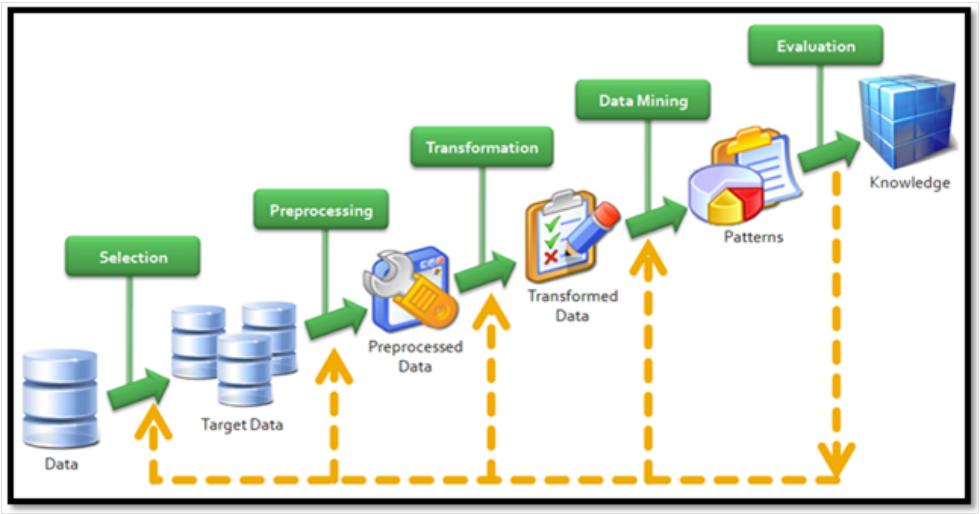

Figura 5.17: Procedimiento KDD generalizado

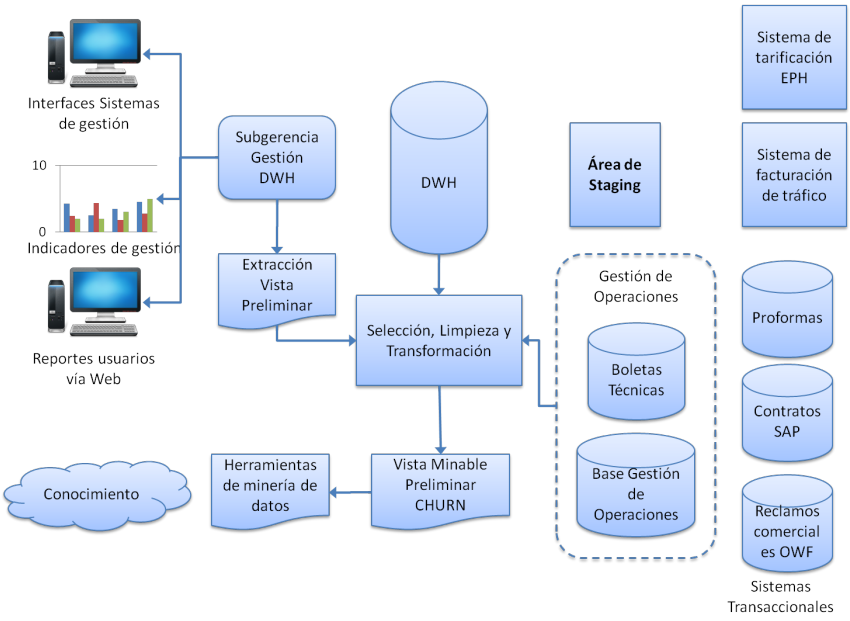

Figura 5.18: Procedimiento KDD aplicado en experiencia anterior

Bibliografía

[1] Broadband stimulus kickstarts ict industry growth del sitio: 2010 ict market review & forecast extraído el 29 de octubre del 2010 fuente: http://www.tiaonline.org/market_intelligence/mrf/webinar/tia_broadband_webinar_20100319_final.pdf.

[2] The difference between segmentation and clustering. extraído el 16 de julio del 2010. fuente: http://zyxo.wordpress.com/2010/07/17/ the-difference-between-segmentation-and-clustering/.

[3] Estadísticas de inversión y empleo, sitio: Subtel (subsecretaría de telecomunicaciones). extraído el 20 de octubre del 2010. fuente: http://www.subtel.cl/prontus_subtel/site/artic/20100608/asocfile/20100608122246/1_series_inversion_y_empleo_dic09_191010_v1.xls.

[4] Estadísticas de reclamos recibidos por el dpto. gestión de reclamos de la subtel. sitio: Subtel. extraído el 20 de octubre del 2010. fuente: http://www.subtel.cl/prontus_oirs/site/artic/20100503/asocfile/20100503154918/ estadisticas_reclamos_04_10_10.pdf.

