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

ESTIMACIÓN DEL RIESGO ASOCIADO A POTENCIALES CLIENTES JÓVENES EN EL NEGOCIO DE SEGUROS DE AUTOMÓVILES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PAMELA INÉS ROMERO VÁSQUEZ

PROFESOR GUÍA:
ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN
LUIS ABURTO LAFOURCADE
CRISTIAN MATURANA

SANTIAGO DE CHILE
2015

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Pamela Inés Romero Vásquez
FECHA: 02/03/2015
PROFESOR GUÍA: Alejandra Puente Chandía

ESTIMACIÓN DEL RIESGO ASOCIADO A POTENCIALES CLIENTES JÓVENES EN EL NEGOCIO DE SEGUROS DE AUTOMÓVILES

El presente proyecto se enmarca en el problema de la existencia de un importante grado de incertidumbre al momento de tarificar las primas asociadas a cada individuo en el negocio de los seguros de automóviles, específicamente, de los clientes jóvenes. Esta incertidumbre viene dada por el desconocimiento de su comportamiento de conducción, lo que finalmente lleva a imponer altas primas para este segmento.

De esta forma, el objetivo es estimar el nivel de riesgo adecuado para cada individuo perteneciente al segmento joven, y consecuentemente, la prima asociada a éste. Así, se desarrollan dos modelos de predicción, el primero estima la probabilidad de siniestro de cada persona, mientras que el segundo, el costo medio asociado, lo que es una medida de la gravedad de los incidentes que reporte cada uno de ellos.

Para estimar la propensión de siniestro se utilizan dos modelos, una regresión logística y un árbol de decisión C5.0, cuyos resultados son bastante similares; sin embargo, el modelo *logit* supera levemente al árbol de decisión. Por otro lado, para la estimación del costo medio se utiliza un algoritmo C5.0. Debido a que los resultados obtenidos en la segunda predicción no entregan valor adicional al obtenido con la estimación de la probabilidad, se decide no utilizarlo en el modelo final.

Un punto importante dentro del proyecto era analizar la relación entre el comportamiento de conducción y el rendimiento escolar, donde se concluye que este último es un indicador del grupo socioeconómico de cada individuo. De este modo, al realizar una regresión logística sin estas variables, y al obtener resultados similares, se opta por este último modelo, sin la PSU.

Se realiza un análisis de escenarios, donde se asignan diferentes primas a cada grupo de riesgo creado. De esta forma, se logra una mayor diferenciación y distribución del segmento objetivo, lo que conlleva a mayores utilidades para la compañía.

Finalmente, como propuestas y trabajos futuros, la principal recomendación es incorporar otras fuentes de datos que ayudarían a estimar de mejor forma el nivel de riesgo del segmento estudiado.

Agradecimientos

Al encontrarme en este periodo de mi vida, donde se cierra una gran etapa y comienza una totalmente nueva, agradezco todos los momentos vividos con grandes y hermosas personas, quienes me han dejado innumerables recuerdos y enseñanzas.

Agradezco a mi familia, por estar siempre conmigo, por amarme y apoyarme incondicionalmente; especialmente a mis padres, con quienes sé que puedo contar en todo momento y que siempre han dado todo por mí. A mi mamá, por ser la mejor en todo lo que hace, y por ser de todo por mí y mis hermanas; profesora, enfermera, cocinera, artista, arquitecta claramente… hasta ingeniera. A mi papá, por el sacrificio infinito que ha hecho y que hace día a día por nosotras, por su ayuda y apoyo en los momentos más complicados, muchas gracias.

Le agradezco a mi Willy, por su infinito amor y apoyo durante todo el tiempo juntos, y más. Gracias por estar conmigo en las buenas y en las malas, por ayudarme a enfrentar mis miedos y superar los obstáculos que se han presentado en el camino. Gracias por la inmensa fe que tienes en mí.

A mis amigos con los que he vivido increíbles e inolvidables momentos durante el transcurso de este periodo. A mi mejor dupla por supuesto, Cataleli, le agradezco los buenos momentos vividos, las risas y secretos compartidos. Gracias por venir corriendo cuando te necesito, gracias por los chocolates y por estar siempre conmigo. A las Culis y Carillo, por siempre poner el toque de alegría en todo momento, gracias por los consejos entregados, y por estar siempre para mí, las quiero mucho.

A mis profesores, quienes han sabido guiarme por el camino para llevar este trabajo a buen término. Les agradezco su consejo y apoyo entregado en el transcurso de este periodo, tan importante para mí.

Gracias a todas aquellas personas que estuvieron presentes en el transcurso de esta etapa, a todas aquellas personas que me dieron una palabra de aliento, un consejo, un abrazo amigo, en fin, a todos los que estuvieron para mí, de alguna forma u otra, durante este periodo de mi vida.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

1. Introducción

1.1 Antecedentes Generales

El presente estudio es realizado en Penta Security, compañía de seguros generales, la cual ofrece una gran variedad de productos y servicios, los que se clasifican en Incendio y Terremoto, Vehículos, Cascos, Transporte, Seguro Obligatorio de Accidentes Personales (SOAP) y Otros Ramos. Específicamente, el proyecto se enmarca en el negocio de seguros de vehículos motorizados.

Actualmente, Penta Security se enfrenta a la problemática de cómo abordar el riesgo asociado a clientes con un importante grado de incertidumbre en cuanto a su comportamiento de conducción. Específicamente de los que forman parte del grupo etario comprendido entre los 18 y 33 años aproximadamente.

Esto se debe a que los modelos que la empresa utiliza hoy en día se basan esencialmente en información relacionada al historial de conducción, y dado que los individuos jóvenes no tienen gran cantidad de información relacionada a dichos hábitos, los modelos actuales no son muy efectivos al momento de estimar su nivel de riesgo.

Por consiguiente, al no contar con la información ni con los métodos necesarios para caracterizar y diferenciar a los clientes jóvenes, no existe una metodología de alta confiabilidad que permita tomar las decisiones de tarificación apropiadas. En consecuencia, la compañía, al enfrentarse a situaciones de alto riesgo como la ya mencionada, acostumbra a cobrar un cargo adicional a la prima base. De este modo, la dificultad de diferenciación entre los clientes del grupo joven se ve reflejada en tarifas elevadas y prácticamente homogéneas para todo el grupo.

Como ejemplo, en la Figura 1 se puede observar los resultados obtenidos de un estudio realizado por ComparaOnline[1], plataforma que ofrece información sobre distintas opciones presentes en el mercado, tanto en la industria de los seguros, como en la de las finanzas y telecomunicaciones.

Dicho estudio utilizó más de 300.000 cotizaciones realizadas durante el año 2013, y considera ocho compañías aseguradoras: Penta Security, AIG, RSA, Zenit Seguros, Consorcio, Mapfre, HDI y Chilena Consolidada. Se puede notar que en el caso del modelo Toyota Yaris – vehículo más cotizado durante el año 2013 – una persona menor de 30 años paga hasta un 30% más en promedio que alguien de mayor edad (considerando un deducible de 5 UF). En el caso del modelo Hyundai Accent, el incremento es de un 28%, y para el modelo Kia Morning, de un 26%.

[1] www.comparaonline.cl

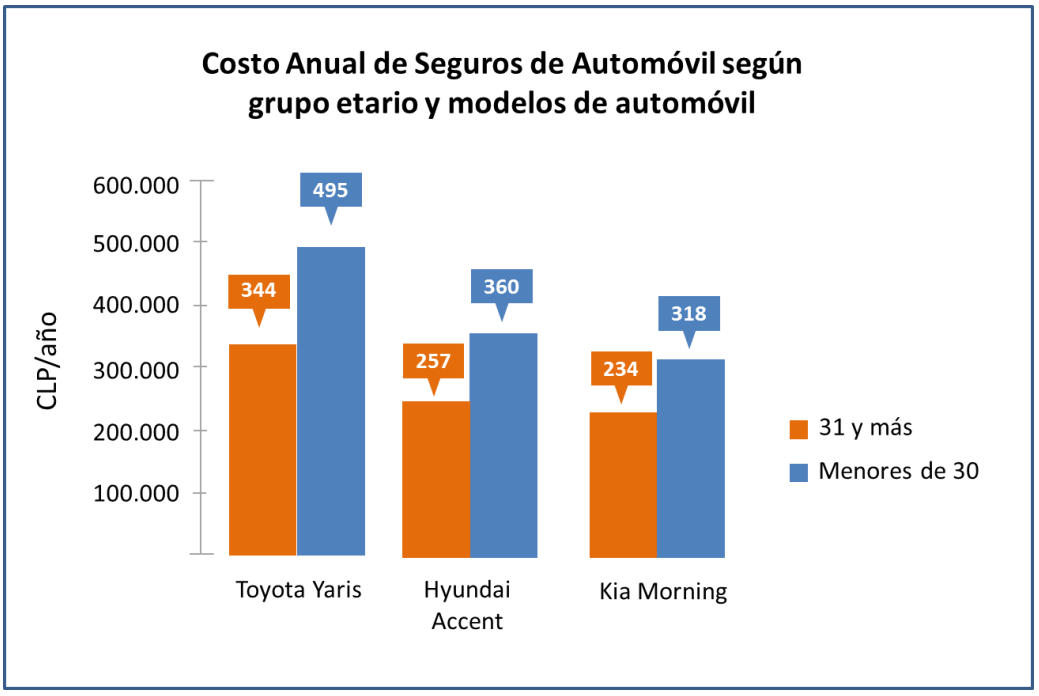

Figura 1. Costo anual de Seguros de Automóvil según grupo etario y modelo de automóvil. Fuente: www.comparaonline.cl

Estas mayores primas impuestas a dicho segmento se deben a la alta tasa de siniestralidad[2] que presentan históricamente estos clientes. En la Figura 2 se muestra la evolución mensual de dicha tasa por grupo etario[3], donde se puede observar que aquellos clientes con 22 años o menos presentan el mayor nivel de siniestralidad durante todos los periodos, y este se encuentra siempre sobre el 100%, es decir que se utiliza toda su prima pagada (y más) en cubrir los costos de sus siniestros. El siguiente grupo con mayor tasa de siniestralidad son los clientes entre 22 y 33 años, la cual ronda entre un 70% y un 80% aproximadamente. Para los clientes de mayor edad la tasa de siniestralidad es menor, siendo los clientes mayores de 67 años aquellos con la menor tasa.

[2] La tasa de siniestralidad es el porcentaje de la prima destinado a pagar los siniestros del individuo, es decir, el porcentaje del costo de siniestro respecto a la prima anual pagada por el asegurado.

[3] Estadísticas obtenidas de un estudio realizado por Penta Security, utilizando datos de los clientes capturados por uno de sus corredores masivos.

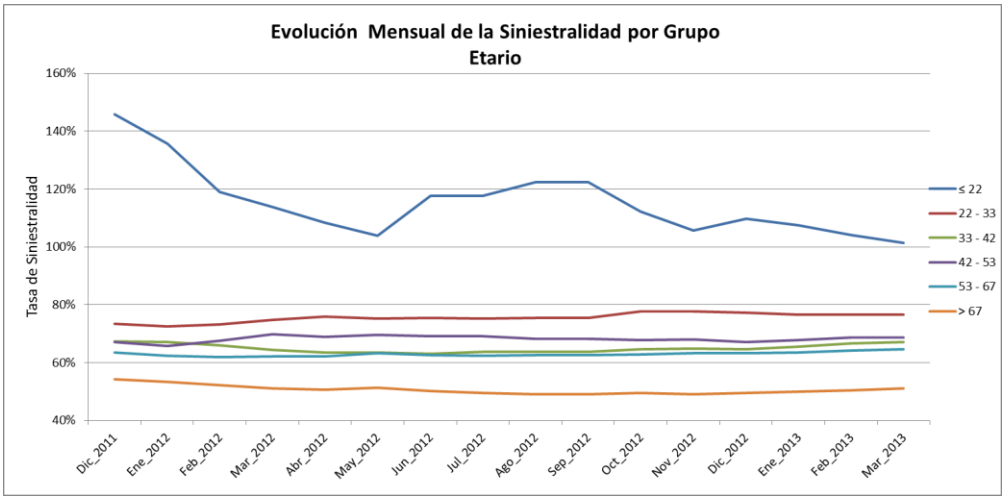

Figura 2. Evolución Mensual de la Tasa de Siniestralidad por Grupo Etario.

Finalmente, aunque los modelos existentes actualmente en la industria de los seguros de automóviles asignan al segmento joven un nivel de riesgo elevado, en la práctica se observa que no todos los individuos pertenecientes a este grupo se comportan del mismo modo. En la Figura 3 se muestra que existe una alta proporción de individuos, en cada rango de edad, que no han presentado siniestros en los últimos cuatro años[4]. De esta forma, al imponer altas primas para la mayoría de los individuos del segmento bajo estudio, no se logra capturar todo el potencial de clientes que podrían ser rentables para la empresa.

Es por esto que es importante realizar un estudio que genere perfiles de riesgo dentro del grupo, con el objetivo de discriminar entre ellos y así ofrecer una prima acorde a sus características.

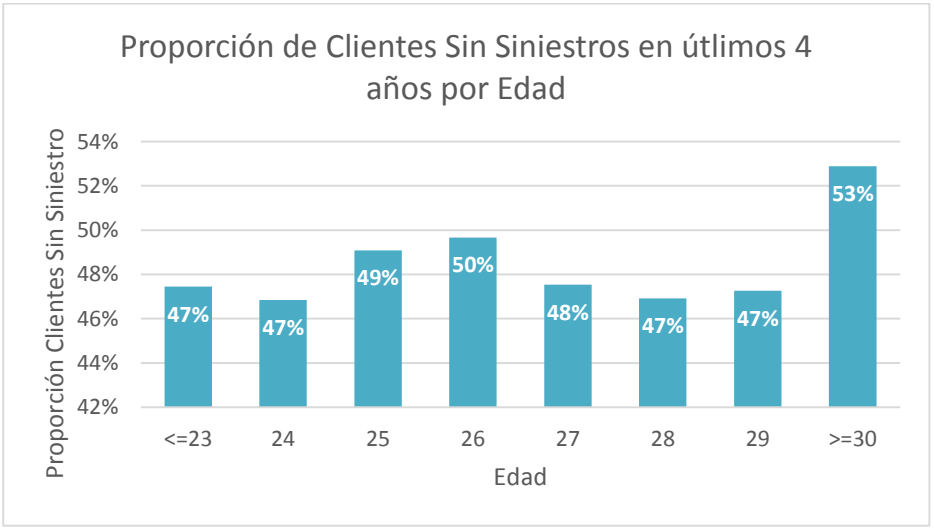

Figura 3. Proporción de Clientes Sin Siniestro en los últimos 4 años, por Edad.

[4] Datos derivados de la base de datos final a utilizar en el estudio.

2. Descripción del Proyecto y Justificación

2.1 Descripción del Proyecto

El problema que enfrenta la empresa hoy en día es que no se conoce el nivel de riesgo adecuado para los individuos pertenecientes al grupo etario comprendido entre los 18 y 33 años, principalmente debido a que dichos clientes no tienen un extenso historial de conducción. De esta forma, se les cobran altas primas con la finalidad de evitarlos.

Así, con el fin de que la empresa pueda tomar decisiones basadas en las características y diferencias de los individuos pertenecientes al grupo joven, y de aumentar sus utilidades por medio de la captación de nuevos clientes rentables para la compañía, se estimará el nivel de riesgo de cada persona perteneciente a dicho segmento, basándose en atributos relacionados a su comportamiento de conducción, características demográficas y rendimiento escolar.

Para esto último se trabajará con variables asociadas a los puntajes obtenidos en la PSU[5], y se analizará la supuesta relación entre el comportamiento de conducción y dichos resultados escolares.

Específicamente, se estimará tanto la probabilidad de que cada individuo tenga uno o más siniestros en el año, como el costo medio del o los siniestros asociados.

