



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

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA MEJORAR EL PROCESO DE CONTROL DE GESTIÓN EN ENTEL

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CLEMENTE ANTONIO MARTÍNEZ ÁLVAREZ

PROFESOR GUÍA:

SEBASTIÁN A. RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN EVALUADORA:

FELIPE AGUILERA VALENZUELA

ANDRÉS FELIPE CHACÓN SANDOVAL

EDUARDO DURÁN NARDECCHIA

SANTIAGO DE CHILE
OCTUBRE, 2012

Resumen

El presente estudio se enfoca en el análisis de ingresos no percibidos en la empresa de telecomunicaciones ENTEL, dentro del proceso de provisión de servicios privados de telefonía, internet y comunicaciones a los clientes de mercados no residenciales. Dicho proceso es controlado mediante indicadores de gestión, obtenidos a partir de la transformación de datos de clientes y servicios. La generación de estos indicadores demanda tiempo y esfuerzo por parte de los analistas de la empresa, debido a que es un trabajo realizado en forma manual.

El objetivo principal de esta tesis consiste en reducir el tiempo de cálculo de los indicadores de servicios privados de ENTEL, para lo cual se aplicó modelamiento multidimensional, técnicas de minería de datos y automatización de procesos, y de este modo poder entregar información más oportunamente.

La metodología de este trabajo se basa principalmente en las etapas del proceso conocido como *Knowledge Discovery in Databases (KDD)*, implementadas de acuerdo a la metodología *CRISP-DM*, la cual es usada para el desarrollo de proyectos de minería de datos. Para comenzar, se hizo un levantamiento de las métricas existentes para la gestión de la provisión de servicios. Luego, se seleccionaron y procesaron las fuentes de datos para el estudio de forma automática, almacenando las variables más relevantes en un repositorio multidimensional (*Data Mart*), reduciendo drásticamente el tiempo de cálculo de indicadores y liberando recursos humanos altamente calificados. A partir de lo anterior, se aplicaron técnicas de *clustering* para obtener grupos de elementos con datos de clientes y servicios cuyas características fueran similares, asociándoles un valor de precio según información histórica de consumo. Por último, se generó un modelo de clasificación que asignara, de acuerdo a una medida de similitud, elementos que no habían sido facturados a los grupos previamente definidos, y de esta manera estimar los ingresos no percibidos.

Con ayuda de minería de datos se logró diseñar nuevas métricas para el proceso e identificar a los clientes y servicios más críticos, lo que permite llegar a valores más exactos de los ingresos perdidos en cada segmento, y aplicar estrategias diferenciadas para hacer el cobro a sus clientes.

El trabajo realizado permitió una reducción del tiempo de obtención de los indicadores en un 78%, pasando de un total de 14 horas inicialmente a tan sólo 3 horas, logrando además estimar los ingresos perdidos mensualmente por servicios no facturados en un monto de MM \$ 210, con un error de la estimación menor al 5%. Se espera que, con ayuda de este estudio, la empresa pueda tomar decisiones informadas y mejorar su capacidad de control del proceso de provisión de servicios privados, con el fin de regularizar su flujo de ingreso mensual.

Abstract

This study focuses on the analysis of unperceived incomes in the supply process of private telephone, internet and communication services to business customers on the telecommunication company ENTEL. This process is controlled by using management indicators obtained from processing of customer and services data. The generation of these indicators is a very time-consuming task for business analysts because it's made manually.

The aim of this thesis project consists in reducing the time necessary to calculate the private management service indicators in ENTEL, through use of multidimensional modeling, data mining techniques and process automation in order to yield information in a more appropriate way.

The methodology used in this research is primarily based on the different stages of the Knowledge Discovery in Databases process implemented according to CRISP-DM methodology which is used for development of data mining projects. First of all, information was gathered on existing metrics used for managing supply process of private services. Then, data sources were selected and processed automatically for the study so the most important variables could be stored in a multidimensional repository (Data Mart), thereby reducing the calculating time of indicators and releasing of highly qualified human resources. After that, clustering techniques were applied to obtain groups of elements using data of customers and services which characteristics were similar within a cluster assigning to each one a value obtained from historical consumption data. Finally, a classification model was created for the purpose of assign a new element with services not yet billed to any of the previously defined clusters according to a similarity measure to estimate the foregone revenues.

New metrics were designed to identify the most critical customers and services by using data mining techniques, enabling reach more accurate values of revenue losses for each segment. Therefore, different strategies could be applied in order to charge customers.

Results show a decrease of indicators calculation time in a 78%, from 14 to 3 working hours. In addition, monthly revenues losses because of services not yet billed were estimated at \$ 210 million with an estimation error less than 5%. This work hopes to offer useful results to ENTEL so their analysts make informed decisions and improve its ability to control their services supply process, in order to regularize their monthly income.

Dedicatoria

A mis padres, por darme la oportunidad de vivir. Ustedes han velado desde siempre por mi bienestar y educación, son un pilar fundamental en mi vida. Gracias por su apoyo en todo ámbito y por los valores que me han inculcado. Aprecio todo su esfuerzo y cariño, tanto por mí como por mis hermanos, en gran parte gracias a ustedes soy el hombre que soy ahora. Los amo con mi alma.

A mis hermanos menores: Sebastián, María José y Fernando, sepan que estoy muy feliz y orgulloso por quienes son, tengan la confianza de que ustedes pueden lograr lo que quieran en su vida, cuenten con mi apoyo y ayuda siempre.

A mi abuelo Fernando, quien me enseñó muchísimas cosas que no se aprenden ni en el colegio ni en la universidad, sólo de la experiencia en la vida. A él y a mi hermanito Cristóbal les dedico este logro, sé que desde el cielo me han dado las fuerzas para conseguirlo.

Por último, quiero dedicar esto a todos aquellos que creyeron en mí y me ayudaron de alguna manera durante este proceso, ya sea con una palabra de aliento, un minuto de su tiempo para escucharme, un abrazo cuando estuve desanimado o recomendaciones para la redacción de este trabajo, todo ello me sirvió para poder llegar al final de esta etapa llamada "Universidad", acercándome al objetivo de ser un gran profesional y una persona íntegra.

Clemente Martínez A.

"Da el primer paso con la fe. No tienes por qué ver toda la escalera. Basta con que subas el primer peldaño" (Dr. Martin Luther King Jr).

"Hasta lo más difícil se puede decir de manera simple. Pero es difícil. Hasta lo más simple se puede decir de forma difícil. Y es fácil." (Soya)

Agradecimientos

Detrás de cada sueño siempre hay personas que nos apoyan, y de alguna manera permiten que éstos se cumplan. Por ello, en las siguientes palabras, quiero expresar mis más sinceros agradecimientos a quienes hicieron posible que llegara hasta aquí:

A mi profesor guía, Sebastián Ríos, le agradezco infinitamente la oportunidad de trabajar con usted en un área tan interesante, admiro la forma en que ve y enfrenta los problemas, lo cual ha sido muy alentador para mí. Más que cualquier cosa, gracias por su paciencia durante este tiempo, fue largo el camino recorrido. De igual manera debo agradecer a Luciano, por su apoyo y seguimiento en los avances de la tesis.

Al equipo de personas de ENTEL, en especial a mis tutores en la empresa, Andrés Chacón y Mauricio Aguirre, que me explicaron la realidad de su negocio con la mejor disposición. Incluyo en este agradecimiento a Eduardo Durán, del equipo de Innovación, quien apoyó las ideas y propuestas de alumnos como yo.

A la Facultad de Cs. Físicas y Matemáticas, y al Depto. de Ingeniería Industrial, por haberme brindado los espacios para aprender y desarrollarme. Siento que la decisión de carrera que tomé hace mucho fue la correcta. Gracias por la darme la posibilidad de ser tanto alumno como docente, de pertenecer al equipo de tutoría para ayudar a otros, de organizar iniciativas estudiantiles, de trabajar en proyectos con profesores, de realizar un postgrado y de representar a la facultad en un deporte, la natación.

A mis compañeros de oficina y jefes en LAN que, conociendo mi situación, me valoraron como persona con habilidades y conocimientos, fueron flexibles cuando lo necesité y sabían que en algún momento lograría mi meta.

A mi familia, por estar siempre presentes, espero entiendan el esfuerzo que hice para poder terminar esta tesis la cual dedico a ustedes. A mi polola, Kattina, por su paciencia y comprensión durante el último tiempo, gracias por la ayuda que me brindaste cuando necesité aclarar mis ideas, me sirvió demasiado. Te amo mucho.

A todos y todas quienes de alguna forma aportaron con una palabra, correcciones, me escucharon, o estuvieron ahí cuando los necesité: Francisco, Gastón, Daniela, Diego, Lorena, Isidora, Monique, Alonso y Francisca entre otros, les doy las gracias.

Clemente Martínez A.

Índice de Contenidos

RESUMEN	II
ABSTRACT	III
DEDICATORIA	IV
AGRADECIMIENTOS	V
1. INTRODUCCIÓN	1
1.1 ANTECEDENTES GENERALES	1
1.2 PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN	2
1.3 OBJETIVOS	8
1.3.1 <i>Objetivo general</i>	8
1.3.2 <i>Objetivos específicos</i>	8
1.4 RESULTADOS ESPERADOS	9
1.5 ALCANCES	10
1.6 HIPÓTESIS	11
1.7 CONTRIBUCIÓN	11
1.8 ESTRUCTURA DEL TRABAJO	12
2. MARCO CONCEPTUAL	14
2.1 ANÁLISIS Y MODELAMIENTO MULTIDIMENSIONAL	14
2.1.1 <i>Sistemas de información</i>	15
2.1.2 <i>Sistemas OLTP y OLAP</i>	16
2.1.3 <i>Modelamiento multidimensional</i>	18
2.1.3.1 <i>Cubo multidimensional</i>	20
2.1.3.2 <i>Modelo estrella</i>	21
2.1.4 <i>Diseño de un Data Warehouse</i>	23
2.1.5 <i>Data Mart</i>	25
2.1.6 <i>El proceso KDD para la obtención de conocimiento</i>	26
2.2 MINERÍA DE DATOS	28
2.2.1 <i>Aprendizaje supervisado y no supervisado</i>	28
2.2.2 <i>Métodos de minería de datos</i>	30
2.2.3 <i>Análisis de clusters</i>	31
2.2.4 <i>Medidas de similitud</i>	32
2.2.5 <i>Determinación del número de clusters</i>	34
2.2.6 <i>Validación de clusters</i>	36
2.2.7 <i>Modelos y algoritmos de minería de datos</i>	37
2.2.7.1 <i>K-means</i>	37
2.2.7.2 <i>Árboles de decisión</i>	38
2.2.7.3 <i>Two-Step Cluster</i>	40
2.2.7.4 <i>Redes Neuronales Artificiales (RNA)</i>	40
2.2.7.5 <i>Support Vector Machines (SVM)</i>	42
2.2.8 <i>Cross Validation</i>	44
3. METODOLOGÍA	45

3.1 DESCRIPCIÓN DE METODOLOGÍAS EXISTENTES	45
3.1.1 <i>Compresión del negocio</i>	47
3.1.2 <i>Compresión de los datos</i>	49
3.1.3 <i>Preparación de los datos</i>	50
3.1.4 <i>Modelamiento y evaluación</i>	51
3.1.5 <i>Despliegue del proyecto</i>	52
3.2 METODOLOGÍA UTILIZADA EN ESTE TRABAJO	52
4. SOLUCIÓN PROPUESTA	54
4.1 SELECCIÓN DE LAS FUENTES DE DATOS	54
4.2 DISEÑO DEL MODELO DE DATOS	56
4.3 LIMPIEZA Y PROCESAMIENTO DE LOS DATOS	58
4.4 TRANSFORMACIÓN DE LOS DATOS.....	63
4.5 <i>DATA MINING</i>	66
4.5.1 <i>Selección de atributos</i>	66
4.5.2 <i>Primer experimento: Clustering</i>	67
4.5.3 <i>Segundo experimento: Clasificación</i>	70
4.5.4 <i>Tercer experimento: Asignación de precios</i>	71
4.6 GENERACIÓN DE NUEVOS INDICADORES.....	73
5. RESULTADOS EXPERIMENTALES	74
5.1 <i>Resultados del primer experimento</i>	74
5.2 <i>Resultados del segundo experimento</i>	76
5.3 <i>Resultados del tercer experimento</i>	78
6. DISCUSIÓN Y EVALUACIÓN DE LOS RESULTADOS.....	80
6.1 DATA MART Y AUTOMATIZACIÓN DE PROCESOS	80
6.2 FORMACIÓN DE <i>CLUSTERS</i> DE CLIENTES Y SERVICIOS PRIVADOS.....	81
6.3 CLASIFICACIÓN DE LOS REGISTROS A CADA <i>CLUSTER</i>	84
6.4 VALORIZACIÓN DE CÓDIGOS DE SERVICIO	85
7. CONCLUSIONES	87
8. TRABAJO A FUTURO.....	89
9. REFERENCIAS	90
10. ANEXOS.....	93
10.1 ANEXO 1: REPORTE MENSUAL CON INDICADOR ISP3.....	93
10.2 ANEXO 2: PROCESAMIENTO DE DATOS EN SOFTWARE <i>RAPIDMINER</i>	95
10.2.1 <i>Proceso de clustering de datos</i>	95
10.2.1.1 <i>Selección de los datos</i>	95
10.2.1.2 <i>Algoritmos para obtener cantidad de clusters</i>	95
10.2.2 <i>Proceso de clasificación de datos</i>	96
10.2.2.1 <i>Selección y transformación de datos</i>	96
10.2.2.2 <i>Grupos de clusters</i>	97
10.2.2.3 <i>Aplicación de modelo de clasificación</i>	97
10.2.2.3.1 <i>Entrenamiento y validación del modelo mediante Cross-validation</i>	98
10.2.2.3.2 <i>Resultados obtenidos por el modelo de clasificación</i>	98
10.3 ANEXO 3: IMPORTANCIA DE LOS ATRIBUTOS EN CADA <i>CLUSTER</i>	99

10.3.1 Importancia de los atributos para el conglomerado N° 1.....	99
10.3.2 Importancia de los atributos para el conglomerado N° 2.....	99
10.3.3 Importancia de los atributos para el conglomerado N° 3.....	100
10.3.4 Importancia de los atributos para el conglomerado N° 4.....	100
10.3.5 Importancia de los atributos para el conglomerado N° 5.....	101
10.3.6 Importancia de los atributos para el conglomerado N° 6.....	101
10.4 ANEXO 4: TABLAS DE FRECUENCIA SEGÚN ATRIBUTOS DE CADA CLUSTER	102

Índice de Figuras

FIGURA 1: PARTICIPACIÓN DE MERCADO DE COMPAÑÍAS OPERADORAS EN CHILE AÑO 2010.....	1
FIGURA 2: INGRESOS DE ENTEL EN EL AÑO 2010 POR SEGMENTO DE MERCADO	4
FIGURA 3: ÁREAS INVOLUCRADAS EN EL PROCESO DE PROVISIÓN DE SERVICIOS PRIVADOS.....	4
FIGURA 4: FUENTES DE DATOS PARA LA GENERAR INDICADORES DE SERVICIOS PRIVADOS (ISP).....	6
FIGURA 5: DISTRIBUCIÓN DE TIEMPO PARA GENERAR LOS ISP	7
FIGURA 6: CUBO MULTIDIMENSIONAL PARA LAS VENTAS DE UNA TIENDA DE RETAIL.	20
FIGURA 7: MODELO ESTRELLA PARA LAS VENTAS DE UNA TIENDA DE RETAIL	22
FIGURA 8: ETAPAS DEL PROCESO KDD.....	26
FIGURA 9: ETAPAS DEL ANÁLISIS DE CLUSTERS.....	31
FIGURA 10: DENDOGRAMA PARA DETERMINAR EL NÚMERO DE CLUSTERS.	35
FIGURA 11: ÁRBOL DE DECISIÓN PARA EVALUAR RIESGO DE UN CLIENTE.....	39
FIGURA 12: NEURONAL CON PESOS ASOCIADOS A CADA NODO	41
FIGURA 13: TRANSFORMACIÓN DEL ESPACIO DIMENSIONAL LOS DATOS	43
FIGURA 14: FASES DEL MODELO CRISP-DM.....	45
FIGURA 15: FLUJO DEL PROCESO DE PROVISIÓN DE SERVICIOS PRIVADOS EN ENTEL.	48
FIGURA 16: DIAGRAMA DE LOS PASOS A SEGUIR EN LA METODOLOGÍA DE LA TESIS	53
FIGURA 17: VARIABLES ALMACENADAS EN DATA STAGING AREA	57
FIGURA 18: MODELO ESTRELLA PARA ALMACENAR INDICADOR ISP3.....	58
FIGURA 19: APLICACIÓN EN LENGUAJE JAVA PARA GENERAR INDICADOR MENSUAL ISP3.....	59
FIGURA 20: PASOS A SEGUIR PARA GENERAR INDICADOR MENSUAL ISP3.....	59
FIGURA 21: PROCESAMIENTO DE LAS FUENTES DE DATOS PARA INDICADOR ISP3.....	60
FIGURA 22: CÁLCULO DE INDICADOR ISP3.....	61
FIGURA 23: GRÁFICO DE DISTRIBUCIÓN DE VELOCIDADES EN MEGA BYTES.....	64
FIGURA 24: GRÁFICO DE DISTRIBUCIÓN DE LA VARIABLE VALOR_REDONDEADO	65
FIGURA 25: DENDOGRAMA DEL EXPERIMENTO DE CLUSTERING.....	67
FIGURA 26: INGRESOS ESTIMADOS PARA CÓDIGOS DE SERVICIO NO DECLARADOS EN CCP.....	79
FIGURA 27: TAMAÑO DE LOS CLUSTERS OBTENIDOS	81
FIGURA 28: GRÁFICO DE CONGLOMERADOS SEGÚN LA VARIABLE SEGMENTO	82
FIGURA 29: GRÁFICO DE CONGLOMERADOS SEGÚN LA VARIABLE CATEGORIA.....	82
FIGURA 30: GRÁFICO DE CONGLOMERADOS SEGÚN LA VARIABLE RUBRO	83
FIGURA 31: GRÁFICO DE CONGLOMERADOS SEGÚN LA VARIABLE TIPO_SERVICIO	83
FIGURA 32: GRÁFICO DE CONGLOMERADOS SEGÚN LA VARIABLE VELOCIDAD_RANGO	84
FIGURA 33: ESTIMACIÓN DE INGRESOS V/S VALOR REAL	85
FIGURA 34: COMPARACIÓN DE LA PREDICCIÓN DE PRECIOS PARA MESES FUTUROS.	86

Índice de Tablas

TABLA 1: VALORES PARA EJEMPLIFICAR INDICADOR <i>ISP1</i>	5
TABLA 2: VALORES PARA EJEMPLIFICAR INDICADOR <i>ISP3</i>	5
TABLA 3: DIFERENCIAS ENTRE <i>OLTP</i> Y <i>OLAP</i>	17
TABLA 4: DATOS DE VENTAS DE UNA TIENDA DE RETAIL	18
TABLA 5: VENTAS DE UNA TIENDA DE RETAIL	19
TABLA 6: VENTAS DE UNA TIENDA DE RETAIL, INCLUYENDO TOTALES	19
TABLA 7: MEDIDAS DE SIMILITUD PARA VARIABLES CONTINUAS	33
TABLA 8: ANALOGÍA ENTRE LAS ETAPAS DEL PROCESO <i>KDD</i> Y LAS DE <i>CRISP-DM</i>	46
TABLA 9: FUENTES DE DATOS Y CAMPOS UTILIZADOS PARA EL INDICADOR <i>ISP3</i>	61
TABLA 10: FUENTES DE DATOS ADICIONALES PARA HACER MINERÍA DE DATOS	63
TABLA 11: VALORES PARA LA VARIABLE <i>RANGO_VELOCIDAD</i>	63
TABLA 12: ATRIBUTOS SELECCIONADOS PARA APLICAR MINERÍA DE DATOS	66
TABLA 13: DISTRIBUCIÓN DE CONGLOMERADOS CON ALGORITMO <i>TWO STEP CLUSTER (K=6)</i>	68
TABLA 14: RESULTADOS OBTENIDOS AL APLICAR TABLAS DE CONTINGENCIA ENTRE CLUSTERS	69
TABLA 15: PARÁMETROS UTILIZADOS EN EL ALGORITMO ÁRBOL DE DECISIÓN <i>J4.8</i>	70
TABLA 16: PARÁMETROS UTILIZADOS EN EL ALGORITMO <i>SVM</i>	70
TABLA 17: IMPORTANCIA DE LOS ATRIBUTOS POR CONGLOMERADO	71
TABLA 18: PESOS ASOCIADOS A CADA ATRIBUTO POR CONGLOMERADO	72
TABLA 19: MATRIZ DE CONFUSIÓN PARA MODELO DE CLASIFICACIÓN, USANDO UN ÁRBOL DE DECISIÓN	77
TABLA 20: MATRIZ DE CONFUSIÓN PARA MODELO DE CLASIFICACIÓN, USANDO ALGORITMO <i>SVM</i>	77
TABLA 21: COMPARACIÓN DE RESULTADOS OBTENIDOS CON LOS MODELOS DE CLASIFICACIÓN	77
TABLA 22: MEDIDAS ESTADÍSTICAS PARA EVALUAR VALORIZACIÓN DE PRECIOS	78
TABLA 23: NIVEL DE CONFIANZA EN LA ESTIMACIÓN DE PRECIOS	78
TABLA 24: RESULTADOS PARA LA RECUPERACIÓN DE INGRESOS POTENCIALES	79
TABLA 25: DISTANCIAS PROMEDIO DE LOS ELEMENTOS DE CADA CLUSTER A SU CENTROIDE	81

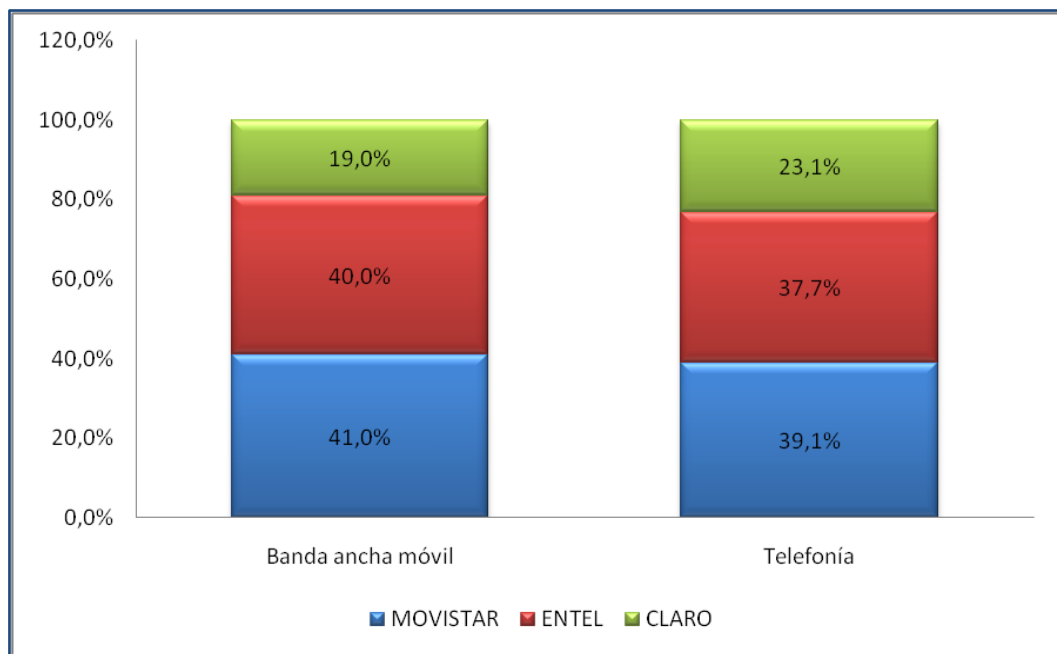
1. Introducción

1.1 Antecedentes generales

En la industria de telecomunicaciones, se genera y almacena una enorme cantidad de datos [21] [34], tanto de sus clientes como de los servicios que se proveen, por ejemplo la telefonía local, telefonía móvil e internet. A partir de estos datos es posible obtener información valiosa y útil para la gestión del negocio. Lo anterior se logra realizando un trabajo de análisis de los datos, pues muchas veces la información se encuentra oculta en ellos.

ENTEL Chile, Empresa Nacional de Telecomunicaciones S.A, es una de las compañías de telecomunicaciones más grandes del país, con fuerte participación de mercado (ver *Figura 1*), que entrega a millones de clientes, tanto personas como empresas, una gran variedad de soluciones de tecnologías de información. Entre el año 2010 y 2011, la empresa decidió rediseñar su organización y trabajar en conjunto con su filial ENTEL PCS, integrando sus negocios de telefonía fija y móvil, con lo cual asumió una complejidad mayor en el manejo de sus clientes.

Figura 1: Participación de mercado de compañías operadoras en Chile año 2010



Fuente: Elaboración propia, basado en informe estadístico anual de SubTel [14].

Para ofrecer el servicio más adecuado a sus clientes, empresas como ENTEL deben segmentar y caracterizar la demanda a través de un análisis de las características de sus clientes y de los productos ofrecidos. Este análisis permite identificar oportunidades de negocio, saber qué clientes y/o productos tienen mayor valor, ver en qué mercados se puede crecer, o ante qué eventualidades hay que estar alerta [2] [26]. Ejemplos de esto último son: la posible fuga de clientes, la baja en el consumo de algún servicio o anomalías en los pagos mensuales.

Sin embargo, dada la inmensa cantidad de clientes y servicios que tiene ENTEL, se cuenta con un gran volumen de datos provenientes desde múltiples fuentes, resultando difícil y a veces tedioso hacer un análisis para obtener información valiosa, más aún de realizarlo de forma manual.

Para facilitar el procesamiento de grandes cantidades de datos como las que tiene ENTEL, se pueden utilizar herramientas de minería de datos. Éstas permiten descubrir patrones no evidentes y ocultos en los datos, ayudando a responder preguntas del negocio que en general consumen o demandan mucho tiempo de los analistas, permitiendo tomar decisiones proactivas y basadas en un conocimiento acabado de la información [21].

Además, con las herramientas de minería de datos se pueden generar modelos para predecir futuras tendencias y comportamientos de los clientes, utilizando información conocida en la empresa como características demográficas, comportamiento de consumo, facturaciones, entre otros.

1.2 Planteamiento del problema y justificación

El desarrollo de esta tesis se enfoca en el estudio de las pérdidas de ingresos de ENTEL debido al no pago de los servicios privados que ofrece, tanto de telefonía como de internet y comunicaciones.

Para lo anterior, se pretende identificar las características que hacen que un cliente empresa demore o deje de pagar alguno de sus servicios. Además, se estimará qué parte de estos ingresos perdidos pueden ser recuperables.

El problema se presenta en el proceso de provisión de los servicios privados que ofrece ENTEL a su cartera de clientes del tipo corporaciones, mayoristas y empresas, de acuerdo al segmento de mercado que pertenezcan. Los servicios privados consisten en un conjunto de servicios de telecomunicaciones no estandarizados, es decir, que se adecúan o ajustan a los requerimientos y necesidades de los clientes.

Los servicios privados representan la mayor fuente de ingresos para la compañía, contando con soluciones de internet, telefonía y transferencia de datos, tales como:

- a) *MPLS*¹, que permite a un cliente intercambiar información a nivel nacional e internacional entre las sucursales, lo cual incluye tráfico de voz, datos y video.
- b) Servicios audiovisuales, para conectar emisoras radiales y estaciones de televisión.
- c) Servicio de conexión entre sucursales bancarias y cajeros automáticos.
- d) Servicios de telefonía e internet banda ancha alámbrica (*ADSL*²), e inalámbrica (*WILL*³ y *Wimax*).
- e) Productos *NGN*⁴ y Trunk IP, con la capacidad de integrar los servicios de telefonía, internet y red de datos a través de un mismo punto de acceso.

Estos servicios se reflejan concretamente en instalaciones técnicas, las cuales deben ser declaradas internamente en el sistema comercial de ENTEL, para que posteriormente se pueda generar el cobro al cliente a través de una factura, y así pague por el servicio que solicitó. Sin embargo, luego de realizar un levantamiento de información con las áreas que monitorean el proceso anterior, se vio que alrededor de 12% de las instalaciones técnicas mensuales ya solicitadas y trabajadas no estaban siendo declaradas oportunamente en los sistemas comerciales de ENTEL; por consiguiente, se realizaba un proceso de facturación tardío o desfasado ó, en otros casos, no se facturaban los servicios prestados, traducándose en una gran fuga de ingresos para la compañía por los volúmenes de dinero que estas instalaciones significan.

Según la memoria anual de ENTEL [22], en el año 2010 la empresa tuvo ingresos totales de MM \$ 289.791, distribuidos entre sus mercados de clientes tal y como se muestra en la *Figura 2*. De éstos, los servicios privados representan una parte importante en mercados como las corporaciones o los negocios mayoristas, manejando volúmenes mensuales de miles de millones de pesos. Por lo tanto, el 12% del total de instalaciones técnicas no declaradas que se mencionó anteriormente puede llegar a significar ingresos de cientos de millones de pesos.

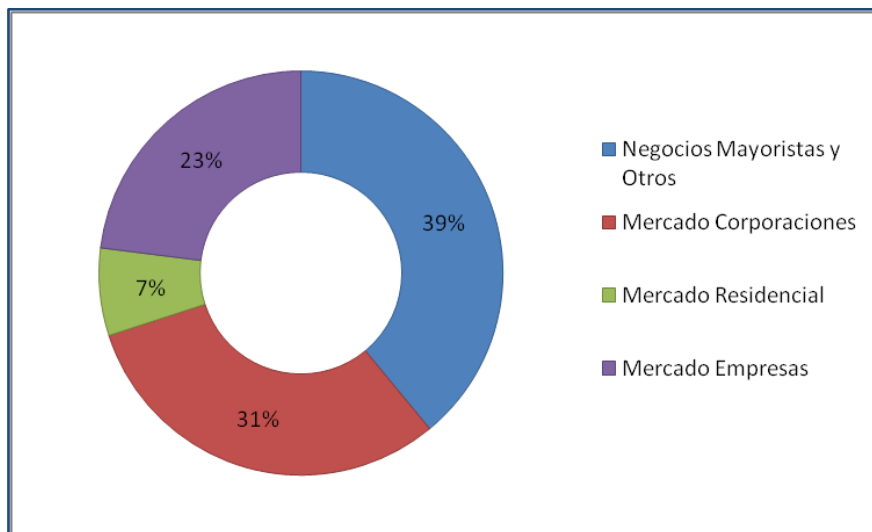
¹Multi Protocol Label Switching.

²Asymmetric Digital Subscriber Line

³WILL o WLL, Wireless IP Local Loop.

⁴Next Generation Networks

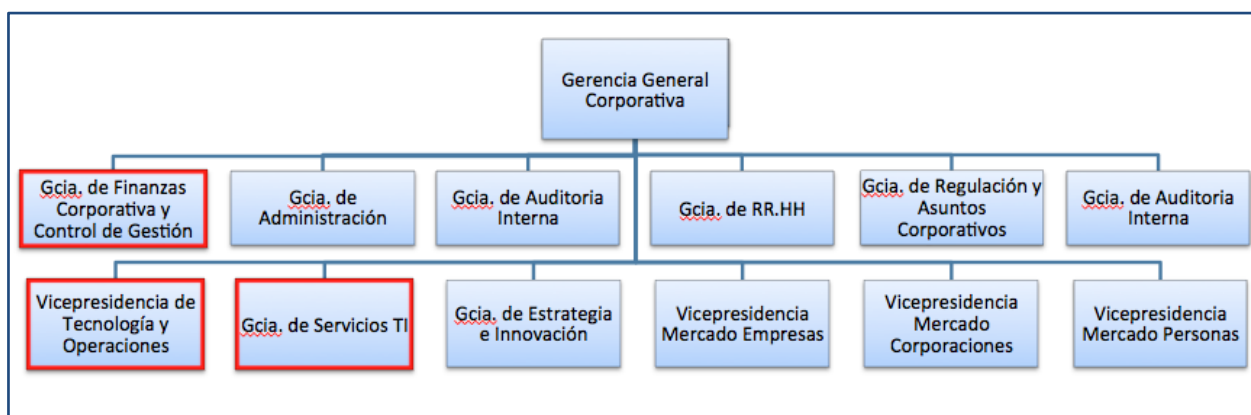
Figura 2: Ingresos de ENTEL en el año 2010 por segmento de mercado



Fuente: Memoria anual 2010 de la empresa ENTEL [22].

Para enfrentar el problema expuesto anteriormente, existe en ENTEL un área llamada “Aseguramiento de Ingresos”, perteneciente a la “Gerencia de Finanzas Corporativa”, la cual se ha encargado de la gestión del proceso de provisión de servicios privados, implementando entre otras medidas el uso de indicadores en etapas críticas del flujo de operación del negocio. Estos indicadores de servicios privados se presentan en reportes mensuales (ver ejemplo en anexo 1), con el fin de medir y controlar la efectividad y cumplimiento de las metas de las áreas involucradas en el proceso de provisión, las cuales se encuentran destacadas en el organigrama de la Figura 3.

Figura 3: Áreas involucradas en el proceso de provisión de servicios privados.



Fuente: Elaboración propia [22]

Existen varios Indicadores de Servicios Privados (ISP) diseñados para hacer gestión en la empresa, con sus correspondientes márgenes de aceptación, de los cuales se destacan los siguientes:

- a) **ISP1:** se usa para validar qué proporción del total de servicios privados vigentes se encuentra incorporada en la base comercial de ENTEL. En el ejemplo de la *Tabla 1*, corresponde al % de los servicios no declarados en el sistema comercial hasta la fecha, es decir, 6.40% del total.

Tabla 1: Valores para ejemplificar indicador ISP1

Mes de control	Total de servicios	Cantidad de servicios declarados	% de servicios declarados	Cantidad de servicios no declarados	% de servicios no declarados
Enero	851	828	97.30%	23	2.70%
Febrero	603	593	98.34%	10	1.66%
Marzo	443	420	94.81%	23	5.19%
Abril	600	593	98.83%	7	1.17%
Mayo	747	734	98.26%	13	1.74%
Junio	551	545	98.91%	6	1.09%
Julio	752	703	93.48%	49	6.52%
Agosto	1088	916	84.19%	172	15.81%
Septiembre	751	645	85.89%	106	14.11%
Total acumulado año 2010	6386	5977	93.59%	409	6.40%

Fuente: Elaboración propia

- b) **ISP2:** sirve para controlar que los contratos de clientes vigentes de la plataforma comercial de ENTEL se estén facturando, y corresponde al % de contratos con facturación emitida en cuenta corriente.
- c) **ISP3:** mide la proporción de los servicios privados instalados cada mes que no se encuentran declarados en la base comercial de ENTEL, y que por lo tanto no se han podido facturar. Como se ve en la *Tabla 2*, corresponde al % de servicios no declarados en el sistema comercial, y se desglosa para cada uno de los segmentos de clientes a quienes atiende la empresa.

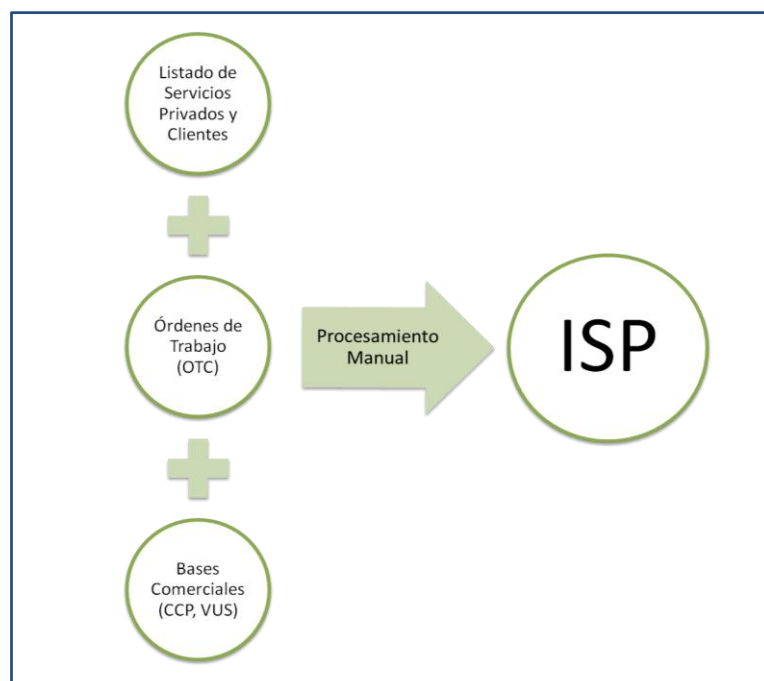
Tabla 2: Valores para ejemplificar indicador ISP3

Segmentos	Total de servicios	Cantidad de servicios declarados	% de servicios declarados	Cantidad de servicios no declarados	% de servicios no declarados (ISP3)b
Corporaciones	232	232	30.89%	0	0.00%
Empresas	315	308	41.01%	7	0.93%
Mayoristas	204	105	13.98%	99	13.18%
Total	751	645	85.89%	106	14.11%

Fuente: Elaboración propia

Los indicadores mencionados se obtienen recopilando y procesando datos desde múltiples fuentes, sistemas y áreas de la empresa. Como se aprecia en la *Figura 4*, se realiza un procesamiento manual de los datos de clientes y servicios, obteniendo como resultado los ISP. Este trabajo consume mucho tiempo y esfuerzo para quienes lo realizan debido a la dispersión y volumen de los datos. Además, se hacen validaciones para evitar errores en los cálculos, los cuales son más probables de ocurrir cuando se trata de un proceso manual.