[5] Ibm compra spss y dispara un 40 por ciento sus títulos en wall street. sitio: Expansión.com. extraído el 5 de julio de 2010. fuente: http://www.expansion.com/2009/07/28/empresas/1248784790.html.

[6] Ict is key driver for global economy. sitio 2010 ict market review & forecast. extraído el 29 de octubre del 2010. fuente: http://www.tiaonline.org/market_intelligence/mrf/ webinar/tia_ict_international_report_webinar_20100428_final.pdf.

[7] Informe anual 2008 subtel, extraído el 26 de octubre del 2010. fuente: http://www.subtel.cl/prontus_subtel/site/artic/20080509/asocfile/20080509130640/ informe_anual_2008_020909_v1.pdf.

[8] ¿qué es la telefonía ip?. sitio: Telefoníaip.uchile.cl. extraído en la fecha: 18 de abril de 2011. fuente: http://www.telefoniaip.uchile.cl/ capacitacion_telefonia.htm.

[9] Ranking de reclamos segundo semestre. sitio: Subtel.cl. extraído el 22 de octubre. fuente: http://www.subtel.cl/prontus_oirs/site/artic/20100503/asocfile/20100503154950/ ranking_reclamos_2do_sem_09.pdf.

[10] Real academia española. extraído el 16 de julio de 2010 desde: www.rae.es.

[11] Telefonía tradicional. sitio: Empresas.entel.cl. extraído el 20 de marzo del 2011. fuente: http://empresas.entel.cl/portalempresas/appmanager/entel/ entel;jsessionid=bfd3tb1kxhxv1zvjkcv7bgglq5hp5dqz65xpmldfrsysdn1yfwrb!2119535786?_nfpb=true&_pagelabel=b4037481263567693469.

[12] Informe modelo de negocio pyme. Informe interno a la compañía, Diciembre 2005.

[13] Rapid-i: Report the future. sitio: Rapid-i.com. extraído el 14 de abril. fuente: http://rapid-i.com/content/view/189/198/. Online, 03 2010.

[14] Sas, the power to know: History. sitio: sas.com. extraído en julio del 2010. fuente: http://www.sas.com/company/about/history.html. Online, 07 2010.

[15] Sas, the power to know: Sas enterprise miner semma. sitio: sas.com. extraído en julio del 2010. fuente: http://www.sas.com/offices/europe/uk/ technologies/analytics/datamining/miner/semma.html. Online, 07 2010.

[16] Rumores. revista trend management., Agosto-Septiembre 2010.

[17] J. Álvarez Menéndez. Minería de datos: Aplicaciones en el sector de las telecomunicaciones. Technical report, Universidad Carlos III, 2008.

[18] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 1994.

[19] S. S. Bharadwaj. Module 1: Segmentation,targeting and positioning. In *Curso : Management Science II*, 2010.

[20] M. Bosch. Clase: Medición y escalas. In *Curso IN58B: Ingeniería en Marketing*, 2010.

[21] J. Burez and D. Van den Poel. Handling class imbalance in customer churn prediction. *Expert Syst. Appl.*, 36:4626–4636, April 2009.

[22] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2:121–167, June 1998.

[23] K. P. R. Chandar M Purna, Laha Arijit. Modeling churn behavior of bank customers using predictive data mining techniques. Technical report, Institute for Development and Research in Banking Technology (IDRBT), 2006.

[24] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[25] K. Coussement and D. Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Syst. Appl.*, 34:313–327, January 2008.

[26] C. M. Cuadras. Distancias estadísticas. Technical report, Universidad de Barcelona, 1989.

[27] Data-Miners. Chapter 6: Decision trees. In -, -.

[28] P. M. Dixon. Bootstrap resampling. In *Encyclopedia of Environmetrics*, 2002.

[29] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *J. of Biomedical Informatics*, 35:352–359, October 2002.

[30] A. Farhangfar, L. Kurgan, and J. Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recogn.*, 41:3692–3705, December 2008.

[31] L. Fausett. *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1994.

[32] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, 2004.

[33] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[34] T. B. Fomby. K-nearest neighbors algorithm: Prediction and classification. Technical report, Southern Methodist University, Dallas, 2008.

[35] Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 124–133, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[36] M. Galván and F. Medina. Imputación de datos: teoría y práctica. Technical report, CEPAL Naciones Unidas, 2007.

[37] P. Gonzalez Juré. Unidad 2: Preparación y evaluación de proyectos de tecnologías de información y comunicaciones. In *Curso: CC60R -Preparación y Evaluación de Proyectos Informáticos 2010,*, 2010.