Para el desarrollo de este proyecto es necesario conocer el funcionamiento del proceso de tarificación actual de la compañía.

Proceso de Tarificación Actual

La tarifa asociada a cada cliente se compone de dos factores, una tarifa base y una política de precios, ya sea un recargo sobre la tarifa base, un descuento, o la misma prima inicial.

La tarifa base de la empresa se calcula en función del tipo, marca, modelo y año del vehículo del asegurado. De esta forma, la prima base varía según el automóvil a cotizar.

La política de precios se construye según las características del cliente, como el año del primer siniestro, número de eventos siniestrales, gravedad de los siniestros, edad, entre otros. Este factor se deriva del modelo actual de estimación de riesgo de los clientes de la empresa.

[5] Prueba de Selección Universitaria.

De forma adicional esta política de precio incluye un factor asociado a la edad del asegurado, de modo que los clientes jóvenes presenten un recargo en caso de que no se les haya asignado uno en el modelo anterior. Luego, a la tarifa a publicar se le agregan las comisiones y gastos administrativos promedios de la compañía.

En resumen, se tiene que:

$$T = T_b \cdot P_p + G$$

Donde T es la tarifa final, T_b es la tarifa base, P_p es la política de precios, y G son las comisiones y gastos administrativos agregados al precio de la prima.

Es importante mencionar que la tarifa asociada a cada cliente está sujeta a evaluación año a año. Según el comportamiento de conducción del individuo y del costo que haya significado para la empresa, se decide la política de precios asociada a este cliente para el año venidero.

2.2 Justificación del Proyecto

El valor del proyecto radica en la mejor asignación y distribución de tarifas asociadas a cada individuo dentro del segmento de interés. Esto se traduciría en los siguientes beneficios para la compañía:

- Capturar nuevos clientes que actualmente están mal catalogados como altamente riesgosos, sin serlo, y que por ende tienen una tarifa muy alta, lo que los desincentiva a contratar un seguro de la compañía.

- Mejorar la rentabilidad reubicando el nivel de riesgo desde un valor inferior a uno superior para aquellas personas que, según este estudio, están en un nivel de riesgo menor al que les correspondería. De esta forma, al cobrarles una tarifa acorde a su riesgo, se pueden dar dos posibilidades: la primera, dichos clientes se van o no escogen esta compañía, los cuales significaban pérdidas para ésta; de este modo, se lograría incrementar la rentabilidad de la empresa. Y la segunda alternativa es que estos clientes paguen de todos modos la prima impuesta, a pesar de su incremento, lo que deja de significar un costo para la empresa ya que se está cobrando una prima acorde a su riesgo.

Como resultado de lo anterior, se lograría incrementar los ingresos y los márgenes comerciales de Penta.

3. Objetivos

3.1 Objetivo General

Estimar el nivel de riesgo de los potenciales clientes de seguros de automóviles de Penta Security pertenecientes al grupo etario joven.

3.2 Objetivos Específicos

- Identificar las principales variables que influyen en el nivel de riesgo de los clientes pertenecientes al segmento objetivo.

- Estimar el nivel de riesgo del segmento objetivo, considerando las variables descriptivas relevantes seleccionadas.

- Analizar y determinar la relación entre las variables de rendimiento escolar y el comportamiento de conducción de los clientes.

- Clasificar a los prospectos de clientes en segmentos de riesgo generados a partir del modelo desarrollado.

4. Alcances

Los alcances de este trabajo de memoria se limitarán a los siguientes aspectos:

- Se trabajará con el segmento de potenciales clientes pertenecientes al grupo etario comprendido entre los 18 y los 35 años que hayan rendido la PSU.

- Se trabajará con la línea de negocio de Seguros de Vehículos Motorizados, dentro del área de Seguros Generales de Penta Security.

- Se utilizará como fuente de información las bases de datos del Registro Nacional de Vehículos Motorizados (RNVM) del año 2013, y el Sistema de Siniestros de Seguros Generales (SISGEN) desde el año 2011 al 2014.

- Se trabajará con datos de variables asociadas al rendimiento escolar (PSU y NEM[6]) desde el año 2004 al 2014.

- El trabajo realizado no contempla la integración e implementación del modelo desarrollado en los actuales sistemas de la compañía.

[6] Promedio de Notas de Enseñanza Media.

5. Marco Conceptual

Para estimar el nivel de riesgo de los potenciales clientes en el grupo de interés se pronosticará tanto la probabilidad de que cada individuo tenga uno o más siniestros en el año 2014, así como su costo medio asociado. De esta manera, se entrega una medida de gravedad de los siniestros de cada persona, en conjunto con su propensión a incurrir en un accidente automovilístico.

Para el desarrollo del proyecto se seguirá la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), un modelo de procesos que provee un enfoque estructurado para llevar a cabo un proyecto de minería de datos. Esta metodología se divide en seis etapas, las cuales no siguen necesariamente una secuencia lineal. Estas son:

- Comprensión del negocio: Esta fase se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, para luego convertir este conocimiento en una definición del problema de minería de datos en conjunto con sus objetivos.

- Comprensión de los datos: Esta etapa incluye la recopilación e integración de los datos a utilizar, la familiarización con la data, identificación de problemas de calidad de las fuentes de información, y el descubrimiento de las primeras relaciones o señales relevantes para el estudio.

- Preparación de los datos: Esta fase cubre todas las actividades necesarias para construir el conjunto de datos. Se incluyen tareas como la selección de atributos, así como también la transformación y limpieza de los datos.

- Modelamiento: Se seleccionan y aplican técnicas de modelamiento, con el objetivo de encontrar patrones "ocultos" o poco evidentes en los datos. En esta etapa se considera tanto la selección del modelo, como su diseño, construcción y evaluación.

- Evaluación: En esta fase se evalúan los resultados obtenidos del o los modelos construidos desde una perspectiva de negocio. Así, se evalúa el valor del modelo para éste.

- Despliegue: Esta etapa consiste en la implementación del modelo creado, integrándolos en las tareas de toma de decisiones de la organización. En este caso, esta fase final se encuentra fuera del alcance del proyecto de título.

En la Figura 4 se muestra un esquema de la metodología CRISP-DM, donde se observa la no linealidad entre las distintas etapas del proceso, y la naturaleza cíclica de los proyectos de minería de datos.

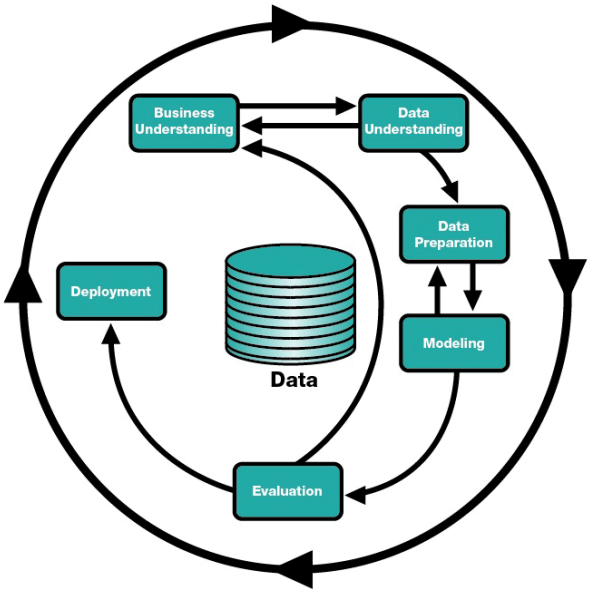

Figura 4. Metodología CRISP-DM (Cross Industry Standard Process for Data Mining).

5.1 Técnicas de Minería de Datos a Desarrollar

5.1.1 Regresión Logística

La regresión logística es un tipo de análisis de regresión que se utiliza para predecir el resultado de una variable categórica (que puede adoptar un número limitado de categorías), ya sea con dos clases (dicotómica) o más. De esta forma, se habla de una regresión logística binaria cuando la variable a predecir es dicotómica, y de una regresión logística multinomial cuando tiene más de dos categorías.

En el caso de estudio se hará uso de una regresión logística binaria, para así estimar la probabilidad de que cada individuo se vea enfrentado a uno o más siniestros en el año 2014.

5.1.1.1 Regresión Logística Binaria

Como se mencionó anteriormente, en una regresión logística binaria se intenta predecir una variable con dos categorías. En este caso, presencia o ausencia de siniestro en el periodo bajo estudio, en función de una o más variables explicativas, las cuales pueden ser cuantitativas o cualitativas.

La ecuación que describe a la regresión logística binaria viene dada por:

$$P(y_i = 1 \mid x_i) = \frac{exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}$$

Donde Y es la variable a estimar, x_{ij} es la i-ésima observación de la j-ésima variable explicativa dentro de las k utilizadas, y β es el vector de coeficientes o regresores del modelo.

La estimación del valor de los coeficientes en una regresión logística se lleva a cabo mediante el método de la Máxima Verosimilitud (o MLE).

Es importante mencionar también que en este tipo de regresiones no resulta posible interpretar directamente las estimaciones de los coeficientes, debido a la no linealidad del modelo, por lo que se utilizan otro tipo de medidas para comprender la magnitud del efecto de los regresores en la variable objetivo. La más utilizada es el *Odds*, que calcula la probabilidad de tener siniestro sobre la de no tenerlo. Formalmente:

$$Odds = \frac{P(y_i = 1 \mid x_i)}{1 - P(y_i = 1 \mid x_i)} = exp\left(\beta_0 + \sum \beta_j x_{ij}\right)$$

Si se realiza una transformación tomando logaritmos neperianos a ambos lados de la ecuación, se puede observar que se obtiene una regresión lineal:

$$ln(Odds) = ln\left(\frac{P(y_i = 1 \mid x_i)}{1 - P(y_i = 1 \mid x_i)}\right) = \beta_0 + \sum \beta_j x_{ij}$$

En esta última expresión se ve a la izquierda el llamado *logit*, es decir, el logaritmo natural del odds de la variable dependiente.

Para evaluar la bondad de ajuste de una regresión logística no se utiliza el conocido estadístico R^2 de la regresión lineal, ya que los coeficientes estimados en la regresión logística se generan mediante el método de Máxima Verosimilitud[7], el cual no es calculado con el objetivo de minimizar la varianza como en el método MCO. De todas formas, para evaluar la bondad de ajuste de un modelo logístico, se han creados varios pseudo-R^2. En este caso se hará uso del R^2 de Nagelkerke, cuyo valor varía entre 0 y 1:

$$R^2_{Nagelkerke} = \frac{1 - \left(\frac{L(M_{intercept})}{L(M_{full})}\right)^{2/N}}{1 - L\left(M_{intercept}\right)^{2/N}}$$

[7] En el caso de la regresión lineal se utiliza el método MCO (Mínimos Cuadrados Ordinarios).

Donde $L(M_{intercept})$ es el valor de la máxima verosimilitud del modelo considerando tan sólo el intercepto, y $L(M_{full})$ la del modelo completo (considerando las variables independientes), N es el tamaño de la muestra.

Además, existen pruebas de bondad de ajuste entre los datos esperados (o teóricos) y los observados (o reales). En este caso se hará uso del test de Hosmer – Lemeshow.

Test de Hosmer – Lemeshow

El Test de Hosmer – Lemeshow es una prueba para evaluar la bondad de ajuste de un modelo de regresión logística binaria, en éste se evalúa la distancia entre lo observado y lo esperado.

La prueba consiste en dividir el recorrido de valores de la variable dependiente en una serie de intervalos. La idea es contar intervalo por intervalo el valor esperado y el observado para cada uno de los dos resultados posibles de la variable dependiente dicotómica (en este caso, si presenta, o no, uno o más siniestros en el año 2014). El valor observado es el real, mientras que el esperado es el entregado por el modelo construido.
De esta forma, las observaciones son divididas en g grupos dependiendo de las probabilidades estimadas; generalmente $g = 10$. En la práctica cada observación tiene una probabilidad predicha diferente, por lo que las probabilidades predichas varían en cada grupo creado. Para calcular las observaciones de éxito ($Y = 1$) esperadas, el test calcula el promedio de las probabilidades esperadas en cada grupo, y la multiplica por el número de observaciones de cada grupo. El test genera los mismos cálculos para $Y = 0$, y luego calcula un estadístico de bondad de ajuste de Pearson:

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g p_g (1 - p_g)}$$

Donde O_g son los eventos observados, E_g los esperados, N_g las observaciones en el grupo g, y p_g las probabilidades predichas en el grupo g. G es el total de grupos creados. El estadístico sigue una distribución χ^2 con $G - 2$ grados de libertad.

La hipótesis nula dice que el modelo propuesto se ajusta a lo observado, por lo tanto si obtenemos un p-valor superior a α (generalmente 0,05) no se rechaza la hipótesis nula, y el modelo se ajusta a la realidad.

5.1.2 Árboles de Decisión

Los árboles de decisión son una técnica de modelación no paramétrica que clasifica las observaciones o casos de una base de datos según distintas variables explicativas, con el fin de predecir una variable de respuesta. Cuando la variable de respuesta toma un número limitado de categorías se habla de árboles de clasificación, mientras que cuando la variable objetivo toma valores continuos se habla de árboles de regresión.

El esquema del árbol de decisión representa una serie de reglas que llevan a predecir la variable objetivo, donde cada nodo interior (o no terminal) corresponde a una de las variables explicativas, y cada rama que sale de estos nodos representan los posibles valores que puede tomar dicho atributo. Los nodos terminales indican la clase en que el modelo clasifica finalmente cada instancia.

En la Figura 5 se puede observar un ejemplo de un árbol de clasificación en el que se intenta predecir si el tiempo es adecuado para jugar tenis o no (ejemplo de Mitchell en [10]).

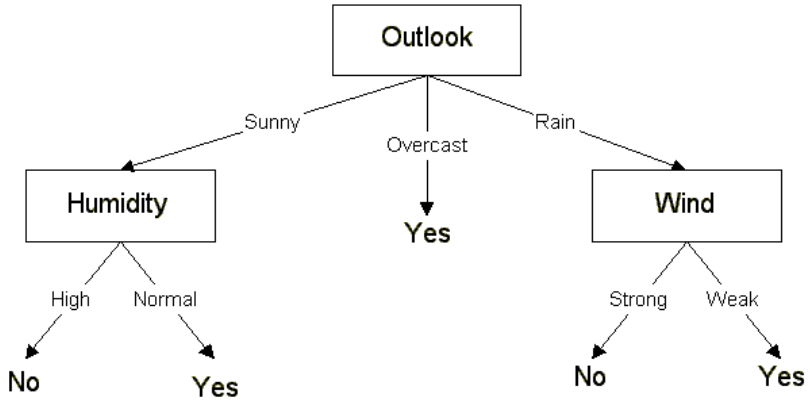

Figura 5. Ejemplo Árbol de Decisión.

La división de cada nodo se realiza según la variable predictora que mejor ayude a clasificar los casos de la base de datos, la cual es escogida según algún test estadístico, así como también se escoge el punto de corte óptimo para cada atributo utilizado. Este proceso se repite usando los datos de entrenamiento asociado a cada nodo generado, hasta que se cumplan los criterios de parada establecidos en el modelo.

Criterios de Corte

Existen distintas medidas para seleccionar el atributo con mayor poder discriminatorio en cada etapa, así como también el punto de corte óptimo de cada variable, si corresponde. En muchas de estas métricas se utiliza el concepto de *entropía*, medida que caracteriza la pureza (o impureza) de un

conjunto de datos. La entropía es una medida de la incertidumbre o desorden dentro de una fuente de información.

Dada una colección de datos S, cuya variable objetivo puede tomar c clases distintas, la entropía del conjunto S se define como:

$$Entropía(S) = \sum_{i=1}^{c} -p_i \, log_2 \, p_i$$

Donde p_i es la proporción de casos que pertenecen a la clase i.

Para el caso de una clasificación binaria:

$$Entropía(S) = -p \, log_2(p) - (1-p) \, log_2(1-p)$$

Donde p corresponde a la proporción de casos positivos o de éxito en el conjunto S.