Figura 4: Fuentes de datos para la generar indicadores de servicios privados (ISP)



Fuente: Elaboración propia

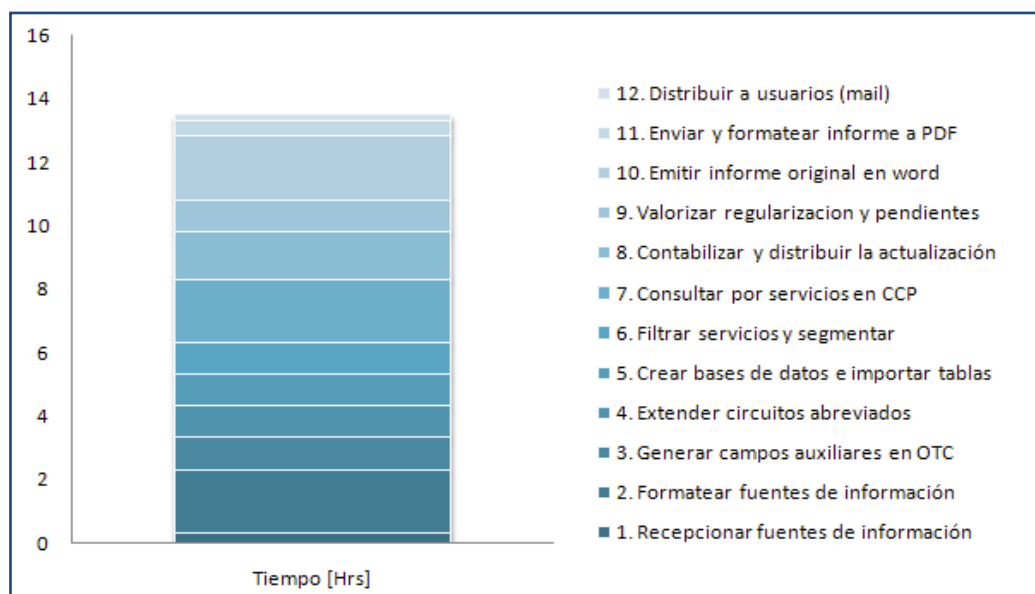
Según se ha estimado por las personas del área de Aseguramiento de Ingresos de ENTEL, el tiempo destinado a la captura y procesamiento de los datos necesarios para la generación de los ISP es bastante, y toma en promedio 2 días completos de jornada laboral cada mes (ver *Figura 5*), tiempo que podrían dedicar estos analistas a otras tareas de control de gestión.

Existen oportunidades de mejora en el proceso actual de ENTEL para la provisión de servicios privados, tanto en el manejo de la información como en el análisis posterior que se realiza de ella:

- a) La gran cantidad de datos que se usan en el proceso de provisión de servicios privados, se puede manejar de una manera eficiente en un repositorio diseñado para la generación de los indicadores de gestión, almacenando sólo la información necesaria para obtener respuestas rápidas a las consultas.

- b) La validación y correctitud de los datos se puede asegurar al automatizar tareas manuales que comúnmente generan errores, al tratar con grandes volúmenes de información.
- c) Adicionalmente, se pueden aplicar técnicas de minería de datos en este repositorio con el fin de estimar los ingresos potenciales no percibidos.
- d) Hacer visible el gap de ingresos en la empresa, realizando un control con indicadores que capturen monetariamente el valor de las pérdidas.

Figura 5: Distribución de tiempo para generar los ISP



Fuente: Elaboración propia

La fuga de ingresos mencionada anteriormente es un tema que no se ha logrado desarrollar y estimar de forma precisa hasta el momento en la empresa. Contar con una valorización de estos ingresos permitiría dimensionar la diferencia entre los ingresos reales y el potencial alcanzable, así como cuánto de esa diferencia es recuperable.

Utilizando herramientas de minería de datos se espera obtener conocimiento del comportamiento de los servicios privados y clientes de ENTEL en base a sus características. Al tener la información bien organizada, los analistas pueden tomar decisiones asertivas basándose en datos confiables, y prever situaciones que ocurran con los clientes o servicios. Con esto se puede, por ejemplo, ofrecer un conjunto de servicios que se ajusten mejor a las necesidades de un tipo de cliente, o también ver los servicios que generan mayor rentabilidad para la empresa y colocar entonces atención a los clientes que los consumen para disminuir fugas de ingresos [2].

1.3 Objetivos

1.3.1 Objetivo general

- ✓ Reducir el tiempo de cálculo de indicadores de servicios privados, utilizados para la gestión de ingresos de ENTEL, y mejorar la oportunidad de la información a través del uso de modelamiento multidimensional y la aplicación de técnicas de minería de datos.

1.3.2 Objetivos específicos

- ✓ Estudiar y documentar el estado del arte en técnicas de minería de datos y en modelamiento multidimensional [16], que se utilizarán para realizar la estimación de los ingresos potenciales para ENTEL.
- ✓ Diseñar e implementar un *Data Mart* [15] para el área de Aseguramiento de Ingresos, como repositorio único y consolidado de datos e indicadores, que permita tener la información disponible para su consulta, usando modelamiento multidimensional.
- ✓ Disminuir el tiempo que tarda el procesamiento y almacenamiento de los datos que se utilizan para el cálculo de métricas e indicadores de gestión de ingresos, y minimizar la cantidad de errores que se pueden producir al realizar un tratamiento manual de los datos.
- ✓ Aplicar algoritmos de minería de datos, en particular de *clustering* [38] y clasificación, para extraer conocimiento desde los datos de clientes y servicios privados de ENTEL, con el fin de estimar ingresos potenciales debido a servicios no facturados.
- ✓ Revisar formas de cálculo de indicadores de servicios privados actuales, crear otros nuevos y mejorar los existentes.

1.4 Resultados Esperados

- ✓ Documentación de minería de datos y modelamiento multidimensional realizada en capítulo 2 de esta tesis.
- ✓ Documentación de la metodología utilizada para el tratamiento y almacenamiento de datos en el *Data Mart*, así como la forma en que se obtiene conocimiento a partir de estos al utilizar minería de datos, escrita en el capítulo 3 de esta tesis.
- ✓ Construcción e implementación de *Data Mart* para el área de Aseguramiento de Ingresos de ENTEL, dejándolo funcionando operativamente en la empresa.
- ✓ Diseñar y construir una aplicación que permita automatizar el procesamiento y almacenamiento de los datos con los cuales se construyen los Indicadores de Servicios Privados en ENTEL, validando la correctitud de los datos.
- ✓ Manual de uso de la aplicación construida, que dé cuenta del software a nivel de diseño y a nivel de uso de la herramienta.
- ✓ Generar un modelo de estimación de ingresos, que permita cuantificar las pérdidas que se producen por concepto de instalaciones técnicas no declaradas en el sistema comercial de ENTEL, descrito en el capítulo 4 de esta tesis, cuyos resultados se presentan en el capítulo 5.
- ✓ Evaluar los algoritmos utilizados en los diferentes modelos de minería de esta tesis, mediante el uso de medidas y criterios de validación, los cuales se presentan en el capítulo 6.
- ✓ Creación de nuevo indicador que permite valorizar las pérdidas de ingresos, explicándose en capítulo 4 de esta tesis.

1.5 Alcances

Este estudio de tesis se basa en información del consumo de servicios privados que provee la empresa de telecomunicaciones ENTEL a sus clientes, en los segmentos de mercado del tipo corporaciones, mayoristas y empresas, dejando fuera de este estudio a todos los clientes del mercado residencial o personas. Se trabajó sólo con uno de los indicadores de servicios privados utilizados en ENTEL, el llamado ISP3, revisando las mediciones mensuales de provisión de servicios privados. Quedan fuera de los análisis los otros indicadores de servicios privados ya mencionados (ISP1 e ISP2).

Los datos que se analizaron consideraron el período comprendido entre Octubre del 2009 hasta Diciembre del 2010. Estas fechas se eligieron debido a que permitieron analizar, comparar y validar los resultados de los indicadores en transcurso del tiempo y además, porque durante ese período se mantuvo una realidad del negocio similar, la cual cambió durante el año 2011 debido a los procesos de fusión con ENTEL PCS.

Las fuentes de datos que se utilizaron se acotan a reportes mensuales y bases de datos, los cuales contienen información de los clientes y de los productos que están asociados a sus cuentas corrientes. Los reportes y archivos se encontraban en formato Excel y en bases de datos Access.

Las variables en estudio se construyeron a partir de los atributos de clientes y servicios, y se consideraron de los siguientes tipos:

- **Demográficas:** en qué lugar se realiza la prestación del servicio, indicando comunas, regiones, etc.
- **Comportamiento de consumo:** categorías y sub-categorías definidas internamente por ENTEL para sus clientes, en relación a sus facturaciones mensuales. Estas categorías difieren entre cada segmento de clientes:
 - a. Corporaciones: capital, cumbre, insigne, preferente o inactivos.
 - b. Mayoristas: capital.
 - c. Empresas: normal, especial, superior o premium.
- **Rubros:** son las áreas o sectores económicos de un cliente, que pueden ser Entel, finanzas & servicios, industria & comercio, gobierno, minería y fuerzas armadas.
- **Tipos de servicio:** por ejemplo, cuando son alámbricos o inalámbricos.

Para este estudio no se consideró el comportamiento de pago de un cliente, lo cual se podría incluir a futuro para identificar clientes buenos y malos, o su tendencia a la fuga.

1.6 Hipótesis

1. Es posible reducir a la mitad el tiempo que tardan los analistas de ENTEL en generar los indicadores de control de ingresos, y al mismo tiempo, mejorar la oportunidad y calidad de la información para la toma de decisiones mediante el uso de *Business Intelligence* y el modelamiento multidimensional.
2. Con ayuda de minería de datos es posible caracterizar, mediante ciertas variables, las preferencias de los clientes por los servicios que ofrece ENTEL, con el fin de posteriormente realizar una segmentación de éstos.
3. Aplicando algoritmos de *clustering* y clasificación, es posible estimar el consumo de un cliente por un servicio⁵ y obtener un valor concreto en dinero de su comportamiento futuro, en base a información de consumo pasada conocida de otros clientes y servicios con características similares.
4. Es posible construir un nuevo indicador de servicios privados, que permita valorizar los ingresos perdidos, a partir de los datos almacenados en un *Data Mart*.

1.7 Contribución

Con este trabajo de tesis se busca generar motivación por ayudar a mejorar los procesos de una empresa, en el ámbito del trabajo de un ingeniero civil industrial, haciendo uso de las tecnologías de información.

Se pretende obtener una aplicación real de modelos de minería de datos, que sirva para el apoyo a la toma de decisiones en una empresa. Se espera entregar nuevas herramientas a los analistas de ENTEL, que sirvan de apoyo al proceso actual de gestión, facilitando el análisis de los datos, liberando recursos y tiempo que se consumen durante las etapas de limpieza y procesamiento de datos, de manera que puedan dedicarlo a tareas que generen mayor valor.

La información se consolidará y almacenará en un repositorio de datos, lo cual facilitará la generación de los indicadores y obtención rápida de respuestas a consultas asociadas al proceso de control de ingresos.

⁵Entendiendo que un servicio es un conjunto de atributos para un producto ofrecido.

Se generará una propuesta de segmentación de los productos (cliente-servicio) con el fin de tener mayor conocimiento de la demanda y tomar mejores decisiones. También, se establecerán nuevos indicadores para la gestión de ingresos, que permitan hacer un mejor seguimiento al proceso de provisión de servicios y muestren con claridad el impacto en la rentabilidad del negocio.

1.8 Estructura del Trabajo

Luego del capítulo introductorio se presentan los próximos capítulos, estructurados de la siguiente forma:

Capítulo 2: Marco conceptual

En este capítulo se presenta el estado del arte en dos ámbitos principales de investigación de esta tesis: análisis & modelamiento multidimensional y algoritmos de minería de datos.

El primero de ellos muestra las principales metodologías utilizadas para hacer modelos multidimensionales, y definiciones o conceptos aplicados en este trabajo para el diseño de un repositorio de datos en el cual se pueda hacer análisis.

El segundo punto, contempla los pasos a seguir para poder aplicar modelos y algoritmos de minería de datos sobre el repositorio previamente, esto con el fin de identificar patrones o relaciones entre los datos. Además, se presentan modelos de aprendizaje (supervisado y no supervisado), algoritmos de clasificación o de *clustering*, algoritmos de regresión y medidas de similitud que permiten comparar los modelos y resultados obtenidos.

Capítulo 3: Metodología

Este capítulo presenta una metodología propuesta con el fin de dar solución a la problemática del capítulo 1. Ésta se basa en la literatura recopilada sobre modelamiento multidimensional y minería de datos, siguiendo los pasos del proceso *Knowledge Discovery in Databases (KDD)* y también de metodologías propuestas para la realización de un proyecto de minería de datos.

Capítulo 4: Solución propuesta

En este capítulo se presenta la solución del trabajo, mostrando cómo se aplicó modelamiento multidimensional al diseñar una estructura capaz de almacenar la información del negocio. Se explica también los procesamientos realizados para tratar los datos, las consideraciones y las fuentes utilizadas.

Además de lo anterior, se muestra el prototipo funcional diseñado para ENTEL, con los indicadores para el control del proceso de provisión de servicios privados. Por último, se exponen los algoritmos de minería de datos con sus parámetros aplicados al problema, los atributos seleccionados y el conjunto de datos en estudio.

Capítulo 5: Resultados experimentales

Este capítulo presenta los resultados obtenidos al clasificar los clientes y productos de ENTEL, valorizando también la pérdida de ingresos debido a los servicios que aun no han sido declarados comercialmente, y entregando un indicador que mida qué tanto de la pérdida es recuperable.

Además de los resultados obtenidos, se muestran las medidas aplicadas para realizar una comparación entre los datos, que permita la discusión y llegar a una conclusión del análisis.

Capítulo 6: Discusión y evaluación de los resultados

En este capítulo se discuten los resultados obtenidos a partir de las pruebas realizadas anteriormente, utilizando los prototipos diseñados y basándose siempre en la metodología establecida para la tesis. Se destacan los puntos principales de los resultados mostrados en el capítulo anterior, su implicancia y el impacto que pueden llegar a tener en los objetivos de este trabajo.

Capítulo 7: Conclusiones y trabajo a futuro

Este capítulo concluye con los principales aprendizajes de este trabajo de tesis, haciendo una revisión general de cada tema tratado, comprobando el cumplimiento de los objetivos propuestos en un comienzo, y los aspectos a destacar de los resultados ya discutidos. Además de lo anterior, se deja en claro los alcances no tratados en este trabajo y que pueden ser tomados a futuro como mejoras.

2. Marco Conceptual

A continuación se muestran los principales puntos con respecto al análisis & modelamiento multidimensional, y a los algoritmos de minería de datos relacionados con el trabajo desarrollado.

En primer lugar se explican conceptos o definiciones utilizados para el diseño de modelos multidimensionales, el uso que se le ha dado en las empresas a través de los sistemas de información, y la importancia que tienen en el apoyo a la toma de decisiones.

Por otra parte, se mencionan algoritmos de minería de datos que se pueden utilizar para obtener conocimiento desde la información disponible en los repositorios de datos. Esta descripción se enfoca en técnicas de clasificación, algoritmos de regresión y algoritmos de aprendizaje.

2.1 Análisis y modelamiento multidimensional

En la actualidad, las empresas se encuentran constantemente condicionadas por factores del medio o negocio en el que estén involucradas, lo que implica una toma de decisiones que pueden ser del tipo tácticas (corto plazo) o estratégicas (mediano a largo plazo). En general, estas decisiones se basan en información histórica, proveniente de distintas fuentes de datos relacionadas directamente con el caso en estudio (ventas históricas, estudio de clientes, análisis de mercado, encuestas, etc.).

Por lo anterior, muchas empresas realizan un almacenamiento de la información de sus transacciones comerciales, ya sea para contar con un registro de su operación o en otros casos con un fin estratégico. Sin embargo, esta recopilación de datos y su análisis posterior puede tornarse un problema cuando aumenta el volumen y la variedad de éstos. Es aquí donde se aplica el uso de metodologías y de herramientas como los sistemas de información y de apoyo a la toma de decisiones.

El análisis multidimensional busca entregar al usuario final una manera fácil de representar la información, con varios componentes dimensionales o atributos, en una estructura común con la cual poder tomar decisiones [30].

2.1.1 Sistemas de información

Un sistema de información (SI) es un conjunto de personas, datos, procesos, funciones, interfaces, redes y tecnologías que interactúan entre sí para apoyar y mejorar las operaciones diarias de la empresa, así como también la toma de decisiones [31].

El objetivo de los SI es transformar los datos en información y conocimiento útil, de tal forma que el negocio pueda solucionar sus problemas. Los SI's se clasifican en 3 tipos: transaccionales, de apoyo a las decisiones y estratégicos [31].

Los SI's Transaccionales u Operacionales, se encargan de automatizar tareas y procesos que se realizan a diario en la empresa, manejando datos del funcionamiento de la organización. Se mantienen grandes cantidades de datos y a un nivel detallado.

A pesar de ser una fuente de datos completa, este tipo de sistema no se utiliza en la toma de decisiones de alto nivel, porque al procesar una gran cantidad de datos tarda en entregar las respuestas y además, no responde a todas las preguntas que puede tener el negocio ya que sólo cuenta con cálculos simples.

Por otra parte, los SI's de Apoyo a la Toma de Decisiones, apoyan al usuario final de un proceso de negocio, entregándole información útil en las cuales basar sus decisiones. A diferencia de los anteriores, este tipo de sistema realiza un pre-procesamiento de los datos transaccionales, por lo que requiere de un menor trabajo al momento de entregar sus respuestas, las cuales son mucho más rápidas. Aplicaciones de estos sistemas se presentan en los modelos de inventario, compra de materiales, programación de producción, etc.

El último tipo son los SI's Estratégico, que apoyan a la empresa en decisiones de alto nivel en el mediano-largo plazo. Estos sistemas se tienen que adaptar a las necesidades del negocio, y se desarrollan de forma incremental adicionando nuevos procesos o funciones a las consideradas en una etapa inicial. Un ejemplo de estos sistemas son los *ERP (Enterprise Resource Planning)*, sistemas modulares que se integran y adaptan a la realidad de la empresa, para gestionar los procesos de negocio y satisfacer las necesidades de información de cada área que los utilice.

Así como los sistemas de información nombrados anteriormente, se han diseñado otras herramientas para el manejo de información orientadas al procesamiento en línea, estos son los llamados OLTP y OLAP, los cuales se explicarán a continuación.

2.1.2 Sistemas OLTP y OLAP

Los sistemas OLTP⁶ surgen a mediados de 1970, con el fin de almacenar grandes volúmenes de información y capturar las transacciones que realizaban los negocios a diario (en tiempo real), con aplicación en sistemas de reservas, entrada de pedidos, sistemas bancarios, control de manufactura, etc. [15]

Los sistemas OLTP trabajan con bases de datos relacionales y están diseñados para almacenar y modificar continuamente datos de la operación diaria de un negocio. Sin embargo, no todos los datos almacenados son relevantes para los reportes o análisis que los usuarios de las empresas buscan, es decir, OLTP es limitado para la toma de decisiones y por lo tanto se requiere de un trabajo adicional sobre los datos para darles un formato o estructura que permita capturar el valor de la información [6].

Las consultas históricas en este tipo de sistemas producen un impacto negativo en la operación, pues su orientación se centra más en la aplicación que en los usuarios. Por ejemplo, si se quiere agrupar la información para obtener las ventas totales de un año, OLTP no es eficiente en los tiempos de respuesta [32].

La arquitectura *Data Warehouse*⁷ fue diseñada específicamente para cubrir la brecha que tienen los sistemas OLTP en la entrega de información para los usuarios, consolidando las fuentes de información en una base única y consistente, en la cual se tiene la información de forma agregada para responder rápidamente a las preguntas del negocio [8].

En el año 1993 aparece el concepto de OLAP⁸, que son sistemas para extraer información relevante del negocio desde bases de datos complejas, analizando la información y entregando una respuesta rápida a las consultas de los usuarios [30].

Los sistemas OLAP trabajan en línea con datos resumidos en una estructura multidimensional, a diferencia de los OLTP que tienen una estructura relacional, por este motivo los OLAP obtienen resultados más ágiles a las consultas [30]. Las principales diferencias entre ambos sistemas se resumen en la *Tabla 3*.

La información en los sistemas OLAP se obtiene desde múltiples fuentes y dispone en variados formatos, ya sea tablas, gráficos o reportes para que la utilicen áreas del negocio como ventas, marketing, control de gestión, etc.

⁶ *Online Transactional Processing*

⁷ Ver capítulo 2.1.4, donde se define como un repositorio de datos consolidado para la toma de decisiones.

⁸ *OnLine Analytical Processing*

Tabla 3: Diferencias entre OLTP y OLAP.

OLTP	OLAP
Orientado a la operación diaria y al funcionamiento de las aplicaciones transaccionales.	Orientado al usuario que toma las decisiones del negocio.
La información se almacena en bases de datos relacionales, con datos normalizados, y se accede principalmente para insertar, modificar o eliminar datos.	La información se almacena en estructuras multidimensionales, y se accede para hacer consultas.
Muchos usuarios acceden a modificar los datos constantemente.	Los datos permanecen estáticos hasta su próxima actualización, los usuarios sólo acceden para lectura de los datos.
El tamaño de la base de datos incrementa rápidamente, por lo cual se le da preferencia a los datos más actuales. Se busca tener la mínima redundancia posible y consistencia en los datos.	El tamaño de la base de datos puede ser muy grande debido a la redundancia de datos. Los datos históricos y actuales son igual de importantes.

Fuente: Elaboración propia [32]

OLAP muestra ser más eficiente que OLTP cuando los usuarios necesitan tomar decisiones basados en datos del negocio. Además, permite generar reportes cambiando los atributos que se quieran ver, realizando los cálculos en línea.

Existen dos formas de clasificar los sistemas OLAP, la diferencia está en el tipo de base de datos que se utiliza [8] [32]:

1. **MOLAP**⁹: Esta implementación almacena información pre-calculada en una base de datos multidimensional, optimizando los tiempos de respuesta.
2. **ROLAP**¹⁰: En este caso la información se guarda a nivel detallado en una base de datos relacional, diferente al modelo entidad-relación, ya que posee tablas des-normalizadas. Algunas representaciones de esto son el *Star Model* y el *Snowflake Model*. La principal ventaja de esta arquitectura es que permite el análisis de una enorme cantidad de datos.

⁹ Multidimensional OnLine Analytical Processing

¹⁰ Relational OnLine Analytical Processing

2.1.3 Modelamiento multidimensional

El modelamiento multidimensional (MMD) es una técnica para modelar lógicamente bases de datos, de tal forma que el usuario comprenda la relación entre los atributos o variables que está analizando. Se implementa en un repositorio de información, que puede ser utilizado para la creación de reportes que requiera el usuario final o para la aplicación de técnicas de extracción de patrones de datos. Se presenta como un apoyo a la arquitectura *Data Warehouse*, ya que ofrece una forma integrada de ver la información [16] [30].

La representación de lo anterior se hace utilizando los conceptos de “dimensiones” y “medidas”, definiendo las “dimensiones” como los atributos, categorías o jerarquías por las que se quiere agrupar o mostrar la información, y las “medidas” como las métricas o valores cuantitativos. En cada una de estas jerarquías se cuenta con un “punto de agregación”, es decir, un valor pre-calculado. Esto lo hace diferente de las herramientas tradicionales OLAP, que necesitan calcular los datos al momento de su consulta [30].

Una aplicación del MMD son las ventas en una tienda de *retail*, donde los atributos producto y zona se pueden tomar como dimensiones, mientras que el atributo ventas se considera como una medida. En la *Tabla 4*, se presentan datos de las ventas de algunos productos en distintas zonas del país.

Tabla 4: Datos de ventas de una tienda de retail

Producto	Zona	Ventas
Zapatos	Norte	200
Zapatos	Centro	100
Zapatos	Sur	50
Poleras	Norte	500
Poleras	Centro	400
Poleras	Sur	200
Camisas	Norte	50
Camisas	Centro	500
Camisas	Sur	300

Fuente: Elaboración propia

Se puede ver que no hay correspondencia uno a uno entre los datos de cada campo, lo cual si ocurre en una tabla relacional. Una manera más clara de ver esta misma información se muestra en la *Tabla 5*, donde se utiliza una matriz con 2 dimensiones: Producto y Zona.

Tabla 5: Ventas de una tienda de retail

Producto\Zona	Norte	Centro	Sur
Zapatos	200	100	50
Poleras	500	400	200
Camisas	50	500	300

Fuente: Elaboración propia.

Si se lleva esto a un lenguaje multidimensional, se puede decir que esta matriz representa las ventas dimensionadas por producto y zona. La representación anterior sirve cuando se requieren responder preguntas del tipo: ¿Cuál fue la venta de zapatos en la zona Centro? o ¿Cuál fue la venta de camisas en la zona Sur?

En el caso que se quieran responder otras preguntas como: ¿Cuáles fueron las ventas totales de zapatos? o ¿Cuáles fueron las ventas totales en la zona Centro?, la tabla anterior no muestra esa información directamente, ya que involucra una agregación de los datos. Para obtener una respuesta rápida a las preguntas anteriores, se pueden incluir los totales y subtotales para cada una de las dimensiones, lo cual se ve en la Tabla 6.

Tabla 6: Ventas de una tienda de retail, incluyendo totales

Producto\Zona	Norte	Centro	Sur	Total
Zapatos	200	100	50	350
Poleras	500	400	200	1100
Camisas	50	500	300	850
Total	750	1000	550	2300

Fuente: Elaboración propia

Cuando las medidas pueden ser agregadas de alguna forma, entonces se les llama dimensiones aditivas. En el ejemplo anterior, las ventas de cada zona se pueden agregar, obteniendo las ventas totales del país. Sin embargo, no siempre tiene sentido agregar las medidas de una tabla, esto ocurre cuando se trabaja con porcentajes.

Si el usuario quisiera adicionar una o más dimensiones a sus consultas, por ejemplo la dimensión tiempo, para preguntar cosas como: ¿Cuáles fueron las ventas de zapatos el año 2010 en la zona Centro?, entonces el modelamiento sería totalmente diferente.

Las interrogantes de los usuarios no siempre logran ser contestadas puesto que ello depende de las dimensiones y medidas que se hayan considerado al hacer el modelo de datos y del nivel de detalle con que se cuenta. Por ejemplo, si los datos se tienen a nivel temporal mensual, no se puede saber ¿Cuál es el día de la semana con mayor consumo? Este nivel mínimo de información necesario para responder las preguntas de los usuarios se denomina granularidad [30].

El modelamiento multidimensional debe ser capaz de proveer la información básica que los usuarios finales necesiten de modo de satisfacer todas sus consultas en diferentes niveles de agregación.

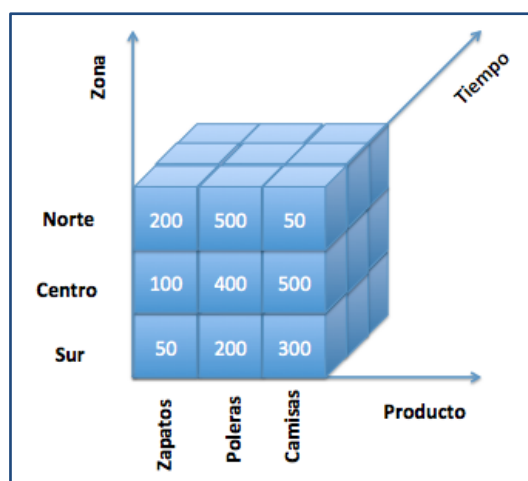
Existen dos técnicas o enfoques básicos para el diseño lógico de bases de datos utilizando el modelamiento multidimensional, y se diferencian en el motor de base de datos que se quiera utilizar. El primero es a través de un cubo multidimensional, y el segundo a través de un modelo de datos relacional, como el modelo estrella. Ambos enfoques tienen ventajas y desventajas, según el uso que se le quiera dar, y se explican en los capítulos siguientes.

2.1.3.1 Cubo multidimensional

Los cubos multidimensionales consisten en una estructura que cuenta con dimensiones y medidas, agrupadas de tal forma que simplifica la formulación de consultas complejas para los usuarios, entregando una rápida respuesta. El cubo requiere de una base multidimensional para su implementación, la cual interactúa con el usuario a través de alguna interfaz gráfica. Se puede representar a través de arreglos multidimensionales.

Tomando el mismo ejemplo de las ventas de una tienda de *retail*, la *Figura 6* muestra el concepto de un cubo multidimensional, considerando las dimensiones tiempo, producto y zona. Cada celda del cubo representa una medida asociada a las dimensiones que se interceptan. Con esta estructura se permite al usuario visualizar por ejemplo las ventas de poleras, en la zona centro para una fecha determinada.

Figura 6: Cubo multidimensional para las ventas de una tienda de retail.



Fuente: Elaboración propia.

En el cubo dimensional es posible realizar distintas operaciones, descritas a continuación [30]:

1. **Pivoting:** rotar el cubo y mostrar una cara en particular.
2. **Slicing:** seleccionar una dimensión del cubo.
3. **Dicing:** seleccionar una o más dimensiones del cubo.
4. **Drill-down:** mostrar o desplegar el detalle del punto de agregación.
5. **Roll-up:** agrupar los datos según el punto de agregación.

2.1.3.2 Modelo estrella

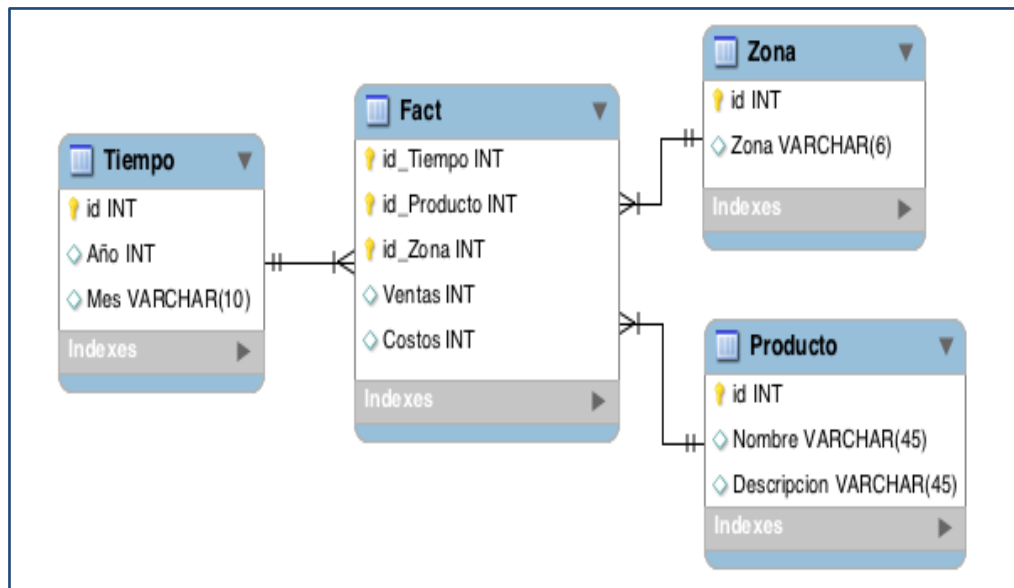
El modelo estrella es una representación de un modelo dimensional en una base de datos relacional, la cual cuenta con una tabla principal (*fact table*) unida o relacionada mediante una llave, compuesta por las llaves principales de otras tablas (*dimensional tables*). Se utiliza el modelo entidad relación (ER) como referencia para su diseño, sin embargo sus reglas de normalización son mucho más flexibles, con el fin de optimizar el acceso a los datos, lo cual lo hace mucho más orientado a la información [15].

Se puede decir que es un modelo simple y simétrico, haciendo fácil su comprensión y la lectura de los datos guardados en él. La ventaja de este tipo de esquemas es la eficiencia de las consultas, ya que la información principal y recurrente está contenida en una única tabla, usando referencias para incluir información adicional [16].

La *Figura 7* muestra un ejemplo de modelo estrella. Utilizando el mismo ejemplo antes expuesto de la tienda de *retail*. Se puede definir una *fact table* con indicadores como las ventas y costos, así también las dimensiones: producto, zona y tiempo. Esta forma de representar la información, permite responder a las preguntas planteadas anteriormente en las que se requiere de variados niveles o dimensiones: ¿Cuáles fueron las ventas de zapatos el año 2010 en la zona Centro?

En la *fact table* se almacenan las medidas del negocio (por ej. las ventas), que en conjunto con las dimensiones (por ej. zona, producto y tiempo) referenciadas desde otras tablas, permiten responder a las preguntas de los usuarios. En esta tabla es importante considerar la granularidad de los datos, vale decir, con qué nivel de detalle se quieren ver las medidas definidas anteriormente, y de acuerdo a eso elegir las dimensiones (por ejemplo visualizar los datos temporalmente en años, meses o días).

Figura 7: Modelo estrella para las ventas de una tienda de retail



Fuente: Elaboración propia.

Cuando las tablas dimensionales incluyen muchos atributos, suelen empeorar los tiempos de respuesta ante las consultas de los usuarios. Para solucionar esto, existe una adaptación del modelo estrella, llamado modelo *snowflake*, cuya metodología propone normalizar las tablas dimensionales. Al usar este tipo de modelamiento se evita cierta redundancia en los datos [30].

Los modelos estrella y *snowflake* son utilizados de preferencia en el diseño y creación de un *Data Mart*¹¹, cuando se trabaja con una base de datos relacional. En el caso de un diseño mucho más complejo, como es el caso de los *Data Warehouses*, una mejor opción consiste conectar varios modelos estrella, con lo cual se conforma una nueva estructura llamada modelo constelación, el que cuenta con varias *fact table* y tablas dimensionales en común [16].

¹¹ Ver capítulo 2.1.5, donde se define un *Data Mart*

2.1.4 Diseño de un *Data Warehouse*

De acuerdo a la definición de W.H. Inmon, la arquitectura *Data Warehouse* consiste en un repositorio estandarizado de información, una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil, diseñada para dar soporte a la toma de decisiones del negocio [15].

Por otro lado, Ralph Kimball plantea otra definición, en la cual indica que es una colección de datos en forma de una base de datos, que guarda y ordena información extraída directamente desde los sistemas operacionales y datos externos [16].

Las definiciones anteriores presentan a los *Data Warehouses* (DW) como una fuente confiable de información orientada al negocio, con datos limpios y consolidados, provenientes desde uno o más sistemas operacionales distintos y almacenados en un repositorio o base de datos al cual pueden acceder los usuarios de manera rápida para hacer sus consultas. Esto resuelve el problema que tienen la mayoría de los sistemas, que sólo responden preguntas muy específicas, y además se tardan mucho tiempo [32].

Un *DW* tiene características relevantes, tal como dice Inmon en su definición, y se detallan en los siguientes puntos [15]:

- 1) **Orientación al negocio:** Se enfocan en los temas o áreas de interés del negocio, almacenando la información necesaria para los análisis que se quiera realizar, lo cual los diferencia de aplicaciones funcionales o sistemas operacionales que almacenan todo tipo de información.
- 2) **Integrado:** Se alimentan desde diversas fuentes de información, se les da un formato estandarizado y se integran en un único repositorio. La importancia de esto es que todo se encuentra consolidado, por lo que no hay que buscar en múltiples fuentes, se tiene una única imagen de la empresa y aumenta la credibilidad de los datos.
- 3) **No volátil:** La actualización de los datos almacenados no se modifica frecuentemente o cambia ocasionalmente, a diferencia de los sistemas operacionales que cambian constantemente. El *DW* almacena información histórica en algún instante del tiempo para que luego sea consultada (“se saca una foto” de lo que ha ocurrido, y los datos se actualizan con la frecuencia que el negocio lo requiera).
- 4) **Variante en el tiempo:** Los datos siempre son relevantes para alguien en algún instante del tiempo, así que los *DW* almacenan datos históricos además de los actuales, mientras que en los sistemas operacionales aquellos más recientes son los que importan, y a veces actualizan o borran los que ya han pasado.