[38] D. N. Gujarati. *Econometría*. McGraw Hill, 2003.

[39] D. Hand. Data mining: Statistics and more. *The American Statistician*, 52:112–118, 1998.

[40] G. Hernández Oliva. Capítulo 4:interpolación y aproximación funciones. In *Curso MA-33A: Cálculo Numérico*, Primavera 2006.

[41] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall. Multiclass alternating decision trees. In *Proceedings of the 13th European Conference on Machine Learning*, ECML '02, pages 161–172, London, UK, 2002. Springer-Verlag.

[42] R. C. Holte, L. E. Acker, and B. W. Porter. Concept learning and the problem of small disjuncts. In *Proceedings of the 11th international joint conference on Artificial intelligence - Volume 1*, pages 813–818, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[43] B. Q. Huang, T. M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, and T. Rashid. A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Syst. Appl.*, 37:3657–3665, May 2010.

[44] G. Huerta. Balanceo de datos para la clasificación de imágenes de galaxia. Technical report, Universidad Politécnica de Puebla, 2010.

[45] C. Hurtado. Arboles de decisión. In *Inteligencia Artificial*, 2009.

[46] R. Irizarry. Statistical learning: Algorithmic and nonparametric approaches. In *Chapter 11: Classification Algorithms and Regression Trees*, 2006.

[47] M. Jaeger and T. D. Nielsen. K nearest neighbor classifier. In *Curso: Data Warehousing and Data Mining*, 2008.

[48] J. Jantzen. Introduction to perceptron networks. Technical report, Technical University of Denmark, 1998.

[49] Y. Jiangsheng. Method of k-nearest neighbors. Technical report, Institute of Computational Linguistics, Peking University, 2002.

[50] T. O. Jones and W. Sasser. Why satisfied customers defect. *Harvard business review*, 73(6):88 – 99, 1995.

[51] J.-J. Jonker, N. Piersma, and D. V. d. Poel. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Working papers of faculty of economics and business administration, ghent university, belgium, Ghent University, Faculty of Economics and Business Administration, 2003.

[52] N. K. Kasabov. *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*. MIT Press, Cambridge, MA, USA, 1st edition, 1996.

[53] Y. Kim and W. N. Street. An intelligent system for customer targeting: a data mining approach. *Decis. Support Syst.*, 37:215–228, May 2004.

[54] B. Krse and P. van der Smagt. *An introduction to Neural Networks*. None, 1996.

[55] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[56] L. Lam and Y. S. Ching. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICSPART A: SYSTEMS AND HUMANS,*, 27:553–568, 1997.

[57] M. A. P. M. Lejeune. Measuring the impact of data mining on churn management. *Internet Research*, 11(5):375–387, 2001.

[58] J. Lévy Mangin and J. Varela Mallou. *Análisis multivariable para las ciencias sociales*. Prentice Hall, 2003.

[59] G. L'Huillier. Auxiliar 6:arboles de decision. In *Curso IN643: Introduccion a la Mineria de Datos*, 2010.

[60] G. L'Huillier. Auxiliar 9: Support vector machines. In *Curso IN643: Minería de Datos*, 2010.

[61] T. Y. Lin. Database mining on derived attributes. In *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing*, TSCTC '02, pages 14–32, London, UK, UK, 2002. Springer-Verlag.

[62] R. J. A. Little and D. B. Rubin. *Statistical Analysis With Missing Data*. Probability and Statistics. Wiley, New Jersey, second edition, 2002.

[63] N. Littlestone and M. Warmuth. The weighted majority algorithm. Technical report, Baskin center for Computer engineering & Information Sciences; Univesidad de California, 1992.

[64] H. Liu and H. Motoda. *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007.

[65] R. Liu. The spss twostep cluster. Technical report, Department of Mathematics University of North Texas, 2003.

[66] D. Lowd and P. Domingos. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 529–536, New York, NY, USA, 2005. ACM.

[67] J. Lu. Predicting customer churn in the telecommunications industry an application of survival analysis modeling using sas. Technical report, Sprint Communications Company, 2001.

[68] C. M. Luque. Clasificadores bayesianos. el algoritmo nave bayes, 2003.

[69] S. Maldonado. Utilización de support vector machines no lineal y selección de atrobutos para credit scoring. Master's thesis, Universidad de Chile, 2007.

[70] H. Maletta. Metodología de análisis de panel de variables categóricas. Technical report, Instituto de Investigación en Ciencias Sociales, 2002.