Los criterios más utilizados son:

- Ganancia de Información: Esta medida es la reducción esperada en la entropía del conjunto de datos a causa de la partición de las instancias según un atributo. Formalmente, la ganancia de información de un atributo A, en un conjunto de datos S, es:

  $$Ganancia\ de\ Información(S, A) = Entropía(S) - \sum_{v \in Dom(A)} \frac{|S_v|}{S} Entropía(S_v)$$

  Donde v son los posibles valores que puede tomar el atributo A.

- Ratio de Ganancia: Dado que el criterio de Ganancia de Información favorece aquellos atributos con una mayor cantidad de valores, se crea el Ratio de Ganancia, el cual penaliza a aquellas variables que toman muchos valores. Para esto se incorpora el término de *Información de Separación* (del inglés *Split Information*), el cual es sensible a qué tan amplia y uniformemente el atributo divide la data. Específicamente,

  $$Separación\ de\ Información(S, A) = -\sum_{i=1}^{c} \frac{|S_i|}{S} \log_2 \frac{|S_i|}{S}$$

  El Ratio de Ganancia de información se define en términos de la medida de Ganancia de Información y el Split Information.

$$Ratio\ de\ Ganancia(S,A) = \frac{Ganancia\ de\ Información(S,A)}{Separación\ de\ Información(S,A)}$$

- Índice de Gini: Es un criterio basado en la impureza que mide la divergencia entre las distribuciones de probabilidad de los valores del atributo objetivo.

$$Índice\ de\ Gini = 1 - \sum_{i=1}^{k} p_i^2$$

Criterios de Parada

La fase de crecimiento del árbol continúa hasta que se activa algún criterio de parada. Las reglas más comunes son:

- Máxima Profundidad: Se alcanza la máxima profundidad del árbol establecida previamente.

- Mínimo de observaciones para nodo padre: El número de casos en el nodo terminal es menor que el número mínimo de casos para los nodos padres.

- Mínimo de observaciones para nodo hijo: Si el nodo fuese dividido, el número de casos en uno o más nodos hijos sería menor que el número mínimo de casos previamente establecidos para estos nodos.

- Pureza mínima: La pureza del nodo es menor que algún límite especificado anteriormente.

5.1.2.1 Árbol de Decisión C5.0

El árbol de decisión C5.0 es una mejora al algoritmo C4.5, el cual, a su vez, es una extensión del algoritmo ID3. De este modo, se comenzará explicando el funcionamiento de este último árbol de decisión.

Árbol de Decisión ID3

El algoritmo ID3 es un método inductivo que busca dentro de un espacio de hipótesis (conjunto de posibles árboles de decisión) aquella que mejor ajusta a los datos de entrenamiento. Esta búsqueda comienza con un árbol vacío y la progresiva búsqueda de los nodos a testear en cada etapa.

Específicamente, se crean árboles de decisión construyéndolos desde la raíz hacia las hojas (*top-down*), comenzando con la pregunta "¿qué atributo debería ser utilizado en la raíz del árbol?". Para contestar a esta pregunta, cada variable es testeada para determinar qué tan bien logra clasificar las

instancias por sí sola. Así es como se crean ramas desde el nodo raíz para cada posible valor del atributo seleccionado, y se clasifican las instancias en el nodo descendiente que corresponde. Luego se repite el proceso utilizando los casos asociados a cada nodo hijo, y se escoge el mejor atributo para testear en dicho punto.

En la Tabla 1 se resume el algoritmo ID3 para un conjunto de entrenamiento T_r, siendo c_i las clases a predecir de la variable objetivo, y A el conjunto de los atributos predictores utilizados.

$\boldsymbol{ID3(T_r, c_i, A)}$

- Crear nodo *raíz* para el árbol
- Si todos los ejemplos en $\boldsymbol{T_r}$ son positivos, regresar el árbol con el único nodo *raíz* etiquetado como $\boldsymbol{c_i}$
- Si todos los ejemplos en $\boldsymbol{T_r}$ son negativos, regresar el árbol con el único nodo *raíz* etiquetado como $\boldsymbol{\bar{c}_i}$
- Si la lista $\boldsymbol{A}$ está vacía, regresar al árbol con el único nodo *raíz* etiquetado como el valor de $\boldsymbol{c_i}$ más frecuente en $\boldsymbol{T_r}$
- En otro caso comenzar,
    - Sea $\boldsymbol{a}$ el atributo en $\boldsymbol{A}$ que mejor clasifica a $\boldsymbol{T_r}$
    - Etiquetar a la *raíz* como $\boldsymbol{a}$
    - Para cada posible valor $\boldsymbol{v_i}$ de $\boldsymbol{a}$ ($\boldsymbol{v_i} \in \boldsymbol{Valores(a)}$)
        - Agregar una rama bajo el nodo *raíz* correspondiente a la prueba $\boldsymbol{a = v_i}$
        - Sea $\boldsymbol{T_{r_{v_i}}}$ el conjunto de ejemplos para los que $\boldsymbol{a = v_i}$
        - Si $\boldsymbol{T_{r_{v_i}}}$ está vacío
            - Agregar debajo de la rama un nodo con el valor de $\boldsymbol{c_i}$ más frecuente en $\boldsymbol{T_r}$ como etiqueta
        - De lo contrario
            - $\boldsymbol{ID3(T_{r_{v_i}}, c_i, A - \{a\})}$
- Terminar
- Regresar *raíz*

Tabla 1. Algoritmo ID3.

Este árbol de decisión utiliza la medida de Ganancia de Información para seleccionar entre los atributos candidatos en cada paso del árbol. Es importante mencionar que los atributos que ya han sido incorporados anteriormente son excluidos, por lo que cada atributo puede participar a lo más una vez dentro del árbol. De tal manera que el proceso de selección de atributos continúa para cada nodo hasta que todas las variables explicativas

hayan sido incluidas en el modelo, o hasta que todas las instancias hayan sido correctamente clasificadas.

Limitaciones

- El algoritmo no realiza retrocesos en la búsqueda, es decir, una vez que se selecciona un atributo a testear en un nivel particular, nunca vuelve a reconsiderar esta decisión. Debido a esto, el algoritmo es susceptible al riesgo de converger a soluciones óptimas locales que no son globalmente óptimas.

- Sólo maneja variables discretas.

- Favorece la elección de variables con una mayor cantidad de categorías.

- Presenta riesgo de sobreajuste, debido a que el árbol crece hasta clasificar todas las instancias correctamente.

- La complejidad crece linealmente con el número de instancias de entrenamiento y exponencialmente con el número de atributos.

Árbol de Decisión C4.5

Como se mencionó anteriormente, el árbol de decisión C4.5 es una extensión del algoritmo ID3, que implementa las siguientes mejoras:

- Soluciones al problema de sobreajuste: Para esto incluye métodos de post-poda. Es decir, el algoritmo permite el sobreajuste de los datos, para luego reemplazar los subárboles por una hoja.

  Para esto se consideran cada uno de los nodos en el árbol como candidatos para la poda y, finalmente, son removidos aquellos nodos en que el árbol podado no se desempeña peor que el original en los datos de validación. De este modo, se van podando nodos de forma iterativa, siempre escogiendo aquellos cuya eliminación incrementa el *Accuracy* (porcentaje de datos que son clasificados correctamente) sobre el conjunto de validación. Este proceso continúa hasta que la poda es "dañina"; es decir, hasta que disminuye el *Accuracy* del conjunto de validación.

- Incorporación de atributos continuos: Cuando se utilizan atributos predictores continuos, el algoritmo define dinámicamente nuevos atributos discretos, particionando la variable en un conjunto de intervalos. Para esto se escoge el punto de corte que proporcione mayor Ganancia de Información.

- Medidas alternativas para la selección de atributos: Dado que el criterio de Ganancia de Información para seleccionar atributos favorece aquellas variables con una mayor cantidad de posibles valores, el algoritmo C4.5 utiliza la medida de Ratio de Ganancia, la cual favorece aquellos atributos que, en igualdad de Ganancia de Información, separen los datos en menos clases.

- Manejo de datos incompletos: El algoritmo C4.5 es capaz de trabajar con datos que presenten valores perdidos, para lo que asigna una probabilidad a cada posible valor del atributo (según las demás instancias), y luego se distribuyen los casos en cada rama (o valor del atributo) según dicha probabilidad.

Árbol de decisión C5.0

El árbol de decisión C5.0, comercialmente disponible, es una extensión al algoritmo C4.5. Algunas de las mejoras son:

- Velocidad: El algoritmo C5.0 es más rápido que el C4.5. La velocidad varía dependiendo del tamaño de la base de datos y de la cantidad de atributos que ésta tenga. En [11] se muestra que el algoritmo C5.0 es desde 3 veces a 22 veces más rápido que el C4.5 para las bases de datos utilizadas en dicho estudio.

- Uso de la memoria: C5.0 es más eficiente en uso de la memoria.

- Árboles de decisión más pequeños: C5.0 obtiene resultados similares al C4.5 con árboles de decisión más pequeños.

- Ponderación: C5.0 permite ponderar los distintos casos y tipos de errores de clasificación.

- Soporte para boosting: Algoritmo para mejorar la clasificación de atributos. Este consiste en definir pesos (o cargas) para las distintas instancias en la base de datos según su importancia, de manera que aquellos casos más relevantes influencien en mayor medida la construcción del árbol de decisión.

- Winnowing: Es una fase de selección de atributos previa a la construcción del modelo, donde se descartan aquellas variables que son sólo marginalmente relevantes.

Aunque no se utilizan todas las nuevas opciones del algoritmo C5.0 frente al C4.5, se utiliza este modelo ya que es el algoritmo que incluye el software SPSS Modeler v.16 (sistema requerido por la compañía para implementar el proyecto).

5.1.2.2 Árbol de Decisión CHAID

El algoritmo CHAID (Chi-Squared Automatic Interaction Detection) es un árbol de decisión que se creó con la intención de detectar interacciones entre las variables independientes, pero que también es muy utilizado como técnica predictiva. CHAID construye árboles no binarios, es decir, que un nodo se puede particionar en más de dos ramas. Este árbol es utilizado tanto con fines de clasificación como de regresión, y puede trabajar tanto con variables independientes categóricas como continuas.

A modo de resumen, el algoritmo procede de la siguiente manera:

- Preparación de predictores: El primer paso es crear predictores categóricos en caso de que estos sean continuos, dividiéndolos en categorías con aproximadamente igual número de observaciones. Para el caso de los predictores categóricos, las clases están "naturalmente" definidas.

- Fusionar categorías: Para cada predictor se determina el par de categorías que son menos diferentes estadísticamente con respecto a la variable dependiente. Cuando la variable dependiente es categórica se utiliza el test Chi-Cuadrado, mientras que cuando la variable dependiente es continua se utiliza el test F. Si el test respectivo para un par de categorías del atributo no es estadísticamente significativo, entonces se fusionan las categorías. De esta forma, se repite el proceso de encontrar el siguiente par de clases con menor diferencia estadística (que ahora puede incluir la categoría previamente fusionada).

  Si el test es significativo para el par de categorías del predictor, entonces (opcionalmente) se puede computar el p-valor ajustado de Bonferroni para el set de categorías del atributo respectivo. El p-valor de Bonferroni es calculado como el p-valor por el multiplicador de Bonferroni, el cual depende de la naturaleza de la variable predictora.

  Suponiendo que la variable predictora tiene I categorías originalmente, las cuales son reducidas a r luego del proceso de fusión, el multiplicador de Bonferroni B es el número de posibles formas en que las I categorías pueden ser fusionadas en r clases. Para $r = I$, $B = 1$. Para $2 \leq r < 1$, se utilizan las siguientes ecuaciones:

$$B = \begin{cases} \binom{I-1}{r-1} & Predictores\ Ordinales \\ \sum_{v=0}^{r-1} (-1)^v \frac{(r-v)^I}{v!\,(r-v)!} & Predictores\ Nominales \end{cases}$$

- Seleccionar la variable de corte: Se escoge la variable explicativa con el menor p-valor ajustado, es decir, la variable que genera la división más significativa. Si el menor p-valor ajustado para cualquier predictor es mayor que algún p-valor máximo definido previamente, entonces no se producirán más divisiones, y en consecuencia el nodo respectivo es un nodo terminal.

5.2 Evaluación de Modelos de Clasificación

5.2.1 Matriz de Confusión

La Matriz de Confusión es una herramienta de visualización que contiene información sobre la clasificación de categorías. En esta matriz las columnas representan el número de predicciones de alguna clase, mientras que las filas representan las instancias de la clase real. El resultado de esta matriz permite identificar si el modelo está clasificando correcta o erróneamente las clases al predecir.

El esquema básico de la Matriz de Confusión se muestra en la Tabla 2.

		Predicho	
Real		Positivo	Negativo
	Positivo	TP	FN
	Negativo	FP	TN

Tabla 2. Esquema Matriz de Confusión.

Donde:

- **TP** es el número de predicciones correctas cuando una instancia es positiva.
- **FP** es el número de predicciones incorrectas cuando una instancia es negativa.
- **FN** es el número de predicciones incorrectas cuando una instancia es positiva.
- **TN** es el número de predicciones correctas cuando una instancia es negativa.

De la Matriz de Confusión se derivan ciertas métricas que ayudan a visualizar la calidad de la clasificación realizada. Específicamente:

- Accuracy (AC): Es la proporción del número total de predicciones que son correctamente clasificadas; es decir:

$$AC = \frac{TP + TN}{TP + FP + FN + TN}$$

- True Positive Rate (TPR) o Sensitivity: Es la proporción de casos positivos que son correctamente identificados. Se utiliza la ecuación:

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate (FPR): Es la proporción de casos negativos que son incorrectamente clasificados como positivos:

$$FPR = \frac{FP}{FP + TN}$$

- True Negative Rate (TNR): Es la proporción de casos negativos que son correctamente clasificados:

$$TNR = \frac{TN}{FP + TN}$$

- False Negative Rate (FNR): Es la proporción de casos positivos que son incorrectamente clasificados como negativos:

$$FNR = \frac{FN}{TP + FN}$$

- Precisión (P): Es la proporción de casos predichos positivos que son correctos; esto es:

$$P = \frac{TP}{TP + FP}$$

5.2.1 Curva de Ganancia

La Curva de Ganancia es una forma gráfica de representar la ganancia de los modelos de clasificación, es decir, cuánto porcentaje de casos de éxito se están clasificando correctamente del total en cada punto de la curva. En general, para graficarla, se crean deciles y se calcula la ganancia acumulada en cada uno de éstos. En el eje X se encuentran los deciles, y en el eje Y se encuentra el porcentaje de casos de éxito (en este caso siniestros) capturados.

De esta forma se compara la Curva de Ganancia generada por el modelo de predicción frente al caso base (o sin modelo) y frente al mejor de los casos, es decir, la curva que se tendría en caso de que se estuvieran clasificando correctamente los casos en cada decil. De este modo, mientras la curva del modelo construido se acerque más a esta última, mejor es el modelo.

La Curva de Ganancia es una técnica muy útil para comparar el desempeño en el pronóstico de distintos modelos. En la Figura 6 se muestra un ejemplo de una Curva de Ganancia, donde la línea verde representa el mejor caso, la azul el modelo, y la roja, la línea base.