Para el diseño y construcción de un *DW*, se han propuesto diferentes metodologías, las cuales tienen en común las siguientes etapas [30]:

- 1) **Analizar las necesidades del usuario final:** se debe entregar información de tal manera que aplicaciones y herramientas de análisis puedan obtener información útil, satisfaciendo así la mayor parte de las necesidades del negocio.
- 2) **Seleccionar las fuentes de información:** una vez que se definió la necesidad de los usuarios, entonces se buscan las fuentes de información que permitan cubrir esto.
- 3) **Desarrollar un modelo lógico de datos:** entre los modelos más utilizados se encuentran el cubo multidimensional y el modelo estrella.
- 4) **Preparar un prototipo para el usuario final:** el objetivo es mostrar al usuario una idea preliminar del *DW* a desarrollar, para ver el cumplimiento de los requerimientos y ajustar el modelo a las especificaciones.
- 5) **Escoger el sistema administrador de bases de datos (SABD):** el tiempo de respuesta a las consultas de los usuarios es una de las variables más importantes para medir el rendimiento de un *DW*. El SABD se debe encargar de cumplir con este tiempo de respuesta utilizando la estructura o el modelo de datos para organizar la información en el *DW*.
- 6) **Diseño físico de la base de datos:** se refiere a la implementación física del modelo lógico de datos, usando la estructura que provea el SABD (relaciones entre las tablas del modelo de datos).
- 7) **Almacenar la información a través de un proceso de extracción, transformación y carga de datos, y luego evaluar el modelo.**
- 8) **Afinar el desempeño obtenido:** dado que el tiempo de respuesta ante las consultas tiene una importancia alta, puede ser necesario realizar modificaciones a la estructura interna de los datos para mejorar el desempeño.

El diseño y construcción de un *DW* es un proceso iterativo, en el cual se repiten varias de las etapas mencionadas anteriormente, por lo tanto requiere de presentación de prototipos y de escuchar al usuario para ajustar el modelo que se quiere implementar, de manera de responder a la mayor parte de las necesidades del negocio.

Una alternativa de solución a la construcción de un *DW* son los *Data Mart (DM)*, donde el diseño se realiza de forma similar, pero a una escala menor que los anteriores, enfocándose en un área específica del negocio.

2.1.5 *Data Mart*

Un *Data Mart* se puede considerar como un *Data Warehouse* más pequeño, el cual se ha diseñado para satisfacer necesidades de un número limitado de usuarios o área específica de la empresa, a diferencia de los *Data Warehouse* donde su implementación se realiza para responder con un alcance de nivel corporativo, atendiendo a varias áreas del negocio a la vez [16] [30].

La ventaja de los *Data Mart* frente a los *Data Warehouses* tiene que ver con su implementación, ya que al ser de menor escala que los anteriores extraen información desde menos fuentes de datos y son más rápidos, fáciles y baratos de construir.

La existencia de los *Data Mart* permite explicar dos enfoques para el diseño lógico de un repositorio de datos [32]:

1. **Enfoque *Top-Down*:** propuesto por W.H. Inmon, consiste en construir un *Data Warehouse* o repositorio centralizado, y a partir de éste alimentar con información a varios *Data Marts*.
2. **Enfoque *Bottom-Up*:** propuesto por R. Kimball, consiste en construir varios *Data Marts* independientes, para posteriormente unirlos y a partir de éstos formar un *Data Warehouse*.

La ventaja que tiene el enfoque *Bottom-Up* en relación al enfoque *Top-Down* es su rapidez de implementación, ya que la construcción del proyecto *Data Warehouse* se va haciendo por etapas a medida que se construyen los *Data Mart* (no es necesario hacer todo de una vez). Se puede decir que diversifica el riesgo de falla en su implementación, y se pueden apreciar resultados o entregables mucho antes que en el caso del enfoque *Top-Down*.

Por otro lado, el enfoque *Bottom-Up* también tiene la desventaja de que necesita asegurar la consistencia entre la información que se obtiene desde los distintos *Data Marts*, por lo tanto si se utiliza este enfoque es necesario pensar en cada uno de los repositorios involucrados, en las dimensiones que tiene cada uno y en cómo se relacionan entre ellos.

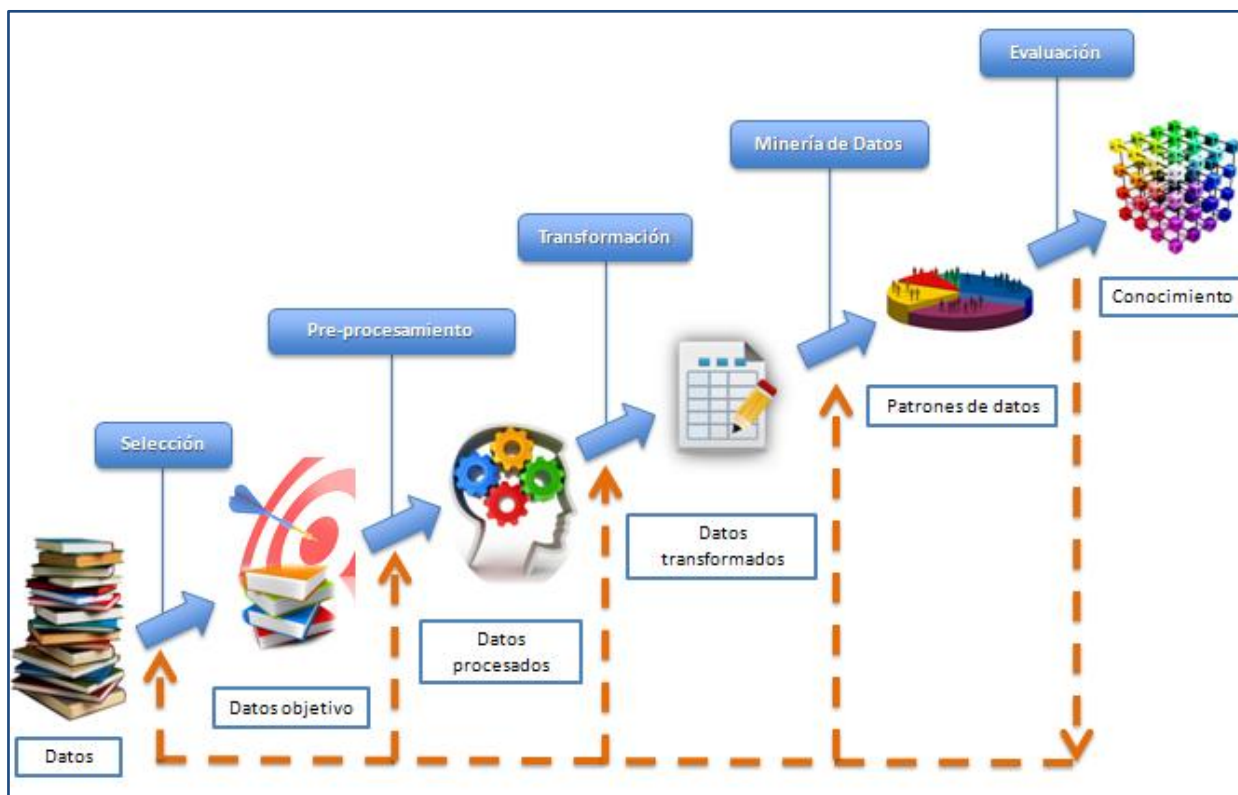
2.1.6 El proceso *KDD* para la obtención de conocimiento

KDD (*Knowledge Discovery in Databases*) es una metodología genérica para encontrar información en un gran conjunto de datos y con ello generar conocimiento. Se define como un proceso no trivial de extracción de información a partir de los datos, la cual se encuentra presente de forma implícita, previamente desconocida y potencialmente útil para el usuario o para el negocio [10] [30].

El objetivo principal de esta metodología es automatizar el procesamiento de los datos, permitiendo a los usuarios dedicar más tiempo a las tareas de análisis y al descubrimiento de relaciones entre los datos.

El *KDD* es un proceso que consta de una serie de etapas consecutivas, y funciona de forma iterativa e interactiva. Iterativa, ya que es posible regresar desde cualquier etapa a una anterior para ajustar los parámetros o supuestos previos, e interactiva pues el usuario experto del negocio tiene que estar presente para aportar con su conocimiento en la preparación de los datos y en la validación de los resultados que se obtengan durante el proceso [11].

Figura 8: Etapas del proceso *KDD*



Fuente: Elaboración propia [10]

Las etapas de este proceso, tal como se muestra en la *Figura 8*, son:

- 1) **Identificación del problema en estudio**, teniendo un objetivo claro para el problema a resolver, entendiendo las metas del proceso y cuáles son las preguntas que se quieren responder.
- 2) **Selección e integración de los datos**, para contar con un conjunto objetivo desde el cual obtener el conocimiento. Se obtienen los datos desde los sistemas operacionales, los cuales pueden venir en diferentes formatos y en algunas oportunidades con errores, por lo cual es importante realizar una etapa de procesamiento.
- 3) **Preparación de los datos (limpieza y pre-procesamiento)**, ya que en general, como se dijo en la etapa anterior, los datos provienen desde varias fuentes y en diferentes formatos. En esta etapa se escogen técnicas y estrategias para corregir errores en el conjunto de datos seleccionado, tratar la información faltante y unificar formatos.
- 4) **Transformación y almacenamiento de los datos**, punto en el que se pueden reducir o agrupar los datos en las características de interés. Se consolida la información y escoge una arquitectura acorde a las necesidades del problema que permita almacenarla, por ejemplo un *Data Mart*.
- 5) **Selección y aplicación de algoritmos de *Data Mining***, utilizando técnicas adecuadas según la hipótesis planteada y el análisis que se quiera hacer. Las técnicas seleccionadas permitirán generar modelos de minería de datos, y con ello descubrir patrones de información implícitos en los datos.
- 6) **Interpretación y evaluación** de los patrones encontrados, identificando los nuevos conocimientos y apoyándose en los expertos del negocio para ver si se pueden tomar acciones con estos resultados. Para interpretarlos, es necesario visualizarlos de diversas formas, validando los patrones y modelos de datos, documentando los procedimientos y consideraciones de manera que se generen propuestas de valor para el negocio.

Las etapas iniciales del proceso *KDD* son muy importantes porque serán la base sobre la cual se hará minería de datos. Si la preparación de los datos no está bien hecha, los resultados obtenidos en los análisis no serán confiables. Por lo tanto, hay que asegurar que se esté trabajando sobre un repositorio bien diseñado, y es por esta razón que la mayor parte de los esfuerzos se emplean en las etapas de selección y preparación de los datos. La relación entre el proceso *KDD* y los *Data Warehouses* se da de forma natural, pues el primero busca contar con datos procesados, limpios y consolidados, mientras que los segundos ofrecen una estructura bien definida en donde almacenar la información con esas características.

2.2 Minería de datos

Data Mining corresponde a una de las etapas del proceso llamado “*Knowledge Discovery in Databases*” (*KDD*) [10] [11] [13] [18] [29]. Está conformado por un conjunto de técnicas y algoritmos que sirven para hacer análisis de conjuntos de datos, extrayendo patrones y relaciones entre ellos, convirtiéndolos en información valiosa y útil para quienes toman las decisiones.

El uso potencial del *Data Mining* en las empresas es identificar nuevas oportunidades de negocio, adaptar los productos ofrecidos o encontrar los clientes más valiosos con el fin de retenerlos, y de esta manera aumentar los ingresos y reducir las pérdidas o costos. Al determinar las características de los buenos clientes (*profiling*), las empresas pueden enfocarse en aquellos de características similares y diseñar productos o servicios acordes a sus necesidades [25].

Dentro de la industria de telecomunicaciones por ejemplo, las áreas de interés donde aplicar minería de datos son: detección de fraude, asignación de recursos para instalaciones o servicios técnicos, análisis de clientes, pronósticos de demanda, proyecciones de crecimiento de la industria y predicción de fallas en la red [34]. Los algoritmos que destacan en estos casos son los de regresión, *clustering* y clasificación.

El uso de minería de datos se debe entender como un apoyo para los analistas, y no reemplaza al conocimiento que tienen los expertos del negocio, ni elimina la necesidad de entender los datos. El *Data Mining* no funciona por sí sólo, ya que los patrones que se encuentren en los datos deben ser interpretados y validados para ver si responden a las consultas del negocio, y si son aplicables en el mundo real.

2.2.1 Aprendizaje supervisado y no supervisado

El objetivo del *Data Mining*, como se ha dicho antes, es producir nuevo conocimiento para que los usuarios del negocio puedan tomar decisiones, a partir de la construcción de un modelo del mundo real y basándose en datos de diversas fuentes. La decisión del modelo de *Data Mining* a utilizar está condicionada por los objetivos del negocio, los cuales se alcanzan al diseñar y probar combinaciones de algoritmos.

Es importante notar que no existe un “mejor” modelo o algoritmo de minería de datos, depende del problema en estudio y de los datos disponibles para decir cuál entrega resultados más confiables.

Los modelos de *Data Mining* se clasifican como predictivos y descriptivos. En el primer caso, se tiene una variable con valor desconocido, y la finalidad es determinarlo. Esta variable se llama respuesta, variable dependiente u objetivo, mientras que aquellas utilizadas para hacer la predicción son los predictores o variables independientes [29].

Los modelos predictivos requieren ser “entrenados”, utilizando un conjunto de datos de entrenamiento cuyo valor de variable objetivo es conocido. La idea es que el modelo entregue resultados en base un aprendizaje, en otras palabras, que se vaya ajustando a la realidad conocida.

A este tipo de modelos se les conoce también como modelos de aprendizaje supervisado, debido a que los valores estimados o calculados son comparados con los resultados conocidos, y por lo tanto se tiene una clara medida del éxito o falla de la predicción [24] [29]. Algunos algoritmos que se utilizan en estos modelos son los de clasificación y las regresiones.

El aprendizaje supervisado se utiliza en problemas en los que se tiene conocimiento del resultado al que se quiere llegar, por ejemplo para la detección de aquellos clientes que son más propensos a la fuga de la empresa.

Por otra parte, se tienen los modelos descriptivos, en los cuales no se cuenta con un resultado conocido para poder guiar a los algoritmos, y por ello se conocen como modelos de aprendizaje no supervisado, donde el modelo se va ajustando de acuerdo a las observaciones o datos entregados, y se recurre muchas veces a argumentos heurísticos para evaluar la calidad de los resultados. Algunos algoritmos que se utilizan en estos modelos son los de *clustering* y las reglas de asociación [29].

El aprendizaje no supervisado, es usado en los casos en que no se tiene conocimiento previo del resultado al que se va a llegar, por ejemplo al segmentar a los clientes en grupos que no hayan sido definidos previamente. Luego que el modelo ya ha sido entrenado, se utiliza una muestra de datos independiente de aquella utilizada para la fase de construcción y entrenamiento del modelo, con la intención de evaluar la capacidad de predicción de éste.

2.2.2 Métodos de minería de datos

Los dos caminos principales del *Data Mining* hacen referencia a la predicción y a la descripción. Para ambos existen una variedad de métodos de minería de datos que se pueden utilizar, con el fin de descubrir conocimiento. Dentro de los métodos predictivos se encuentran la clasificación y regresión, por otra parte en los descriptivos se tienen el *clustering* y las reglas de asociación [10] [29] [36] [37].

La clasificación consiste en “mapear” un elemento dentro de un grupo de datos, de acuerdo a una clase predefinida, en otras palabras indicar a qué conjunto pertenece. Se apunta a identificar las características o atributos que hacen que un elemento se vincule a un grupo siguiendo un patrón de datos. Este último se puede utilizar para predecir cómo se comportarán nuevas instancias.

La regresión es una función que le asigna a un elemento un valor real, utilizando valores existentes para predecir datos futuros. En el caso más simple, la regresión usa técnicas estadísticas como la regresión lineal, sin embargo muchos de los problemas del mundo real no funcionan con proyecciones lineales. Las regresiones se pueden utilizar por ejemplo para predecir comportamiento de la demanda futura, utilizando las ventas o el consumo pasado.

El *clustering* divide el conjunto de datos en grupos que son muy diferentes unos de otros, pero cuyos elementos sean muy similares entre sí. Es un método descriptivo que identifica un grupo de categorías o “*clusters*” para describir los datos. Estas categorías pueden ser exclusivas, jerárquicas o superpuestas. Entre los algoritmos más conocidos de clustering se encuentran *K-means* y las redes neuronales o mapas de Kohonen.

A diferencia de la clasificación, en *clustering* no se sabe cuáles serán los grupos que se formarán según los atributos escogidos, por lo tanto, es necesario que los expertos del negocio interpreten las categorías que se formen y vean si hacen sentido o no. Una vez obtenidos los *clusters* que segmentan los datos, se pueden clasificar otros nuevos.

Es importante distinguir entre las definiciones de *clustering*, segmentación y clasificación [29]: segmentar se refiere a identificar grupos de datos que tienen características comunes, *clustering* es encontrar grupos de datos que no estaban definidos, y clasificar es asignar elementos a un grupo ya definido.

Las reglas de asociación son otro instrumento descriptivo, donde el objetivo es encontrar relaciones significativas entre los datos, utilizando probabilidades de ocurrencia de dos objetos. Un claro ejemplo es el análisis de los artículos o productos de una canasta de compras en una tienda.

2.2.3 Análisis de *clusters*

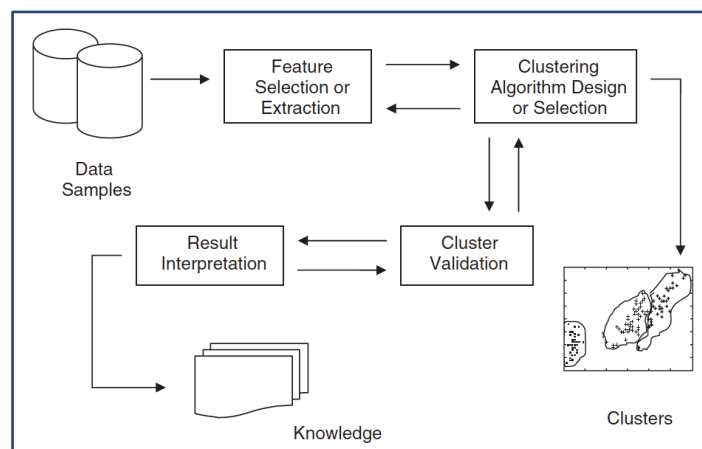
El análisis de *clusters* (o *clustering*), como se explicó en el punto anterior, tiene un objetivo principal asociado a segmentar un conjunto de datos en grupos o *clusters*, de manera que contengan objetos con características similares, pero que a su vez entre los grupos los objetos sean diferentes [20].

Cada *cluster* representa un conjunto de objetos con propiedades o atributos, algunos más importantes considerados al momento de hacer la segmentación. Los métodos de *clustering* intentan agrupar los objetos basándose en una “medida de similitud”, similar a lo que se hace con las medidas de éxito (funciones de costos) en los métodos de aprendizaje supervisado.

El análisis de *clusters*, tal y como se muestra en la *Figura 9*, se compone de las siguientes cuatro etapas [38]:

1. **Selección o extracción de atributos**, donde se escogen los atributos principales con los que se quiere hacer *clustering*.
2. **Selección del algoritmo**, eligiendo el criterio de similitud adecuado.
3. **Validación de los *clusters***, ya que dependiendo del algoritmo y los parámetros utilizados, se pueden obtener cantidad y composición de *clusters* distintos.
4. **Interpretación de los resultados**, con apoyo de expertos del negocio para entregar a los usuarios finales puntos de vista que tengan sentido.

Figura 9: Etapas del análisis de clusters



Fuente: *Clustering (Wiley) [38]*

2.2.4 Medidas de similitud

Para todas las técnicas de *clustering* resulta válido preguntar cómo se determina la proximidad entre los datos, en otras palabras, cómo se mide la distancia entre un par de objetos. Para ello, se explicarán a continuación las medidas de similitud y conceptos utilizados en los análisis de *clusters*.

Las medidas de proximidad o similitud se aplican sobre objetos, los cuales están compuestos por un conjunto de características, generalmente representadas por un vector multidimensional. Sea N el número de objetos, cada uno con M características, entonces el conjunto de datos se puede representar en una matriz de tamaño $(N \times M)$, donde cada fila corresponde a un objeto y cada columna a un atributo [37].

Por cada atributo o variable, dado este conjunto de N objetos, se puede generar una matriz simétrica de $(N \times N)$, llamada matriz de proximidad (S), cuyos elementos S_{ij} representan la comparación (medida de proximidad) entre los objetos x_i y x_j para un mismo atributo (con $i, j \in \{1, \dots, N\}$). Esta matriz cumple con tener todos los elementos no negativos, y además $S_{ii} = 0 ; \forall i = \{1, \dots, N\}$.

Una medida de similitud debe satisfacer las siguientes propiedades [36] [37]:

- 1) **Simetría:** $S_{ij} = S_{ji}$
- 2) **Positividad:** $S_{ij} \geq 0 ; \forall i, j$
- 3) **Desigualdad triangular:** $S_{ij} \leq S_{ik} + S_{kj} ; \forall i, k, j$
- 4) **Reflexividad:** $S_{ij} = 0$ si y solo si $x_i = x_j$

Una variable puede ser clasificada como continua, discreta o binaria dependiendo de los valores que esta tome. Además, poseen la propiedad llamada nivel de medida, lo cual las permite separar en los siguientes tipos [37]:

- a) **Variabes nominales o categóricas:** son atributos que pueden ser representados por números, sin embargo no tienen un orden o significado matemático intrínseco [4]. Por ejemplo, el atributo “color” que puede tomar los valores 1 = rojo, 2 = azul y 3 = verde.
- b) **Variabes ordinales:** son atributos similares a las variables nominales, la diferencia está en que existe un orden o jerarquía en sus valores. Por ejemplo, el atributo “tamaño” con valores 1 = pequeño, 2 = mediano y 3 = grande.

- c) **Variabes cuantitativas:** en este caso, los valores son números con algún significado matemático, por lo tanto si tiene sentido la operación matemática entre los valores, pues entrega un resultado interpretable.

Existen variadas medidas de similitud (ver Tabla 7), y dependiendo del tipo de variable con las que se esté trabajando se utilizará una u otra. En general, las medidas que se asemejan a funciones de distancia se utilizan con variables continuas y cuantitativas, mientras que para variables cualitativas es necesario utilizar otro tipo de medidas de similitud.

Tabla 7: Medidas de similitud para variables continuas

Medida	Fórmula
Distancia de Minkowski	$S_{ij} = \left(\sum_{m=1}^M x_i^m - x_j^m ^n \right)^{\frac{1}{n}}$
Distancia Euclideana	$S_{ij} = \sqrt{\left(\sum_{m=1}^M x_i^m - x_j^m ^2 \right)}$
Distancia City-Block	$S_{ij} = \left(\sum_{m=1}^M x_i^m - x_j^m \right)$
Distancia Sup	$S_{ij} = \max_{1 < j < M} x_i^m - x_j^m $
Similitud Coseno	$S_{ij} = \cos \alpha = \frac{x_i^T x_j}{ x_i x_j }$

Fuente: Clustering (Wiley) [38]

Tomando la definición anterior, se define una medida de similitud entre dos objetos (i, j) como una función monótona creciente de las similitudes entre sus M atributos. Esta función puede considerar ponderadores o factores para darle mayor importancia a un atributo [37]:

$$S_{ij} = S(x_i, x_j) = F(S_{ij}^m) ; m \in \{1, \dots, M\} \quad (2.1)$$

Para variables cuantitativas continuas, la medida más comúnmente utilizada es la distancia Euclideana, caso especial de la familia de métricas llamada “distancia de Minkowski”, y se define como [36] [37]:

$$S_{ij} = \left(\sum_{m=1}^M |x_i^m - x_j^m|^p \right)^{\frac{1}{p}} ; \text{con } p = 2 \quad (2.2)$$

Cuando se trata de variables categóricas, se pueden utilizar otras definiciones, por ejemplo ver si los pares de objetos coinciden o no [4]:

$$S_{ij} = \frac{1}{M} \sum_{m=1}^M S_{ij}^m \quad (2.3)$$

, donde:

$$S_{ij}^m = \begin{cases} 1 & \text{si } x_i = x_j \text{ en el atributo } m \\ 0 & \text{si } x_i \neq x_j \text{ en el atributo } m \end{cases} \quad (2.4)$$

En el caso anterior, si se quieren hacer comparaciones utilizando el concepto de “disimilitud”, esto es posible haciendo uso de la conversión $D_{ij} = 1 - S_{ij}$

2.2.5 Determinación del número de *clusters*

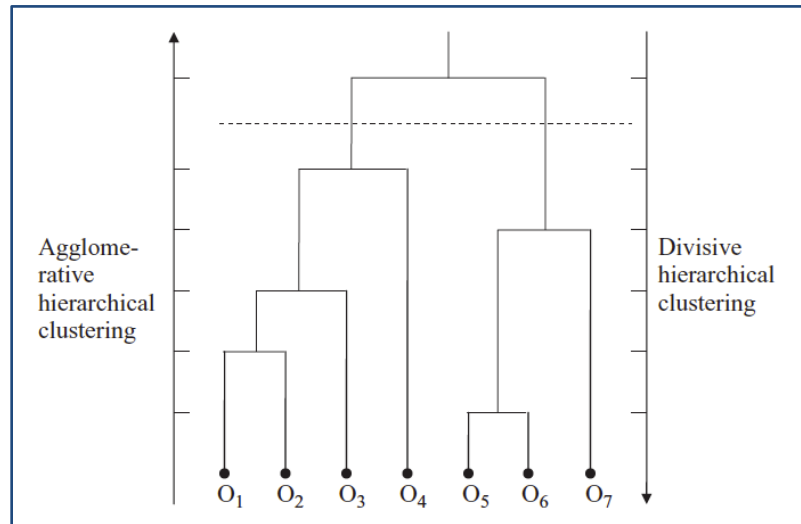
La finalidad del *clustering* es encontrar grupos de datos con características comunes, tarea que no es sencilla pues no se sabe a priori cuántos de ellos se formarán. Esto es un problema fundamental en los problemas de clasificación no supervisada.

El número de *clusters* a determinar se asocia a la letra K , y para obtenerlo existen variadas soluciones basadas en índices y criterios [23] [27], utilizando la naturaleza de los datos o el juicio de un experto del negocio. Según lo anterior, se han definido dos tipos de algoritmos de *clustering* [20]:

- 1) Aquellos a los que se les debe indicar como parámetro el número de *cluster* a obtener para que se pueda aplicar a la muestra de datos. Un ejemplo de esto son los algoritmos de *clustering* particional como el conocido *K-means*.
- 2) Otros donde la cantidad de *clusters* es desconocida, y se obtendrá a partir de la similitud entre los datos, como es el caso de los algoritmos de *clustering* jerárquico. Estos últimos poseen dos métodos o formas de trabajar [38]:
 - a) **Método aglomerativo**, que define cada objeto como un *cluster*, y luego en cada iteración del algoritmo se van mezclando o agrupando aquellos objetos que se asemejan.
 - b) **Método divisivo**, que parte por el conjunto total de objetos como un único *cluster*, y luego en cada iteración del algoritmo efectúa divisiones, separando los datos hasta llegar a un criterio de corte.

Los últimos dos métodos mencionados organizan los datos de forma jerárquica, en base al uso de la matriz de proximidad y criterios de asignación [38]. Los resultados de estos métodos se pueden ver gráficamente en un árbol binario o en un dendograma (ver ejemplo en Figura 10), donde se puede hacer un “corte” en las ramas, obteniendo el número de *clusters*.

Figura 10: Dendograma para determinar el número de *clusters*.



Fuente: *Clustering (Wiley) [38]*

Otra forma de determinar el número de *clusters*, es utilizando algoritmos probabilísticos como *EM (Expectation-Maximization)* [9] [36], el cual busca a través de varias iteraciones la máxima verosimilitud, es decir, la función de distribución que se ajuste mejor a la muestra de datos. Este algoritmo funciona en dos etapas: la primera llamada expectativa (E) donde crea una función de verosimilitud a partir de parámetros estimados hasta ese momento, y luego una segunda etapa de maximización (M), donde se calculan los parámetros que maximizan el resultado de la función, obteniendo probabilidades de pertenencia de los objetos a cada *cluster*.

Una característica importante de este algoritmo es que no toma en cuenta la distorsión dentro de un *cluster*, por lo tanto los resultados obtenidos pueden diferir en relación a los de un algoritmo de *clustering* particional [27].

Cabe destacar que, si el conjunto de datos es homogéneo y no tiene una estructura que permita separarlos en grupos con características similares, entonces el resultado final de la aplicación de algoritmos de *clustering* no tendrá significado relevante, ya sea que se indique el número de *clusters* a obtener inicialmente o no.

2.2.6 Validación de *clusters*

Dependiendo del algoritmo y los parámetros utilizados se pueden obtener cantidad y composición de *clusters* distintos, por lo tanto es importante evaluar la calidad del resultado obtenido. Para esto se han definido índices o medidas, que de acuerdo al tipo de algoritmo y característica a evaluar se han clasificado de la siguiente manera [12] [20]:

- a) **Índices externos o supervisados:** se basan en una estructura o clasificación pre-definida de los datos, y se compara ésta con el resultado obtenido al aplicar el algoritmo de *clustering*. Son llamados externos pues utilizan información que no está presente en la muestra de datos. Algunos índices externos son [38]:
 - i. Rand index (1971)
 - ii. Jaccard Coefficient
 - iii. Fowlkes & Mallows index (1983)

- b) **Índices internos o no supervisados:** no utilizan información externa, y se basan únicamente los datos de la muestra. Estos índices se clasifican adicionalmente de dos maneras, aquellos que miden el nivel de cohesión de los *clusters*, es decir cuán cercanos son los objetos dentro de un mismo conjunto, y aquellos que miden el nivel de separación de los *clusters*, o cuán distintos son los objetos entre un *cluster* y otro. Algunos índices internos son [3] [38]:
 - i. Davies-bouldin index
 - ii. Dunn index

- c) **Índices relativos:** buscan comparar diferentes estructuras de clusterización, es decir, los resultados obtenidos de aplicar varios algoritmos de *clustering* sobre una misma muestra de datos, determinando cual es el que entrega el mejor resultado, por ejemplo utilizando diferentes valores para el parámetro K en algoritmos como *K-means*.

Otra manera de comparar el resultado del *clustering* de los datos al aplicar diferentes valores del parámetro K es utilizando “tablas de contingencia”, las cuales permiten ver y analizar la relación entre los datos de dos grupos diferentes.

2.2.7 Modelos y algoritmos de minería de datos

En este punto se presentan algunos de los modelos y algoritmos de minería de datos utilizados para extraer conocimiento desde los datos. Estos algoritmos son genéricos, pudiendo existir variantes de cada uno, ya que se van adaptando, combinando o incluyendo mejoras dependiendo del tipo de problema en estudio.

Como se dijo anteriormente, no hay un “mejor” algoritmo, sino que existen una variedad de ellos que se pueden ir probando para encontrar el que se ajuste mejor a los datos.

2.2.7.1 K-means

Es un algoritmo de *clustering* particional, uno de los métodos de *clustering* más conocidos y utilizados cuando todas las variables son de tipo cuantitativo. Funciona de forma iterativa, dividiendo óptimamente el conjunto inicial de datos en un número K de *clusters*, el cual se indica como parámetro [36].

El algoritmo *k-means* trabaja realizando los siguientes pasos [38]:

Paso 0: Inicialización del algoritmo: se escogen o determinan los K centroides. Estos pueden elegirse de manera arbitraria, aleatoria o de acuerdo a algún conocimiento previo de los datos.

Paso 1: Asignación de los datos: se forman los *clusters*, asociando cada objeto al centroide “más cercano”, de acuerdo a una medida de proximidad utilizada. La medida más común es la distancia Euclidiana, y lo que hace el algoritmo es minimizar la varianza de los objetos que se encuentran en el mismo *cluster*.

Paso 2: Actualización de los centroides: a partir de la asignación de datos en cada *cluster*, se calcula la media de los valores con lo cual se obtiene un nuevo centroide.

La convergencia del algoritmo es lineal, y se produce cuando las asignaciones de los N objetos ya no sufren cambios, es decir, cuando los centroides no cambian. Este algoritmo realiza $N \times K$ comparaciones en cada iteración.

Tiene la ventaja de ser simple, no le afecta el orden de la muestra y se basa en el análisis de varianzas entre los datos. Sin embargo, tiene algunas desventajas, como su

sensibilidad ante la asignación inicial de los centroides y a la presencia de *outliers*. Para minimizar ese riesgo, se puede realizar un procesamiento previo de los datos con el fin de remover los *outliers*. También, puede ser de utilidad revisar la asignación final de *clusters* por si alguno es muy pequeño o poco significativo.

2.2.7.2 Árboles de decisión

Corresponde a uno de los métodos inductivos de aprendizaje supervisado, el cual realiza divisiones sucesivas del conjunto de datos, utilizando algún criterio de selección, manteniendo organizada su estructura de forma jerárquica, con el fin de maximizar la distancia entre los grupos de datos generados en cada iteración [11] [28].

Son una manera de representar una serie de reglas que llevan hacia una clase o valor de los datos, y se utilizan para examinarlos y realizar predicciones.

Los árboles de decisión poseen una estructura formada por [29]:

- a) **Nodos**, que corresponden a los nombres o identificadores de los atributos que caracterizan al conjunto de datos. El nodo inicial o nodo raíz contiene la muestra total de atributos que definen a los datos.
- b) **Ramas**, representan a las variables de decisión o las condiciones que cumplen los objetos para separarse unos de otros.
- c) **Hojas**, que son finalmente los conjuntos o grupos de datos resultantes de la división que realiza el algoritmo.

El algoritmo realiza una clasificación discreta de los objetos, determinando a qué clase pertenece, mediante la decisión de qué rama escoger. Para esto, se asume que los grupos o clases que se formarán serán disjuntas, es decir, una instancia u objeto no puede pertenecer a dos clases a la vez. Esta misma condición se cumplirá para cada partición o sub-árbol que se forme, característica particular que tienen los árboles de decisión conocida como propiedad exhaustiva [13].

Existen diversos algoritmos de aprendizaje que se pueden utilizar para obtener un árbol de decisión. El algoritmo utilizado puede determinar aspectos como la compatibilidad con el tipo de variables de entrada y salida, el procedimiento de evaluación de la distancia entre los grupos generados en cada división, y también la cantidad de ramas que se obtengan cada vez que un nodo se divide.

Por ejemplo, con respecto al último punto mencionado, si se utiliza el algoritmo llamado *CART* (*Classification And Regression Trees*), se pueden obtener árboles con sólo dos ramas por cada división de los nodos, y es por esta razón que se les llaman árboles binarios [29].

Existen otros algoritmos que se pueden utilizar para la construcción de un árbol de decisión, tales como el algoritmo *CHAID* (*Chi-Squared Automatic Interaction Detection*), el *Quest*, el *C5.0*, etc.

Una de las ventajas de utilizar los árboles de decisión es que funcionan muy bien con variables categóricas, evitando realizar transformaciones de los datos. Otra ventaja de su uso es que permiten interpretar de buena manera las decisiones tomadas por el modelo para sus predicciones, cosa que en algoritmos como las redes neuronales no es posible deducir.

Desventajas de los árboles de decisión es que son sensibles ante pequeños cambios en los datos, y además, dado que las decisiones para clasificar se realizan considerando una variable predictora a la vez, es difícil detectar las posibles relaciones entre los atributos y se pueden llegar a omitir algunos [29].

En la *Figura 11* se presenta un ejemplo de árbol de decisión, para el caso un banco que está evaluando si otorgarle el crédito a un cliente o no. Para ello se consideran atributos de los clientes para ver su nivel de riesgo, y según esto entregarle el crédito.

Figura 11: Árbol de decisión para evaluar riesgo de un cliente



Fuente: Elaboración propia

2.2.7.3 Two-Step Cluster

Este algoritmo de *clustering*, a diferencia de muchos otros, fue diseñado para operar sobre conjuntos de datos muy grandes, y en comparación al algoritmo *K-means* tiene la ventaja de trabajar con variables continuas y categóricas, sin afectar la robustez de sus resultados.

Como su nombre lo dice, el algoritmo realiza dos etapas [19]:

Paso 1: Formación de *pre-clusters*: se subdivide la muestra de datos en varios conjuntos, aplicando un algoritmo no jerárquico llamado *CF-Tree* sobre las variables categóricas o nominales. El algoritmo en esta etapa decide, de acuerdo al criterio de similitud escogido, si un elemento se parece a alguno de los *clusters* ya formados para unirlos a ese grupo, o si debe formar un nuevo *cluster*.

El *CF-Tree* es en un árbol de decisión de altura balanceada, que almacena los subconjuntos de datos en sus nodos.

Paso 2: Agrupación de los datos: a partir de los conjuntos generados en la etapa anterior, se forman nuevos grupos de datos que serán el resultado del *clustering*. Para ello, se utilizan criterios de corte sobre las variables continuas, los cuales pueden ser *AIC* (*Akaike Information Criterion*) o *BIC* (*Bayesian Inference Criterion*) [38].