[71] N. K. Malhotra. *Investigacion De Mercados*. Librisite, 2008.

[72] E. Manso and J. J. Dolado. *Técnicas cuantitativas para la gestión en la Ingeniería del software*. Netbiblio, 2007.

[73] J. M. Marin. Tema 4: Análisis factorial. In *Curso Análisis multivariable*, 2010.

[74] J. Mezo. Variables temporales: series temporales y números índice. In *Curso de Estadística Aplicada a las Ciencias Sociales*, 2010.

[75] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, June 1985.

[76] T. M. Mitchell. Generative and discriminative classifiers: Naive bayes and logistic regression. In *Machine Learning*, 2010.

[77] S. Molina. Aplicación de técnicas de minería de datos para predicción del churn de clientes en una empresa de telecomunicaciones. Master's thesis, Escuel de Ingeniería de la Pontificia Universidad Católica de Chile, 2009.

[78] R. Montoya. Clase 10: Segmentation, targeting, & positioning (stp). In *Curso IN58a: Introducción al Marketing*, 2009.

[79] G. J. Myatt. *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. Wiley-Interscience, 2006.

[80] R. Nisbet, J. Elder, and G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, 2009.

[81] J. R. Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1 edition, January 1993.

[82] J. A. O. Ramírez. *Patrones de comportamiento temporal en modelos semicualitativos con restricciones*. PhD thesis, Universidad de Sevilla, 2000.

[83] J. Ramo and A. Franco. Análisis factorial. In *Curso: Análisis multivariante I, de la Universidad Carlos III de Madrid*, 2010.

[84] M. Reyes. Modelos de segmentación. In *Curso IN58B: Ingeniería en Marketing*, 2009.

[85] M. Richeldi and A. Perrucci. Analyzing churn of customers.

[86] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, pages 732–741, 2010.

[87] G. Ríos. Métodos basados en casos y en vecindad. In *Curso CC52A: Inteligencia Artificial*, 2009.

[88] L. Rokach and O. Maimon. *Data Mining With Decision Trees:Theory And Applications*. World Scientific Publishing, 2008.

[89] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[90] J. Rossat, J. Larsen, D. Ruta, and M. Wawrzynosek. Customer loyalty, a literature review and analysis. Technical report, UNIPEDE, 1998.

[91] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–590, 1976.

[92] G. Ruiz Merino. Curso de estadística básica. sesiones iii y iv. Online, Agosto 2010. http://www.google.cl/url?sa=t&source=web&cd=1&ved=0CBYQFjAA&url=http % 3A %2Fups %2FSesion_III_IV.ppt&ei=TJDLTMqfN4-q8AbwiPHDAQ&usg= AFQjCNFN-juIBJKNVgbb4Z0yhFBlwl4WjUw&cad=rja.

[93] C. Rygielski, J.-C. Wang, and D. C. Yen. Data mining techniques for customer relationship management. *Technology in Society*, 24:483 – 502, 2002.

[94] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. Technical report, IEEE, 2003.

[95] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[96] J. Shao. Cold deck and ratio imputation. *Survey Methodology*, 26:79–85, 2000.

[97] D. A. Simovici and C. Djeraba. *Mathematical Tools for Data Mining: Set Theory, Partial Orders, Combinatorics*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[98] S. Skrivanek. The use of dummy variables in regression analysis. Technical report, MoreSteam.com, 2009.

[99] D. B. Suits. Use of dummy variables in regression equations. *Journal of the American Statistical Association , No. 280*, Vol. 52:548–551, 1957.

[100] S. Thilagamani and N. Shanthi. Literature survey on enhancing cluster quality. *International Journal on Computer Science and Engineering*, 2:1–4, 1999-2002.

[101] B. Thompson. The loyalty connection: Secrets to customer retention and increased profits. Technical report, RightNow Technologies, 2005.

[102] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[103] S. Varas. Modelamiento en bases de datos. In *Curso IN72K: Tecnologías de Información y Rediseño de Procesos*, 2005.

[104] J. Wang, editor. *Data mining: opportunities and challenges*. IGI Publishing, Hershey, PA, USA, 2003.

[105] L. Wang. *Support Vector Machines: Theory and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[106] G. I. Webb. Multiboosting: A technique for combining boosting and wagging. *Mach. Learn.*, 40:159–196, August 2000.

[107] G. M. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.*, 6:7–19, June 2004.

[108] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, 2. edition, 2005.

[109] C. Wusst. La lealtad de los clientes y su medición. Technical report, ESMESAC, ESTUDIOS DE MERCADO, 2000.

[110] R. Xu and D. Wunsch. *Clustering*. John Wiley and Sons, 2009.