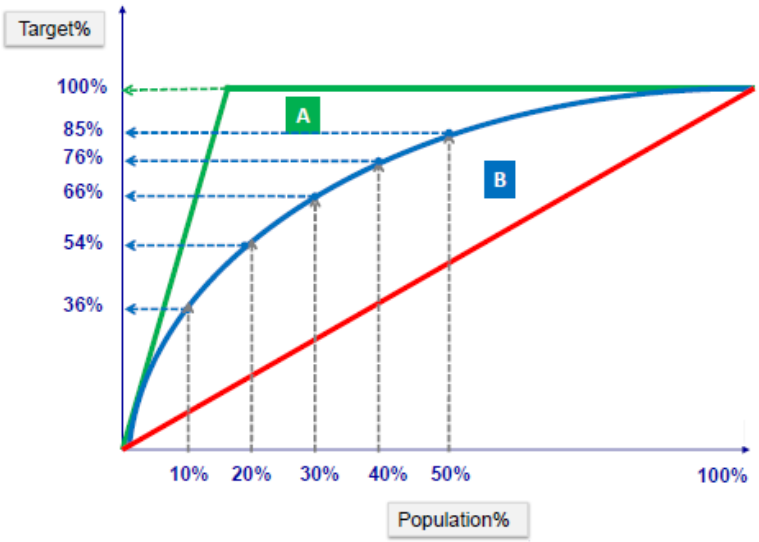

Figura 6. Ejemplo Curva de Ganancia.

5.2.2 Curva ROC (Receiver Operating Characteristic)

Este tipo de curvas es otro método gráfico utilizado para examinar el desempeño de un modelo de clasificación. En la curva ROC se representa la tasa de *False Positive* en el eje X y la tasa de *True Positive* en el eje Y.

El punto (0,1) es el óptimo de clasificación, teniéndose en este caso todas las instancias clasificadas correctamente. El punto (0,0) representa un clasificador que predice todos los casos negativos, mientras que el punto (1,1) predice que todos los casos serán positivos. Finalmente, el punto (1,0) es el peor clasificador pues se equivoca en todas las clasificaciones. En la Figura 7 se muestra un ejemplo de una curva ROC.

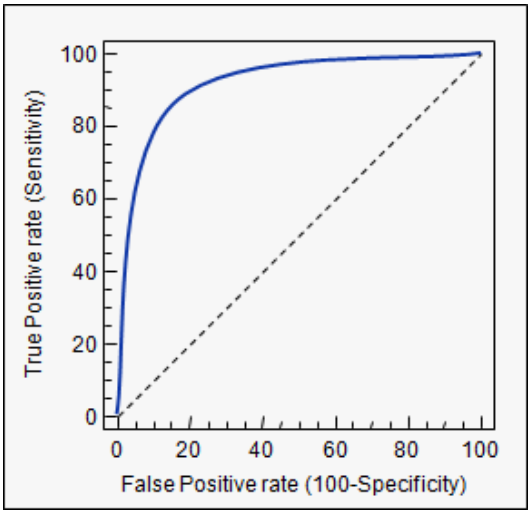

Figura 7. Ejemplo Curva ROC.

AUC (Área bajo la curva ROC)

El área bajo la curva ROC se usa como una medida de la calidad de un clasificador probabilístico y se calcula como sigue:

$$A = \int_0^1 \frac{TP}{P}\, d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP\, dFP$$

Un clasificador perfecto tiene un AUC de 1, de manera que mientras más cercana a 1 sea el área bajo la curva, mejor será su calidad.

6. Metodología

Para el desarrollo del proyecto se sigue la metodología CRISP-DM, siguiendo las distintas fases estándares de dicho proceso.

6.1 Comprensión del negocio

El propósito del presente proyecto es resolver la problemática que tiene hoy en día la empresa para lograr diferenciar entre los posibles clientes pertenecientes al segmento joven. De esta forma, la compañía implementa cargos adicionales a la prima base, los cuales varían desde un 10% a un 140%, dependiendo de los resultados del modelo actual. Es importante señalar que los recargos menores a un 70% se ofrecen solamente a aquellas personas entre 34 y 35 años.

Así, el objetivo del proyecto es estimar el nivel de riesgo de cada individuo perteneciente al segmento de interés, para de esta forma poder ofrecer una prima adecuada al nivel de riesgo de cada cliente, y así, obtener ganancias tanto por la incorporación de nuevos clientes con un bajo nivel de riesgo, como por la disminución de clientes con un alto nivel de riesgo.

Para esto se estimará tanto la probabilidad de que cada individuo tenga uno o más siniestros, como el costo medio de siniestro asociados a cada persona.

6.2 Comprensión de los datos

6.2.1 Confección y Limpieza de la base de datos

Los datos a utilizar son generados por la integración de tres bases de datos. Específicamente:

- RNVM (Registro Nacional de Vehículos Motorizados): Cuenta con información de los vehículos que se encuentran en circulación en el país en el año 2013, y con información demográfica de los dueños de éstos.

- SISGEN (Sistema de Siniestros de Seguros Generales): Sistema administrado por la Asociación de Aseguradores de Chile A.G. que cuenta con información relativa a los siniestros denunciados a las distintas compañías aseguradoras de Chile. Se utilizarán los registros de los últimos cuatro años de siniestros, es decir, desde el 2011 al 2014.

- PSU: Cuenta con información de los puntajes obtenidos a nivel nacional en la Prueba de Selección Universitaria (PSU) y en las Notas de Enseñanza Media (NEM) entre los años 2004 y 2014.

Las variables con las que cuenta inicialmente cada base de datos se muestran en el Anexo A. Algunas de estas son eliminadas debido a su poca utilidad (como el nombre del cliente) o debido a los pocos datos existentes en el atributo (como el color del vehículo).

De forma previa a la integración de las bases de datos se analiza la calidad de cada una de éstas por separado. La cantidad de registros iniciales de cada fuente de datos se muestra en la Tabla 3.

Base de Datos	Cantidad de Registros
RNVM	6.401.742
SISGEN	1.145.864
PSU	2.346.489

Tabla 3. Cantidad de registros iniciales por fuente de datos.

El RNVM es la base de datos con mayores problemas de calidad, por lo que se realizan distintos tratamientos para intentar mejorar los datos en la mayor medida posible. Para empezar se eliminan los registros duplicados y aquellos casos que probablemente pertenecen a rubros de empresa. De esta forma, se obtienen 5.438.144 registros. Además, se recalcula la cantidad de vehículos por individuo, debido a que esta variable tiene una gran cantidad de datos "sucios" (como valores negativos, valores nulos, y registros que no coinciden con la realidad). Adicionalmente, debido a errores en las variables relacionadas a la comuna y región de los individuos, se realiza un cruce con una tabla externa de comunas, distritos y regiones de Chile. Aun así, cabe destacar que incluso después de la imputación de valores realizada, dichos atributos presentan una gran cantidad de datos faltantes (específicamente, cerca de un 23,6% de los 5.438.144 casos). De todas formas, se conservan dichas variables para no perder registros de interés antes de la integración con las demás bases de datos.

El SISGEN es una fuente bastante limpia, por lo que luego de analizarla en esta etapa, tan sólo se obtiene un registro duplicado, el cual es eliminado de la base de datos.

En cuanto a los datos de la PSU, para aquellas personas que rinden la prueba más de una vez, se mantienen los registros de la última PSU rendida y se eliminan los demás. Luego de este proceso se obtienen 1.867.944 registros.

Una vez efectuada la primera limpieza a las distintas fuentes de datos, se procede a realizar la integración de éstas, la Figura 8 muestra un esquema de dicho proceso. Para confeccionar la base de datos a utilizar se comienza cruzando el RNVM con el SISGEN, logrando obtener información relacionada a los siniestros de cada individuo. Dado que el SISGEN sólo registra los datos de aquellas personas que han tenido algún siniestro dentro de los últimos cuatro

años, se asume que los individuos que entregan valores nulos al realizar la integración, no sufren siniestros durante el periodo. Seguidamente, se integran los datos resultantes con la base PSU de aquellas personas entre 18 y 35 años.

De este modo, se obtienen los registros de todas aquellas personas que han rendido la prueba (dentro del segmento joven) y que tienen algún vehículo en el año 2013, en conjunto con su historia siniestral. Así se obtiene una base de datos de 265.466 registros, donde cada uno de ellos es una póliza-ítem, es decir, duplas conformadas por cada individuo (o RUT) en conjunto con su vehículo (o patente asociada). Esto significa que si una persona tiene más de un automóvil, tendrá tantos registros como vehículos posea.

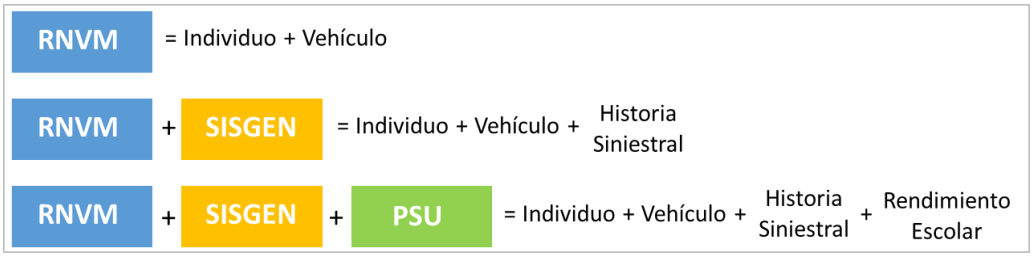

Figura 8. Esquema Integración Bases de Datos.

Es importante mencionar que al realizar el primer cruce (del RNVM con el SISGEN) y, al asumir que los individuos que no se encuentran en esta última base son personas que no han presentado siniestros en el periodo, se está incurriendo en un sesgo muestral. Esto se debe a que dentro del grupo de personas que se está suponiendo que no tienen siniestros, se están considerando tres tipos de casos:

- Individuos que tienen seguro y que efectivamente no presentan siniestros.
- Individuos que tienen seguro, que han tenido siniestros y que no han sido declarados.
- Individuos que no tienen seguro, y que por lo tanto no se tiene certeza de si presentan siniestros o no.

Sin embargo, actualmente la compañía no tiene información para verificar si una persona pertenece a alguno de estos grupos. Es por esto que para evitar este sesgo en la mayor medida posible se realiza un cruce de la base de datos anterior con los individuos que son o han sido clientes de Seguros Falabella[8]. De esta forma, se logra resolver el primer y el tercer punto, pero sigue

[8] Seguros Falabella es un corredor masivo de distintas compañías de seguro, entre ellas, de Penta Security. Se consiguieron los datos de dicha empresa, debido a que Penta Security prácticamente no tiene clientes jóvenes.

existiendo cierto nivel de riesgo debido a que no se logra resolver el sesgo generado por el segundo punto.

De este modo, se obtienen 25.243 registros, lo que corresponde a un 10,08% de la base de datos general.

6.2.2 Análisis Descriptivo

De forma previa a la transformación de los datos se realiza un análisis descriptivo preliminar para familiarizarse mejor con la data disponible.

La base de datos a utilizar cuenta con 25.243 registros, donde las variables a predecir son:

- Tiene o no tiene siniestros en el 2014: En la Figura 9 se puede observar que un 22,1% de los individuos tienen uno o más siniestros en el año 2014 (lo que equivale a 5.578 personas), mientras que un 77,9% no presenta siniestros (19.665 personas).

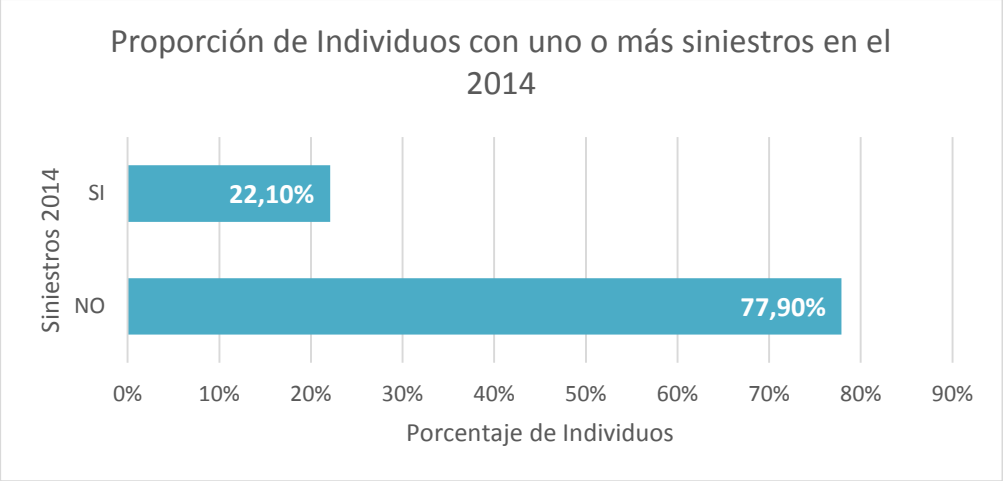

Figura 9. Proporción de individuos con uno o más siniestros en el año 2014.

- Costo Medio 2014: En este caso se va a predecir el valor del costo medio para cada individuo en caso de que la persona haya tenido siniestros en el año. Esto ya que la predicción de la probabilidad entrega los casos en que una persona no presenta siniestros, por lo tanto no es necesario estimarlo nuevamente. Para esto se categoriza la variable como se muestra en la Figura 10. El primer grupo corresponde a 1.505 personas, el segundo a 1.751 y el tercero a 2.267.

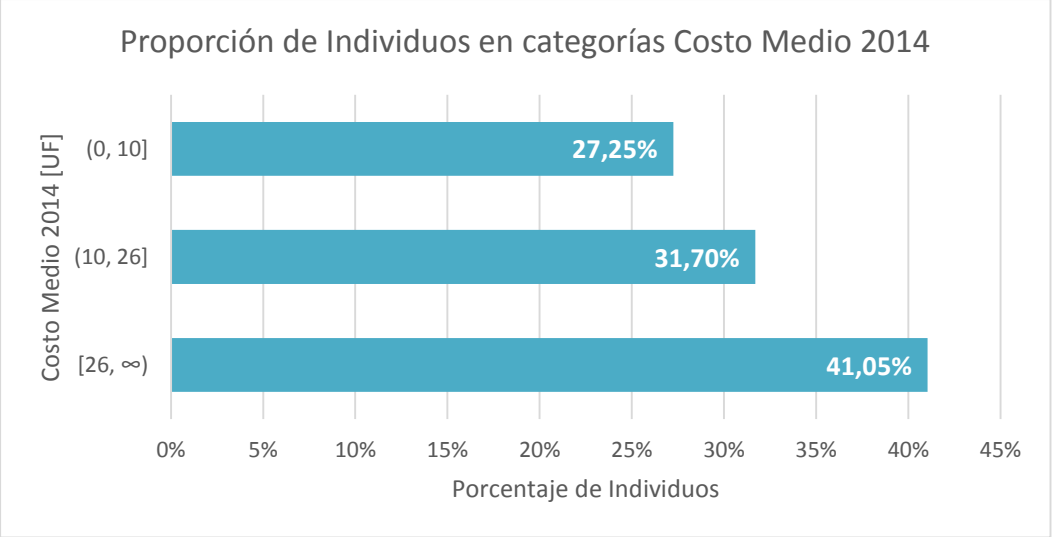

Figura 10. Proporción de individuos por categoría de Costo Medio en el año 2014.

Para pronosticar dichas variables se utilizan predictores asociados a características demográficas, de rendimiento escolar (PSU) y al historial de conducción de cada persona. A continuación se analizan algunas de las variables de cada uno de los grupos.

Variables Demográficas

- Sector Geográfico: En la Tabla 4 se muestra la cantidad de pólizas contratadas por región, y en cuántas de ellas ocurre uno o más siniestros. Se puede observar que la Región Metropolitana es el sector con mayor cantidad de pólizas; sin embargo, no es la que presenta la mayor tasa de siniestros. De hecho, las regiones de Tarapacá y Antofagasta son las que presentan la mayor tasa.