2.2.7.4 Redes Neuronales Artificiales (RNA)

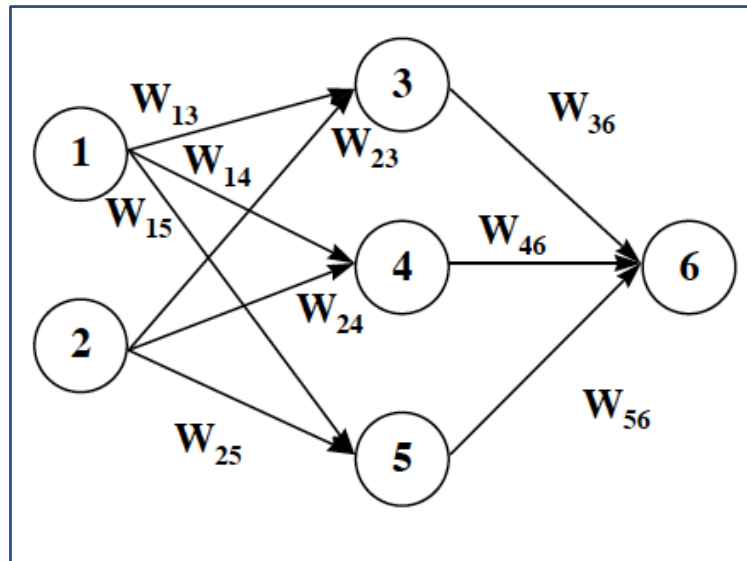
Las redes neuronales son sistemas de procesamiento de datos, cuya estructura y diseño se basa en el proceso natural del funcionamiento del cerebro. Son muy interesantes de estudiar, ya que permiten modelar eficientemente problemas complejos, en los cuales se cuenta con muchas variables predictoras [30] [32].

Las redes neuronales se pueden utilizar tanto para problemas de clasificación (cuando se utilizan variables categóricas), como para problemas de regresión (cuando se cuenta con variables continuas).

La estructura de una red neuronal está compuesta por un set de nodos interconectados, transmitiendo “señales” a través de sus conexiones, y cada una de esas conexiones o enlaces tiene un peso asociado [29].

La red neuronal funciona “aprendiendo de ejemplos”, por lo que necesita un conjunto de datos iniciales, presentados en una capa de entrada, donde cada nodo de esa capa corresponde a una variable predictora. Estos nodos iniciales se conectan con otros nodos en una capa oculta, donde se les aplica una “función de activación” utilizando los pesos que tienen asociadas las conexiones entre los nodos. La *Figura 12* muestra un ejemplo de red, con nodos interconectados y pesos asociados.

Figura 12: Neuronal con pesos asociados a cada nodo



Fuente: “Introduction to Data Mining and Knowledge Discovery in Databases” [29]

Los resultados obtenidos mediante la función de activación se traspasan de un nodo a otro hacia una capa de salida, que consiste en una o más variables de respuesta del modelo.

En la *Figura 12*, el valor que se le entrega al nodo 6 corresponde a una composición de los valores ponderados obtenidos desde el nodo 1 y 2, es decir:

$$W_{14} * ValorNodo_1 + W_{24} * ValorNodo_2 \quad (2.5)$$

Los pesos en la red son parámetros desconocidos del modelo, y se estiman a través del aprendizaje y métodos de entrenamiento. Existen varios métodos de entrenamiento, uno de los más conocidos es el llamado "*feed-forward backpropagation*", y se realiza en dos pasos [29]:

Paso 1: Feed-forward: el valor de los nodos de salida se calcula en base a los nodos de entrada y pesos iniciales de la red, combinados en la capa oculta de la red.

Paso 2: *Backpropagation*: se calcula el error según el valor obtenido y el valor esperado, luego se utiliza el error para asociarlo a los nodos de la capa oculta, proporcionalmente a sus pesos. Este proceso permite ir ajustando los pesos de las conexiones en la red con el fin de reducir el error.

El proceso anterior se aplica sobre un conjunto de datos de entrenamiento, y se realizará reiterativamente en etapas o “épocas”, hasta que el error sea aceptable, es decir, el resultado que entrega el modelo es correcto. Esto puede ser a veces un problema en estos modelos si se sobre-ajustan mucho a los datos.

La ventaja más relevante de las redes neuronales es su buen funcionamiento predictivo, ya que tienen una alta tolerancia a los datos anómalos y capturan muy bien la relación entre las variables y el resultado entregado [29]. Sin embargo, a diferencia de modelos de clasificación como los árboles de decisión, es difícil interpretar las decisiones tomadas por el modelo para llegar a sus resultados, ya que funciona como una “caja negra”.

Existe un tipo especial de red neuronal que no utiliza aprendizaje supervisado, sino otro tipo de aprendizaje llamado aprendizaje competitivo, pues en este modelo las neuronas de la red compiten entre ellas para ser “activadas”, mientras que el resto de las neuronas son inhibidas [33].

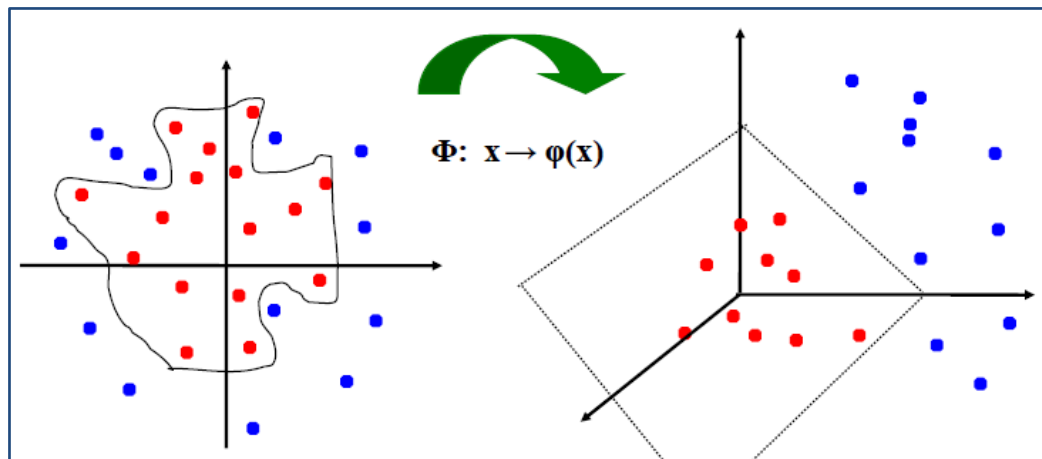
2.2.7.5 Support Vector Machines (SVM)

Es uno de los métodos de minería de datos más robustos y precisos, cuyo uso se ha vuelto muy popular para resolver problemas de clasificación y regresión. Su objetivo es encontrar la mejor función para clasificar un conjunto de datos, encontrando los hiperplanos que mejor dividan la muestra, maximizando el grado de separación entre las clases generadas [30].

La ventaja de utilizar este método con respecto a las redes neuronales, es que resuelve el problema que tienen estas últimas de llegar a soluciones locales, ya que *SVM* siempre encuentra una solución global luego de haber sido entrenado, y es menos propenso a sobre ajustarse a los datos [33].

Esta técnica se basa en transformar los datos de entrada, desde un espacio de baja dimensión hacia uno dimensionalmente mayor (*ver Figura 13*). Lo anterior se realiza a partir de la elección de una función de *kernel* (polinomial), que se aplica sobre los datos, buscando los parámetros del modelo a través de programación cuadrática. La medida utilizada para encontrar la mejor función es maximizar el margen o la distancia entre los objetos de dos clases [5] [24].

Figura 13: Transformación del espacio dimensional los datos



Fuente: Apuntes del curso Web Mining [33]

Lo que busca este método es encontrar el vector de pesos w y el de parámetros b , tales que resuelvan el siguiente problema de optimización [5]:

$$\max_{y_i(w * x_i + b) \geq 1 ; \forall x_i} \frac{2}{\|w\|} \quad (2.6)$$

s.a

$$(w * x_i + b) \geq 1 ; \forall x_i \text{ donde } y_i = 1 \quad (2.7)$$

$$(w * x_i + b) \leq -1 ; \forall x_i \text{ donde } y_i = -1 \quad (2.8)$$

El supuesto que tiene este modelo es que los datos se generan a partir de una función de distribución de probabilidad conjunta desconocida, y la meta es encontrar, a partir de los datos de la muestra la función que los clasifique correctamente.

A pesar que el espacio sobre el cual se está trabajando tiene una dimensionalidad alta, y por lo tanto tendría muchos parámetros que estimar, SVM no se ve afectado pues el objetivo fundamental es la flexibilidad del modelo y no la estimación de los parámetros.

El objetivo final es obtener un clasificador lo más flexible posible, aumentando la complejidad del modelo y reduciendo al mínimo la suma de los errores de los datos utilizados para el entrenamiento.

2.2.8 Cross Validation

Para poder validar el comportamiento de los modelos generados con los diferentes algoritmos de minería de datos explicados anteriormente, se puede utilizar la técnica llamada validación cruzada. Esta técnica funciona de la siguiente manera [35]:

- a) Se escoge un número fijo de particiones en las cuales se dividirá la muestra de datos en partes iguales, correspondiente a la cantidad de validaciones que se quiera realizar.
- b) En cada iteración, se escoge una de las particiones generadas para utilizarla como datos de prueba, mientras que el resto de los datos será usado para entrenamiento del modelo de minería de datos que se quiere validar.
- c) El paso anterior se repite con cada una de las particiones, es decir, se aplica el mismo análisis de forma iterativa a la muestra total de datos.
- d) El resultado final es un promedio de los resultados obtenidos con cada partición, generando un vector de *performance* con indicadores agregados que permiten estimar qué tan bien funcionará el modelo de minería de datos ante nuevas instancias.

3. Metodología

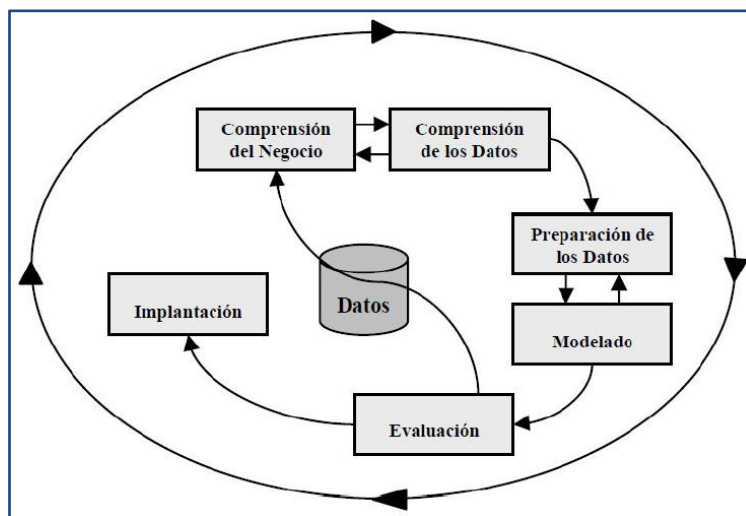
En este capítulo se explica la metodología que se utilizó para mejorar el proceso de control de gestión de ingresos en ENTEL, aplicando técnicas de minería de datos. En primer lugar, se describen las consideraciones a tener para aplicar esta metodología, y luego se ven en detalle cada una de esas etapas, desde el tratamiento y procesamiento de los datos, hasta la selección de los modelos y algoritmos de *Data Mining*.

3.1 Descripción de metodologías existentes

Una metodología de trabajo conocida es la llamada *CRISP-DM* (*Cross Industry Standard Process for Data Mining*) [7], que pretende estandarizar los proyectos de minería de datos, cuyo enfoque es obtener el mejor provecho del uso de *Data Mining* al entender de la manera más completa posible el negocio y el problema que se desea resolver. Lo anterior permite hacer una correcta recolección de datos e interpretar bien los resultados de los análisis, alcanzando los objetivos que se hayan propuesto.

CRISP-DM organiza el desarrollo de un proyecto de *Data Mining* en una serie de fases o etapas, con tareas generales y específicas que permitan cumplir con los objetivos del proyecto. Estas fases funcionan de manera cíclica e iterativa, pudiendo regresar desde alguna fase a otra anterior. En la *Figura 14* se presenta un diagrama con las fases propuestas por esta metodología.

Figura 14: Fases del modelo CRISP-DM



Fuente: "CRISP-DM 1.0: Step-by-step Data Mining guide" [7]

Los pasos a seguir en la metodología *CRISP-DM* se resumen en los siguientes cuatro puntos [1] [7]: comprensión del negocio, comprensión y preparación de los datos, modelamiento y evaluación, y despliegue del proyecto.

Se puede decir que las etapas de esta metodología están relacionadas de alguna manera con las del proceso *KDD*, incluso se puede llegar a considerar *CRISP-DM* como una implementación del proceso *KDD*. La analogía entre los pasos que se presentan en ambos casos se puede ver en la siguiente tabla:

Tabla 8: Analogía entre las etapas del proceso KDD y las de CRISP-DM

<i>KDD</i>	<i>CRISP-DM</i>
Identificación del problema en estudio	Comprensión del negocio
Selección e integración de los datos	Comprensión de los datos
Limpieza y pre-procesamiento de los datos	
Transformación de los datos	Preparación de los datos
Selección y aplicación de <i>Data Mining</i>	Modelamiento y evaluación
Interpretación y evaluación	
Post KDD	Despliegue del proyecto

Fuente: Elaboración propia, basado en paper KDD, SEMMA AND CRISP-DM [1]

Además de metodologías como *CRISP-DM*, enfocadas en proyectos de *Data Mining*, existen metodologías para el diseño e implementación de repositorios de datos, ya sean *Data Warehouses* o *Data Marts*, sobre los cuales aplicar técnicas de *Data Mining* para extraer conocimiento de los datos que almacenan.

Una metodología posible es la propuesta por *R. Kimball* [16] [30], la cual consiste en que el proyecto se realice por partes, dividiéndolo para diseñar y construir varios *Data Marts* que en conjunto formen un *Data Warehouse*. Esta manera de trabajar asegura resultados rápidos al ir realizando cada una de estas partes en un tiempo menor al que tomaría desarrollar el proyecto completo de una vez.

En particular, para el diseño y construcción de un *Data Warehouse*, las metodologías existentes tienen en común las siguientes etapas [30]:

1.- Analizar las necesidades del usuario final: se debe entregar información de tal manera que aplicaciones y herramientas de análisis puedan obtener información útil, satisfaciendo así la mayor parte de las necesidades del negocio.

2.- Seleccionar las fuentes de información: una vez que se definió la necesidad de los usuarios, entonces se buscan las fuentes de información que permitan cubrir esto.

3.- Desarrollar un modelo lógico de datos: entre los modelos más utilizados se encuentran el cubo multidimensional y el modelo estrella.

4.- Preparar un prototipo para el usuario final: el objetivo es mostrar al usuario una idea preliminar del *Data Warehouse* a desarrollar, para ver el cumplimiento de los requerimientos y ajustar el modelo a las especificaciones.

5.- Escoger el sistema administrador de bases de datos (SABD): el tiempo de respuesta a las consultas de los usuarios es una de las variables más importantes para medir el rendimiento de un *Data Warehouse*. El SABD se debe encargar de cumplir con este tiempo de respuesta utilizando la estructura o el modelo de datos para organizar la información en el *Data Warehouse*.

6.- Diseño físico de la base de datos: se refiere a la implementación física del modelo lógico de datos, usando la estructura que provea el SABD (relaciones entre las tablas del modelo de datos).

7.- Almacenar la información a través de un proceso de extracción, transformación y carga de datos, y luego evaluar el modelo.

8.- Afinar el desempeño obtenido: dado que el tiempo de respuesta ante las consultas tiene una importancia alta, puede ser necesario realizar modificaciones a la estructura interna de los datos para mejorar el desempeño.

3.1.1 Compresión del negocio

Lo primero es comprender o definir el problema del negocio, lo cual es quizás el paso más importante de la metodología pues permite entender los objetivos y requisitos que tendrá el proyecto. Si no se comprende bien el problema a resolver, no sabremos qué algoritmo de *Data Mining* utilizar ni cómo sacarle provecho a las técnicas y herramientas disponibles.

Para lo anterior, se trató con los expertos del área en estudio de la empresa, con el fin de tener una visión o perspectiva empresarial, entendiendo qué es lo que hacían actualmente y cuáles eran sus necesidades.

El problema detectado se encuentra inmerso en un proceso de negocio de la empresa (ver *Figura 15*), la provisión de Servicios Privados. En este proceso se pueden identificar o vislumbrar problemas de gestión que tienen alusión al control interno que realiza la empresa, con enfoque en que se declaren en los sistemas comerciales de ENTEL todos los servicios privados que ya hayan sido instalados, ya que si estos

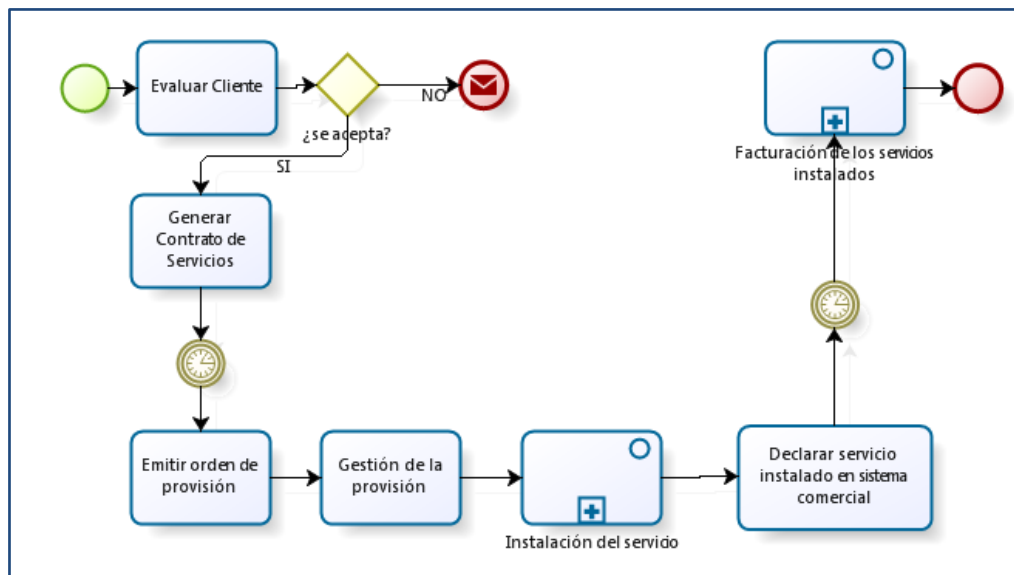
servicios no se encuentran declarados en el sistema, entonces no se puede continuar con el proceso de facturación y por lo tanto percibir ingresos por el trabajo realizado.

Como todo proceso de negocio, cuenta con un conjunto de compromisos, acciones y decisiones necesarias para satisfacer los requerimientos de sus clientes. Las actividades presentadas en este proceso son medibles y estructuradas, de tal forma que puedan producir un resultado específico.

El área de Aseguramiento de Ingresos se ha encargado de implementar el uso de indicadores en etapas críticas de este proceso, generando informes destinados a medir y controlar a todas las áreas involucradas, utilizando información obtenida desde múltiples fuentes de datos.

Existen necesidades y oportunidades en este proceso, las que se pretenden transformar a soluciones que puedan ser implantadas en ENTEL haciendo uso de sistemas y tecnologías de información, generando valor y conocimiento para la organización [6]. Encontrar estas oportunidades significa entender las operaciones que realiza actualmente la empresa y ver en qué punto existen posibilidades de cambio.

Figura 15: Flujo del proceso de provisión de servicios privados en ENTEL.



Fuente: Elaboración Propia

Los principales desafíos encontrados fueron:

- a) Existencia de una fuga de ingresos, debido a instalaciones de servicios técnicos no declarados comercialmente y por consiguiente no facturados cuando corresponde.
- b) Los indicadores utilizados en el área de Aseguramiento de Ingresos no permiten medir las pérdidas monetarias por instalaciones no declaradas. A pesar de que miden el cumplimiento de las metas de las áreas que trabajan en la provisión de servicios, no sirven para dimensionar los volúmenes de dinero que se están manejando, pues no consideran en su diseño variables asociadas a los ingresos.
- c) El procesamiento de los datos se realiza manualmente, consumiendo mucho tiempo y recursos para el área que realiza las tareas de gestión.

3.1.2 Comprensión de los datos

Teniendo definido el problema y el proceso de negocio, se comienza la búsqueda de los datos a los cuales aplicar minería de datos. Para lo anterior se realiza un levantamiento de la información y de las variables que se utilizarán para la generación de los indicadores del proceso.

Este punto comprende una selección de las fuentes y recolección inicial de los datos, para luego identificar su calidad y establecer las relaciones entre ellos. Esta es una de las fases de la metodología *CRISP-DM* que demanda más tiempo y esfuerzo para la realización del proyecto de *Data Mining*.

Las principales tareas que se realizan en esta etapa son [7]:

- a) **Recolección de datos**, teniendo claro desde qué lugar fueron obtenidos.
- b) **Descripción de los datos**, estableciendo los volúmenes de información con que se trabajará, la cantidad de registros, y los significados de cada campo o variable y los formatos en los que se encuentran.
- c) **Exploración de los datos**, indicando una estructura general de la información, comprobar frecuencia y distribución de los datos.
- d) **Verificación de la calidad de los datos**, determinando la consistencia de los valores, comprobando existencia de datos nulos y fuera de rango, identificando irregularidades para asegurar la completitud y exactitud de los datos.

3.1.3 Preparación de los datos

Una vez que se cuenta con toda la información necesaria y que se han hecho los análisis y validaciones necesarias para asegurar la calidad de los datos, comienza la etapa de preparación de los datos para adaptarlos a las técnicas de *Data Mining* que se utilicen posteriormente.

Esta etapa considera tareas generales de selección de datos, posible limpieza de la información, generación de variables adicionales, cambios en los formatos e integración de las diversas fuentes de datos [17].

La preparación de los datos está relacionada con el modelamiento y la aplicación de técnicas de *Data Mining*, puesto que dependiendo de la técnica escogida se requerirá que los datos se encuentren procesados de una u otra forma.

Los pasos a seguir en esta fase son los siguientes [7]:

- a) **Selección de datos**, escogiendo un subconjunto de los datos recopilados en la etapa anterior.
- b) **Limpieza de los datos**, preparándolos para la fase de modelación, ya sea aplicando técnicas de normalización, discretización de campos numéricos, tratamiento de valores nulos, etc. Este paso es uno de los que mayor tiempo consume dentro de la preparación de los datos.
- c) **Estructuración de los datos**, con lo cual se pueden generar nuevos atributos a partir de los existentes o transformar valores de los atributos con que se cuenta.
- d) **Integración de los datos**, lo que permite agrupar tablas o campos que se encuentren relacionadas, definiendo una estructura que las pueda contener.
- e) **Formateo de los datos**, que consiste en transformar los datos sin modificar su significado, para que se puedan ajustar a las técnicas de *Data Mining* que se quiera utilizar.

3.1.4 Modelamiento y evaluación

En esta fase se escogen las técnicas de modelado que sean más apropiadas para el problema a resolver. Una vez escogida la técnica se debe asegurar que se disponga de los datos necesarios y en el formato que la técnica lo necesite, lo cual se realizó en la etapa anterior.

Elegir una técnica de *Data Mining* para aplicar a un problema no es sencillo, pues debe cumplir los requisitos del problema en estudio y se tiene que tener un conocimiento del funcionamiento de la técnica, por ejemplo de los parámetros que se tengan que utilizar o ajustar.

Para la evaluación de los modelos obtenidos se debe determinar previamente algún método de validación que permita establecer el grado de confiabilidad de los resultados del modelo.

En resumen, en este punto de modelamiento y evaluación se realizan los siguientes pasos [7]:

- a) **Selección de la técnica de modelado**, escogiendo la más apropiada para el tipo de problema, en base al objetivo principal del proyecto. Por ejemplo, si el problema fuera de clasificación se pueden utilizar árboles de decisión o redes neuronales, en cambio si el problema fuera de *clustering* se pueden utilizar otras técnicas como *K-means*.
- b) **Generación del plan de prueba**, diseñando un procedimiento para probar y validar el modelo a obtener. En general, se separa el conjunto de datos en dos: una parte de los datos destinada a entrenamiento del modelo y otra parte que será utilizada para las pruebas.
- c) **Construcción del modelo**, a partir de la técnica de modelado escogida, se aplica sobre el conjunto de datos para generar uno o más modelos. En este punto se van ajustando los parámetros de la técnica seleccionada de forma iterativa para obtener mejores resultados.
- d) **Evaluación del modelo**, interpretando los modelos en base al conocimiento existente y los criterios de éxito ya establecidos.
- e) **Evaluación de los resultados**, en base a los objetivos del negocio y basándose en juicio experto.
- f) **Proceso de revisión**, que consiste en calificar el proceso completo de minería de datos realizado y ver si existen puntos a mejorar.

3.1.5 Despliegue del proyecto

Desde las etapas anteriores se puede obtener conocimiento a partir de los datos inicialmente seleccionados para el problema. Con el modelo ya construido y validado, este conocimiento se transforma en acciones o recomendaciones a realizar dentro del proceso de negocio. Las acciones van desde la aplicación del modelo a diferentes y nuevos conjuntos de datos, hasta cambios en los procedimientos o formas de trabajar.

Generalmente, los proyectos de *Data Mining* no concluyen con la implementación del modelo, ya que el usuario que lo emplee necesita tener documentados los resultados obtenidos y las consideraciones que se tomaron para el desarrollo del proyecto, para que a futuro comparta este conocimiento con otras personas [7].

3.2 Metodología utilizada en este trabajo

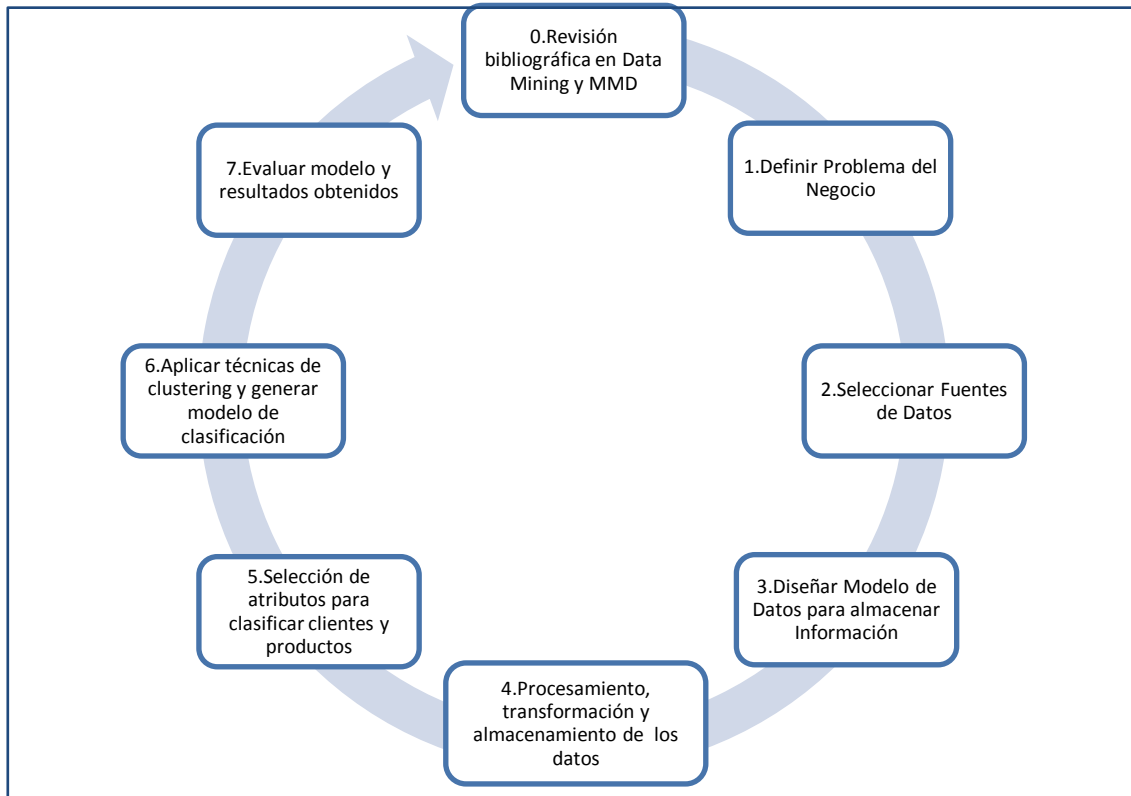
La metodología de este trabajo se basa principalmente en los pasos del proceso conocido como *Knowledge Discovery in Databases (KDD)*, y en la aplicación de *Business Intelligence*¹². Aunque el *KDD* no es una metodología propiamente tal, existen trabajos científicos y libros donde se utiliza como si lo fuera [30].

Una de las fases del proceso *KDD* es la minería de datos, etapa de análisis que permite extraer conocimiento desde los datos. Se espera que éste sea útil para la toma de decisiones, entendible por la persona que interpretará los resultados, y válido para que asegure que la información será correcta durante el proceso de selección y tratamiento de los datos. Esta tesis considera las etapas del proceso *KDD*, implementadas en base a la metodología *CRISP-DM* [1] [7], la cual permite desarrollar un proyecto de *Data Mining* aplicado a un caso real. De acuerdo a lo anterior, se formularon los pasos a seguir detallados en la *Figura 16*, adaptándose al problema de gestión de ENTEL.

Lo primero que se hizo fue revisar la literatura existente en conceptos de *Data mining* y modelamiento multidimensional. Además, se trabajó en conjunto con los expertos del negocio para entender sus necesidades para ver la forma de entregarles una solución al problema del negocio. A continuación, se realizó un levantamiento de la información existente y de las fuentes de datos que pudieran ser de utilidad para resolver el problema.

¹² Disciplina que le da a las empresas la capacidad de descubrir y utilizar información que ya poseen, y convertirla en conocimiento, utilizando para ello tecnologías de información [32].

Figura 16: Diagrama de los pasos a seguir en la metodología de la tesis



Fuente: Elaboración propia

Una vez que se contó con las fuentes de datos necesarias, se trabajó en el diseño de una estructura que fuera capaz de almacenar esa información, considerando las relaciones existentes entre los datos. Para esto, se pensó en el modelamiento multidimensional y en particular, en el uso de un modelo estrella.

Los datos provenían de diferentes fuentes, por lo que fue necesario realizar un procesamiento para dejarlos en un formato adecuado, revisándolos, corrigiendo inconsistencias y transformándolos, para luego almacenarlos en un repositorio [17].

Con los datos consolidados y dispuestos en un repositorio, se escogieron las variables con las cuales aplicar técnicas de minería de datos. Primero se usaron técnicas de *clustering* para formar conjuntos o grupos con características similares, y luego se utilizaron técnicas de clasificación para generar modelos que asignaran los datos existentes o nuevos a los conjuntos definidos previamente.

Finalmente, se verificó que los resultados obtenidos fueran satisfactorios, usando medidas para validar que los modelos funcionaran correctamente y además, corroborando con los expertos del negocio que los análisis tuvieran sentido.

4. Solución Propuesta

A continuación se propone una solución al problema planteado en esta tesis. Para ello se aplican los pasos de la metodología explicada en el capítulo anterior. Primero se describe el diseño del repositorio de datos donde se almacena la información, utilizando para ello metodología de modelos multidimensionales. Luego se continúa con la aplicación de técnicas de *Data Mining*.

Finalmente, se muestra como se puede estimar o valorizar los ingresos potenciales para ENTEL por los servicios que no tiene declarados en su sistema comercial, llevando estos números a un indicador con el cual hacer gestión.

4.1 Selección de las fuentes de datos

Teniendo en cuenta que el problema del negocio ya fue identificado en los capítulos anteriores, se prosigue con la descripción de los datos a utilizar y del procesamiento que es necesario realizar a cada uno.

El conjunto de datos proviene de diferentes archivos, los cuales actualmente son recopilados por el área de aseguramiento de ingresos. A continuación se describen las fuentes de información utilizadas:

- 1) Ordenes Terminadas (OTT/OTC):** es un reporte mensual que detalla las órdenes de trabajo asociadas a todos los productos de la Compañía, y se alimenta desde los distintos sistemas de provisión de ENTEL. En particular, contiene las órdenes de instalación de los servicios técnicos (privados y no privados). Es un archivo en formato Excel con 4000 registros en promedio, tomando en cuenta las OTC desde Agosto 2009 hasta Mayo 2010.
- 2) Cliente Contrato Privado (CCP):** es la base de datos comercial de la compañía, donde se registra la información de los clientes de servicios privados. En esta base se almacena información contractual, relacionada a servicios comerciales y técnicos. Se dispone de un archivo en formato Access con los registros históricos de todas las instalaciones técnicas ingresadas al sistema.
- 3) Vista Unificada de Servicios (VUS):** la Base de Servicios o Vista Unificada de Servicios corresponde a la base de datos operativa de la compañía, que registra información de los servicios técnicos con su respectivo estado operativo. Es un

archivo en formato Access, con más de 900.000 registros que sirve para conocer el último estado en que se encuentra una instalación técnica.

- 4) **Cartera de clientes:** corresponde a un archivo Excel con el listado de clientes de la compañía, diferenciados según el segmento de mercado que Entel ha definido (corporación, mayorista o empresa).
- 5) **Listado de servicios:** es el conjunto de servicios técnicos de continuidad operacional TI, tanto privados como no privados, generados a partir de la combinación de dos tipos de servicios (TS1/TS2) asociados a un código. Se mantienen en un archivo Excel que cuenta con 700 tipos de servicios.
- 6) **Estado de servicios:** es un archivo Excel con el listado de “estados” que puede tener un servicio, importante para saber si un servicio se encuentra vigente o no. Se incluye la sigla correspondiente al estado, una glosa descriptiva y su vigencia.

Los archivos mencionados se utilizaron para el cálculo de los indicadores de servicios privados, en particular se trabajó con el indicador mensual ISP3. Este indicador entrega una visión del cumplimiento y la eficiencia de la gestión que se realiza mensualmente en el proceso de provisión de servicios privados, específicamente en la incorporación de estos servicios en la plataforma comercial de ENTEL (CCP), mostrando cuántos de ellos se han declarado oportunamente, y se calcula de la siguiente manera:

$$ISP3_{s,t} = \frac{Pendientes_{s,t}}{Total_{s,t}} \quad (4.1)$$

s = Segmento de mercado del cliente {corporaciones, mayoristas, empresas}.

t = Mes y año de control del indicador.

$Pendientes_{(s,t)}$ = Cantidad de servicios privados instalados del segmentos en el mes de control t , y que no han sido declarados en el CCP en el mes siguiente.

$Total_{s,t}$ = Total de servicios privados instalados del segmento s en el mes de control t .

El proceso se realiza con un desfase de dos meses de tiempo, ya que se revisan las órdenes de trabajo con las instalaciones de un mes completo, y se da un plazo de un mes más para que se ingresen todas ellas en el sistema comercial de la empresa. Por ejemplo, si se quiere obtener el indicador ISP3 para el mes de Abril del año 2010, se consideran todos los servicios privados instalados durante ese mes, los cuales se revisan a finales del mes de Mayo, obteniéndose los resultados de este indicador a principios del mes de Junio.

La fase de minería de datos necesita de atributos adicionales que no se encuentran en las fuentes utilizadas en el diseño del *Data Mart*, por lo tanto se tuvo que complementar la información disponible con datos obtenidos desde el repositorio de gestión de la empresa:

1. **Rubros:** es un archivo Excel con el listado de clientes y de las actividades o sectores económicos asociados a cada uno de ellos.
2. **Modelo de segmentación de empresas:** es el modelo utilizado en ENTEL para separar o clasificar a las empresas en categorías, según características de consumo de cada cliente. Se cuenta con un archivo Excel con el listado de clientes de la compañía y la categoría correspondiente.
3. **DATAKOM CCP:** archivo generado a partir del CCP, que contiene el registro de todos los servicios técnicos y sus montos facturados, correspondientes a servicios de continuidad operacional TI para un período determinado. Se cuenta con reportes mensuales en formato Excel con el detalle de las facturaciones desde Octubre 2009 hasta Diciembre 2010.
4. **Tipo de conexión:** existen diferentes tipos de conexión o “tecnologías” asociados a un producto entregado por ENTEL, diferenciándose en la calidad y capacidad entregada. Este archivo Excel indica la tecnología utilizada para las instalaciones técnicas ya realizadas.
5. **Lugares:** es un archivo Excel que contiene el listado de lugares o puntos desde y hacia los que se puede hacer la provisión de un servicio, con su correspondiente comuna y región.

4.2 Diseño del modelo de datos

La información obtenida desde las fuentes de datos se almacenó en un *Data Mart*, que consta de un repositorio temporal llamado *Data Staging Area (DSA)*, en el cual se puede hacer el tratamiento y limpieza de los datos, y de una base de datos basada en un modelo estrella. El motor de datos que soporta esto está en mysql, usando el software de gestión de base de datos *EasyPHP*.

Además de la base de datos *DSA*, se creó otra base cuya estructura fuera basada en el modelamiento estrella, capaz de almacenar la información histórica para la generación del indicador ISP3. El modelo está compuesto por dos tipos de tablas relacionadas, las tablas *fact* y las tablas de dimensiones.

Figura 17: Variables almacenadas en Data Staging Area

Tabla	Columna	Dato Tipo
SERVICIO	ID_SERVICIO	INT
	TS1	VARCHAR(45)
	TS2	VARCHAR(45)
	MACRO_FAMILIA	VARCHAR(45)
	TIPO_SERVICIO	VARCHAR(45)
ESTADO_SERVICIO	ESTADO	VARCHAR(45)
	GLOSA	VARCHAR(60)
	ESTADO_RETIRO	VARCHAR(45)
	VIGENCIA	VARCHAR(45)
ORDEN_TERMINADA	ANO	INT
	MES	VARCHAR(45)
	COD_SERV_TEC	VARCHAR(45)
	RUT_DV	VARCHAR(45)
	TS1	VARCHAR(45)
	TS2	VARCHAR(45)
	UNIDAD_COMERCIAL	VARCHAR(45)
	TIPO_WORKFLOW	VARCHAR(45)
	NUM_OTC	VARCHAR(45)
	TIPO_SERVICIO	VARCHAR(45)
	EN_CCP	VARCHAR(2)
	EN_CCP_ACTUAL	VARCHAR(2)
	ANO_CCP	INT
	MES_CCP	VARCHAR(45)
	ESTADO_SER	VARCHAR(45)
	SEGMENTO	VARCHAR(45)
	EXTREMOA	VARCHAR(45)
EXTREMOB	VARCHAR(45)	
URGENCIA_PROYECTO	VARCHAR(45)	
DENTRO_PLAZO	VARCHAR(2)	
CLIENTE	RUT	VARCHAR(45)
	DV	VARCHAR(1)
	RUT_DV	VARCHAR(45)
	RAZON_SOCIAL	VARCHAR(80)
	AREA_ORGANIZACIONAL	VARCHAR(45)
CATEGORIA	VARCHAR(45)	
SEGMENTO	VARCHAR(45)	
RUBRO	VARCHAR(45)	
CCP	ID_SERV_TEC	VARCHAR(45)
	RUT	VARCHAR(45)
	DV	VARCHAR(1)
	ANO	INT
MES	VARCHAR(45)	

Fuente: Elaboración propia.

Las tablas *fact* contienen las medidas o indicadores del negocio, y las tablas de dimensiones son los diferentes niveles en los cuales un indicador puede ser visualizado. A continuación se muestran las tablas dimensionales que se utilizaron:

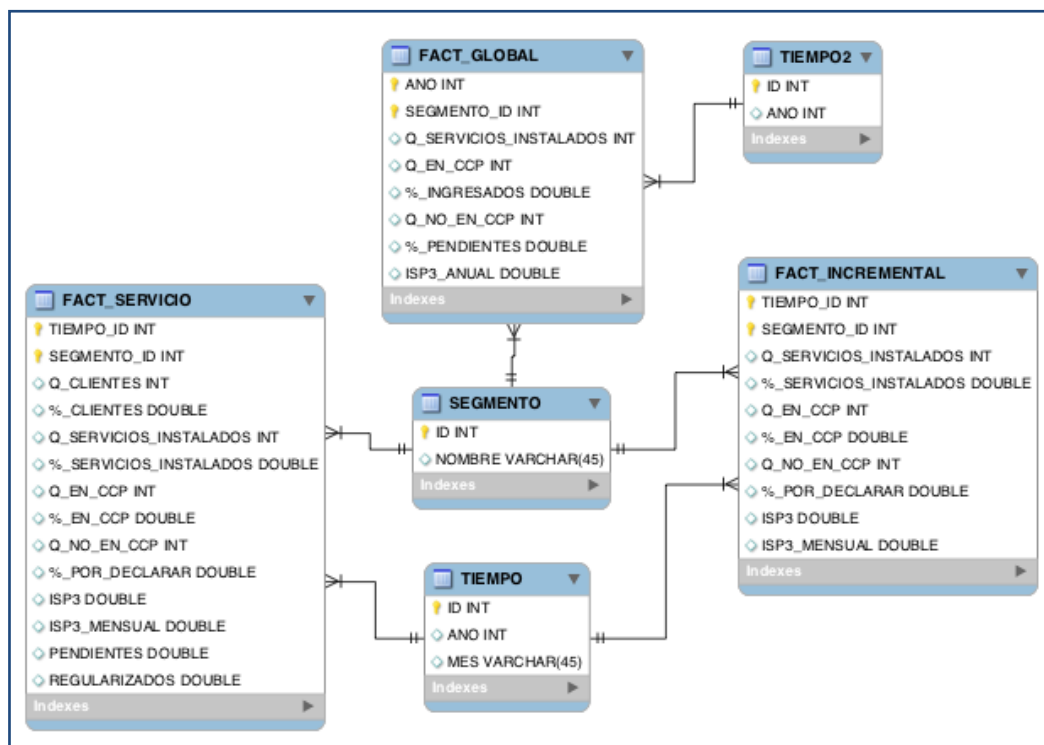
1. **Segmento:** contiene las 3 categorías de mercado definidas por ENTEL para sus clientes, que son corporaciones, mayoristas o empresas.
2. **Tiempo:** esta tabla tiene los años y los nombres de los meses
3. **Tiempo2:** esta tabla sólo contiene años y un indicador asociado.

Se crearon 3 tablas *fact* que se relacionan con las tablas dimensionales que se explican a continuación:

4. **Fact_servicio:** es una tabla que contiene la información necesaria para la creación del indicador ISP3, almacena información histórica de los indicadores generados en cada fecha de control, en otras palabras, guarda una foto o registro de los indicadores.

5. **Fact_incremental:** esta tabla tiene una estructura similar a la anterior, y su función es tener una actualización de los indicadores pasados luego de la regularización de las instalaciones no declaradas.
6. **Fact_global:** en esta tabla se muestra un resumen de los datos acumulados durante el año y el indicador ISP3 calculado en base a eso.

Figura 18: Modelo estrella para almacenar indicador ISP3



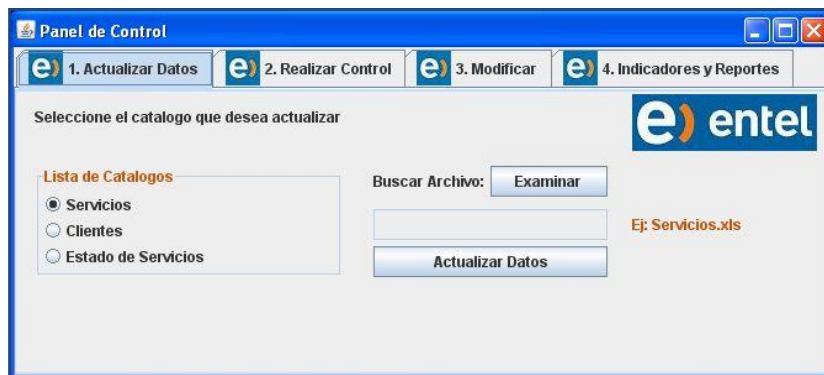
Fuente: Elaboración propia

4.3 Limpieza y procesamiento de los datos

La limpieza y el procesamiento de los datos son fundamentales en el desarrollo del proyecto de *Data Mining*, pues la calidad de los resultados que se obtendrán con los algoritmos de minería de datos aplicados depende de ello [17].

Para procesar las fuentes de datos que se usan en el cálculo del indicador ISP3, se diseñó una aplicación en lenguaje *Java* (ver Figura 19), la cual realiza las tareas de limpieza, procesamiento y carga de la información hacia una base de datos de forma semi-automática.

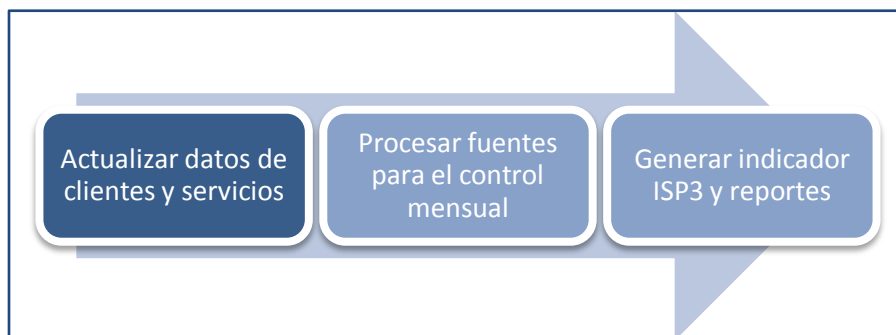
Figura 19: Aplicación en lenguaje Java para generar indicador mensual ISP3



Fuente: Elaboración propia

La aplicación en Java cuenta con diferentes pestañas para que el usuario pueda generar el indicador mensual ISP3, siguiendo los pasos mostrados en la Figura 20.

Figura 20: Pasos a seguir para generar indicador mensual ISP3



Fuente: Elaboración Propia

- 1. Actualizar datos de clientes y servicios:** el primer paso que realiza la aplicación es actualizar los listados de clientes y servicios. Es importante mantener actualizados estos listados, ya que de ello depende que un cliente o servicio aparezca en los análisis posteriores. Los archivos que se procesan en este paso son: cartera de clientes, listado de servicios y estado de servicios.
- 2. Procesar fuentes para el control mensual:** el segundo paso del proceso es cargar y procesar las siguientes fuentes de datos para el mes que corresponda. Como se muestra en la Figura 21, las fuentes que se utilizan son: archivo de órdenes terminadas, base de datos CCP y base de datos VUS.

Figura 21: Procesamiento de las fuentes de datos para indicador ISP3



Fuente: Elaboración propia

- a) **Ordenes terminadas:** es un archivo con el detalle de las instalaciones técnicas del mes, del cual se seleccionaron sólo aquellos registros donde el trabajo fuera una “instalación” o “instalación por evento”, y que además se encontrara en estado “terminado”. Con esto se redujo el conjunto de datos desde 4.000 a 900 registros en promedio por mes.

Luego de tener filtrados los datos, se validó el formato del código de servicio, que es el identificador de cada instalación. Para ello se utilizaron las siguientes reglas de verificación definidas por el negocio:

- ✓ Si tenía menos de 7 caracteres, entonces se revisaron las variables número de circuito inicial y final. Se debía repetir la línea del archivo tantas veces como números consecutivos existieran entre el circuito inicial y el final. Se agregaron ceros a la izquierda del circuito hasta que tuviera 4 caracteres, y luego se concatenó el código de servicio con el circuito.
- ✓ Si tenía 7 o más caracteres, se dejó igual.

- b) **Base de datos CCP:** es un archivo que tiene información acumulada en el tiempo, por lo tanto no se carga completamente en el *DSA*, sino que se compara la última versión de este archivo con los datos guardados en el *DSA* para guardar sólo las diferencias.

- c) **Base de datos VUS:** es un archivo que contiene la información más actual de los servicios asociados a una instalación, y se utiliza para validar con respecto a la información que viene desde el archivo de órdenes terminadas.

3. Generar indicador ISP3 y reportes: el último paso de este proceso es el cálculo del indicador en base a la información cargada anteriormente. Se escoge la fecha del mes de control, con esto la aplicación calcula y guarda los resultados en la base de datos. Adicionalmente, se generan reportes con los resultados del indicador gracias al uso de la *API Jasper Report* (ver ejemplo en anexo 1).

Figura 22: Cálculo de indicador ISP3

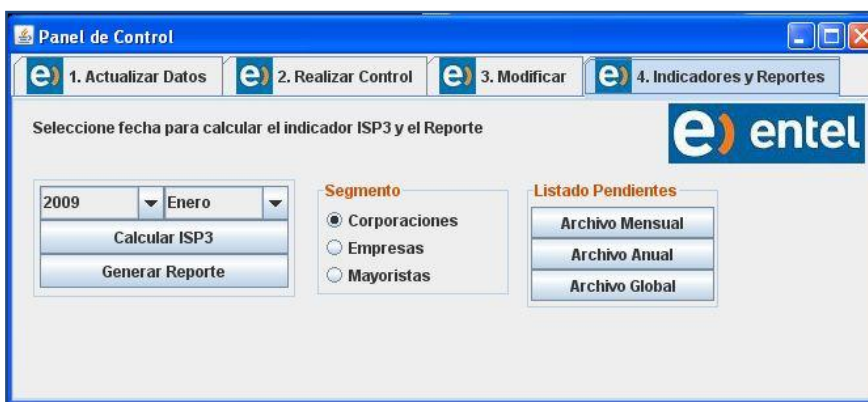


Tabla 9: Fuentes de datos y campos utilizados para el indicador ISP3

Fuente de Datos	Atributos o campos	Descripción
Cartera de Clientes	Rut y DV	Rut y dígito verificador de un cliente
	Razón Social	Nombre del cliente
	Segmento	Segmento de mercado (corporaciones, mayoristas o empresas)
	Categoría	Clasificación de clientes definida por ENTEL
Listado de Servicios	Producto	Nombre del servicio
	TS1	Código asociado al tipo de servicio 1
	TS2	Código asociado al tipo de servicio 2
Estado de Servicios	Estado	Código asociado al estado de un servicio
	Vigencia	Si el servicio se encuentra vigente o no
	Glosa	Descripción del estado de un servicio
Órdenes Terminadas	Código de servicio	Identificador de una instalación técnica
	Circuito inicial y final	Sufijo asociado al código de servicio para indicar los números correlativos entre el circuito inicial y final
	Rut y DV	Rut y dígito verificador de un cliente
	Unidad Comercial	Área a la cual se asocia la venta de un servicio
	Tipo de Workflow	Sistema utilizado para registrar la instalación
	Número de OTC	Código asociado a la orden de trabajo
	Extremo A y B	Lugares desde y hacia el cual se provee un servicio
	Urgencia	Prioridad del proyecto (Express o Normal)
Base de datos CCP	Dentro plazo	Si la instalación fue realizada en plazo o no
	Código de servicio	Identificador de una instalación técnica
Base de datos VUS	Rut y DV	Rut y dígito verificador de un cliente
	Código de servicio	Identificador de una instalación técnica
	Estado del servicio	Código asociado al estado de un servicio
	TS1	Código asociado al tipo de servicio 1
	TS2	Código asociado al tipo de servicio 2

Adicionalmente, se trabajó con otras fuentes de datos, no utilizadas para el cálculo del indicador ISP3, pero necesario para el trabajo de investigación en la etapa de minería de datos:

- a) **Rubros:** se realizó una limpieza de forma manual del listado de rubros, estandarizando los nombres y revisando sus valores con los expertos del negocio. Además, se consideró la distribución de los datos, para eliminar aquellos rubros que tuvieran una baja frecuencia (menor al 1% del total), asignándolos a la categoría “OTROS”.
- b) **Modelo segmentación de empresas:** los datos obtenidos en este archivo para la variable categoría sólo aplican a los clientes del segmento empresas. Para los clientes corporaciones y mayoristas se obtuvieron los valores desde el archivo “cartera de clientes”.
- c) **DATAKOM CCP:** se filtran los registros según el campo cuota/periodo para obtener sólo los montos correspondientes a la facturación del mes. Con esto se buscan los montos asociados a un código de servicio.
- d) **Tipo de conexión:** Este archivo indica el tipo de tecnología utilizada para una instalación técnica, que puede ser cobre, fibra óptica, wimax, etc. En el caso que no se indicara la tecnología utilizada se asignó el valor “NO TIENE”. Las velocidades están asociadas al servicio de internet, que puede estar en unidades de Kilobytes (Kb) o Megabytes (Mb). Para tener todo en una misma unidad, se optó por dejar los valores en Megabytes y redondearlos.
- e) **Lugares:** En la fase de minería de datos se trabajó con datos obtenidos desde el *Data Mart* diseñado para el área de Aseguramiento de Ingresos, seleccionando la información de clientes y de servicios privados instalados, que en conjunto totalizaban 7926 registros.

Adicionalmente, se crearon las variables “ESTA_EN_CCP”, la cual permite distinguir si un servicio se encuentra declarado en el sistema comercial de ENTEL o no, y la variable “TIENE_VALOR”, que identifica aquellos códigos de servicio que registraban facturaciones en el archivo DATAKOM.

Luego de reunir todas las fuentes de datos necesarias para la investigación, se continuó con la etapa de transformación de los datos, en la cual se definen las variables finales para generar los modelos de *Data Mining*.

Tabla 10: Fuentes de datos adicionales para hacer minería de datos

Fuente de Datos	Atributos o campos	Descripción
Rubros	Rut y DV	Rut y dígito verificador de un cliente
	Rubro	Sector o actividad económica
Modelo de Segmentación de Empresas	Rut y DV	Rut y dígito verificador de un cliente
	Categoría	Clasificación de clientes definida por ENTEL
Datacom CCP	Número de Circuito	Código de servicio técnico
	DCO_Monto	Monto facturado del servicio
	DCB_cuota/período	Mes correspondiente a la facturación
Tipo de Conexión	Código de Servicio	Identificador de una instalación técnica
	Wimax/Cobre	Indica, cuando corresponda, si el servicio utiliza tecnología de cobre o wimax
	Velocidad	Velocidad de transferencia de los servicios de telecomunicación medida en Mb
Lugares	Lugar	Nombre de los lugares de provisión de servicios
	Comuna	Comuna a la cual pertenece un lugar
	Región	Región a la cual pertenece un lugar

4.4 Transformación de los datos

Se tuvo que realizar transformaciones a algunas de las variables escogidas anteriormente, para modificarlas y crear otras nuevas a partir de ellas. Las modificaciones realizadas fueron:

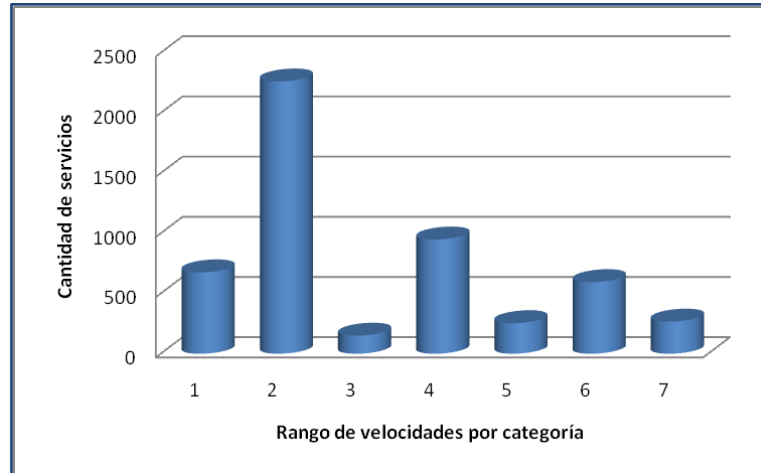
- En el caso de la variable VELOCIDAD, se decidió discretizar los valores que tomaba en rangos de velocidades, con el fin de disminuir la dispersión de esos datos. Los intervalos fueron definidos de acuerdo a un árbol de decisión *CHAID*, lo cual permitió segmentar el conjunto de valores usando un test de chi-cuadrado. Con esto, los rangos quedaron como se indica en la siguiente tabla:

Tabla 11: Valores para la variable RANGO_VELOCIDAD

Rango	Valores de velocidad [Mb]
1	[0 , 1.5]
2	(1.5 , 2]
3	(2 , 8]
4	(8 , 10]
5	(10 , 12]
6	(12 , 16]
7	> 16

Fuente: Elaboración propia

Figura 23: Gráfico de distribución de velocidades en Mega bytes



Fuente: Elaboración propia

- b) Se creó la variable VALOR_SERVICIO, que representa la tendencia en el consumo de un cliente por un servicio privado instalado. Se formó a partir de las facturaciones mensuales presentes en los archivos “DATACOM CCP”. Esta variable se calculó para cada código de servicio de la siguiente manera:
- I. Se buscaron todas las facturaciones disponibles en un plazo de 12 meses, desde Octubre de 2009 hasta Septiembre de 2010.
 - II. Se contó la cantidad de facturaciones existentes en esos meses. Si sólo se contaba con un valor, se utilizaba ése como el precio para valorizar el código de servicio. En cambio, si se tenía más de uno, se definió un rango de aceptación para escoger los valores a considerar en la estimación, eliminando de aquellos datos anómalos o fuera de rango.
 - III. El rango de aceptación se definió como un intervalo compuesto por una cota superior e inferior. Este intervalo se obtiene utilizando una desviación con respecto al promedio ponderado de las facturaciones.
 - IV. En el promedio ponderado se le dio mayor importancia a los valores más recientes, considerando el efecto de cambios en el consumo. Sea F_i^j la facturación número i que tuvo el código de servicio j durante el período de 12 meses, y sean N el total de facturaciones, entonces los pesos normalizados asociados a cada facturación se obtienen así:

$$P_i = 2 * \frac{i}{N*(N+1)} \quad ; i = \{1, \dots, N\} \quad (4.2)$$

$$\text{Promedio ponderado del código de servicio } j = \sum_{i=1}^N P_i * F_i^j \quad (4.3)$$

- V. Para la desviación se usó el estadístico MAD^{13} , una medida robusta de la variabilidad o dispersión, definida como la mediana de las desviaciones absolutas de los datos con respecto a la mediana de los mismos.

$$MAD^j = mediana (|F_i^j - mediana(F_1^j, \dots, F_N^j)|) \quad (4.4)$$

- VI. Utilizando lo anterior, se definen las cotas del rango de aceptación como:

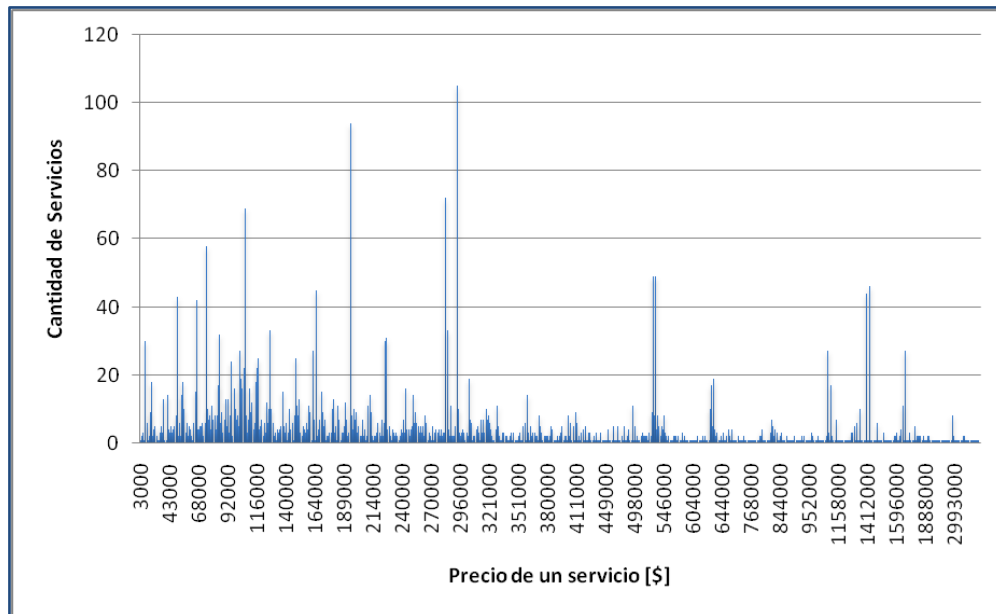
$$Sup = Promedio Ponderado + MAD \quad (4.5)$$

$$Inf = Promedio Ponderado - MAD \quad (4.6)$$

$$\text{Rango de aceptación} = F \in [Inf, Sup] \quad (4.7)$$

- VII. Todos los valores de facturación dentro de ese rango se promediaron, con lo cual se obtuvo una estimación del precio. En los casos en que ningún valor cayera dentro del rango definido, se utilizó simplemente el promedio ponderado de los datos existentes.
- VIII. Finalmente se redondearon los precios obtenidos en la variable VALOR_SERVICIO a la unidad de mil para tener una mejor distribución de los datos, creando con ello la variable llamada VALOR_REDONDEADO.

Figura 24: Gráfico de distribución de la variable VALOR_REDONDEADO



Fuente: Elaboración propia

¹³Median Absolute Deviation

- c) Se creó la variable llamada TIPO_CONEXION, en base a los valores de la tecnología utilizada en cada servicio privado, indicados tanto en el archivo “Tipo de conexión” como en el “Listado de servicios”. Esta nueva variable puede tomar los siguientes valores:

$$TIPO\ CONEXION = \begin{cases} ALAMBRICA & \text{si es COBRE o FIBRA OPTICA} \\ INALAMBRICA & \text{si es WIMAX} \\ NO\ APLICA & \text{si NO TIENE} \end{cases} \quad (4.8)$$

- d) Se crearon las variables COMUNA_A y COMUNA_B, que corresponden a la comuna asociada a las variables EXTREMO_A y EXTREMO_B respectivamente, utilizando para ello el archivo llamado “Lugares”.

4.5 Data Mining

En la fase de minería de datos se realizaron tres experimentos. El primero de ellos fue aplicar técnicas de *clustering* para formar grupos de clientes y servicios privados con características similares. El segundo experimento consistió en clasificar los datos que no tuvieran un precio, asignándolos a uno de los *clusters* formados en el experimento anterior. El último experimento fue valorizar los códigos de servicio clasificados y asignados a los *clusters*, buscando los que se parecían más en de un mismo conjunto.

4.5.1 Selección de atributos

En total se seleccionaron 17 atributos para realizar minería de datos, los cuales se presentan en la siguiente tabla:

Tabla 12: Atributos seleccionados para aplicar minería de datos

Nº	Nombre del atributo	9	DENTRO_PLAZO
1	CODIGO_SERVICIO	10	COMUNA_A
2	TS1 TS2	11	COMUNA_B
3	TIPO_SERVICIO	12	SEGMENTO
4	ESTADO_SERVICIO	13	CATEGORIA
5	VELOCIDAD_RANGO	14	RUBRO
6	TIPO_CONEXION	15	VALOR_REDONDEADO
7	UNIDAD_COMERCIAL	16	TIENE_VALOR
8	URGENCIA	17	ESTA_EN_CCP

Fuente: Elaboración propia

4.5.2 Primer experimento: *Clustering*

Lo primero que se hizo fue separar las muestras de datos para armar el modelo de *clustering*, utilizando la herramienta de minería de datos *RapidMiner* (ver diagramas de procesos en anexo 2).

La muestra consistió en los registros que tenían un valor de precio asignado (3420 en total), utilizando para ello el filtro de la variable *TIENE_VALOR* = 1. Además se incluyeron aquellos en que la variable *ESTA_EN_CCP* = NO (298 en total). Con lo anterior, el set de datos a los cuales se le aplicó *clustering* fue de 3718 registros.

Lo siguiente fue determinar el número de *clusters* que se iban a generar. Para esto se utilizaron dos métodos de *clustering* jerárquico [27]:

- *EM (Expectation-Maximization)*
- Aglomerativo

El primero de los métodos se basa en probabilidades a partir de la muestra de datos, indicando que el número de *clusters* era 9. El segundo método entrega sus resultados en forma de un dendograma (ver Figura 25), en el que visualmente se realiza un corte en la cantidad de 6 *clusters*. Se decidió que el número a utilizar en los siguientes pasos del proceso sería este último, y se compararían posteriormente los efectos de ir aumentando ese número.

Figura 25: Dendograma del experimento de *clustering*



Fuente: Resultado obtenido desde el software *RapidMiner*

Luego de lo anterior, se trabajaron los datos en la herramienta *SPSS* para aplicar algoritmos de *clustering*, en particular se usó el algoritmo *TwoStepCluster*, escogiendo al vecino más lejano. La medida de distancia que usa el algoritmo es log-verosimilitud, y el criterio para formar los conglomerados es el criterio bayesiano de *Schwarz* (BIC).

Se utilizó el algoritmo *Two Step Cluster* ya que, como se explica en el capítulo 2, permite trabajar con variables tanto categóricas como nominales, a diferencia de otros algoritmos como *K-means* que sólo trabajan con variables numéricas. El atributo *VALOR_REDONDEADO* no fue considerado para el algoritmo, ya que la idea fue evitar considerar el efecto del precio en la formación de los *clusters*, y sólo basarse en las características de los clientes y servicios. El resultado obtenido de esta segmentación se muestra en la *Tabla 13*.

Tabla 13: Distribución de conglomerados con algoritmo Two Step Cluster (K = 6)

<i>Cluster</i>	Número de elementos	% del total
1	1111	29,9 %
2	578	15,5 %
3	366	9,8 %
4	755	20,3 %
5	473	12,7 %
6	435	11,7 %
Total	3718	100 %

Fuente: Elaboración propia

Para asegurarse que el número de *clusters* determinados tenía sentido, se comparó el resultado que entregaba el algoritmo con $K = 7$, $K = 8$ y $K = 9$. La comparación se realizó a través de tablas de contingencia. Las tablas de contingencia se utilizan para ver la relación entre dos o más variables, en este caso para ver la relación entre el *cluster* asignado inicialmente y el que se forma al utilizar un K mayor.

Se revisaron los atributos que provocaban la nueva separación, resumiendo los resultados en la *Tabla 14*. Estos resultados se analizaron con los expertos del negocio, y se vio que al aumentar el número de conglomerados se especificaban aún más las características de un grupo sin embargo, para el problema en estudio esto no aportaba mayor información de la que ya se tenía.

Tabla 14: Resultados obtenidos al aplicar tablas de contingencia entre clusters

Tabla de contingencia	Cluster afectado	Clusters formados	Semejanzas	Diferencias
K = 6 y K = 7	1	1 y 2	Son servicios de acceso a red MPLS (ACCE-MPET)	El <i>cluster</i> 1 sólo provee servicio en Santiago, el <i>cluster</i> 2 en el resto de las regiones.
K = 7 y K = 8	5	5 y 6	Son servicios de tipo dedicados.	El <i>cluster</i> 6 sólo tiene en su mayoría servicios con conexión de tipo alámbrica.
K = 8 y K = 9	7	7 y 9	Clientes del segmento empresa con servicios privados	El <i>cluster</i> 7 tiene servicios de internet TRUNK IP.

Fuente: Elaboración propia

Para la evaluación de la calidad de los conjuntos obtenidos, se utilizó uno de los índices mencionados en el capítulo 2. El índice escogido fue el *Davies-Bouldin* [3] [38], función que relaciona las distancias entre los elementos de un *cluster* con la separación entre los centroides de dos de ellos. Sea n el número de *clusters* que se quiere validar y c_i el centroide del *cluster* i , entonces:

$$Davies\ Bouldin\ index\ (DB) = \frac{1}{n} \sum_{i=1}^n \max_{i \neq j} \left\{ \frac{\bar{S}(c_i) + \bar{S}(c_j)}{S(c_i, c_j)} \right\} \quad (4.9)$$

Donde $S(x, y)$ representa la medida de distancia entre los elementos x e y , mientras que $\bar{S}(c_i)$ es el promedio de las distancias de los elementos del *cluster* i a su centroide. De acuerdo a lo anterior, un índice DB bajo muestra que el algoritmo de *clustering* formó grupos compactos y de alta similitud entre sus elementos, y a su vez los grupos están alejados unos de otros.

4.5.3 Segundo experimento: Clasificación

Esta etapa consistió en generar modelos de clasificación, que permitieran asignar elementos a los *clusters* formados en el experimento anterior. Se trabajó con algoritmos que entregaran resultados rápidos, como árboles de decisión y SVM. Los datos fueron los mismos del experimento anterior (3718 registros), que ahora contaban con una “etiqueta” indicando el *cluster* de pertenencia.

Para aplicar los algoritmos de clasificación, se realizó un tratamiento de los datos con la herramienta *RapidMiner* (ver diagramas en anexo 2), transformando las variables nominales a binomiales con valores numéricos (0 y 1). Se usaron 13 atributos regulares para clasificar, además del CODIGO_SERVICIO que identifica los registros.

La muestra de datos se separó en dos: un conjunto para entrenamiento del modelo, y otro para la validación posterior (30% del total con precio asignado). Cada modelo se entrenó y probó con la técnica *Cross Validation*, realizando un total de 10 iteraciones, que corresponden a las particiones en que se divide el conjunto de datos para validar.

El primer modelo se construyó con un árbol de decisión J4.8, el cual utiliza medidas de información para ir separando los elementos. El segundo modelo utilizó como clasificador el algoritmo SVM. Los parámetros escogidos para cada uno de ellos se muestran en *Tabla 15* y *Tabla 16* respectivamente.

Tabla 15: Parámetros utilizados en el algoritmo árbol de decisión J4.8

Parámetro	Valor
Criterio	<i>Gain ratio</i>
Tamaño mínimo para la división	4
Tamaño mínimo de las hojas	2
Ganancia mínima	0,1
Profundidad máxima	20
Confianza	0,25
Número de podas	3

Tabla 16: Parámetros utilizados en el algoritmo SVM

Parámetro	Valor
Tipo de SVM	C-SVC
Tipo de Kernel	<i>rbf</i>
Gamma	0,05
C	100
Tamaño del caché	80
Epsilon	0,001

4.5.4 Tercer experimento: Asignación de precios

El último experimento consistió en valorizar los códigos de servicio que fueron clasificados en el experimento anterior. Para ello, se tuvo que buscar dentro de un mismo *cluster* aquellos elementos que se asemejaban más en sus atributos al elemento clasificado, es decir, se utilizó una medida de similitud.

Lo primero que se hizo en esta etapa fue asociarle un peso a los atributos, revisando para ello los gráficos de importancia de cada uno a nivel de conglomerado, los cuales se obtuvieron en el experimento 1 con la herramienta *SPSS* (ver detalle en anexo 3). El resumen de estas relevancias se muestra en la *Tabla 17*, donde 1 significa que el atributo es el más importante, y 13 el menos.

Tabla 17: Importancia de los atributos por conglomerado

Variable \ Cluster	C1	C2	C3	C4	C5	C6
TS1TS2	1	1	2	1	1	5
VELOCIDAD_RANGO	2	3	6	10	6	7
TIPO_CONEXION	12	13	12	12	12	10
TIPO_SERVICIO	11	12	9	2	9	12
ESTADO_SERVICIO	10	2	10	11	10	11
UNIDAD_COMERCIAL	3	5	1	5	2	3
URGENCIA	8	10	11	7	13	8
DENTRO_PLAZO	13	11	13	13	11	13
COMUNA_A	9	8	5	9	8	9
COMUNA_B	7	9	4	8	7	6
SEGMENTO	5	6	8	3	3	2
CATEGORIA	6	7	7	4	4	1
RUBRO	4	4	3	6	5	4

Fuente: Elaboración propia.

Para llevar este ranking de variables a pesos, se aplicó una función monótona decreciente $f(x)$ y luego se normalizaron los valores. Sea x_{ij} la importancia del atributo i en el *cluster* j , entonces el peso W_{ij} asociado a ese atributo se calcula como sigue:

$$W_{ij} = \frac{f(x_{ij})}{\sum_{i=1}^{13} f(x_{ij})} \quad ; \quad f(x) = \frac{1}{\sqrt{x}} \quad (4.10)$$

A partir de los pesos W_{ij} calculados anteriormente, los que se pueden apreciar en la *Tabla 18*, se compararon los elementos clasificados a través de una medida de similitud. Sea k el código de servicio clasificado, y m uno de los datos que ya forman parte de un cluster, entonces la similitud entre ambos se calcula de la siguiente manera:

$$S_{km}^j = \sum_{i=1}^{13} W_{ij} * \delta_{km}^j \quad (4.11)$$

$$\delta_{km}^j = \begin{cases} 1 & \text{si el elemento } k \text{ tiene el mismo valor que } m \text{ en el atributo } j \\ 0 & \text{si no} \end{cases} \quad (4.12)$$

Tabla 18: Pesos asociados a cada atributo por conglomerado

Variable \ Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
TS1TS2	0,170	0,170	0,120	0,170	0,170	0,076
VELOCIDAD_RANGO	0,120	0,098	0,069	0,054	0,069	0,064
TIPO_CONEXION	0,049	0,047	0,049	0,049	0,049	0,054
TIPO_SERVICIO	0,051	0,049	0,057	0,120	0,057	0,049
ESTADO_SERVICIO	0,054	0,120	0,054	0,051	0,054	0,051
UNIDAD_COMERCIAL	0,098	0,076	0,170	0,076	0,120	0,098
URGENCIA	0,060	0,054	0,051	0,064	0,047	0,060
DENTRO_PLAZO	0,047	0,051	0,047	0,047	0,051	0,047
COMUNA_A	0,057	0,060	0,076	0,057	0,060	0,057
COMUNA_B	0,064	0,057	0,085	0,060	0,064	0,069
SEGMENTO	0,076	0,069	0,060	0,098	0,098	0,120
CATEGORIA	0,069	0,064	0,064	0,085	0,085	0,170
RUBRO	0,085	0,085	0,098	0,069	0,076	0,085

Fuente: Elaboración propia

Lo que se hizo fue encontrar aquellos registros que tuvieran la mayor similitud con respecto al elemento por valorizar, y utilizando la variable VALOR_REDONDEADO se le asoció un precio.

La estimación se logró a través de la combinación de los precios de estos registros, obteniendo un único valor (π), usando funciones como el promedio y la mediana. Se compararon los resultados entregados por cada una, quedándose con aquella función que se ajustaba mejor estadísticamente.