Región	Cantidad de pólizas	Casos con siniestro	% de siniestros	Media de S2014_BIN		Test de Proporciones	
				0	1	Prueba χ^2	p-valor
1	108	30	27,8%	0,004	0,005	2,033	0,154
2	243	60	24,7%	0,009	0,011	0,959	0,327
3	80	16	20,0%	0,003	0,003	0,205	0,651
4	153	38	24,8%	0,006	0,007	0,671	0,413
5	703	138	19,6%	0,029	0,025	2,557	0,110
6	179	29	16,2%	0,008	0,005	3,641	**0,056**
7	216	44	20,4%	0,009	0,005	0,377	0,539
8	672	152	22,6%	0,029	0,027	0,109	0,741
9	316	72	22,8%	0,012	0,013	0,088	0,767
10	246	36	14,6%	0,011	0,006	8,040	**0,005**
11	23	3	13,0%	0,001	0,001	1,096	0,295
12	141	15	10,6%	0,006	0,003	10,816	**0,001**
13	6876	1577	22,9%	0,269	0,283	3,852	**0,050**

14	149	24	16,1%	0,006	0,004	3,124	**0,077**
15	34	4	11,8%	0,002	0,001	2,111	0,146
0	15104	3340	22,1%	0,598	0,599	0,006	0,940
Total	25243	5578	22,1%				

Tabla 4. Cantidad de pólizas y Siniestros 2014 por Sector Geográfico, y Test de Proporciones para cada Sector.

Además, en el lado derecho de la Tabla 4 se muestran los resultados de un test de proporciones, en el cual se prueba si es que existen diferencias significativas en la media de la variable objetivo (presencia o ausencia de siniestros en el año 2014) según cada una de las regiones. Asumiendo un nivel de confianza del 90% se obtiene que las diferencias son significativas en cinco de ellas: regiones 6, 10, 12, 13 (Región Metropolitana) y 14.

En la Figura 11 se muestra el porcentaje de siniestros en el año 2014 por sector geográfico, agrupados en zona norte, centro y sur, donde se puede ver que la zona norte (específicamente las regiones 1 y 2) es la que presenta mayor tasa de siniestralidad. De todas formas, del test de proporciones se obtiene que la zona centro y algunas de las regiones de la zona sur son las que presentan mayor influencia en la tasa de siniestro.

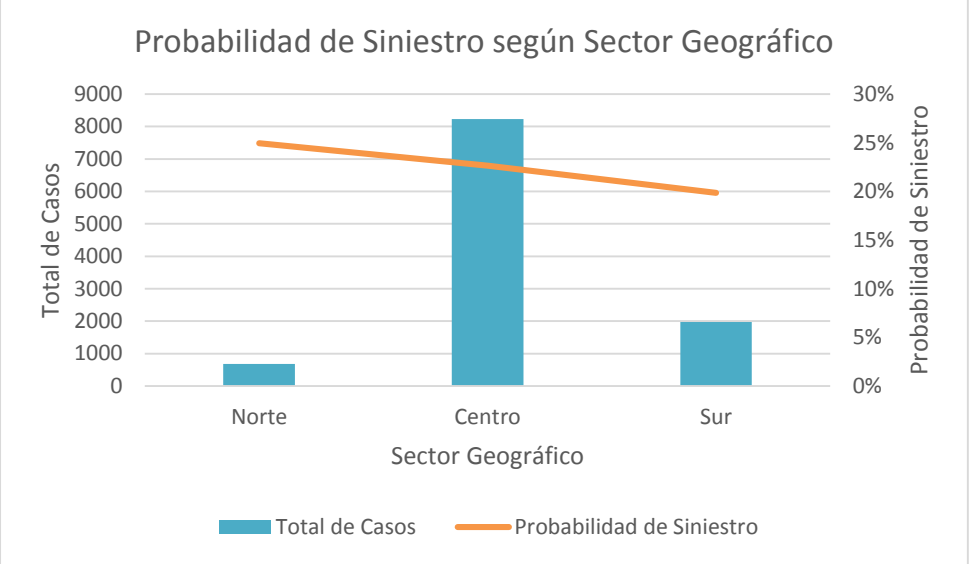

Figura 11. Probabilidad de Siniestro según Sector Geográfico.

- Rango Etario: Esta variable es muy importante para la compañía dado que hoy en día es una de las características más significativas por las que tarifican a los potenciales clientes. En la Figura 12 se puede observar que los individuos entre los 18 y 25 años presentan un porcentaje de siniestros mayor que los individuos con 26 años o más.

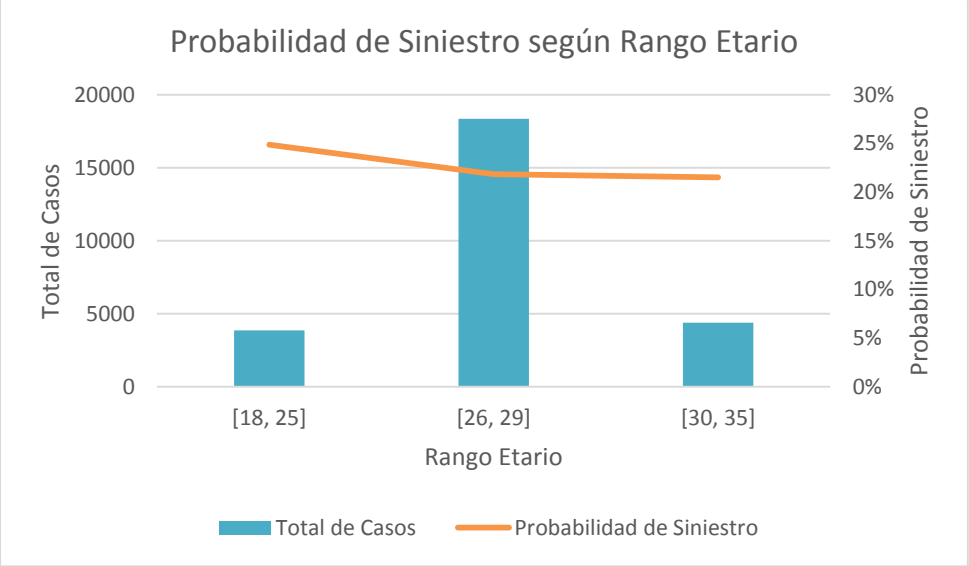

Figura 12. Probabilidad de Siniestro según Grupo Etario.

- Género: Contrario a lo que se cree habitualmente, el género parece no influir en la tasa de siniestros. Como se muestra en la Figura 13 el porcentaje de siniestros no varía de un grupo a otro.

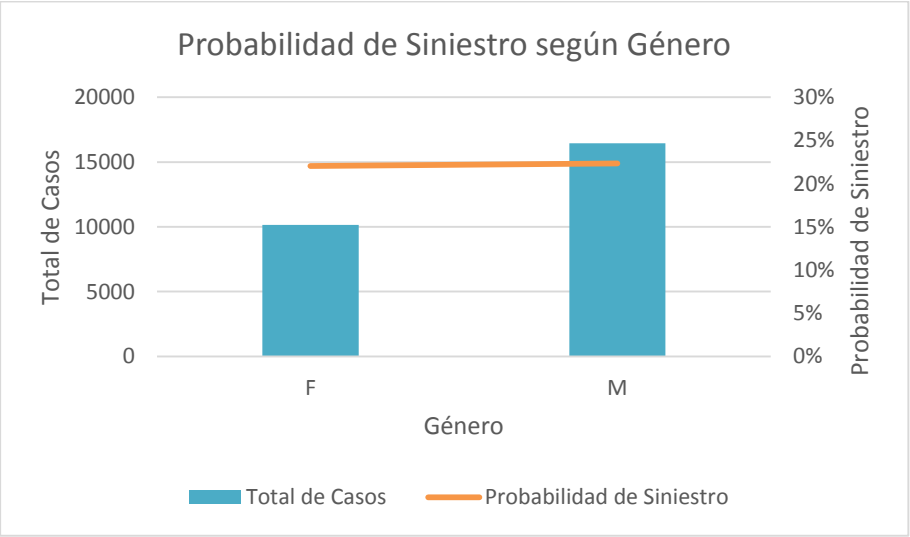

Figura 13. Probabilidad de Siniestro según Género.

- Vehículo: En la Figura 14 se muestra la probabilidad de siniestro según características asociadas al vehículo de la persona. Específicamente, la cantidad de vehículos que posee el individuo en el año anterior al siniestro (en este caso 2013) parece influir en la tasa de siniestros, especialmente para aquellas personas con 3 o más ítems. También se puede observar que la antigüedad del vehículo parece influir en la probabilidad de siniestro, de modo que mientras éste tiene más años disminuye esta tasa. En el último gráfico se muestra la probabilidad de siniestro según el tipo de vehículo,

donde se muestra que, en general, las camionetas son más propensas a presentar accidentes, mientras que las motocicletas menos.

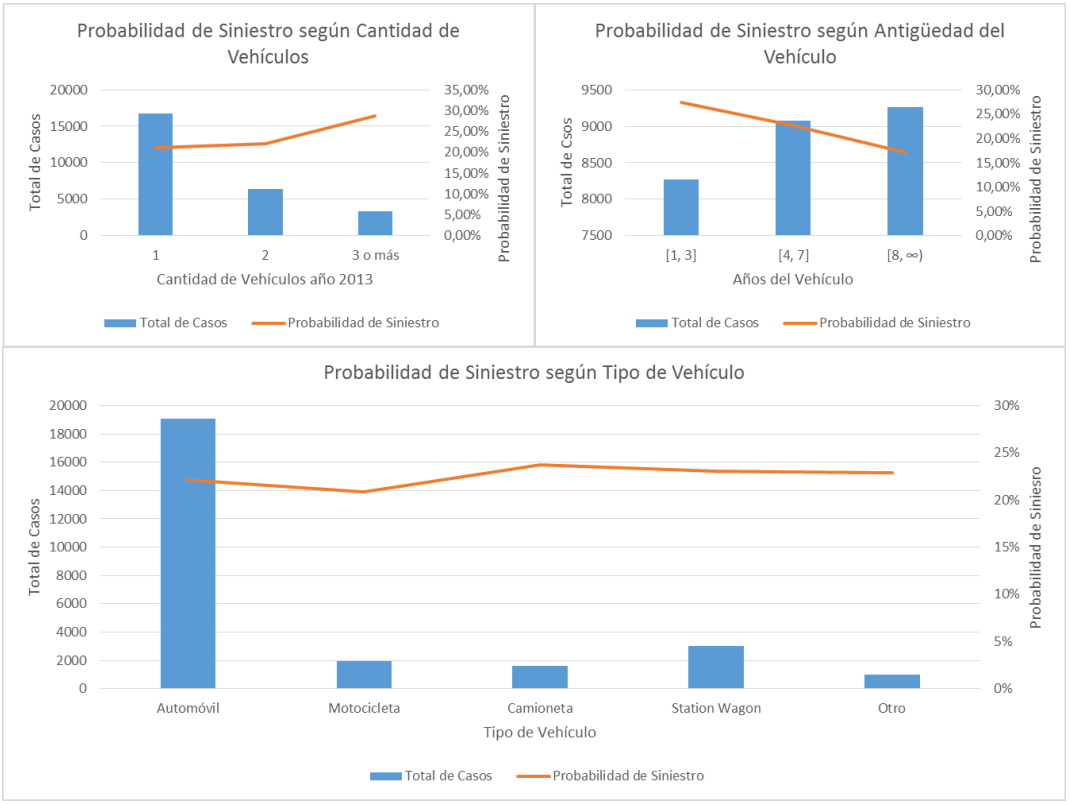

Figura 14. Probabilidad de Siniestro según Cantidad, Antigüedad y Tipo de Vehículo.

Variables de Rendimiento Escolar

- PSU: En la Figura 15 se muestra la tasa de siniestros del año a predecir (2014) dado distintas variables de la PSU, específicamente, Puntaje Matemáticas, Lenguaje, NEM y promedio entre los puntajes de Lenguaje y Matemáticas. Se puede observar que en general, al contrario de la intuición de la compañía, a mayor puntaje PSU la tasa de siniestros aumenta.

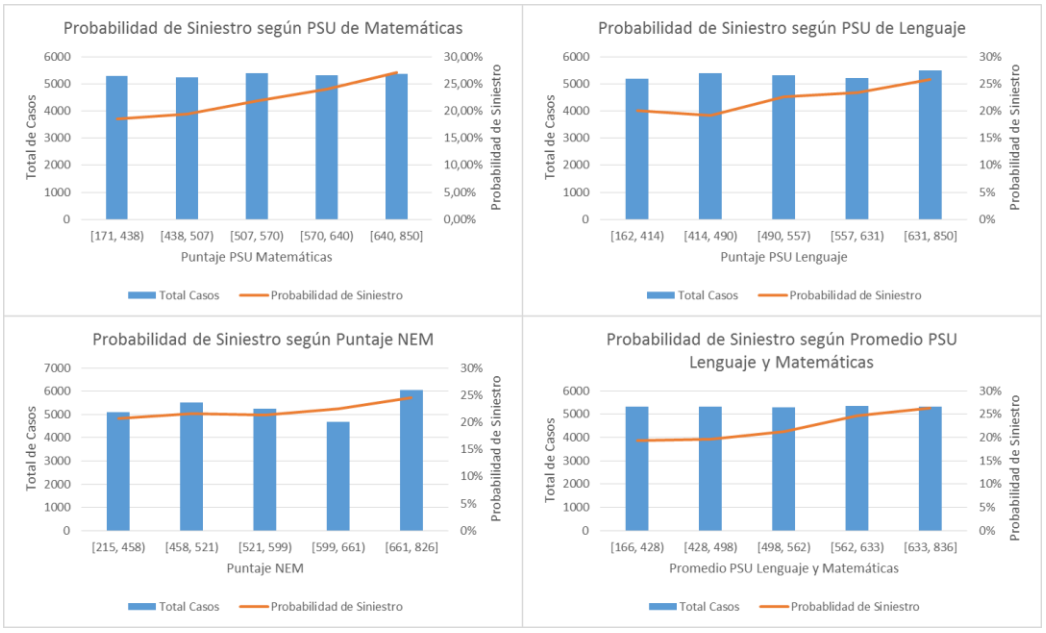

Figura 15. Probabilidad de Siniestro según Variables PSU.

De todas formas, esto se puede deber a que esta variable puede ser un indicador del grupo socioeconómico al que pertenece el individuo, dado que en general quienes pertenecen a un grupo socioeconómico alto, tienen mayor puntaje en la prueba. Esto último se ve reflejado en la Figura 16, generada a partir de estadísticas entregadas por el DEMRE[9] en [21], en la cual se puede observar que los estudiantes provenientes de colegios particulares pagados obtienen un mayor puntaje en las distintas pruebas, entre un 27% a 30% más que en los colegios municipales. Es por esto que más adelante, en la etapa de preparación de los datos, se crearán variables para intentar disminuir este posible efecto.

[9] Departamento de Evaluación, Medición y Registro Educacional.

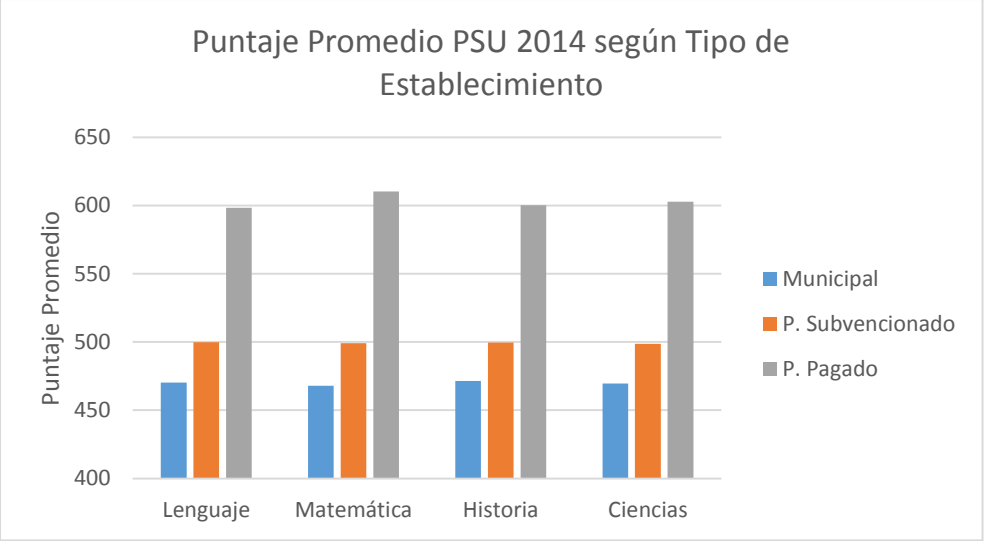

Figura 16. Puntaje Promedio PSU 2014 para las distintas pruebas según Tipo de Establecimiento.