4.6 Generación de nuevos indicadores

Con la estimación de precios generada en las etapas previas para cada código de servicio privado instalado, se definieron nuevos indicadores para ayudar al control de gestión de ENTEL.

El primer indicador se creó a partir del ISP3, asignándole un precio a cada servicio privado instalado, valorizando la cantidad de pendientes mensuales. La forma de calcularlo se muestra a continuación:

$$\lambda_{s,t}^i = \begin{cases} 1 & \text{si el cod. de serv. privado } i \text{ del segmento } s \text{ y mes de control } t \text{ ya fue declarado} \\ 0 & \text{si no} \end{cases}$$

π_i = Precio estimado para el código de servicio privado i

s = Segmento de mercado del cliente {corporaciones, mayoristas, empresas}.

t = Mes y año de control del indicador.

$Total_{s,t}$ = Total de servicios privados instalados del segmento s en el mes de control t .

$$\text{ISP3 Valorizado}_{s,t} = \frac{\sum_{i=1}^N (1 - \lambda_{s,t}^i) * \pi_i}{Total_{s,t}} \quad (4.13)$$

El segundo indicador tiene como finalidad mostrar el porcentaje de ingresos potencialmente recuperables. Según los expertos del negocio, existe un límite de meses (T_{max}) que pueden pasar entre la instalación de un servicio hasta su facturación, y transcurrido ese tiempo ENTEL no puede exigir o reclamar al cliente por los cobros que no realizó en su debido momento. Esta restricción es considerada en este nuevo indicador llamado Ingresos Potencialmente Recuperables (IPR). Sea t_i la fecha de instalación del servicio privado i , entonces:

$\Delta_t^i = t - t_i$ = Cantidad de meses entre la fecha de instalación del servicio i y una fecha de referencia t .

T_{max} = Tiempo máximo permitido de recuperación de ingresos.

$$\text{IPR} = \frac{\text{Total de Ingresos Potencialmente Recuperables}}{\text{Total de Ingresos por servicios no declarados}} = \frac{\sum_{i=1}^N \min\{\Delta_t^i; T_{max}\} * \pi_i}{\sum_{i=1}^N \Delta_t^i * \pi_i} \quad (4.14)$$

Del total de ingresos potencialmente recuperables, el negocio puede tener expectativas menores al máximo disponible, para ello se puede reemplazar el parámetro T_{max} por una variable móvil θ con valores enteros, tal que $1 \leq \theta \leq T_{max}$.

5. Resultados Experimentales

A continuación se presentan los resultados obtenidos en los tres experimentos explicados anteriormente. En primer lugar se muestran los *cluster* que se formaron, explicando las características relevantes de cada uno, luego siguen los modelos de clasificación y la validación de cada uno para ver cuál es el que se ajusta mejor al problema y que entrega resultados más confiables. Por último se muestra la estimación de precios, valorizando los servicios privados que aun no han sido declarados comercialmente en ENTEL, y con ello estimando la pérdida recuperable a través de un indicador de gestión.

5.1 Resultados del primer experimento

Como se mostró en el capítulo anterior, mediante algoritmos de *clustering* se pudieron formar 6 grupos con los datos de clientes y servicios privados. Estos grupos son descritos de acuerdo a tablas de frecuencias de los atributos utilizados para la formación de los *clusters* (*ver detalle en anexo 4*).

A continuación se presentan las características principales que se destacan en cada uno de los conglomerados.

5.1.1 Descripción del *cluster* N° 1

Está compuesto en su totalidad por instalaciones de accesos alámbricos a la red utilizando MPLS (ACCE-MPET). En su mayor parte, las velocidades utilizadas son medias-altas (12 a 16 Mb). Este servicio se entrega de preferencia a clientes ENTEL del segmento de mercado mayorista, ubicados en las comunas de Santiago y Valparaíso.

5.1.2 Descripción del *cluster* N° 2

Se caracteriza por contar con servicios de tramas¹⁴ para telefonía privada con velocidad de 2 Mb (TFPR-MIC). Este servicio es contratado por clientes del segmento de mercado mayorista, ubicados en comunas grandes del país como Santiago, Antofagasta y Concepción.

¹⁴ Unidad de envío de datos a través de “paquetes”, utilizado en los modelos de redes de telefonía e internet.

5.1.3 Descripción del *cluster* N° 3

Este grupo es principalmente de clientes mayoristas, que solicitan circuitos virtuales de datos a través de una red de comunicación *Frame Relay*¹⁵, tecnología utilizada para transmitir tanto voz como datos a una mediana velocidad y con un bajo costo.

5.1.4 Descripción del *cluster* N° 4

A diferencia de los conglomerados anteriores, este conjunto está definido por clientes del segmento de mercado empresas, los cuales solicitan servicios dedicados de internet con una alta urgencia por su instalación. Dentro de estos servicios se encuentran aquellos de conexión por cable (INTE-MPLS, INTE-NGN) u otros vía inalámbrica (INTE-WIMAX).

5.1.5 Descripción del *cluster* N° 5

Al igual que el cluster anterior, en este también se cuenta con clientes empresas, sin embargo consumen servicios privados como accesos de red. Un servicio particular de este grupo es el acceso a internet utilizando la tecnología TRUNK IP (ACCE-TKIN) y se proveen en su mayor parte en las comunas de la región metropolitana.

5.1.6 Descripción del *cluster* N° 6

Este último grupo tiene como característica principal que sus clientes son del segmento de mercado corporaciones, y que solicitan instalaciones de servicios privados con una alta urgencia.

¹⁵ *Frame-mode Bearer Service* permite la retransmisión de tramas para redes de circuito virtual.

5.2 Resultados del segundo experimento

En este experimento se utilizaron dos modelos de clasificación: el primero consideró como algoritmo un árbol de decisión, y el segundo SVM. Para cada uno de ellos se obtuvo una matriz de confusión (ver *Tabla 19* y *Tabla 20*), y a partir de éstas se obtuvieron las medidas con las cuales evaluar la calidad de los modelos. A continuación se definen cada una de estas medidas [20]:

Predicción (observación)	Clases actuales (expectativa)	
	TP	FP
	FN	TN
Totales	P = TP + FN	N = TN + FP

- *TP (true-positive)*: resultado correcto.
- *TN (true-negative)*: ausencia correcta de resultado
- *FP (false-positive)*: resultado inesperado.
- *FN (false-negative)*: resultado erróneo

1) **Accuracy**: se refiere al nivel de certeza del modelo, y se calcula como la proporción del total de predicciones que fueron correctas:

$$Accuracy = \frac{TP+TN}{P+N} \quad (5.1)$$

2) **Recall**: es la proporción de casos positivos que fueron identificados correctamente, en otras palabras, la probabilidad que un objeto de una clase sea clasificado correctamente en esa misma:

$$Recall = \frac{TP}{P} \quad (5.2)$$

3) **Precision**: es la probabilidad de que una predicción efectivamente corresponda con su valor real:

$$Precision = \frac{TP}{TP+FP} \quad (5.3)$$

4) **F – measure (F1)**: es una medida geométrica que combina las medidas *precision* & *recall* para evaluar de forma más realista la certeza del modelo.

$$F - measure = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \quad (5.4)$$

5) **Error rate**: corresponde a la proporción de casos incorrectamente predichos:

$$Error\ rate = \frac{FN+FP}{P+N} \quad (5.5)$$

Tabla 19: Matriz de confusión para modelo de clasificación, usando un árbol de decisión

Predicción\Clase actual	C1	C2	C3	C4	C5	C6	Precision
P1	1111	1	0	0	0	0	99,91%
P2	0	577	3	0	0	0	99,48%
P3	0	0	362	0	0	2	99,45%
P4	0	0	0	755	1	0	99,87%
P5	0	0	0	0	468	2	99,57%
P6	0	0	1	0	4	431	98,85%
Recall	100,00%	99,83%	98,91%	100,00%	98,94%	99,08%	
F	99,95%	99,65%	99,18%	99,93%	99,25%	98,96%	

Fuente: Elaboración propia

Tabla 20: Matriz de confusión para modelo de clasificación, usando algoritmo SVM

Predicción\Clase actual	C1	C2	C3	C4	C5	C6	Precision
P1	1109	1	1	0	0	0	99,82%
P2	1	577	1	0	0	0	99,65%
P3	1	0	364	0	0	0	99,73%
P4	0	0	0	755	1	0	99,87%
P5	0	0	0	0	471	2	99,58%
P6	0	0	0	0	1	433	99,77%
Recall	99,82%	99,83%	99,45%	100,00%	99,58%	99,54%	
F	99,82%	99,74%	99,59%	99,93%	99,58%	99,65%	

Fuente: Elaboración propia

La comparación entre ambos modelos se presenta en la siguiente tabla, mostrando los resultados para las medidas explicadas anteriormente. Se puede apreciar que ambos modelos son muy precisos en sus resultados, y que el segundo de ellos entrega un nivel de error menor.

Tabla 21: Comparación de resultados obtenidos con los modelos de clasificación

Medida	Modelo 1 (árbol de decisión)	Modelo 2 (SVM)
Accuracy	99,62%	99,76%
Error rate	0,36%	0,25%
F (promedio)	99,49%	99,71%

Fuente: Elaboración propia

5.3 Resultados del tercer experimento

En este último experimento se valorizaron los códigos de servicio clasificados a cada *cluster* en la fase previa. Se usaron dos muestras de datos: una que ya tenía precios, con el fin de comparar los resultados con el valor real, y otra sin precios para reflejar la valorización de códigos de servicios no declarados en el sistema comercial de ENTEL.

La primera muestra se compuso de 1026 registros con precio, que contabilizaban un total de \$ 410.940.000 en ingresos. Se comparan los resultados obtenidos de aplicar las funciones promedio y mediana a los precios de elementos con características similares, con el fin de estimar un precio único. Estos resultados se presentan en la *Tabla 22*.

Tabla 22: Medidas estadísticas para evaluar valorización de precios

Medidas Estadísticas	Promedio de los datos	Mediana de los datos
NRMSD (Normalized Root Mean Square Deviation)	3,69%	3,80%
MAPE (Mean Absolute Percentage Error)	15,39%	10,28%
Number of matches	672	757
Accuracy (Number of matches / Total)	65,50%	73,78%
Total estimated income	\$ 412.105.258	\$ 400.188.500
Accuracy below the average error	79,84%	82,75%
Income weighted error	12,79%	10,33%
Overestimated income	\$ 26.854.797	\$ 15.841.000
Underestimated income	\$ 25.689.539	\$ 26.592.500

Fuente: Elaboración propia

Se tabularon además los niveles de confianza de la estimación, según el error que se podía esperar en cada caso (ver *Tabla 23*).

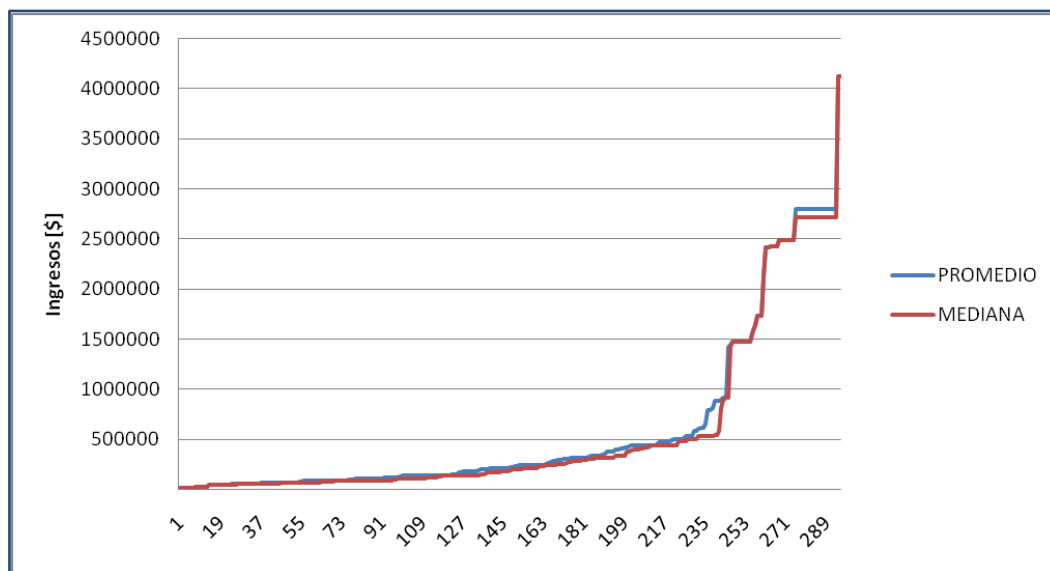
Tabla 23: Nivel de confianza en la estimación de precios

Error esperado de la estimación	Promedio de los datos	Mediana de los datos
1%	67,93%	77,39%
2%	68,91%	78,65%
3%	70,08%	78,95%
4%	71,64%	79,53%
5%	72,32%	80,21%
6%	73,49%	80,99%
7%	74,37%	81,48%
8%	74,66%	81,87%
9%	74,95%	82,26%
10%	75,24%	82,75%

Fuente: Elaboración propia

La segunda muestra de datos, correspondiente a 298 registros que no se encontraban declarados para su facturación, pudo ser valorizada contabilizando un total en ingresos de \$ 210.899.500 utilizando la mediana de los datos, y \$ 220.632.941 si se consideraba el promedio de ellos.

Figura 26: Ingresos estimados para códigos de servicio no declarados en CCP



Fuente: Elaboración propia

Con los valores anteriores para cada código de servicio, se evaluó la capacidad de recuperación de ingresos. Para este análisis, se consideró el mes de Septiembre del año 2010 como fecha de referencia para el nuevo indicador IPR. El total de ingresos perdidos que se estimaron a esa fecha asciende a MM \$ 1.764, de los cuales potencialmente se podría recuperar el 68%, equivalentes a MM \$ 1.195. Se sensibilizaron los resultados con el parámetro θ , que indica los meses de ingresos que se esperan recuperar, y se presentan en la siguiente tabla:

Tabla 24: Resultados para la recuperación de ingresos potenciales

Meses de ingresos que se esperan recuperar (θ)	Total de ingresos no recuperables (pérdida)	Total de ingresos que se espera recuperar	% recuperable del total de ingresos perdidos
$\theta = 1$	\$ 1.553.517.500	\$ 210.899.500	12%
$\theta = 2$	\$ 1.342.618.000	\$ 421.799.000	24%
$\theta = 3$	\$ 1.131.718.500	\$ 632.698.500	36%
$\theta = 4$	\$ 920.819.000	\$ 843.598.000	48%
$\theta = 5$	\$ 735.369.000	\$ 1.029.048.000	58%
$\theta = 6 (T_{max})$	\$ 568.955.500	\$ 1.195.461.500	68%

Fuente: Elaboración propia

6. Discusión y evaluación de los resultados

En lo que sigue se hace un análisis de los resultados obtenidos en los experimentos presentados en el capítulo anterior. Las discusiones se basarán en la formación de los *clusters*, en la correcta clasificación de los elementos a cada uno de ellos, en los precios estimados para los códigos de servicio, y en los indicadores definidos que permiten hacer seguimiento a las pérdidas de ingresos. Adicionalmente, se hace mención a las mejoras realizadas al proceso de control de gestión que realiza el área de Aseguramiento de Ingresos de ENTEL.

6.1 Data mart y automatización de procesos

A través del modelamiento multidimensional se diseñó un repositorio de datos que permitió almacenar información relevante para el proceso de gestión de ingresos que realiza ENTEL.

Este repositorio de datos permite mantener un registro de la información histórica de los servicios privados, accesible de forma rápida para responder las consultas del negocio.

Se diseñó una aplicación para la automatización de tareas y carga de información en el *Data Mart*, lo cual ayudó en gran parte a disminuir los tiempos de procesamiento de las fuentes de datos, evitando errores que se producen por el manejo manual de la información, y generando fácilmente los reportes mensuales que contenían los indicadores de servicios privados.

Como se mencionó en el capítulo 1, el procesamiento de la información tomaba alrededor de 2 días de jornada laboral¹⁶ para los analistas del área de Aseguramiento de Ingresos de ENTEL (aproximadamente 14 horas), lo cual se pudo reducir a tan sólo 3 horas, vale decir, hubo un ahorro en tiempo de 11 horas (78% del total).

El uso operativo de la aplicación y del *Data Mart* se mantuvo por más de un año en la empresa, desde el mes de Septiembre del 2010 hasta Diciembre del 2011. Luego de esa fecha hubieron varios cambios en la realidad del negocio, lo que implicó una modificación tanto de los indicadores que utilizaba el área de Aseguramiento de Ingresos, como software por lo tanto no se cumplían los requerimientos del sistema y las definiciones iniciales del negocio no aplicaban.

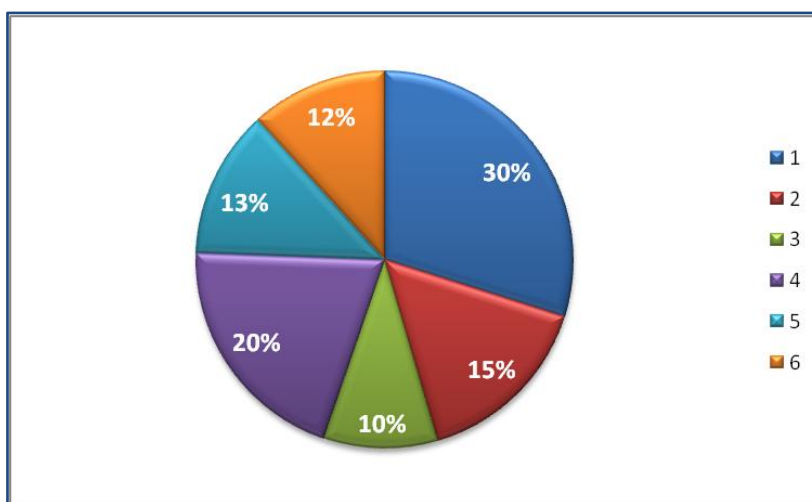
¹⁶ Considerando una jornada de trabajo 8 horas diarias

6.2 Formación de *clusters* de clientes y servicios privados

Una de las propuestas de esta tesis fue extraer conocimiento a partir de los datos disponibles de clientes y servicios privados de ENTEL. Para ello se aplicaron técnicas de *clustering* que permitieran formar grupos cuyos objetos fueran similares entre sí, y que entre grupos los objetos fueran lo más diferentes posibles.

Se formaron 6 *clusters* bien definidos, compactos y de tamaño uniforme, sólo el primero de ellos supera mayormente en cantidad de registros al resto como se aprecia en la siguiente figura:

Figura 27: Tamaño de los *clusters* obtenidos



Fuente: Elaboración propia

De acuerdo al índice de *Davies-Bouldin* calculado ($DB = 1.43$), utilizado para evaluar la calidad de los *clusters*, se pudieron formar grupos muy compactos, con una alta similitud entre sus elementos (*clusters* 1, 2 y 3), y otros que no lo son tanto (*clusters* 4, 5 y 6).

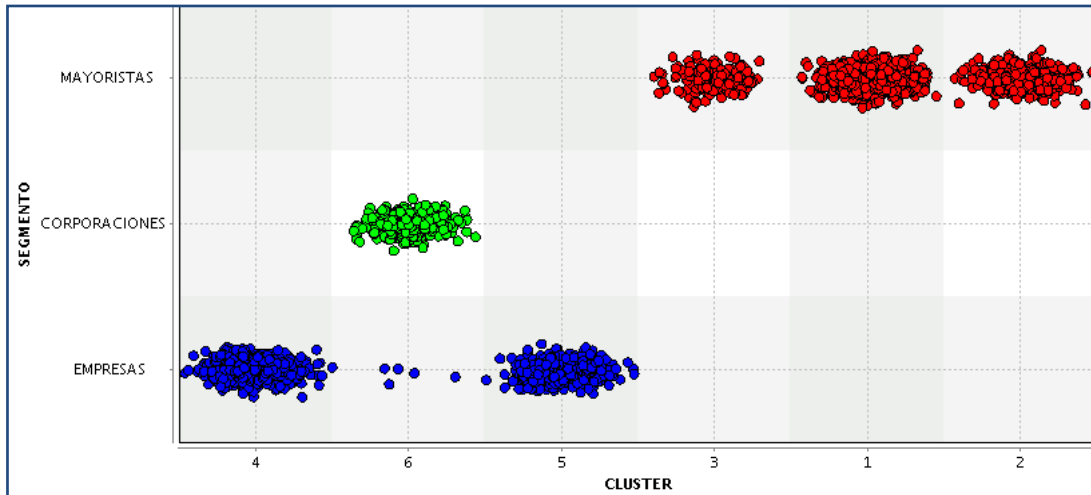
Tabla 25: Distancias promedio de los elementos de cada *cluster* a su centroide

Cluster	Distancia promedio al centroide
1	0.21
2	0.16
3	0.26
4	0.38
5	0.39
6	0.45

Fuente: Elaboración propia

Como se vio en el capítulo anterior, hay características que se destacan en cada *cluster* y que permiten inferir a cuál de ellos debiera pertenecer un nuevo objeto si se tuviera que clasificar. Lo primero que se puede notar es que los clientes del segmento de mercado corporaciones se presentan únicamente en el cluster nº 6 (ver *Figura 28*).

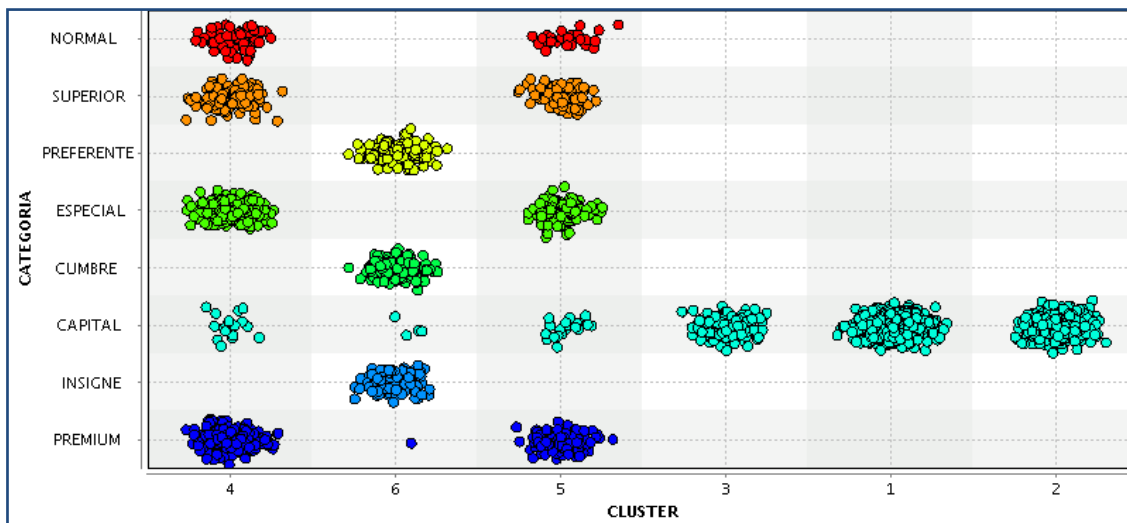
Figura 28: Gráfico de conglomerados según la variable SEGMENTO



Fuente: Resultado obtenido desde el software RapidMiner

Dentro de cada segmento existe una diferenciación en categorías definidas por ENTEL, con ello se puede apreciar la distribución de un tipo de cliente en un grupo.

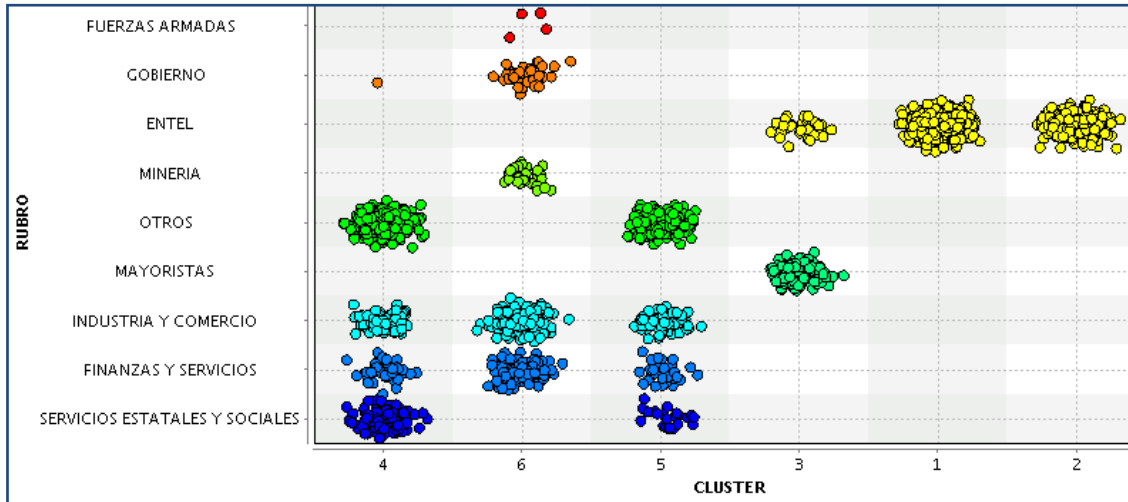
Figura 29: Gráfico de conglomerados según la variable CATEGORIA



Fuente: Resultado obtenido desde el software RapidMiner

Otra característica importante es el rubro o sector económico en que se desempeñan los clientes, por ejemplo aquellos clientes del sector minero sólo se presentan en el *cluster* n° 6, y lo mismo ocurre con los del rubro mayorista en el *cluster* n° 3.

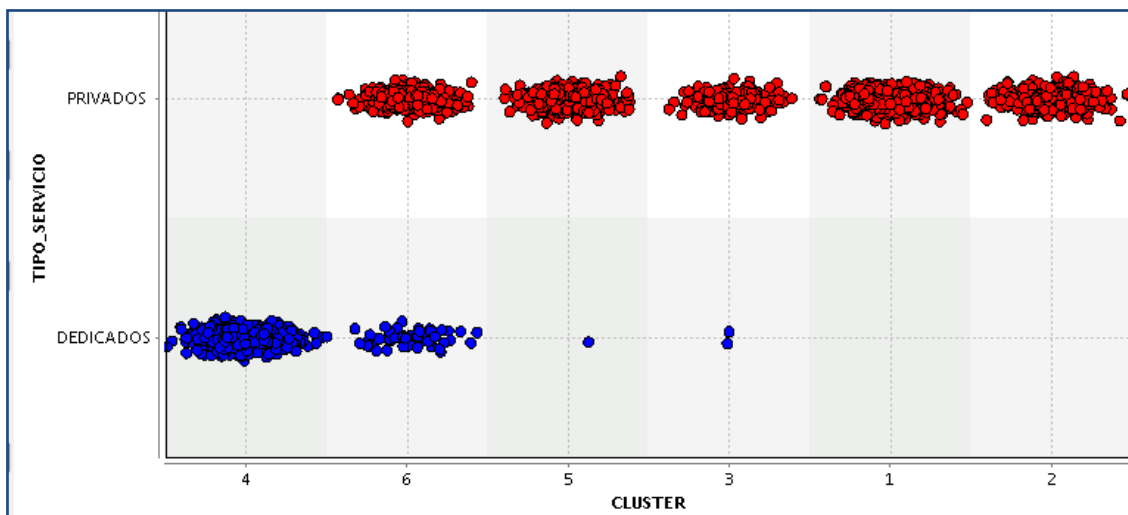
Figura 30: Gráfico de conglomerados según la variable RUBRO



Fuente: Resultado obtenido desde el software RapidMiner

Además de las características de los clientes, también se presentan los atributos de los servicios. Una característica de ellos es el tipo de servicio entregado, por ejemplo aquellos que son dedicados, que se presentan en su mayoría en el cluster n° 4, y en segundo lugar en el n° 6.

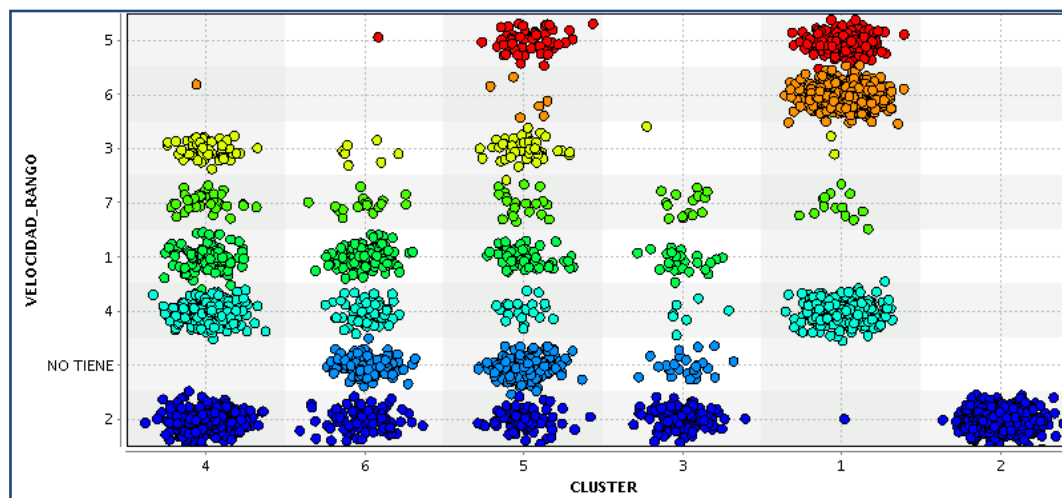
Figura 31: Gráfico de conglomerados según la variable TIPO_SERVICIO



Fuente: Resultado obtenido desde el software RapidMiner

Un último atributo de los servicios es la velocidad de transmisión de datos asociada a los planes que ofrece ENTEL, las cuales se han segmentado en rangos. Se visualiza en la siguiente figura que sólo en el cluster n° 1 se presentan aquellos servicios del rango de velocidades entre 12 y 16 Mb.

Figura 32: Gráfico de conglomerados según la variable VELOCIDAD_RANGO



Fuente: Resultado obtenido desde el software RapidMiner

6.3 Clasificación de los registros a cada *cluster*

De acuerdo a lo que se presentó en el capítulo anterior, se ve que el modelo que utilizó como algoritmo SVM se comportó mejor que aquel donde se usó como clasificador un árbol de decisión.

El modelo entregó muy buenos resultados (99,76% de *accuracy*), obteniendo un clasificador muy preciso, que clasifica de correctamente a los nuevos elementos en los *clusters* ya definidos. Además, en base a la medida F, se afirma que el modelo tiene una alta precisión, pues su valor es muy cercano al 100%.

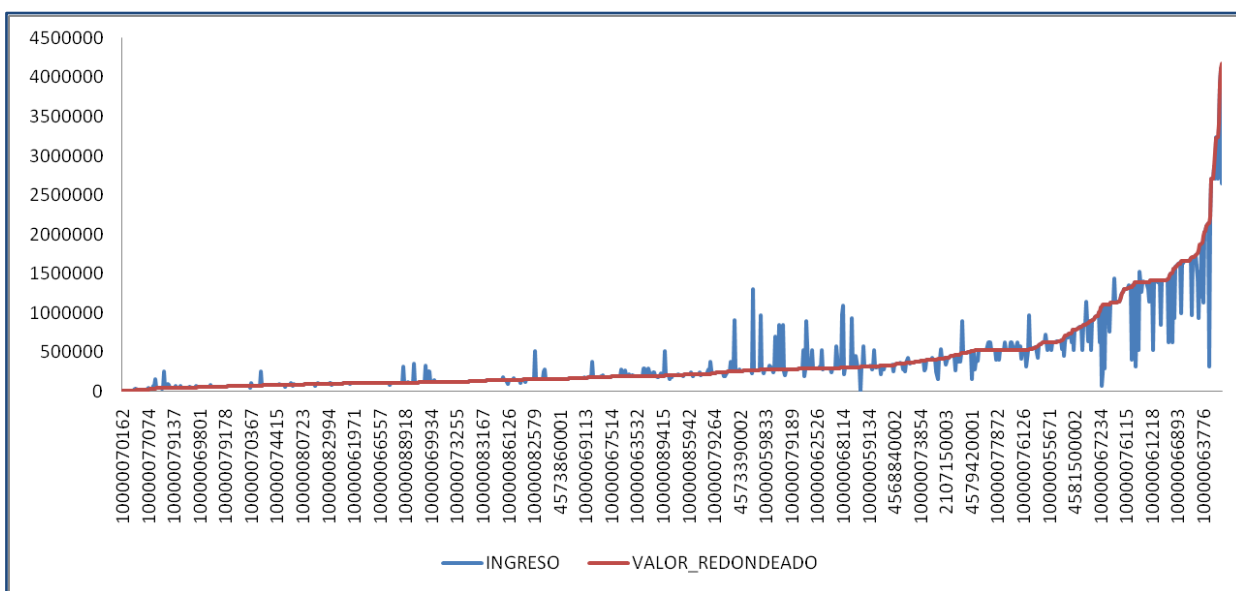
6.4 Valorización de códigos de servicio

Para la estimación de precios se usaron dos funciones para agrupar los valores de precios de aquellos registros que tenían características comunes. Las opciones fueron el promedio y la mediana, y de acuerdo a las estadísticas obtenidas aquella que entregó mejores resultados fue la segunda en cuanto a la cantidad de aciertos en la predicción, al error absoluto y al porcentaje de la muestra que se presentaba con un error aceptable.

En comparación con la mediana, el promedio de los datos sobrestima mucho más los valores con respecto al dato real. De un total de MM \$ 410 en ingresos, un 69% de ellos se pueden explicar con absoluta certeza (MM \$ 282).

Considerando un margen de error del 5% en la estimación, se alcanzó un nivel de confianza de un 80% de la muestra. La fracción anterior representa un 76% del total de los ingresos (MM \$ 310), con errores bajo el margen establecido. La *Figura 33* refleja el resultado de la estimación (línea azul), en relación al valor real de los códigos de servicio (línea roja).

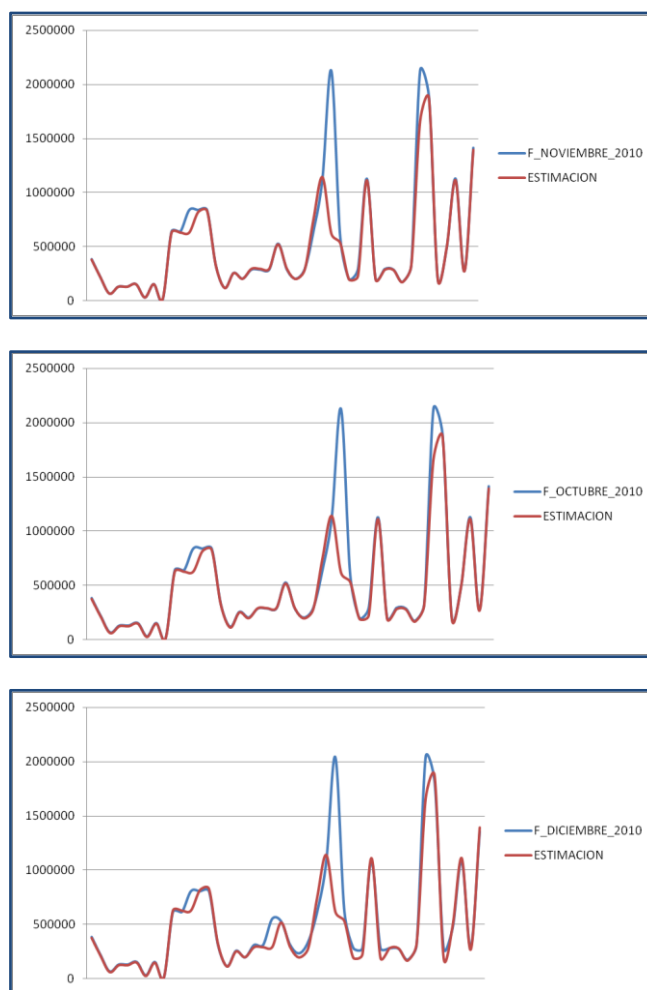
Figura 33: Estimación de ingresos v/s valor real



Fuente: Elaboración propia

Como última validación, se realizó una comparación con datos reales a futuro con respecto a los meses considerados para la predicción, es decir, para aquellos códigos de servicio cuyas facturaciones se realizaron en Octubre, Noviembre y Diciembre del año 2010. Con lo anterior, se comprobó que la estimación mantuvo un comportamiento similar al explicado anteriormente (nivel de confianza del 80% de los datos con errores bajo el 5%). En esta oportunidad se dio que, para una anticipación de 1 y 2 meses en la predicción, el 84,1% de los casos se comportó de esa manera, mientras que para una anticipación de 3 meses hubo una baja en ese porcentaje para el mismo nivel de error, disminuyendo a un 59,1%, el cual sigue siendo un número importante.

Figura 34: Comparación de la predicción de precios para meses futuros.



Fuente: Elaboración propia

La valorización de los códigos de servicio aun no facturados es una gran ayuda para ENTEL, pues hasta el momento no se cuenta con una forma precisa de estimar los ingresos potenciales. Si la empresa toma medidas para regularizar los casos de servicios no declarados para su facturación, se podrían percibir ingresos del orden de cientos de millones (sólo mes de ingresos no facturados contabilizan MM \$ 210).