Variables de Historia Siniestral

- Siniestros por año: Claramente las variables asociadas al historial de conducción deben tener influencia en la tasa de siniestro actual. En la Figura 17 se muestra la probabilidad de accidente según la cantidad de siniestros anuales que ha tenido el individuo en el pasado (entre los años 2011 al 2013), en la cual se puede observar que alguien con mayor siniestros por año tiene mayor probabilidad de sufrir un accidente en la actualidad.

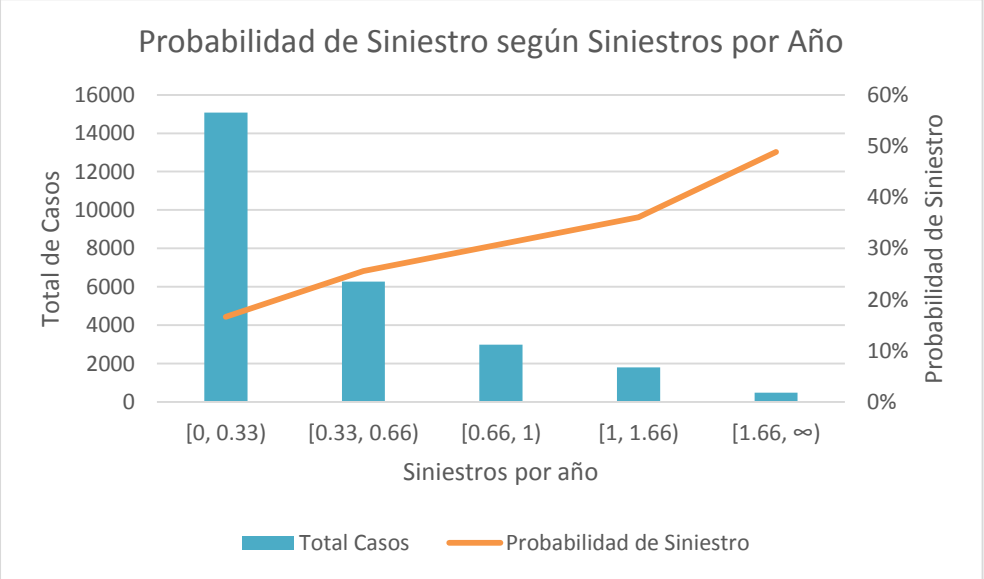

Figura 17. Probabilidad de Siniestro según Siniestros por Año.

- Costo Medio: El costo medio es el promedio del costo por accidente que ha tenido la persona en el pasado (entre el 2011 y el 2013). En la Figura 18 se puede observar que a mayor costo medio existe una mayor probabilidad de sufrir un accidente nuevamente; sin embargo, desde un punto (en este caso desde las 5 UF) el aumento del costo medio no conlleva un importante aumento en la tasa de siniestro.

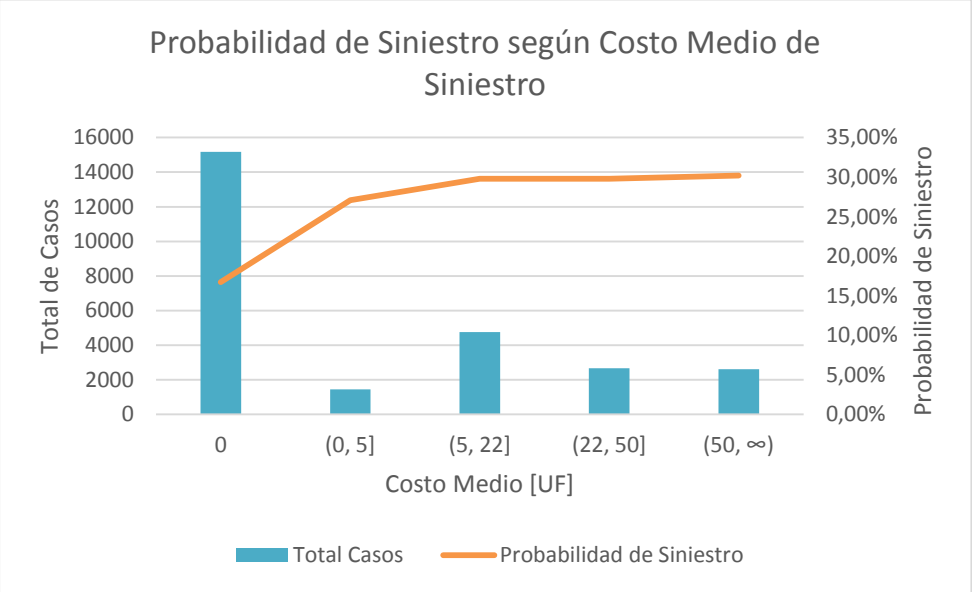

Figura 18. Probabilidad de Siniestro según Costo Medio.

6.3 Preparación de los datos

6.3.1 Transformación de Variables

En esta etapa se generan las variables que serán utilizadas en la fase de modelamiento. Los atributos se transforman según distintas funciones, se categorizan o estandarizan según se necesite.

Algunas de las transformaciones realizadas son:

- Sector: Para crear esta variable se realizan cálculos previos que ayuden a diferenciar las comunas según su nivel socioeconómico. Para esto se calcula el ratio entre el ingreso per cápita de la comuna y el ingreso per cápita del país. De esta forma, se crean 5 grupos según distintos cortes de dicho ratio, donde los primeros grupos representan comunas de un nivel socioeconómico bajo, y los últimos, de un nivel alto.

- PSU estandarizada: Esta variable intenta disminuir el efecto del grupo socioeconómico en las variables relacionadas a la PSU. Para esto se crean

atributos estandarizados (nem_est, mat_est y $leng\_est$), los cuales son calculados para cada individuo i como:

$$psu\_est_i = \frac{psu_i}{prom\_psu_{comuna_i}}$$

Es decir, que para cada persona se divide su puntaje PSU por el puntaje promedio de su comuna. Así, alguien que tiene un valor psu_est sobre 1 se encuentra sobre la media de su comuna, mientras que si tiene bajo 1 se encuentra bajo la media.

- Marca: Para categorizar esta variable se recurre a una tabla que utiliza la compañía actualmente para clasificar las marcas, la cual tiene 6 categorías de marca (A1, A2, B, C, D y E) según la tasación promedio de cada una de ellas. Sin embargo, estas categorías son reducidas a 3, más una categoría de "Otras marcas". La primera categoría la conforman los grupos A1 y A2; la segunda, el grupo B; y la tercera, los grupos C, D y E. Para realizar esta categorización se ejecutaron regresiones, agrupándose aquellas clases que presentaban un efecto similar en la variable objetivo.

- Costo medio: El costo medio de cada póliza i entre los años 2011 al 2013 se calcula como:

$$CM_i = \frac{\sum_t M_{i,t}}{\sum_t S_{i,t}}$$

Donde $M_{i,t}$ es el monto de los siniestros del individuo i en el año t, y $S_{i,t}$ es la cantidad de siniestros del individuo i en el año t. Como esta variable toma valores entre 0 y 998, para el caso de la regresión logística se tomó el logaritmo de dicha variable, específicamente:

$$logCM_i = \log(CM_i + 1)$$

A continuación se realiza un análisis y tratamiento de la calidad de los datos. Específicamente, se tratan los casos atípicos y extremos presentes en las distintas variables, ya que podrían distorsionar los datos en la fase de modelamiento. Además se realiza una última integración con una fuente externa. Algunas de las limpiezas realizadas son:

- Se descartan los casos de vehículos con más de 40 años de antigüedad, ya que muchos de éstos no se encuentran en circulación, y por lo tanto no forman parte del grupo de interés. En este caso se eliminan 85 registros, es decir un 0,33% de la base de datos.

- Se descartan los casos de individuos que poseen más de 5 vehículos. Estos son considerados extremos ya que probablemente se trate de personas con micro empresas (ya sea de taxis, flete, etc.) y que por lo tanto no están dentro del segmento de interés. En este caso se descartan 258 registros, lo que equivale a un 1,02% de la base de datos.

- Se descartan los casos de individuos que han tenido su primer siniestro hace más de 15 años. Esto simplemente porque se encuentran muy desviados de la media y podrían distorsionar los resultados en la etapa de modelado. Además, estos corresponden tan sólo a 11 registros de la base de datos original.

- Se descartan los casos extremos del costo medio (del 2011 al 2013) que superan las 365 UF, con lo que se eliminan 74 registros, lo que corresponde a un 0,29% de la base de datos.

- Se descartan los individuos que presentan más de 2,5 siniestros por año, lo que implica eliminar 58 registros de la base de datos, es decir, un 0,22% de ésta.

- Se descartan los casos extremos de la variable objetivo costo medio del 2014 de aquellos casos que superan las 300 UF, con lo que se eliminan 87 registros, lo que corresponde a un 0,34% de la base de datos.

- Ya que se cree que la variable relacionada a la tasación del vehículo es importante para la estimación de ambas variables objetivo, especialmente para el costo medio, y dado que dicho atributo contiene muchos datos nulos, se crea una integración con la base de datos del SII[10] de Tasación de Vehículos al año 2014. De esta forma, 1.932 casos son eliminados de la base de datos (los que siguen teniendo datos nulos después del cruce), lo que corresponde a un 7,65% de la base de datos original.

Finalmente se obtiene una base de datos con 22.756 registros, es decir, un 90,14% de la base original, de los cuales 4.870 casos presentan al menos un siniestro en el 2014.

Así, luego de realizar las distintas transformaciones y limpieza de atributos se obtienen las variables a utilizar en el proceso de modelamiento. La Tabla 5 muestra dichas variables, su descripción y los valores que éstas toman.

[10] Servicio de Impuestos Internos.

Variable		Descripción	Tipo de Variable	Valores
s2014_bin		Tiene o no tiene siniestros en el 2014	Binaria	
	0	No tiene siniestros en el 2014		
	1	Tiene al menos 1 siniestro en el 2014		
cm2014		Costo Medio 2014 [UF]	Continua	[0,300]
cm2014_g		Costo Medio 2014 categorizada [UF]	Nominal	
	1	(0,10]		
	2	(10,26]		
	3	(26, ∞)		
edad_g		Edad categorizada	Nominal	
	1	[18,25]		
	2	[26,29]		
	3	[30,35]		
sector		Sector del individuo	Nominal	
	1	Norte		
	2	Centro – Grupo 1		
	3	Centro – Grupo 2		
	4	Centro – Grupo 3		
	5	Centro – Grupo 4		
	6	Centro – Grupo 5		
	7	Sur		
	8	Sin Sector		
genero		Género del individuo	Binaria	
	0	Hombre		
	1	Mujer		
anosveh_g		Años del vehículo categorizada	Nominal	
	1	[1,3]		
	2	[4,7]		
	3	[8,∞)		
tipo_veh		Tipo de vehículo	Nominal	
	1	Automóvil		
	2	Motocicleta		
	3	Camioneta		
	4	Station Wagon		
	5	Otro		
marca_g		Marca del vehículo categorizada	Nominal	
	1	Caras		
	2	Medias		
	3	Baratas		
	4	Otra		
tasacion		Tasación del vehículo al año 2014 [Millones]	Continua	[0.05, 41.7]
items_g		Cantidad de vehículos en el año 2013	Nominal	
	1	1 vehículo		
	2	2 vehículos		
	3	3 o más vehículos		
nem		Puntaje NEM	Continua	[236,826]
leng		Puntaje PSU Lenguaje	Continua	[159,850]
mat		Puntaje PSU Matemáticas	Continua	[163,850]
nem_log		$\ln(nem)$	Continua	[5.712, 6.717]
leng_log		$\ln(leng)$	Continua	[5.425, 6.745]

mat_log		$\ln(mat)$	Continua	[5.489, 6.745]
nem_est		Puntaje NEM estandarizado	Continua	[0.443, 1.521]
leng_est		Puntaje PSU Lenguaje estandarizado	Continua	[0.302, 1.699]
mat_est		Puntaje PSU Matemáticas estandarizado	Continua	[0.336, 1.682]
cant_psu		Cantidad de veces que rinde la PSU	Binaria	
	1	Una vez		
	2	Más de 1 vez		
nem_sup		Si el puntaje NEM está sobre la media	Binaria	
	1	Sobre la media		
	0	Bajo la media		
leng_sup		Si el puntaje PSU Lenguaje está sobre la media	Binaria	
	1	Sobre la media		
	0	Bajo la media		
mat_sup		Si el puntaje PSU Matemáticas está sobre la media	Binaria	
	1	Sobre la media		
	0	Bajo la media		
sinporano		$Siniestros\ por\ año = \frac{\sum s_{i,t}}{3}$ de los últimos 3 años (2011 al 2013)	Continua	[0, 2.333]
cm		$Costo\ medio = \frac{\sum M_{i,t}}{\sum s_{i,t}}$ de los últimos 3 años (2011 al 2013) [UF]	Continua	[0, 365]
cm_log		$\ln(cm)$	Continua	[0, 5.903]
anossin1		Años desde el primer siniestro	Continua	

Tabla 5. Variables generadas en el proceso de transformación.

6.3.2 Selección de Variables

Para seleccionar las variables a utilizar se llevan a cabo distintos métodos con la finalidad de comparar los resultados entre éstos.

1. Análisis de Probabilidades Condicionales

En una primera instancia se realiza un análisis exploratorio del poder predictivo de cada variable. Para este efecto, se lleva a cabo un análisis de probabilidades condicionales[11], el cual tiene como finalidad analizar si es que existen diferencias según los valores que tomen las variables explicativas que ayuden a diferenciar entre aquellos que tienen y no tienen siniestros en el 2014.

En otras palabras, se está analizando la probabilidad de siniestro dado los distintos valores de cada variable explicativa. Para los predictores continuos

[11] Como se realiza por Pereira en [2].

se crearon cuantiles (en la mayoría de los casos quintiles) para analizar cómo cambia la probabilidad según distintos valores de la variable.

Un ejemplo del análisis realizado para cada variable se muestra en la Tabla 6, en este caso para el atributo años del vehículo. En cada tabla se calcula la probabilidad de tener uno o más siniestros dado el valor de la variable (o intervalo), además de la probabilidad general de tener siniestros (21,4%). Luego se calcula el ratio entre estas dos probabilidades, para así analizar cuánto influye cada intervalo en el aumento o disminución de la probabilidad sobre la media. Luego, se calcula el logaritmo natural de este ratio con el objetivo de normalizar y hacer comparables los valores de cada categoría. Seguidamente se toma valor absoluto de esta medida y se multiplica por la proporción de casos en cada intervalo. Luego, al sumar estos valores se obtiene una probabilidad absoluta (esquina derecha inferior de la tabla), que indica el nivel de dependencia entre la variable estudiada y la variable objetivo.

Años Vehículo	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[1,3]	7030	2439	9469	0.257577358	1.203579128	0.185299724	0.185299724	42%
[4,7]	5489	1460	6949	0.210102173	0.98174231	-0.018426419	0.018426419	31%
[8,75]	5367	971	6338	0.153202903	0.715869664	-0.334257162	0.334257162	28%
Total General	17886	4870	22756	0.214009492	1			**18%**

Tabla 6. Probabilidades Condicionales – Variable Años del Vehículo.

De esta forma, se realiza un ranking según la importancia de cada variable[12], el que se puede observar en la Tabla 7.

Ranking	Variable	% importancia
1	Anos 1° Sin	27.7%
2	Costo Medio	27.4%
3	Sin por año	27.4%
4	Tasación	20.1%
5	Años Vehículo	17.6%
6	Matemáticas	11.8%
7	Mat sup	11.4%
8	Mat Estand	10.8%
9	Lenguaje	10.2%
10	Leng Estand	9.3%
11	Leng sup	9.2%
12	NEM	6.2%
13	NEM Estand	5.8%
14	Sector	4.9%
15	Marca	4.6%

[12] Los análisis de probabilidades condicionales de todas las variables independientes se encuentran en el Anexo B.