7. Conclusiones

En este trabajo se trataron los problemas asociados a la fuga de ingresos de una empresa de telecomunicaciones, definiendo para ello una metodología de trabajo con la cual diseñar e implementar mejoras en el proceso de control de gestión. En este sentido se cumplieron todos los objetivos planteados, obteniéndose los resultados asociados a cada uno de ellos:

- ✓ **El estado del arte en técnicas de minería de datos y en modelamiento multidimensional fue estudiado y revisado**, con el fin de entender sobre los temas implicados en este trabajo. La investigación fue documentada en los capítulos 2 y 3 de este documento, donde se describió tanto la metodología de trabajo escogida, como los principales conceptos, algoritmos y técnicas asociadas al ámbito de esta tesis.
- ✓ **Se diseñó, construyó e implementó un *Data Mart* para el área de Aseguramiento de Ingresos de ENTEL**, basado en los requerimientos del negocio, con el fin de consolidar la información de indicadores de servicios privados. Este repositorio de datos se mantuvo operativo por más de 1 año en la empresa, desde su implementación en Septiembre del año 2010 hasta Diciembre del año 2011.
- ✓ **Se disminuyó en un 78% el tiempo de procesamiento y almacenamiento de los datos utilizados para el cálculo de indicadores de servicios privados**, equivalente a 11 horas de jornada laboral de un trabajador, mediante la automatización del proceso de generación de indicadores en reportes mensuales. Lo anterior significa un ahorro importante para la empresa y una mejora en la disponibilidad de la información. **Además, se ha minimizado el riesgo de contar con errores en este proceso**, pues se han incluido validaciones durante el procesamiento de los datos y eliminado tareas que se realizaban de forma manual, con el fin de asegurar la calidad de la información.
- ✓ **Se estimaron MM \$ 210 en ingresos potenciales mensuales debido a servicios no facturados, mediante el desarrollo de un modelo de minería de datos, el cual extrae conocimiento desde los datos de clientes y servicios privados de ENTEL, utilizando técnicas de *clustering* y clasificación**. La estimación de estos ingresos presenta resultados confiables, asegurando que un 80% de los casos tenga un error de un 5% o menos con respecto a su valor real.
- ✓ **Se revisó la forma en que se calculaban los indicadores de servicios privados, proponiendo una mejora que consistió en agregar la variable precio al indicador ISP3**. Adicionalmente, se creó un nuevo indicador que permite medir los ingresos potencialmente recuperables (IPR).

La metodología utilizada en esta tesis se basó principalmente en los pasos o etapas del proceso llamado *Knowledge Discovery in Databases (KDD)*, implementadas según se indica en la metodología *CRISP-DM*, la cual está diseñada para el desarrollo de proyectos de *Data Mining*. En primer lugar se realizó una selección de las fuentes de datos, luego fue necesaria una limpieza y tratamiento ellos, dejándolos en el formato adecuado para poder aplicar algoritmos de *Data Mining*.

Uno de los aspectos a destacar del uso de minería de datos es que debe tomarse como un apoyo para los analistas, y no reemplaza el conocimiento que ellos tienen del negocio. Los resultados de los modelos de *Data Mining* no funcionan por sí solos, y es necesario que se interpreten y validen con los expertos del negocio.

Se ha aplicado con éxito modelamiento multidimensional para construir un sistema de información, basado en un *Data Mart*, que almacena la información necesaria para la creación de indicadores de gestión en el área de Aseguramiento de Ingresos de ENTEL. Asimismo, se ha hecho uso de técnicas de minería de datos para identificar relaciones entre los datos de clientes y servicios, generando conocimiento para la empresa a través de la caracterización de las preferencias de cada uno. Esto último abre muchas posibilidades, pues los clientes se pueden tratar diferenciadamente a través de una segmentación, ofreciéndoles servicios acordes a sus necesidades.

Hasta ahora, la estimación de los ingresos perdidos por concepto de servicios no facturados no se había logrado desarrollar de forma precisa dentro de la empresa. En este trabajo se llega a valores razonables para esa estimación, a través del supuesto que clientes con características similares pueden tener un comportamiento de consumo parecido. Con ese supuesto, se generó un modelo de minería de datos para, en primer lugar, agrupar clientes y servicios en *clusters* y luego, clasificar los elementos con ingresos desconocidos en estos conjuntos de manera de determinarles un valor a partir de la información histórica de consumo del resto, utilizando una medida de similitud basada en la importancia de los atributos de los elementos de un mismo cluster.

Se detectó que la fuga de ingresos en la empresa se produce debido a que no se controla bien el proceso de provisión de servicios, por lo tanto es necesario continuar con las medidas existentes que sirven para monitorear lo que está pasando, y hacer seguimiento a los segmentos de clientes y servicios más críticos, utilizando los resultados que se desprenden de este trabajo, de manera que se acelere el proceso de facturación de los servicios ya instalados, y se establezca el flujo mensual de ingresos.

Como se ha visto, existen problemas no menores en el proceso de provisión de servicios privados de ENTEL y en la gestión de ingresos del mismo proceso, asociados en su mayor parte al tiempo que tardan los analistas en procesar la información que les permite tomar decisiones. Con ayuda de las tecnologías existentes, se demostró que se pueden lograr mejoras sustanciales en la oportunidad y la calidad de esta información.

8. Trabajo a futuro

Un punto a considerar como mejora a futuro es el tema del respaldo de la información en el proceso de generación de indicadores de servicios privados. Esto no fue considerado dentro de los alcances de este trabajo, sin embargo también es importante contar con alguna medida de recuperación de datos históricos ante alguna eventualidad que ocurra en la empresa.

Otro aspecto que puede ser de utilidad es la selección de las variables para el estudio, pues por ser un agente externo a la empresa se tuvo algunas limitantes en cuanto al acceso a la totalidad de la información, por lo tanto para una persona interna de ENTEL y con apoyo de expertos del negocio es más fácil que pueda acceder a otros sistemas y tener una visión más general del negocio, agregando variables que quizás no se consideraron en este estudio.

Como última recomendación, es muy necesario que la empresa destine recursos y que genere políticas de limpieza de los datos que almacenan, para corregir inconsistencias o completar información faltante, ya que la calidad de los resultados obtenidos en los análisis que se realicen dependerán de la calidad de las fuentes utilicen (principio de *garbage in – garbage out*).

9. Referencias

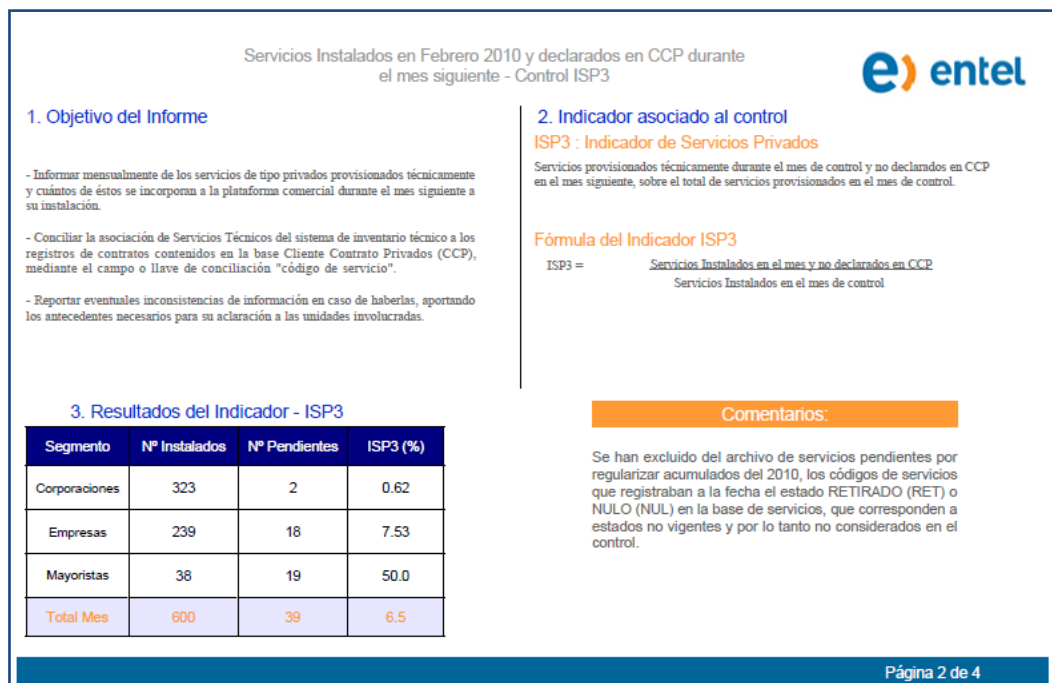
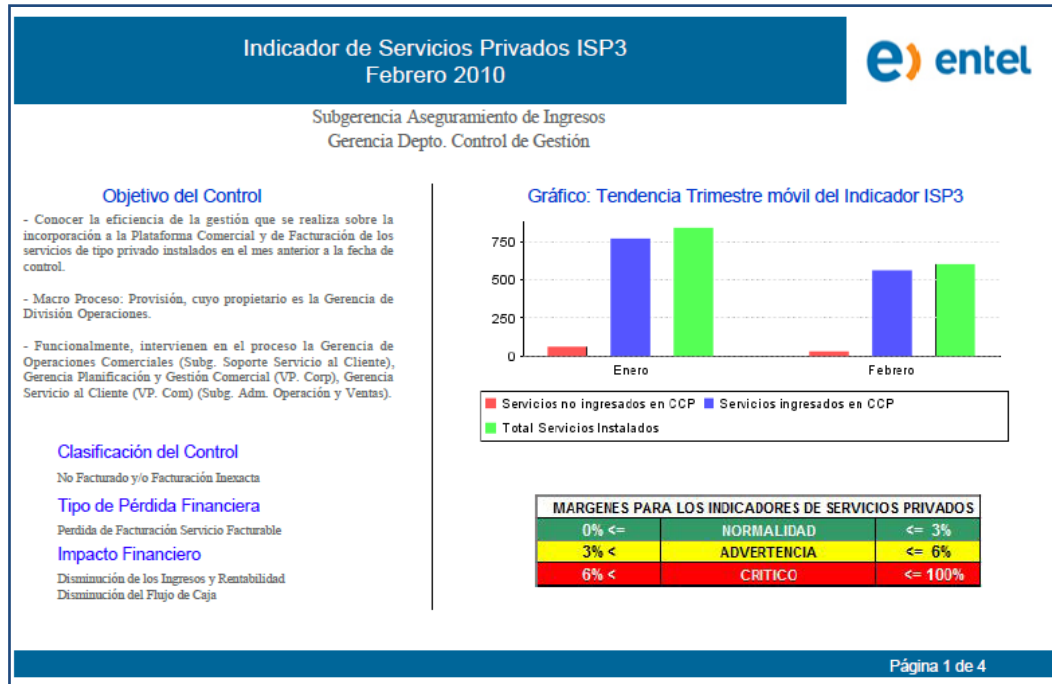
- [1] A. AZEVEDO and M. SANTOS. 2008. *KDD, SEMMA and CRISP-DM: A parallel overview*. IADIS European Conference Data Mining.
- [2] M. J. A. BERRY and G. LINOFF. 1997. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. 2^o ed. Wiley Publishing.
- [3] N. BOLSHAKOVA and F. AZUAJE. 2002. *Cluster validation techniques for genome expression data*. Technical report. Department of Computer Science, Trinity College Dublin.
- [4] S. BORIAH, V. CHANDOLA and V. KUMAR. 2008. *Similarity Measures for Categorical Data: A Comparative Evaluation*. SIAM Data Mining Conference.
- [5] C. J. C. BURGESS. 1998. *A Tutorial on Support Vector Machines for Pattern Recognition*. En: *Data Mining and Knowledge Discovery*. Kluwer Academic Publishers. pp. 121-167.
- [6] J. F. CHANG. 2006. *Business Process Management Systems: Strategy and Implementation*. Auerbach Publications.
- [7] P. CHAPMAN et al. 2000. *CRISP-DM 1.0: Step-by-step Data Mining Guide*. <http://www.crisp-dm.org/CRISPWP-0800.pdf>
- [8] S. CHAUDHURI and U. DAYAL. 1997. *An overview of Data Warehousing and OLAP technology*. ACM Sigmod Rec. pp. 65-74.
- [9] F. DELLAERT. 2002. *The Expectation Maximization Algorithm*. Technical report. College of Computing, Georgia Institute of Technology.
- [10] U. M. FAYYAD, G. PIATETSKY-SHAPIRO and P. SMYTH. 1996. *From data mining to knowledge discovery in Databases: an overview*. Ai Magazine. pp. 37-54.
- [11] E. GERVILLA, R. JIMENEZ, J. J. MONTAÑO, A. SESÉ, B. CAJAL and A. PALMER. 2009. *The methodology of Data Mining: An application to alcohol consumption in teenagers*. Adicciones. pp. 65-80.
- [12] M. HALKIDI, Y. BATISTAKIS, M. VAZIRGIANNIS. 2001. *On clustering validation techniques*. Journal of Intelligent Information Systems. Kluwer Academic Publishers. pp. 107-145.
- [13] J. HERNANDEZ, M. J. RAMIREZ, C. FERRI. 2007. *Introducción a la minería de datos*. Pearson. Prentice Hall.

- [14] *Informe estadístico anual de SubTel Chile*. 2011. <http://www.subtel.gob.cl/>
- [15] W. H. INMON. 2005. *Building the data warehouse*. 4^o ed. Wiley Publishing.
- [16] R. KIMBALL and M. ROSS. 2002. *The Data Warehouse Toolkit: The complete guide to dimensional modeling*. 2^o ed. Wiley Publishing.
- [17] R. KIMBALL and J. CASERTA. 2004. *The Data Warehouse ETL Toolkit: practical techniques for extracting, cleaning, conforming and delivering data*. Wiley Publishing.
- [18] G. L'HUILLIER. 2010. *Apuntes del curso IN643: Introducción a la Minería de Datos*. Depto. de Ingeniería Industrial, Universidad de Chile.
- [19] R. LIU. 2004. *The SPSS two-step cluster*. Technical report. Department of Mathematics, University of North Texas.
- [20] R. MATTISON. 1997. *Cluster Analysis: Basic Concepts and Algorithms*. Artech House, Inc.
- [21] R. MATTISON. 1997. *Data Warehousing and Data Mining for Telecommunications*. Artech House, Inc.
- [22] *Memoria Anual ENTEL*. 2010. <http://www.entel.cl/inversionistas/memorias.html>
- [23] G. W. MILLIGAN and M. C. COOPER. 1985. *An examination of procedures for determining the number of clusters in a data set*. *Psychometrika*, vol. 50. pp. 159-179.
- [24] T. M. MITCHELL. 1997. *Machine Learning*. McGraw-Hill, Inc.
- [25] K. NG and H. LIU. 2000. *Customer Retention via Data Mining*. En: *Artificial Intelligence Review*. Kluwer Academic Publishers. pp. 569-590.
- [26] E. W. T. NGAI, L. XIU and D. C. K. CHAU. 2009. *Application of data mining techniques in customer relationship management: A literature review and classification*. En: *Expert Systems with Applications*, vol. 36, issue 2, part 2. pp. 2592-2602.
- [27] D. T. PHAM, S. S. DIMOV and C. D. NGUYEN. 2005. *Selection of K in K-means clustering*. In *Proceedings of the ImechE, part C*. Journal of Mechanical Engineering Science.
- [28] L. ROKACH and O. Z. MAIMON. 2008. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing.
- [29] TWO CROWS CORPORATION. 1999. *Introduction to Data Mining and Knowledge Discovery*. 3^o ed. Two Crows Corporation.
- [30] J. D. VELASQUEZ and V. PALADE. 2008. *Adaptive Web site: A Knowledge Extraction from Web Data Approach*. IOS Press.

- [31] J. D. VELASQUEZ. 2008. *Apuntes del curso IN55A: Sistemas de Información Administrativos*. Depto. de Ingeniería Industrial, Universidad de Chile.
- [32] J. D. VELASQUEZ. 2009. *Apuntes del curso IN830: Data Warehousing*. Depto. de Ingeniería Industrial, Universidad de Chile.
- [33] J. D. VELASQUEZ. 2009. *Apuntes del curso IN831: Web Mining*. Depto. de Ingeniería Industrial, Universidad de Chile.
- [34] G. M. WEISS. 2005. *Data Mining in Telecommunications*. En: *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers. pp. 1189-1201.
- [35] I. H. WITTEN and E. FRANK. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.
- [36] X. WU et al. 2007. *Top 10 algorithms in data mining*. Springer-Verlag.
- [37] R. XU and D. C. WUNSCH II. 2005. *Survey of Clustering Algorithms*. *IEEE Transactions on neural networks*.
- [38] R. XU and D. C. WUNSCH II. 2009. *Clustering*. Wiley Publishing.

10. Anexos

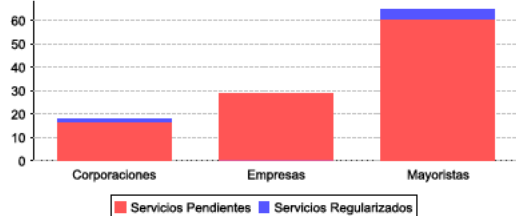
10.1 Anexo 1: Reporte mensual con indicador ISP3



4. Servicios con OTC de instalación finalizadas para los meses controlados del 2010

Mes de Control	Total Servicios	Total Servicios Declarados en CCP						Total Servicios Por Declarar en CCP									
		Global del Mes		Corporaciones		Empresas		Mayoristas		Global del Mes		Corporaciones		Empresas		Mayoristas	
		Q	%	Q	%	Q	%	Q	%	Q	%	Q	%	Q	%	Q	%
Enero	839	771	91.9	484	57.69	216	25.74	71	16.57	68	8.1	15	1.79	11	1.31	42	5.01
Febrero	600	561	93.5	321	53.5	221	36.83	19	9.67	39	6.5	2	0.33	18	3.0	19	3.17
Total Acumulado 2010	1439	1332	92.56	805		437		90		107	7.43	17		29		61	

Gráfico: Inconsistencias detectadas en el mes de control



Comentarios:

El control ISP3 ha detectado mes a mes una cantidad de servicios vigentes cuyas instalaciones, a pesar de estar finalizadas no han sido declaradas en CCP transcurrido 1 mes desde el término de su instalación.

Del total de casos de códigos de servicio pendientes, tras ser notificados por los controles ISP3, se ha regularizado en parte la situación.

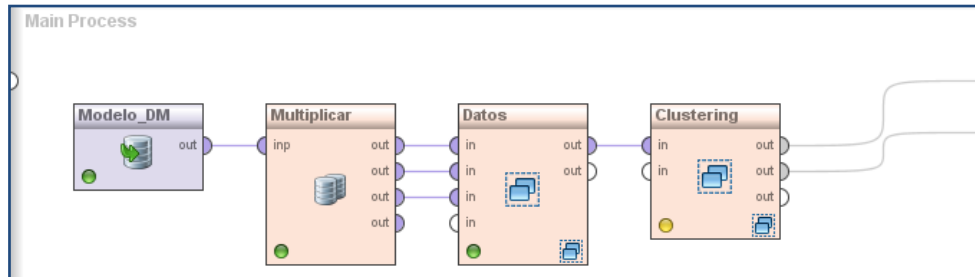
Los datos se presentan en la tabla "Gestión acumulada de regularización"

Tabla: Gestión anual acumulada de regularización

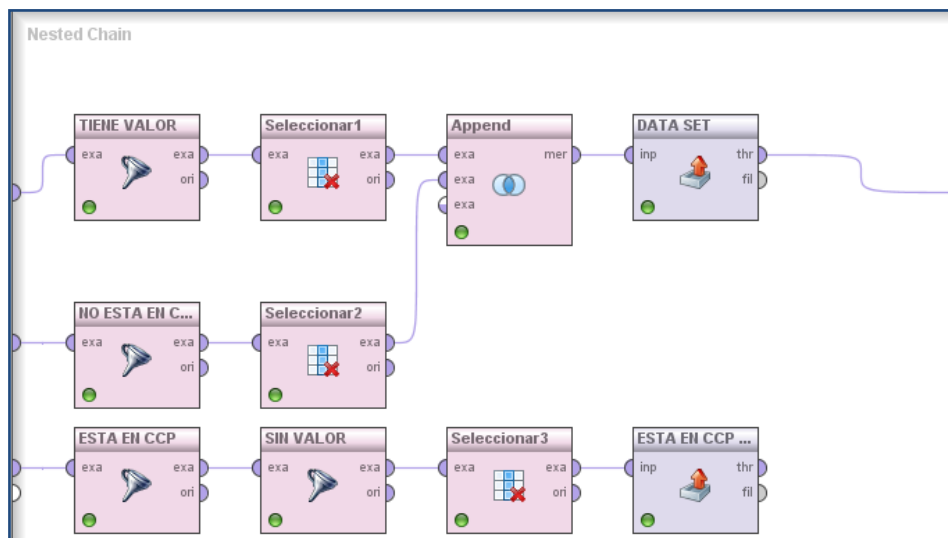
Segmento	Nº Ingresados	Nº Pendientes	Nº Regularizados
Corporaciones	804	17	1
Empresas	437	29	0
Mayoristas	86	61	4
Total Mes	1327	107	5

10.2 Anexo 2: Procesamiento de datos en software *RapidMiner*

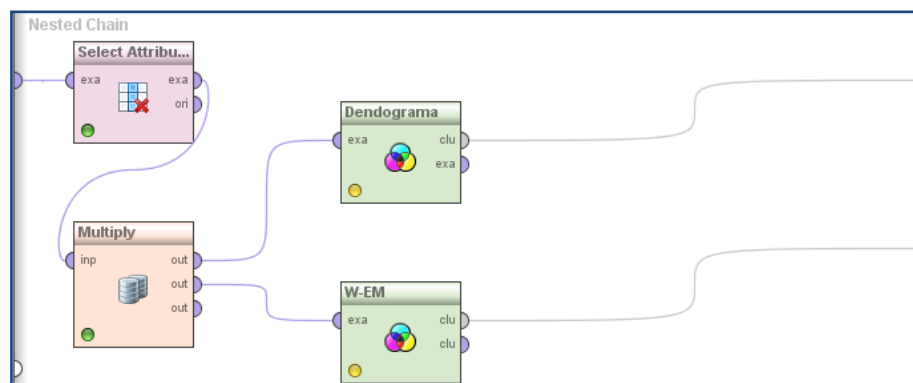
10.2.1 Proceso de clustering de datos



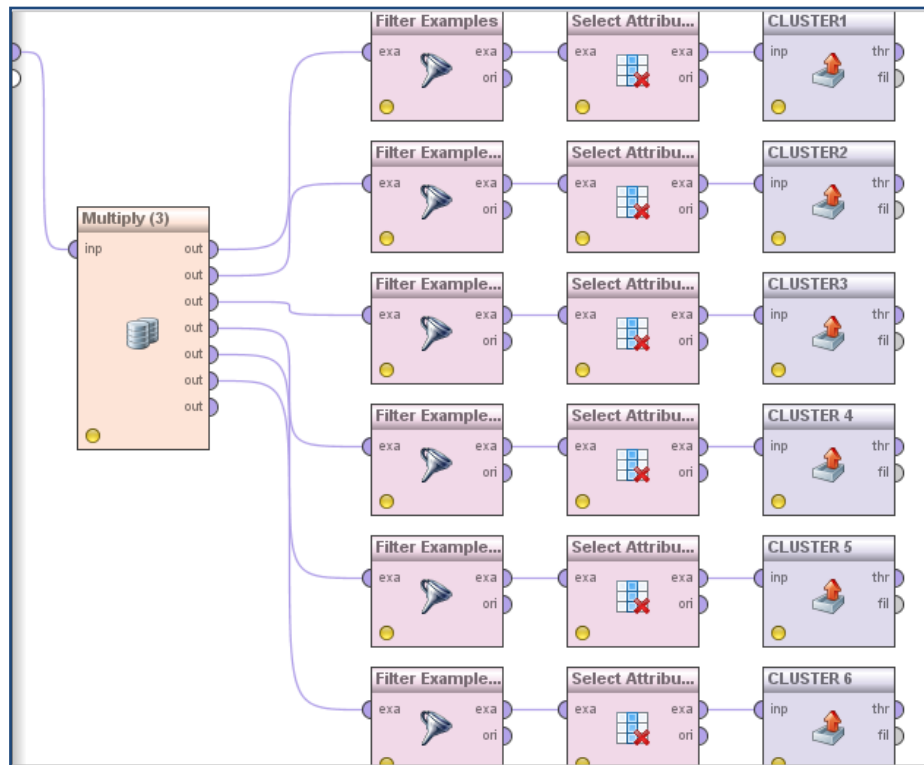
10.2.1.1 Selección de los datos



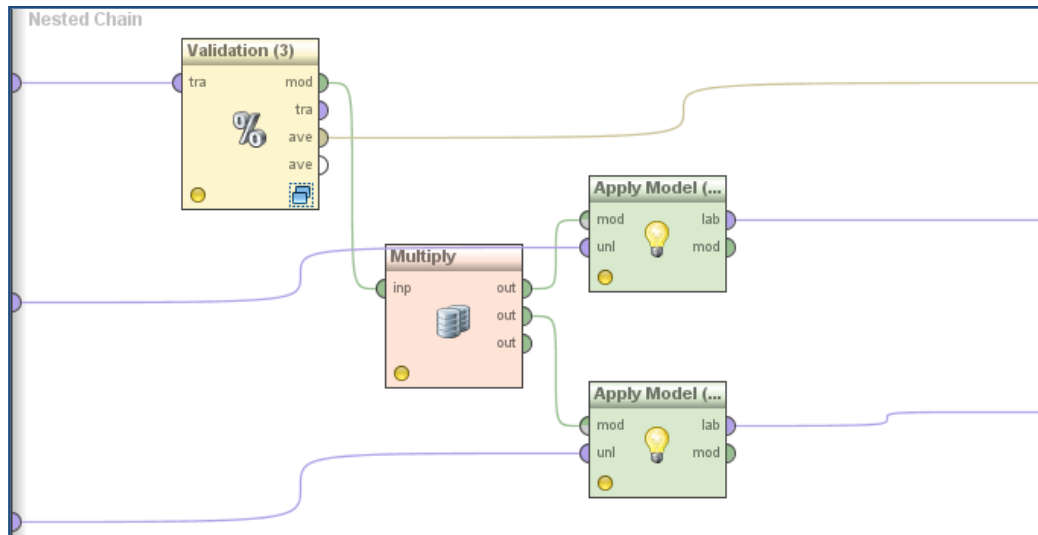
10.2.1.2 Algoritmos para obtener cantidad de clusters



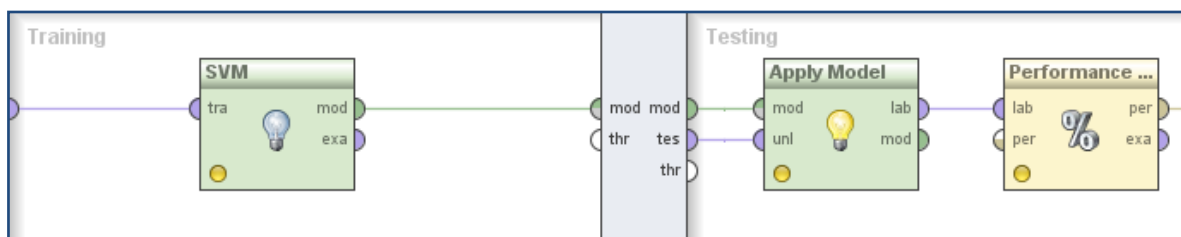
10.2.2.2 Grupos de clusters



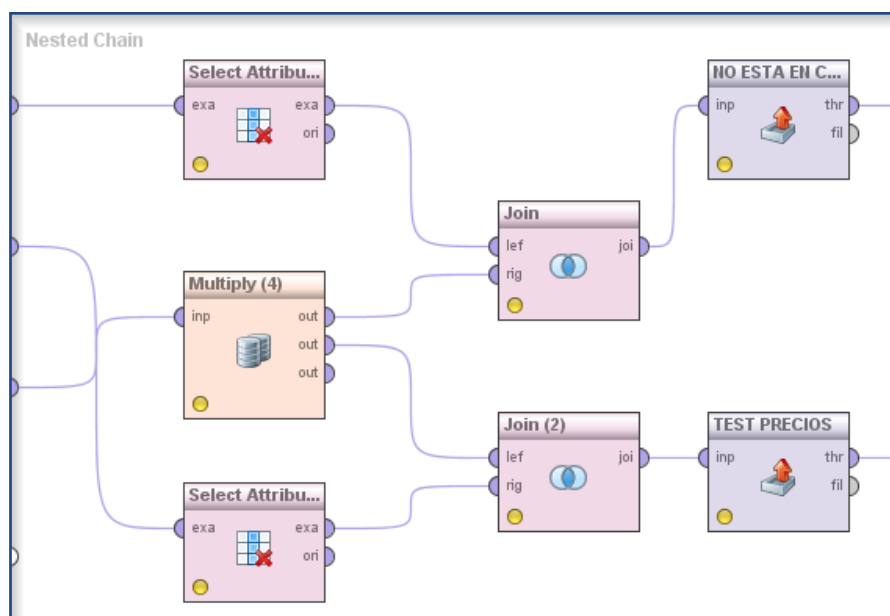
10.2.2.3 Aplicación de modelo de clasificación



10.2.2.3.1 Entrenamiento y validación del modelo mediante Cross-validation

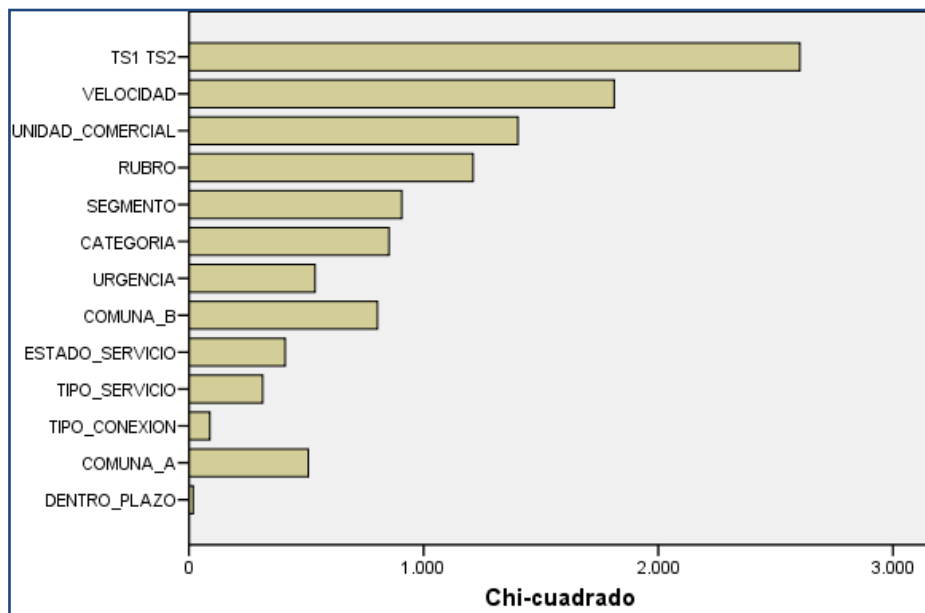


10.2.2.4 Resultados obtenidos por el modelo de clasificación

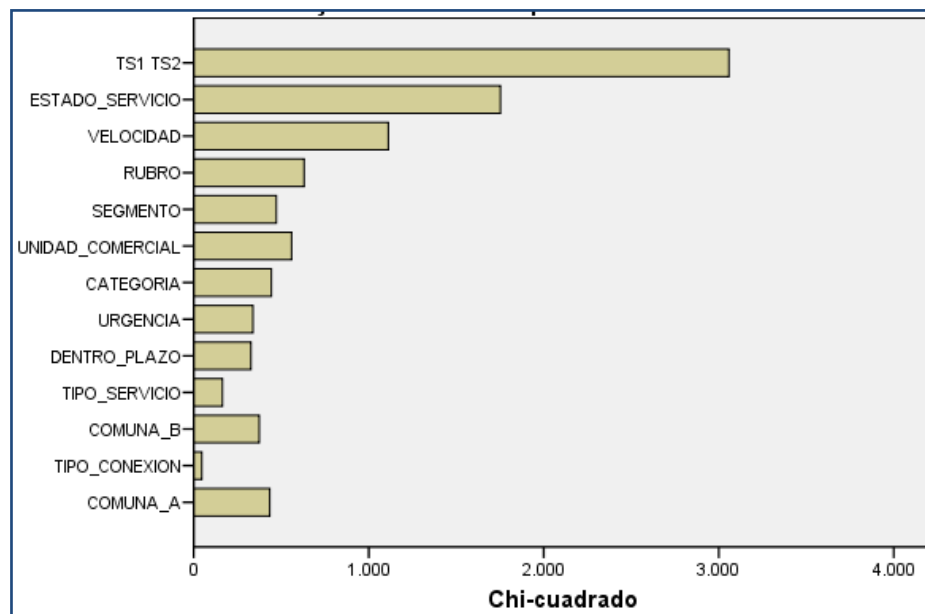


10.3 Anexo 3: Importancia de los atributos en cada *cluster*

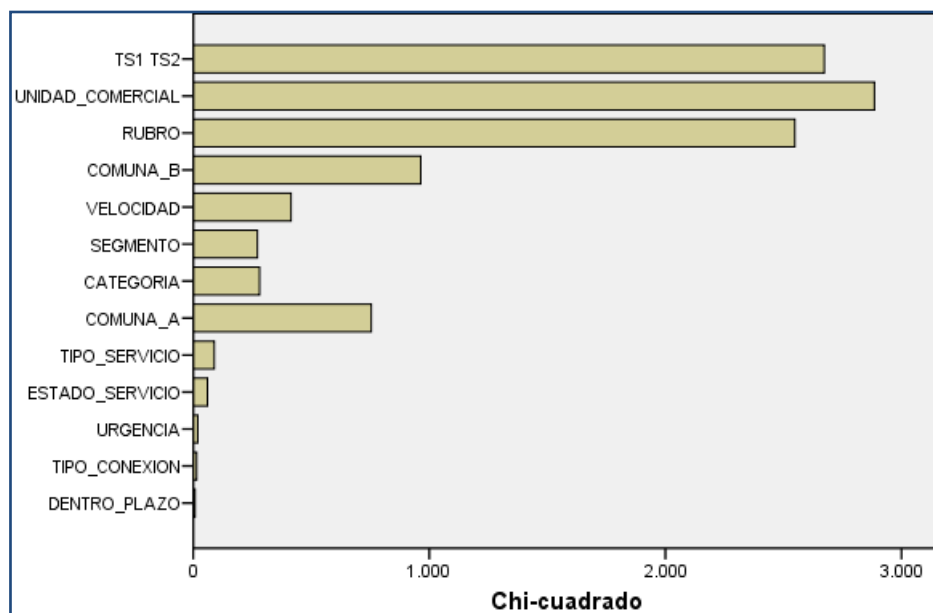
10.3.1 Importancia de los atributos para el conglomerado N° 1



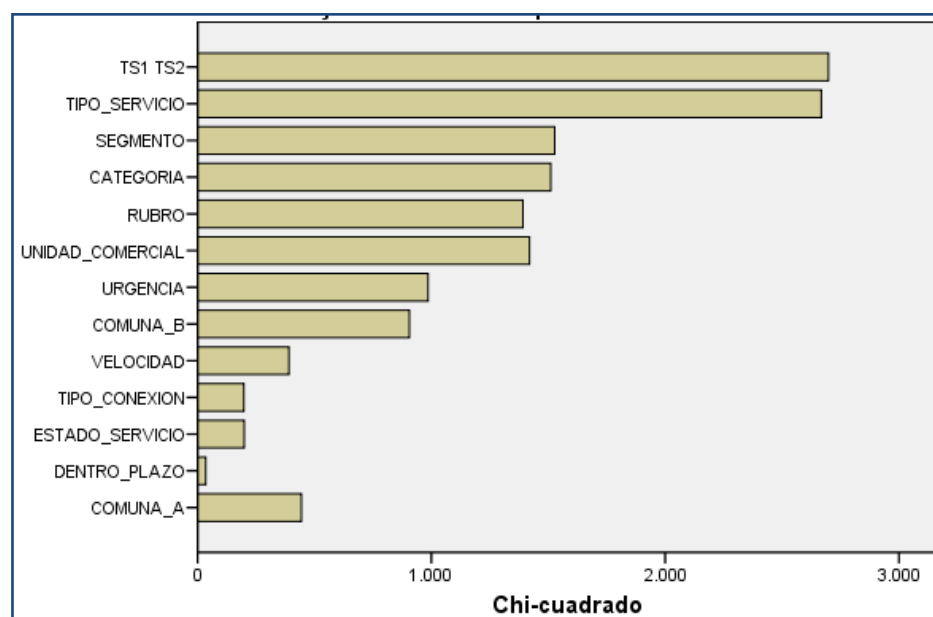
10.3.2 Importancia de los atributos para el conglomerado N° 2