16	NEM sup	4.6%
17	Ítems	4.4%
18	Edad	3.0%
19	Tipo Vehículo	1.0%
20	Género	0.5%
21	Cant PSU	0.2%

Tabla 7. Ranking Variables Independientes – Análisis de Probabilidades Condicionales.

2. Algoritmo CHAID

Como segundo método se utiliza el algoritmo CHAID para analizar la importancia de cada una de las variables a considerar en el modelo. La Figura 19 muestra la importancia normalizada de los distintos predictores según este árbol de decisión. Se puede observar que los atributos más importantes son aquellos relacionados al historial de conducción y al valor y costo medio del automóvil, seguido por las variables asociadas al rendimiento académico.

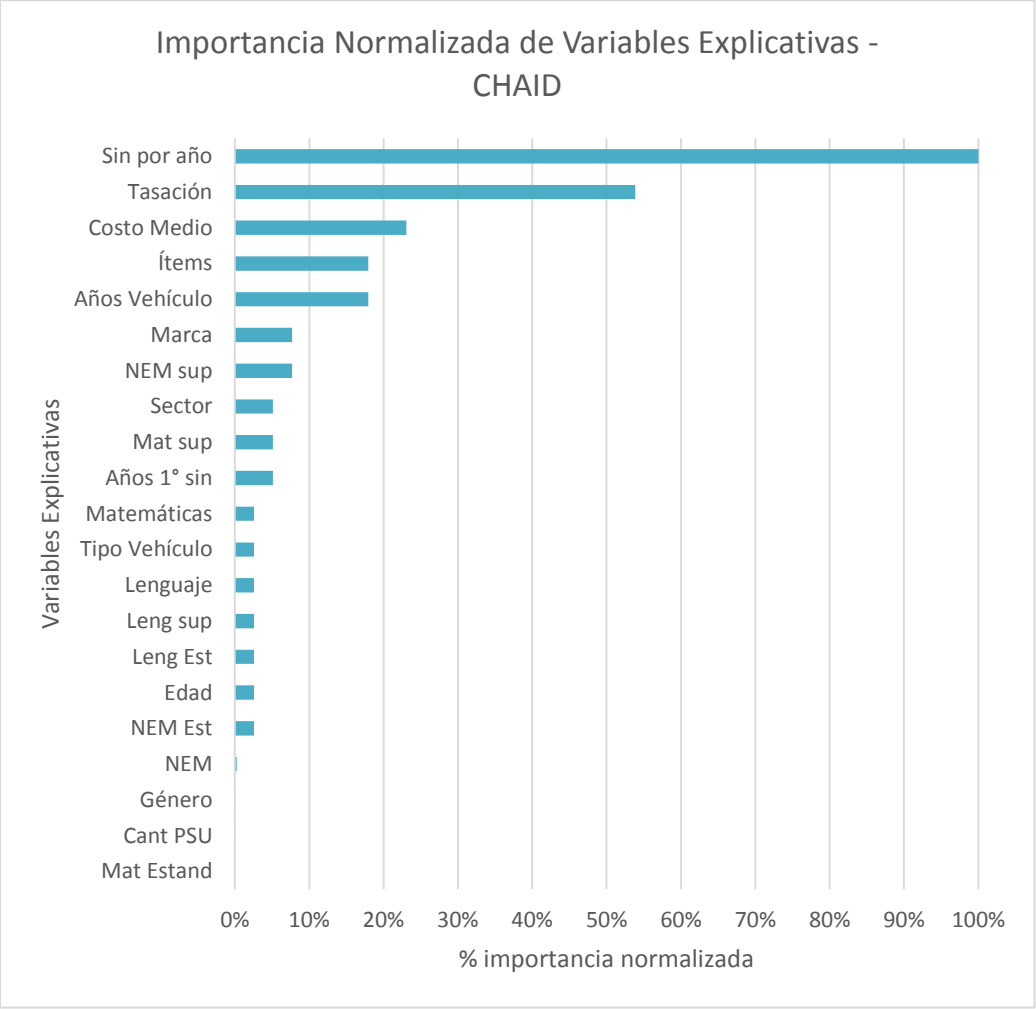

Figura 19. Importancia normalizada de las variables explicativas según método CHAID.

Comparando ambos análisis de importancia se puede observar que en general ambos coinciden en las variables con mayor y menor relevancia; sin embargo, aquellas variables con importancia media (como muchas de las variables PSU) no coinciden exactamente en el orden de importancia.

3. Comparación de Medias

A continuación se realizan pruebas de comparación de medias para las distintas variables explicativas con respecto a la variable objetivo. Este es un método para analizar si es que los valores de una característica en particular difieren al agruparlas en dos o más grupos; en este caso, si presenta o no presenta siniestros en el último año.

En el caso de las variables continuas se utiliza una prueba de comparación de medias (en este caso, un test t de Student), y para las categóricas se realiza un test de proporciones (utilizando la distribución χ^2)[13]. Los resultados de dichas pruebas se muestran en la Tabla 8, en la cual se puede observar que las variables género, cantidad de pruebas rendidas y tipo de vehículo no son significativas, es decir, que la diferencia de la media entre aquellos que tienen siniestros y aquellos que no tienen, no es significativa.

De esta forma, se descartan las variables anteriormente mencionadas: género, cantidad de pruebas PSU rendidas y tipo de vehículo.

Variable	Media de S2014_BIN		Test Comparación de Medias	
	0	1	Prueba (t o χ^2)	p-valor
Edad				
[18,25]	0,253	0,271	7,604	0,006
[26,29]	0,670	0,660	1,691	0,194
[30,35]	0,077	0,068	5,044	0,025
Sector Ingreso				
1	0,132	0,112	16,046	0,000
2	0,133	0,131	0,056	0,818
3	0,030	0,035	3,799	0,051
4	0,028	0,039	16,015	0,000
5	0,076	0,082	2,080	0,149
Sin Sector	0,601	0,601	0,000	0,998
Género				
Mujer	0,380	0,387	1,152	0,283
Años Vehículo				
[1,3]	0,410	0,516	198,841	0,000
[4,7]	0,292	0,279	3,574	0,059
[8, ∞)	0,297	0,204	187,922	0,000
Tipo Vehículo				

[13] En el Anexo C se puede encontrar información relacionada al funcionamiento de ambas pruebas, test de comparación de medias y test de proporciones, utilizadas.

Automóvil	0,760	0,761	0,039	0,844
Motocicleta	0,080	0,074	2,313	0,128
Camioneta	0,010	0,008	2,269	0,132
Station Wagon	0,119	0,125	1,733	0,188
Otro	0,031	0,032	0,063	0,802
Marca				
Caras	0,029	0,049	55,314	0,000
Medias	0,191	0,205	5,634	0,018
Baratas	0,591	0,550	29,956	0,000
Otras	0,189	0,195	1,171	0,279
Tasación	4,82	5,74	273,832	0,000
Cantidad Vehículos				
1	0,675	0,637	27,487	0,000
2	0,230	0,229	0,063	0,802
3 o más	0,095	0,134	71,739	0,000
NEM	556,914	566,596	30,187	0,000
PSU Lenguaje	521,365	535,902	63,071	0,000
PSU Matemáticas	534,386	554,695	121,198	0,000
NEM Estandarizado	0,997	1,013	25,058	0,000
Lenguaje Estandarizado	0,993	1,027	105,443	0,000
Matemáticas Estandarizado	0,995	1,019	50,830	0,000
NEM sobe media				
Sobre media	0,481	0,510	15,005	0,000
Lenguaje sobre media				
Sobre media	0,487	0,564	102,704	0,000
Matemáticas sobre media				
Sobre media	0,496	0,550	49,288	0,000
Cantidad de PSU				
1 vez	0,845	0,842	0,267	0,605
Siniestros por año	0,226	0,389	711,988	0,000
Costo Medio	16,107	23,703	113,286	0,000
Años 1er siniestro	1,564	2,378	400,822	0,000

Tabla 8. Test de Comparación de Medias y Test de Proporciones.

4. Test de Multicolinealidad

Debido a que existe una alta correlación entre algunos de los atributos estudiados, especialmente entre las variables de la PSU, se realiza un test de multicolinealidad[14].

En el Anexo E se encuentra un extracto de la matriz de correlación[15]. Se puede observar que las variables PSU y PSU estandarizada están fuertemente correlacionadas, sobre 0,97; y los puntajes de Matemáticas y Lenguaje tienen

[14] Más información sobre el test de multicolinealidad se puede encontrar en el Anexo D.
[15] Se muestra la correlación entre las variables continuas.

una correlación de 0,76, lo que se considera relativamente alto para la magnitud de las demás correlaciones en la base de datos.

En la Tabla 9 se muestra el factor de inflación de la varianza (VIF), para lo cual se asume que si el promedio o alguno de los factores de cada variable es mayor que 10, se está en presencia de multicolinealidad. En este análisis también se observa que las variables de la PSU están muy correlacionadas entre sí, por lo que es redundante considerar todas estas variables.

Para analizar cuáles son los atributos que entregan mayor información se realizan diferentes regresiones considerando en cada una de ellas tan sólo una de las variables correlacionadas (en este caso, la variable PSU con su respectiva variable estandarizada), y se escoge aquella que entrega el mayor R^2 de Nagelkerke.

Así, luego de realizar las regresiones se obtiene que las variables de rendimiento escolar sin estandarizar entregan mayor información que las estandarizadas, por lo que se eliminan dichos atributos y se obtienen variables explicativas no correlacionadas entre sí, y un factor promedio de inflación de la varianza de 2,78.

Variable	Inicial		Final	
	VIF	1/VIF	VIF	1/VIF
Edad				
[26,29]	1,25	0,801	1,25	0,801
[30,35]	1,23	0,811	1,23	0,811
Años Vehículo				
[2,4]	2,19	0,456	2,19	0,456
[5,7]	2,36	0,423	2,36	0,423
[8, ∞)	3,17	0,316	3,17	0,316
Marca				
Medias	6,88	0,145	6,87	0,145
Baratas	10,69	0,093	10,68	0,093
Otras	7,65	0,131	7,64	0,131
Tasación	1,85	0,539	1,85	0,540
Cantidad Vehículos				
2	1,09	0,921	1,08	0,922
3 o más	1,11	0,902	1,11	0,902
NEM	37,11	0,027	1,60	0,624
Lenguaje	37,37	0,027	2,44	0,409
Matemática	37,53	0,027	2,52	0,397
NEM Estandarizado	34,61	0,028	-	-
Lenguaje Estandarizado	39,28	0,025	-	-
Matemática Estandarizado	38,90	0,025	-	-
NEM sobre media	2,91	0,344	2,90	0,344
Lenguaje sobre media	3,07	0,326	2,91	0,343

Matemática sobre media	2,79	0,358	2,70	0,370
Siniestros por año	1,86	0,538	1,85	0,539
Costo Medio	1,20	0,831	1,20	0,831
Años 1er siniestro	1,97	0,508	1,96	0,509
Sector				
1	1.41	0.708	1,19	0,840
2	1.25	0.803	1,17	0,857
3	1.06	0.942	1,04	0,960
4	1.38	0.724	1,10	0,907
5	1.62	0.615	1,21	0,827
VIF Promedio	**10,17**		**2,78**	

Tabla 9. Test de Multicolinearidad (Antes y Después de la Eliminación de Variables Correlacionadas).

5. Selección final

Por último, se realiza la elección de las variables finalmente a utilizar. Para esto los atributos se introducen uno a uno en el modelo, siguiendo el orden de importancia descrito en los puntos 1 y 2 de este apartado. De esta forma, se escogen aquellas variables que son significativas con un 95% de confianza para el modelo.

Así, se seleccionan siete predictores, específicamente:

- Siniestros por año
- Años transcurridos desde el primer siniestro
- Antigüedad del vehículo
- Tasación del vehículo
- Cantidad de vehículos
- Edad
- Puntaje PSU de Matemáticas

6.4 Modelamiento

6.4.1 Probabilidad de Siniestro

Para la etapa de modelamiento se dividió la data en partición de entrenamiento y prueba, en un 80% y 20% respectivamente. La Tabla 10 muestra las especificaciones de cada una de las particiones.

Partición	Número de Registros	Proporción de Registros	% de Siniestros
Entrenamiento	18.130	80%	21,37%
Prueba	4.626	20%	21,53%
Total	22.756	100%	21,40%

Tabla 10. Especificaciones Partición de Entrenamiento y Prueba – Estimación de Probabilidad.

Regresión Logística

A continuación se construye el modelo de regresión logística tomando como objetivo la variable binaria $s2014\_bin$, y las variables independientes escogidas en la etapa anterior.

En la Tabla 11 se muestra la variable objetivo y las variables predictoras utilizadas finalmente en el modelo de regresión logística.

Rol	Variable	Tipo
Objetivo	Siniestro 2014	Binaria
Predictor	Siniestros por año	Continua
	Años 1er siniestro	Nominal
	Años Vehículo	Nominal
	Tasación	Nominal
	Cantidad de Vehículos	Nominal
	Edad	Nominal
	PSU Matemáticas	Continua

Tabla 11. Variables (dependiente e independientes) utilizadas en la regresión logística.

Como este modelo utiliza variables nominales con más de dos categorías se procede a crear variables ficticias o *dummy*, donde una de las clases toma la función de categoría de referencia, con lo que se generan $n-1$ variables binarias ficticias[16].

De esta forma, luego de ejecutar el modelo de regresión logística se obtienen los coeficientes asociados a cada variable, los que se muestran en la Tabla 12, en conjunto con su significancia (individual y global).

[16] La codificación de las variables nominales se puede observar en el Anexo F.

Variable	Beta	Error Std.	Sig.	Exp(Beta)
Siniestros por año	0,553	0,069	0,000	1,739
Años 1er siniestro			0,000	
2: 0 < X ≤ 2	0,333	0,060	0,000	1,396
3: 2 < X ≤ 3	-0,022	0,087	0,798	0,978
4: X ≥ 4	0,277	0,068	0,000	1,319
Años vehículo			0,000	
2: 1 < X ≤ 4	-0,260	0,052	0,000	0,771
3: 4 < X ≤ 7	-0,432	0,061	0,000	0,649
4 X > 7	-0,620	0,069	0,000	0,538
Tasación			0,000	
2: 3,3 < X ≤ 4,3	0,028	0,064	0,662	1,028
3: 4,3 < X ≤ 6,7	0,166	0,063	0,000	1,180
4 X > 6,7	0,289	0,070	0,000	1,335
Cantidad de Vehículos			0,000	
2: 2 vehículos	0,117	0,045	0,010	1,124
3: 3 o más vehículos	0,423	0,065	0,000	1,526
Edad			0,040	
2: 25 < X ≤ 29	-0,102	0,043	0,017	0,903
3: 29 < X ≤ 35	-0,135	0,079	0,088	0,874
PSU Matemáticas	0,496	0,083	0,000	1,642
Intercepto	-4,500	0,522	0,000	0,011

Tabla 12. Variables en la Regresión Logística.

El modelo presenta un R^2 de Nagelkerke de un orden muy bajo, de 0,071, es decir que este sólo logra explicar un 7% de la variabilidad de la variable dependiente. Esto se debe probablemente a la existencia de múltiples otros factores que influyen en el comportamiento de conducción de los individuos que no se están considerando, tanto factores propios de la persona, como factores externos.

Para analizar la bondad de ajuste del modelo se realizó un test de Hosmer y Lemeshow, [14]. En la Tabla 13 se muestra la tabla de contingencia de dicha prueba, la que con un p-valor de 0,673 lleva a aceptar la hipótesis nula de ajuste del pronóstico.