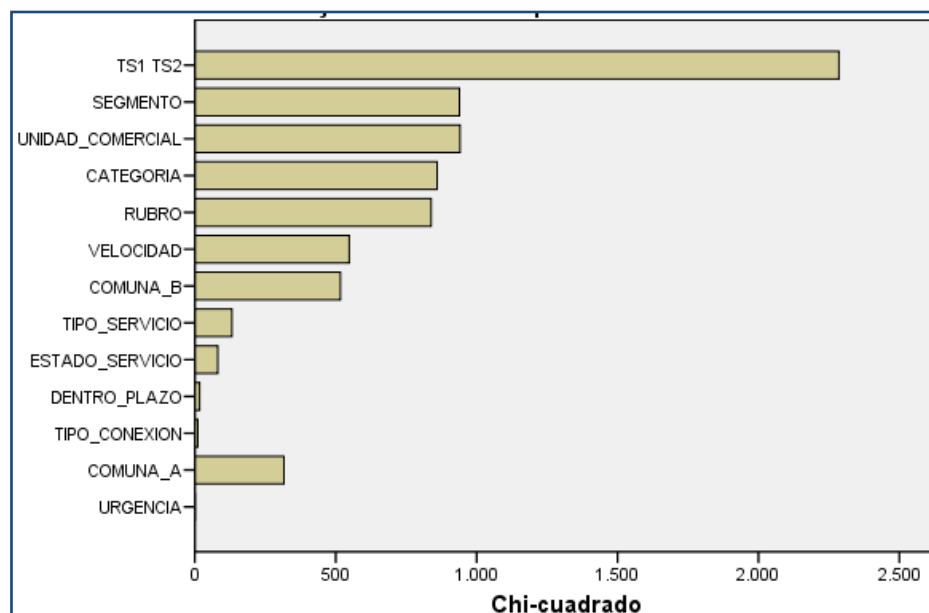
10.3.3 Importancia de los atributos para el conglomerado N° 3



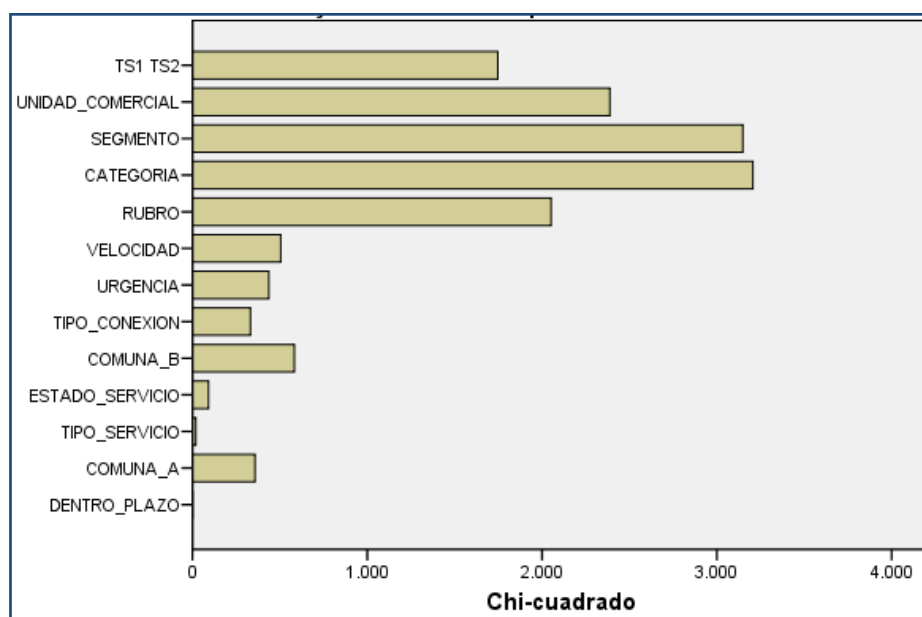
10.3.4 Importancia de los atributos para el conglomerado N° 4



10.3.5 Importancia de los atributos para el conglomerado N° 5



10.3.6 Importancia de los atributos para el conglomerado N° 6



10.4 Anexo 4: Tablas de frecuencia según atributos de cada *cluster*

UNIDAD_COMERCIAL	C1	C2	C3	C4	C5	C6	Total
VAN	0	0	0	21	6	10	37
VANCAL	0	0	0	3	1	4	8
VAR	0	0	0	9	0	1	10
VCAFN	0	0	1	3	5	256	265
VCAGF	0	0	0	0	3	9	12
VCAIC	0	0	0	0	1	31	32
VCARN	0	0	0	0	1	10	11
VCH	0	0	0	22	4	0	26
VCO	0	0	0	22	10	15	47
VCP	0	0	0	13	8	0	21
VCVFN	0	0	0	0	3	0	3
VCVGF	0	0	0	0	0	2	2
VCY	0	0	0	0	2	0	2
VEMP	0	0	0	497	372	50	919
VIQ	0	0	0	9	11	1	21
VLS	0	0	0	33	14	10	57
VMAY	1	48	364	0	0	1	414
VOS	0	0	0	2	1	9	12
VPA	0	0	0	1	3	0	4
VPCS	1110	530	1	0	0	0	1641
VPM	0	0	0	8	3	2	13
VRN	0	0	0	52	12	0	64
VTA	0	0	0	7	0	0	7
VTE	0	0	0	5	1	0	6
VVD	0	0	0	8	3	21	32
VVP	0	0	0	40	9	3	52

VELOCIDAD	C1	C2	C3	C4	C5	C6	Total
1	0	0	29	90	58	110	287
2	1	578	155	373	75	90	1272
3	14	0	3	50	51	7	125
4	339	0	8	207	22	56	632
5	173	0	0	0	64	1	238
6	570	0	0	1	6	0	577
7	14	0	71	34	22	20	161
NO TIENE	0	0	100	0	175	151	426

TIPO_CONEXION	C1	C2	C3	C4	C5	C6	Total
ALAMBRICA	1111	578	357	643	428	326	3443
INALAMBRICA	0	0	3	112	20	36	171
NO APLICA	0	0	6	0	25	73	104

TIPO_SERVICIO	C1	C2	C3	C4	C5	C6	Total
DEDICADOS	0	0	6	755	1	58	820
PRIVADOS	1111	578	360	0	472	377	2898

ESTADO_SERVICIO	C1	C2	C3	C4	C5	C6	Total
ENI	0	344	0	0	0	0	344
ENR	12	7	2	5	0	0	26
MOD	241	5	9	11	12	6	284
NUL	0	0	0	0	0	1	1
RET	0	0	2	15	7	7	31
SER	858	222	353	703	454	421	3011
SUS	0	0	0	21	0	0	21

URGENCIA	C1	C2	C3	C4	C5	C6	Total
EXPRESS	39	1	95	697	176	372	1380
NORMAL	1072	577	271	58	297	63	2338

DENTRO_PLAZO	C1	C2	C3	C4	C5	C6	Total
NO	519	523	168	319	206	234	1969
SI	592	55	198	436	267	201	1749

SEGMENTO	C1	C2	C3	C4	C5	C6	Total
CORPORACIONES	0	0	8	0	3	430	441
EMPRESAS	0	0	0	755	470	5	1230
MAYORISTAS	1111	578	358	0	0	0	2047

CATEGORIA	C1	C2	C3	C4	C5	C6	Total
CAPITAL	1111	578	366	16	25	8	2104
CUMBRE	0	0	0	0	0	164	164
ESPECIAL	0	0	0	244	159	0	403
INSIGNE	0	0	0	0	0	122	122
NORMAL	0	0	0	64	32	0	96
PREFERENTE	0	0	0	0	0	140	140
PREMIUM	0	0	0	252	148	1	401
SUPERIOR	0	0	0	179	109	0	288

RUBRO	C1	C2	C3	C4	C5	C6	Total
ENTEL	1111	578	83	0	3	4	1779
FINANZAS Y SERVICIOS	0	0	0	53	34	149	236
FUERZAS ARMADAS	0	0	0	0	0	7	7
GOBIERNO	0	0	0	1	0	49	50
INDUSTRIA Y COMERCIO	0	0	0	76	71	197	344
MAYORISTAS	0	0	283	0	0	0	283
MINERIA	0	0	0	0	0	29	29
OTROS	0	0	0	471	343	0	814
SERVICIOS ESTATALES Y SOCIALES	0	0	0	154	22	0	176

TS1 TS2	C1	C2	C3	C4	C5	C6	Total
ACCE ATM	0	0	0	0	0	1	1
ACCE BBGE	0	0	2	0	0	0	2
ACCE FO	0	0	0	0	1	7	8
ACCE METH	0	0	11	0	0	0	11
ACCE MPAD	0	0	0	0	10	29	39
ACCE MPAR	0	0	0	0	1	0	1
ACCE MPET	1111	1	0	0	0	0	1112
ACCE MPFR	0	0	0	0	0	1	1
ACCE MPGD	0	0	7	0	32	34	73
ACCE MPLI	0	0	2	0	1	1	4
ACCE MPLS	0	0	55	0	47	62	164
ACCE MPNG	0	0	2	0	51	47	100
ACCE MPVS	0	0	0	0	2	19	21
ACCE MPWI	0	0	0	0	7	11	18
ACCE MS	0	0	2	0	0	0	2
ACCE MSIB	0	0	0	0	1	1	2
ACCE MSIR	0	0	1	0	0	0	1
ACCE MVS2	0	0	0	0	0	1	1
ACCE TKDT	0	0	0	0	0	1	1
ACCE TKIN	0	0	0	0	137	1	138
ACCE VSAT	0	0	1	0	0	0	1
ACCL PEXT	0	0	0	0	1	0	1
ACIR ADSL	0	0	0	0	2	0	2
ACIR GDSL	0	0	20	0	0	0	20
ACTK DTIN	0	0	0	0	2	0	2
ADR LWAN	0	0	0	0	2	5	7
ADR STUD	0	0	0	0	0	6	6
ARRI CETF	0	0	0	0	0	3	3
ARRI FO	0	0	25	0	0	0	25
ARRI LP	0	0	1	0	0	0	1
CEHZ DT	0	0	0	0	2	0	2
DTPR FR	0	0	0	0	1	0	1
DTPR FRPP	0	0	0	0	1	1	2
DTPR IEE	0	0	12	0	0	0	12
DTPR MIC	0	2	114	0	0	2	118
DTPR PP	0	0	13	0	1	13	27
DTPR PVC	0	0	0	0	2	0	2
DTPR PVCB	0	0	0	0	1	0	1
DTPR PVCR	0	0	17	0	1	0	18
ENLA FO	0	0	2	0	0	2	4
HOUS INFR	0	0	2	0	0	0	2
IBS PP	0	0	2	0	0	0	2

TS1 TS2	C1	C2	C3	C4	C5	C6	Total
INTE ADS	0	0	0	8	1	0	9
INTE FR	0	0	0	1	0	1	2
INTE GDSL	0	0	0	117	0	11	128
INTE ITIP	0	0	8	0	0	0	8
INTE MOTI	0	0	0	0	0	1	1
INTE MPLS	0	0	0	321	0	32	353
INTE MS	0	0	0	2	0	0	2
INTE NGN	0	0	1	199	0	12	212
INTE PP	0	0	5	0	0	0	5
INTE WMAX	0	0	0	107	0	1	108
OUTF REXT	0	0	0	0	0	1	1
PRAP WIFI	0	0	0	0	8	2	10
PRAX TFIP	0	0	0	0	0	12	12
PRGW TFIP	0	0	0	0	0	1	1
PRND CETF	0	0	0	0	2	0	2
PRND ROUT	0	0	0	0	0	4	4
PRND SWCH	0	0	0	0	4	34	38
PRND TFIP	0	0	0	0	15	6	21
PVC GPRS	0	0	0	0	0	38	38
PVEI MULT	0	0	0	0	0	9	9
SCES HOR	0	0	0	0	0	1	1
TFPR ANX	0	0	0	0	1	1	2
TFPR MIC	0	575	2	0	6	2	585
TFPR PP	0	0	1	0	1	4	6
TRAN ETH	0	0	3	0	0	1	4
TRAN IDS3	0	0	4	0	0	0	4
TRAN IE1	0	0	1	0	0	0	1
TRAN IS16	0	0	1	0	0	0	1
TRAN IST1	0	0	36	0	0	0	36
TV OMNI	0	0	1	0	0	0	1
TV PP	0	0	2	0	0	0	2
TV STRX	0	0	0	0	0	1	1
VCON MULT	0	0	0	0	0	1	1
VLAN BBGE	0	0	4	0	0	0	4
VLAN GPRS	0	0	0	0	0	1	1
VLAN METH	0	0	2	0	0	0	2
VPNE MPLS	0	0	0	0	1	0	1
VPNI MPLS	0	0	4	0	125	3	132
VTAP WIFI	0	0	0	0	2	0	2
VTEQ GRAL	0	0	0	0	1	5	6
VTEQ TELE	0	0	0	0	0	1	1
VTET TF	0	0	0	0	0	1	1

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
ALGARROBO	7	0	0	0	0	0	7
ALHUE	2	0	0	1	0	0	3
ALTO HOSPICIO	1	1	0	0	2	3	7
ANCUD	2	0	2	0	0	0	4
ANDACOLLO	0	2	0	0	0	0	2
ANGOL	1	0	2	2	0	1	6
ANTOFAGASTA	25	33	7	27	8	22	122
ANTUCO	0	1	0	0	0	1	2
ARAUCO	2	0	0	0	0	0	2
ARICA	14	22	10	10	2	5	63
AYSEN	3	0	0	0	0	1	4
BOLIVIA	0	0	5	0	0	0	5
BUIN	2	2	0	0	2	1	7
BULNES	0	0	2	0	0	0	2
CALAMA	18	12	2	9	2	7	50
CALBUCO	0	1	0	0	0	0	1
CALDERA	0	0	0	1	0	3	4
CALERA	2	1	0	0	0	0	3
CALERA DE TANGO	6	2	0	0	0	0	8
CALLE LARGA	2	0	0	0	0	0	2
CANELA	2	0	1	0	0	0	3
CANETE	3	3	0	1	0	0	7
CARAHUE	1	0	0	3	0	0	4
CARTAGENA	3	1	0	0	0	0	4
CASABLANCA	2	2	0	1	1	0	6
CASTRO	4	0	2	4	0	4	14
CATEMU	1	0	0	0	0	0	1
CAUQUENES	1	0	3	1	0	3	8
CERRILLOS	5	2	4	5	10	5	31
CERRO NAVIA	7	1	0	1	0	0	9
CHAITEN	0	4	0	0	0	0	4
CHANARAL	2	0	0	1	0	0	3
CHANCO	1	0	0	0	0	0	1
CHEPICA	0	0	1	0	0	0	1
CHIGUAYANTE	1	0	4	0	0	1	6
CHILE CHICO	0	0	0	0	0	1	1
CHILLAN	16	9	3	9	3	4	44
CHIMBARONGO	1	0	0	0	0	2	3
CHOLCHOL	1	0	0	0	0	0	1
CHONCHI	1	0	0	0	0	0	1
CISNES	1	1	0	0	0	1	3
COCHRANE	0	0	0	0	0	1	1

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
CODEGUA	1	0	0	0	0	0	1
COELEMU	1	0	0	2	0	0	3
COIHUECO	2	1	0	0	0	0	3
COINCO	2	0	0	0	0	0	2
COLBUN	3	0	0	0	0	0	3
COLCHANE	0	1	0	0	0	0	1
COLINA	11	10	2	4	12	0	39
COLLIPULLI	1	0	0	0	0	0	1
COMBARBALA	1	1	0	0	0	0	2
CONCEPCION	18	20	1	16	9	11	75
CONCHALI	3	0	2	1	0	2	8
CONCON	2	0	0	1	0	2	5
CONSTITUCION	6	2	3	1	0	3	15
CONTULMO	1	0	0	0	0	0	1
COPIAPO	12	2	6	12	8	3	43
COQUIMBO	18	5	1	12	3	2	41
CORONEL	6	6	0	2	1	1	16
CORRAL	0	1	0	0	0	0	1
COYHAIQUE	4	1	6	0	4	3	18
CUNCO	1	0	0	0	0	1	2
CURACAUTIN	1	1	1	0	0	0	3
CURACAVI	1	2	0	0	0	0	3
CURACO DE VELEZ	0	0	0	1	0	0	1
CURANILAHUE	0	0	0	0	0	1	1
CURARREHUE	1	0	0	0	0	0	1
CUREPTO	0	0	0	6	0	0	6
CURICO	14	1	7	11	2	4	39
DIEGO DE ALMAGRO	1	5	2	2	0	2	12
DONIHUE	2	0	0	0	0	0	2
EL BOSQUE	6	0	0	0	0	0	6
EL CARMEN	1	0	0	2	0	0	3
EL MONTE	1	1	0	0	0	0	2
EL QUISCO	0	2	1	0	0	0	3
EL TABO	3	2	0	0	0	0	5
EMPEDRADO	0	0	1	0	0	0	1
ESTACION CENTRAL	5	3	2	1	1	2	14
FLORIDA	1	0	1	0	0	1	3
FRANCIA	0	0	1	0	0	0	1
FREIRE	1	0	0	0	0	0	1
FRESIA	1	0	0	0	0	0	1
FRUTILLAR	1	0	0	0	0	0	1
FUTALEUFU	0	2	0	0	0	0	2

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
FUTRONO	0	1	0	0	0	0	1
GALVARINO	1	1	0	0	0	0	2
GUAITECAS	0	0	0	0	0	1	1
HOLANDA	0	0	0	0	0	1	1
HUALANE	0	1	0	0	0	0	1
HUALPEN	2	2	1	0	0	2	7
HUARA	0	2	0	1	0	0	3
HUASCO	2	0	0	0	2	2	6
HUECHURABA	7	3	13	13	14	4	54
ILLAPEL	3	0	2	1	0	0	6
INDEPENDENCIA	4	9	2	6	3	1	25
IQUIQUE	13	21	3	7	14	11	69
ISLA DE MAIPO	4	1	0	1	0	0	6
ISLA DE PASCUA	0	0	1	0	0	0	1
JUAN FERNANDEZ	0	0	0	0	0	1	1
LA CISTERNA	1	1	4	1	1	0	8
LA CRUZ	0	1	0	0	0	0	1
LA ESTRELLA	0	1	0	0	0	0	1
LA FLORIDA	19	7	2	3	2	4	37
LA GRANJA	4	1	0	0	0	0	5
LA HIGUERA	1	0	0	1	0	0	2
LA LIGUA	0	0	2	2	1	0	5
LA PINTANA	8	0	0	3	0	0	11
LA REINA	12	4	1	7	5	1	30
LA SERENA	21	8	6	9	7	12	63
LA UNION	1	1	2	0	0	0	4
LAGO RANCO	1	2	0	0	0	1	4
LAJA	1	1	0	1	0	0	3
LAMPA	6	3	1	10	6	0	26
LANCO	1	0	0	1	0	0	2
LAS CABRAS	3	0	0	0	0	0	3
LAS CONDES	35	24	17	75	51	27	229
LAUTARO	0	0	0	2	0	2	4
LEBU	1	1	0	0	0	0	2
LICANTEN	1	3	0	2	0	1	7
LIMACHE	1	0	0	0	0	1	2
LINARES	5	5	2	1	0	5	18
LITUECHE	1	0	0	0	0	0	1
LLANQUIHUE	1	0	0	0	1	0	2
LO BARNECHEA	16	9	1	0	1	0	27
LO ESPEJO	2	0	0	1	0	0	3
LO PRADO	3	3	0	11	0	1	18

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
LOLOL	1	0	0	0	0	0	1
LONGAVI	1	0	0	0	0	0	1
LONQUIMAY	1	0	0	0	0	1	2
LOS ALAMOS	1	0	0	0	0	0	1
LOS ANDES	7	3	0	10	0	7	27
LOS ANGELES	14	2	0	6	1	10	33
LOS MUERMOS	1	0	0	0	0	0	1
LOS VILOS	2	0	0	1	2	0	5
LOTA	1	1	0	0	0	1	3
MACHALI	2	2	0	0	0	0	4
MACUL	1	0	0	6	3	5	15
MAFIL	1	0	0	0	0	0	1
MAIPU	25	10	0	0	1	0	36
MARCHIHUE	1	3	0	0	0	1	5
MARIA ELENA	1	2	1	0	0	1	5
MARIA PINTO	1	0	0	1	0	0	2
MARIQUINA	1	1	0	2	0	2	6
MAULE	0	0	0	1	0	0	1
MAULLIN	1	1	0	0	0	0	2
MEJILLONES	0	3	1	3	1	6	14
MELIPILLA	7	1	0	1	0	0	9
MEXICO	0	0	0	0	1	0	1
MOLINA	3	0	0	1	0	0	4
MONTE PATRIA	2	0	0	0	0	0	2
MOSTAZAL	1	2	0	1	0	0	4
MULCHEN	1	0	0	0	0	0	1
NACIMIENTO	2	2	0	0	0	2	6
NANCAGUA	2	0	1	0	0	1	4
NATALES	2	2	0	1	1	0	6
NEGRETE	1	0	0	0	0	0	1
NOGALES	1	0	0	0	0	0	1
NUEVA IMPERIAL	0	0	0	0	1	1	2
NUNOA	24	8	0	10	9	10	61
OLIVAR	1	0	0	3	0	0	4
OSORNO	19	3	0	1	1	4	28
OVALLE	2	3	0	2	1	2	10
PADRE HURTADO	2	0	0	0	0	0	2
PADRE LAS CASAS	3	1	0	1	1	0	6
PAIGUANO	1	0	0	0	0	0	1
PAINE	1	1	1	1	0	1	5
PALENA	0	0	0	0	0	1	1
PANGUIPULLI	2	0	0	1	0	3	6

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
PANQUEHUE	0	1	0	0	0	0	1
PAPUDO	2	1	0	0	1	0	4
PAREDONES	1	0	0	0	0	0	1
PARRAL	0	1	2	1	2	0	6
PEDRO AGUIRRE CERRA	5	1	0	3	1	0	10
PELLUHUE	1	0	0	0	0	0	1
PEMUCO	1	0	0	0	0	1	2
PENAFLORES	4	0	0	0	0	0	4
PENALOEN	10	4	2	2	2	2	22
PENCAHUE	1	0	0	0	0	0	1
PENCO	2	2	0	0	0	1	5
PERALILLO	1	0	0	0	0	0	1
PETORCA	2	0	0	0	0	0	2
PICA	3	3	3	1	0	1	11
PICHIDEGUA	2	0	0	3	1	0	6
PICHILEMU	0	1	0	2	0	0	3
PINTO	1	0	0	1	0	1	3
PIRQUE	3	1	0	0	0	0	4
PORTEZUELO	0	1	0	0	0	0	1
PORVENIR	2	0	0	0	0	0	2
POZO ALMONTE	0	5	0	0	0	2	7
PRIMAVERA	0	0	0	0	0	1	1
PROVIDENCIA	29	41	18	53	55	11	207
PUCHUNCAVI	6	1	0	0	0	0	7
PUCON	9	1	0	2	1	0	13
PUDAHUEL	9	10	0	14	8	7	48
PUENTE ALTO	14	6	2	2	1	5	30
PUERTO MONTT	16	6	4	5	8	8	47
PUERTO VARAS	3	2	2	0	0	1	8
PUNITAQUI	2	0	0	0	0	0	2
PUNTA ARENAS	8	6	8	0	5	15	42
PUQUELDON	1	0	0	0	0	0	1
PUREN	1	1	0	0	0	1	3
PUTAENDO	1	1	0	0	0	0	2
PUTRE	1	0	1	0	0	1	3
QUEILEN	0	0	0	1	0	0	1
QUELLON	1	2	0	0	0	1	4
QUILACO	1	0	0	0	0	0	1
QUILICURA	10	5	0	14	15	5	49
QUILLON	2	0	0	0	0	1	3
QUILLOTA	4	0	1	5	0	2	12
QUILPUE	9	6	0	0	0	2	17

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
QUINCHAO	1	0	0	1	0	0	2
QUINTA DE TILCOCO	1	0	0	0	0	0	1
QUINTA NORMAL	4	0	0	3	2	1	10
QUINTERO	6	0	0	1	0	0	7
QUIRIHUE	1	0	0	0	0	0	1
RANCAGUA	16	6	10	16	8	2	58
RANQUIL	1	3	0	1	0	0	5
RAUCO	2	0	0	0	0	0	2
RECOLETA	9	3	0	6	1	0	19
RENAICO	1	0	0	0	0	0	1
RENCA	10	3	0	1	1	2	17
RENGO	2	0	3	0	1	0	6
REQUINOA	2	1	0	5	0	1	9
RETIRO	0	0	0	1	0	0	1
RIO CLARO	2	0	0	0	0	0	2
RIO NEGRO	1	0	0	1	0	1	3
ROMERAL	1	0	0	0	0	0	1
SAGRADA FAMILIA	1	0	0	1	0	0	2
SALAMANCA	4	3	0	6	3	1	17
SAN ANTONIO	10	3	3	2	2	3	23
SAN BERNARDO	10	7	1	53	9	1	81
SAN CARLOS	1	0	1	1	0	0	3
SAN CLEMENTE	1	0	0	0	0	0	1
SAN FELIPE	4	1	0	10	2	2	19
SAN FERNANDO	2	0	3	4	1	1	11
SAN GREGORIO	1	1	0	0	0	1	3
SAN IGNACIO	0	0	0	0	0	1	1
SAN JAVIER	3	1	0	1	1	0	6
SAN JOAQUIN	3	0	3	3	3	1	13
SAN JOSE DE MAIPO	4	3	0	0	0	0	7
SAN JUAN DE LA COSTA	1	0	0	0	0	0	1
SAN MIGUEL	4	0	1	16	6	2	29
SAN NICOLAS	0	0	0	1	0	0	1
SAN PABLO	1	0	0	1	0	0	2
SAN PEDRO	1	1	1	0	0	1	4
SAN PEDRO DE ATACAMA	1	2	0	8	2	3	16
SAN PEDRO DE LA PAZ	4	7	1	0	0	1	13
SAN RAFAEL	1	0	0	0	0	0	1
SAN RAMON	3	0	0	0	0	0	3
SAN VICENTE	0	0	0	5	1	1	7
SANTA BARBARA	1	0	0	1	0	0	2
SANTA CRUZ	1	0	1	5	0	0	7

COMUNA_A	C1	C2	C3	C4	C5	C6	Total
SANTA JUANA	1	0	0	0	0	0	1
SANTA MARIA	0	0	0	1	0	0	1
SANTIAGO	55	32	69	65	88	49	358
SANTO DOMINGO	6	1	0	0	0	0	7
SIERRA GORDA	3	5	0	1	0	2	11
TALAGANTE	1	0	0	0	0	0	1
TALCA	16	2	21	4	2	3	48
TALCAHUANO	10	2	2	4	3	8	29
TALTAL	0	3	0	0	0	0	3
TEMUCO	14	2	1	6	5	7	35
TENO	3	1	0	0	0	1	5
TEODORO SCHMIDT	0	0	0	0	0	1	1
TIERRA AMARILLA	0	2	0	2	1	0	5
TILTIL	4	1	0	0	0	2	7
TIRUA	0	0	1	0	0	0	1
TOCOPILLA	1	3	2	1	0	0	7
TOLTEN	1	0	0	0	0	0	1
TOME	6	6	0	0	0	1	13
TRAIQUEN	1	0	0	2	0	0	3
TUCAPEL	1	0	0	0	0	0	1
USA	0	0	32	0	0	0	32
VALDIVIA	11	1	0	2	0	5	19
VALLENAR	1	1	3	1	0	0	6
VALPARAISO	20	6	8	18	3	9	64
VICHUQUEN	1	1	0	0	2	0	4
VICTORIA	1	1	3	0	2	0	7
VICUNA	1	0	0	0	0	0	1
VILCUN	1	0	0	0	0	0	1
VILLA ALEGRE	2	0	0	0	0	0	2
VILLA ALEMANA	3	1	1	0	0	0	5
VILLARRICA	5	1	0	0	0	1	7
VINA DEL MAR	39	6	1	12	9	6	73
VITACURA	15	7	2	11	14	4	53
YERBAS BUENAS	1	0	0	0	0	0	1
YUMBEL	1	0	0	0	0	0	1
YUNGAY	1	1	0	0	0	0	2
ZAPALLAR	1	2	0	0	0	0	3

COMUNA_B	C1	C2	C3	C4	C5	C6	Total
ALTO HOSPICIO	0	0	0	0	1	3	4
ANCUD	0	0	1	0	0	1	2
ANGOL	0	0	2	2	0	1	5
ANTOFAGASTA	67	73	5	27	8	23	203
ARGENTINA	0	0	0	0	0	1	1
ARICA	0	0	4	10	2	5	21
AYSEN	0	0	0	0	0	1	1
BOLIVIA	0	0	11	0	0	0	11
BUIN	0	0	0	0	2	1	3
CALAMA	0	0	0	9	2	5	16
CALDERA	0	0	0	1	0	3	4
CANETE	0	0	0	1	0	0	1
CARAHUE	0	0	0	3	0	0	3
CASABLANCA	0	0	0	1	0	0	1
CASTRO	0	0	1	4	0	3	8
CAUQUENES	0	0	2	1	0	3	6
CERRILLOS	0	0	2	4	6	3	15
CERRO NAVIA	0	0	0	11	0	0	11
CHANARAL	0	0	0	1	0	0	1
CHILLAN	1	16	16	10	3	7	53
COELEMU	0	0	0	2	0	0	2
COLINA	0	0	2	13	18	0	33
CONCEPCION	162	54	1	16	9	8	250
CONCHALI	0	0	0	0	0	1	1
CONCON	0	0	0	1	0	2	3
CONSTITUCION	0	0	1	1	0	3	5
COPIAPO	0	3	0	13	10	3	29
COQUIMBO	0	0	1	11	2	4	18
CORONEL	0	0	0	2	1	2	5
COYHAIQUE	0	1	1	0	4	3	9
CURACO DE VELEZ	0	0	0	1	0	0	1
CURANILAHUE	0	0	0	0	0	1	1
CUREPTO	0	0	0	6	0	0	6
CURICO	1	1	5	9	2	3	21
DIEGO DE ALMAGRO	0	0	1	2	0	0	3
ERCILLA	0	0	0	1	0	0	1
ESTACION CENTRAL	0	0	0	1	1	0	2
HUALPEN	0	0	0	0	0	1	1
HUARA	0	0	0	1	0	0	1
HUASCO	0	0	0	0	2	1	3
HUECHURABA	0	0	46	13	12	4	75
ILLAPEL	0	0	1	1	0	0	2

COMUNA_B	C1	C2	C3	C4	C5	C6	Total
INDEPENDENCIA	0	0	1	5	3	0	9
IQUIQUE	32	53	5	7	14	12	123
ISLA DE PASCUA	0	0	1	0	0	0	1
LA CISTERNA	0	0	8	1	0	0	9
LA FLORIDA	0	0	2	2	3	2	9
LA LIGUA	0	0	1	1	0	0	2
LA REINA	0	0	0	5	3	1	9
LA SERENA	59	20	4	9	7	9	108
LA UNION	0	0	1	0	0	0	1
LAGO RANCO	0	0	0	0	0	1	1
LAJA	0	0	0	1	0	0	1
LAS CABRAS	0	0	0	1	0	0	1
LAS CONDES	1	1	10	83	50	22	167
LAUTARO	0	0	0	1	0	2	3
LEBU	0	0	1	0	0	0	1
LICANTEN	0	0	0	2	0	1	3
LIMACHE	0	0	0	0	0	1	1
LINARES	1	0	1	1	0	4	7
LLANQUIHUE	0	0	0	0	1	0	1
LO BARNECHEA	0	0	1	0	1	0	2
LO ESPEJO	0	0	0	1	0	0	1
LONQUIMAY	0	0	0	0	0	1	1
LOS ANDES	0	0	1	9	0	5	15
LOS ANGELES	0	0	0	6	1	5	12
LOS VILOS	0	0	0	1	2	0	3
MACUL	0	0	0	0	1	2	3
MAIPU	0	0	1	1	4	0	6
MARIA PINTO	0	0	0	1	0	0	1
MARIQUINA	0	0	0	2	0	2	4
MEJILLONES	0	0	0	3	1	3	7
MELIPILLA	0	0	0	1	0	0	1
MOLINA	0	0	0	1	0	0	1
MOSTAZAL	0	0	0	1	0	0	1
NACIMIENTO	0	0	0	0	0	2	2
NATALES	0	0	0	1	1	0	2
NUEVA IMPERIAL	0	0	0	0	1	1	2
NUNOA	0	0	0	15	14	6	35
OSORNO	0	0	2	1	1	3	7
OVALLE	0	0	1	3	1	2	7
PADRE LAS CASAS	0	0	0	1	1	0	2
PAINE	0	0	0	1	0	0	1
PANGUIPULLI	0	0	0	2	0	3	5

COMUNA_B	C1	C2	C3	C4	C5	C6	Total
PAPUDO	0	0	0	1	1	0	2
PARRAL	0	0	1	2	2	0	5
PENALOEN	0	0	1	0	0	1	2
PENCO	0	0	0	0	0	1	1
PERU	0	0	1	0	0	0	1
PICA	0	0	0	1	0	1	2
PICHIDEGUA	0	0	0	3	1	1	5
PICHILEMU	0	0	0	2	0	0	2
PLACILLA	0	0	0	0	0	1	1
POZO ALMONTE	0	0	0	0	1	1	2
PRIMAVERA	0	0	0	0	0	1	1
PROVIDENCIA	0	0	10	45	53	9	117
PUCON	0	0	0	2	1	0	3
PUDAHUEL	0	0	0	12	6	4	22
PUENTE ALTO	0	0	2	2	1	3	8
PUERTO MONTT	80	28	8	5	6	6	133
PUERTO VARAS	0	0	1	0	0	1	2
PUNTA ARENAS	12	9	3	0	4	15	43
PUREN	0	0	0	0	0	1	1
PUTRE	0	0	0	0	0	1	1
QUEILEN	0	0	0	1	0	0	1
QUILICURA	1	0	0	19	20	9	49
QUILLON	0	0	0	1	0	1	2
QUILLOTA	0	0	0	5	0	2	7
QUILPUE	0	0	0	0	0	1	1
QUINCHAO	0	0	0	1	0	0	1
QUINTA NORMAL	0	0	0	1	1	0	2
QUINTERO	0	0	0	1	0	0	1
RANCAGUA	115	22	15	15	8	3	178
RECOLETA	0	0	0	2	1	0	3
RENCA	0	0	0	3	0	1	4
RENGO	0	0	1	0	1	0	2
REQUINOA	0	0	0	9	0	0	9
RIO NEGRO	0	0	0	1	0	1	2
SALAMANCA	0	0	2	6	3	1	12
SAN ANTONIO	0	6	2	2	2	3	15
SAN BERNARDO	1	0	0	56	10	1	68
SAN CARLOS	0	0	0	1	0	0	1
SAN FELIPE	0	0	0	9	2	1	12
SAN FERNANDO	0	0	3	4	1	1	9
SAN GREGORIO	0	0	0	0	0	1	1
SAN IGNACIO	0	0	0	2	0	1	3

COMUNA_B	C1	C2	C3	C4	C5	C6	Total
SAN JAVIER	0	0	0	1	1	0	2
SAN JOAQUIN	0	0	1	7	9	4	21
SAN MIGUEL	0	0	0	14	1	2	17
SAN NICOLAS	0	0	0	1	0	0	1
SAN PABLO	0	0	0	1	0	0	1
SAN PEDRO	0	0	4	0	2	21	27
SAN PEDRO DE ATACAMA	0	0	0	8	2	3	13
SAN VICENTE	0	0	0	5	1	1	7
SANTA BARBARA	0	0	0	1	0	0	1
SANTA CRUZ	0	0	1	5	0	0	6
SANTA MARIA	0	0	0	3	0	0	3
SANTIAGO	436	240	83	79	97	97	1032
SIERRA GORDA	0	0	0	1	0	0	1
TALAGANTE	0	0	0	1	0	0	1
TALCA	0	9	20	5	2	3	39
TALCAHUANO	0	0	4	4	3	8	19
TEMUCO	0	7	1	6	5	8	27
TEODORO SCHMIDT	0	0	0	0	0	1	1
TIERRA AMARILLA	0	0	0	1	0	0	1
TILTIL	0	0	3	0	0	1	4
TOCOPILLA	0	0	1	1	0	0	2
TOME	0	0	0	0	0	1	1
TRAIGUEN	0	0	0	2	0	0	2
TUCAPEL	0	0	0	0	0	1	1
USA	0	0	34	0	0	0	34
VALDIVIA	0	3	0	2	0	5	10
VALLENAR	0	0	2	2	0	1	5
VALPARAISO	142	32	13	17	5	5	214
VICHUQUEN	0	0	0	0	1	0	1
VICTORIA	0	0	3	0	2	0	5
VILLA ALEMANA	0	0	1	0	0	0	1
VILLARRICA	0	0	0	0	0	1	1
VINA DEL MAR	0	0	1	13	9	7	30
VITACURA	0	0	0	9	12	4	25
ZAPALLAR	0	0	0	3	1	1	5