Decil	s1_bin=0		s1_bin=1		Total
	Observado	Esperado	Observado	Esperado	
1	1.625	1.615,672	188	197,328	1.813
2	1.593	1575,927	220	237,073	1.813
3	1.533	1.541,111	280	271,889	1.813
4	1.506	1.509,301	307	303,699	1.813
5	1.466	1.477,334	347	335,666	1.813
6	1.432	1.437,974	381	375,026	1.813
7	1.390	1.392,749	423	420,251	1.813
8	1.339	1.340,295	474	472,705	1.813
9	1.247	1.268,066	566	544,934	1.813
10	1.143	1.115,571	670	697,429	1.813

Tabla 13. Tabla de Contingencia para el Test de Bondad de Ajuste de Hosmer y Lemeshow.

A continuación se realiza un análisis de las variables en el modelo:

Siniestros por año: Esta variable muestra que alguien con mayor cantidad de siniestros por año es más propensa a incurrir en un accidente nuevamente. Intuitivamente, esta variable es útil para aquellas personas que tienen más años de experiencia de conducción, ya que alguien que tiene un historial de ser una persona propensa a los accidentes, probablemente incurrirá en uno nuevamente; mientras una que no lo es, no. Sin embargo, para aquellos individuos que están entrando al sistema recientemente, y que por lo tanto no tienen historial, esta variable no afecta en la probabilidad, lo que es lógico ya que no se puede saber a priori su nivel de riesgo. Para estas personas es necesario considerar otras variables que logren diferenciar entre los que tienen y no tienen siniestros, como las descritas a continuación.

Años transcurridos desde el primer siniestro: Las personas que no han tenido siniestros anteriormente son las que presentan la menor probabilidad de accidente. Luego, aquellos que han vivido su primer siniestro recientemente, hace 2 años o menos, son los que tienen mayor probabilidad de incurrir en un siniestro nuevamente. Por otro lado, los individuos que tuvieron su primer accidente hace 4 o más años son menos propensos a accidentes, aunque no con una gran diferencia, ya que el coeficiente asociado a ambos grupos no son muy distintos (0,056 de diferencia).

Tasación: Se puede observar que un vehículo con un precio más alto es más propenso a tener un accidente que uno con un precio bajo. Esto se puede deber al tipo de personas que conducen automóviles de mayor valor económico dentro del segmento joven. Probablemente estos individuos son personas que pertenecen al grupo socioeconómico medio-alto, y que por algún motivo son más descuidados que los demás. Esto se puede deber a que el pago de la prima tiene menor impacto sobre sus ingresos, o el de su familia. En la Figura 20 se puede observar el incremento en la tasa de siniestros cuando se aumenta el valor del vehículo.

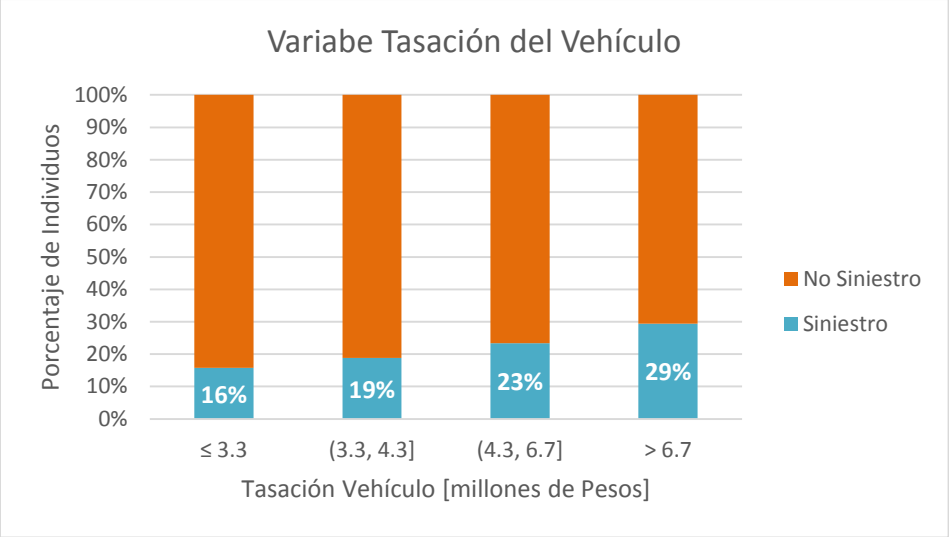

Figura 20. Análisis Variable Tasación del Vehículo.

Años del vehículo: Esta variable indica que los conductores de automóviles más nuevos son más propensos a incurrir en un siniestro que aquellos con vehículos antiguos. Esto podría estar ligado a la variable anterior (tasación), ya que, en general, los vehículos antiguos tienen un menor valor económico; por lo tanto, esta variable también podría ser un indicador del grupo socioeconómico al que pertenece el individuo. En la Tabla 14 se observa cómo varía la tasa de siniestro de un grupo a otro.

Años Vehículo	Siniestro=0	Siniestro=1	Total	% Total General	Tasa Siniestros
1 año	2.722	1.061	3.783	16,62%	28,0%
Entre 2 y 4 años	5.815	1.830	7.645	33,60%	23,9%
Entre 5 y 7 años	3.982	1.008	4.990	21,93%	20,2%
8 años o más	5.367	971	6.338	27,85%	15,3%
Total	17.886	4.870	22.756		

Tabla 14. Análisis Variable Años del Vehículo.

Cantidad de vehículos: En este caso se puede observar que a mayor cantidad de vehículos existe una mayor probabilidad de tener un siniestro. Además, si es que el individuo posee tres o más vehículos, la probabilidad de tener un accidente sobre la de no tenerlo es de 3 veces es a 2 (odds de 1,526) comparado con las probabilidades de siniestro para aquellos con un automóvil. Para el caso de aquellos individuos con dos vehículos, este ratio es de 1,124. De todas formas, como se observa en la Figura 21, el grupo con tres o más vehículos es el que posee menor cantidad de individuos, un 9% específicamente.

Esta relación se puede deber a que, en general, las personas que tienen dos o más vehículos comparten los automóviles con otras personas (ya sea hijo, pareja, etc.), lo que intuitivamente debería influir de forma positiva en la probabilidad de siniestro. Lamentablemente esto no se puede comprobar ya que no existen datos para saber este tipo de información. Otra explicación es que alguien con una mayor cantidad de vehículos, en general, es una persona de un estrato socioeconómico alto y, como se mencionó anteriormente, este tipo de personas parece tener menos cuidado con su o sus automóviles.

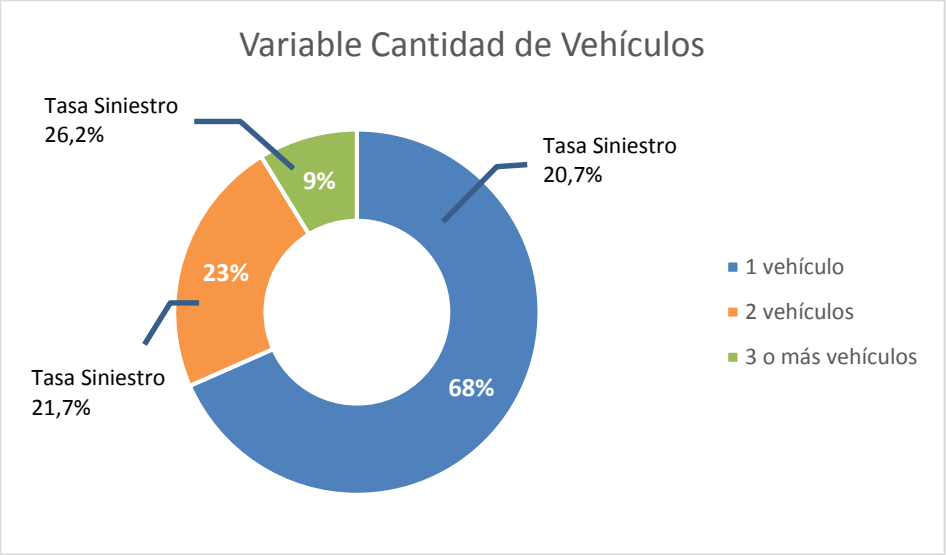

Figura 21. Análisis Variable Cantidad de Vehículos.

Edad: Esta variable indica que a mayor edad hay una menor probabilidad de incurrir en un siniestro, lo que tiene mucho sentido ya que las personas de menor edad, en general, tienen menos experiencia de manejo, y por lo tanto son más propensos a cometer errores de conducción.

PSU: Se puede observar que la PSU de Matemáticas influye positivamente en la probabilidad de siniestro, lo que quiere decir que a mayor puntaje PSU hay una mayor propensión a tener un accidente.

Esto último se puede deber a la alta relación existente entre el puntaje PSU y el nivel socioeconómico de la persona, ya que actualmente en Chile aquellos individuos pertenecientes a un grupo socioeconómico alto obtienen mayores puntajes en la prueba, mientras que los pertenecientes a uno bajo, obtienen menor puntajes. Esto se ve reflejado en la Figura 22, donde se muestra la evolución del puntaje promedio por nivel socioeconómico[17].

[17] Fuente: Mineduc [22].

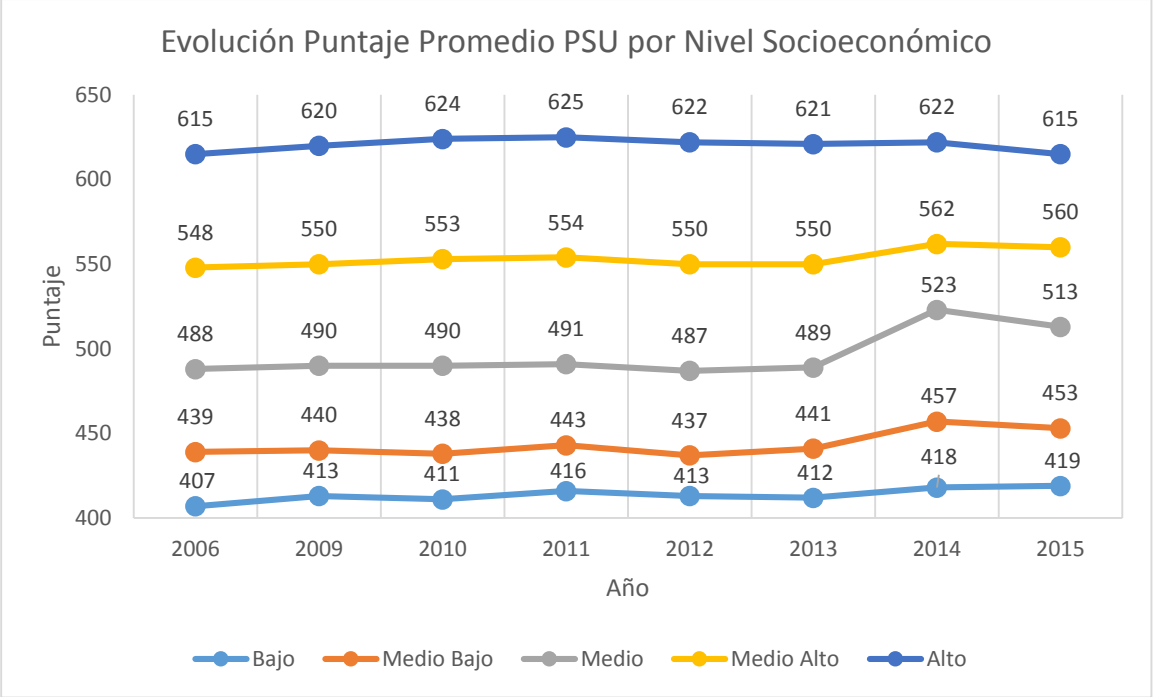

Figura 22. Evolución Puntaje Promedio PSU por Nivel Socioeconómico.

De esta forma, la variable PSU tiene un efecto importante en la probabilidad de siniestro, ya que ésta sería un buen indicador del estrato socioeconómico al que pertenece la persona, y éste último influye en el comportamiento de conducción de los individuos.

De todos modos, a la empresa le convendría obtener los datos asociados al grupo socioeconómico de las personas, ya que esta relación existente entre la PSU y el nivel socioeconómico puede cambiar en el tiempo; por lo mismo también es conveniente verificar cada cierto periodo su utilidad para el objetivo de predicción.

Árbol de Decisión C5.0

Para la realización del árbol de decisión C5.0 también se introducen las variables una a una según la importancia de éstas. Los predictores que se utilizan finalmente se muestran en la Tabla 15. En este caso los atributos son ingresados como variables continuas, ya que el algoritmo C5.0 decide el mejor corte bajo su criterio.

Rol	Variable	Tipo
Objetivo	Siniestro 2014	Binaria
Predictor	Siniestros por año	Continua
	Años 1er siniestro	Continua
	Costo Medio	Continua
	Edad	Continua
	Tasación	Continua
	Años Vehículo	Continua
	Cantidad de Vehículos	Nominal
	PSU Matemáticas	Continua

Tabla 15. Variables (dependiente e independientes) utilizadas en el árbol de decisión C5.0.

El árbol de decisión generado se muestra en el Anexo G, donde se puede observar que las conclusiones coinciden con las de la regresión logística.

Se puede observar que la primera variable escogida por el modelo son los siniestros por año, donde aquellas personas que no han sufrido siniestros anteriormente tienen una baja probabilidad de siniestro (de un 16%).

Luego, para alguien con menos de 0,7 siniestros al año, y con automóviles por un valor superior a los 5,6 millones de pesos, existe una probabilidad de accidente de un 31%, mientras que para alguien con un automóvil bajo ese valor, existe un 23% de probabilidades. Y, si además aquellas personas que tienen automóviles con menor valor económico han presentado siniestros recientemente, existe una mayor propensión a incurrir nuevamente en uno (de un 28%).

También se puede observar que, al igual que en el caso de la regresión logística, aquellas personas que presentan un mayor puntaje PSU de Matemáticas, tienen mayores probabilidades de siniestro. Sin embargo, esta variable parece no ser muy importante según este modelo, ya que se encuentra en los extremos inferiores del árbol.

Por otro lado, para aquellas personas que tienen más de 0,7 siniestros al año, y que además tienen un costo medio bajo las 9 UF existe un 18% de probabilidades de accidente, mientras que aquellos con un costo medio sobre las 9 UF presentan un 35% de probabilidades. Además, en este caso, aquellas personas con 3 o más automóviles presentan un 48% de probabilidades de sufrir un accidente. Y, si aquellos individuos tienen menos de 27 años dicha probabilidad aumenta a un 61%.

Evaluación y Comparación de Modelos

Para evaluar y comparar ambos modelos de estimación de la probabilidad de siniestro se utilizan las siguientes medidas:

- Curva de Ganancias: Se comparan los resultados obtenidos en la curva de ganancias de ambos modelos. Para esto se analizan las ganancias acumuladas en los distintos deciles.

- Matriz de Confusión: Se comparan los aciertos en cada una de las clases, para los casos de éxito o de siniestro (*True Positive*) y los de no siniestro (*True Negative*), así como también el *Accuracy* (o porcentaje de aciertos totales) en cada modelo. Para esto se analizan distintos puntos de corte, y se escoge el que entregue mejores resultados en cada caso.

- Curva ROC y AUC: Además se estudian las diferencias entre las curvas ROC de ambos modelos, y su consecuente área bajo la curva (o AUC).

Es importante mencionar que estas evaluaciones se realizan sobre la partición de prueba, es decir, el 20% escogido anteriormente, que corresponde a 4.626 casos.

Curva de Ganancia

Para graficar la curva de ganancia acumulada se ordenan todos los casos dentro de la partición de entrenamiento según la probabilidad estimada (en forma descendente). Luego se dividen los casos totales en deciles y se calcula la cantidad acumulada de casos con siniestros en cada decil.

La Figura 23 muestra la curva de ganancia acumulada de cada modelo, donde se puede observar que ambos entregan resultados muy similares. Sin embargo, la regresión logística entrega resultados levemente mejores que el árbol de decisión C5.0.

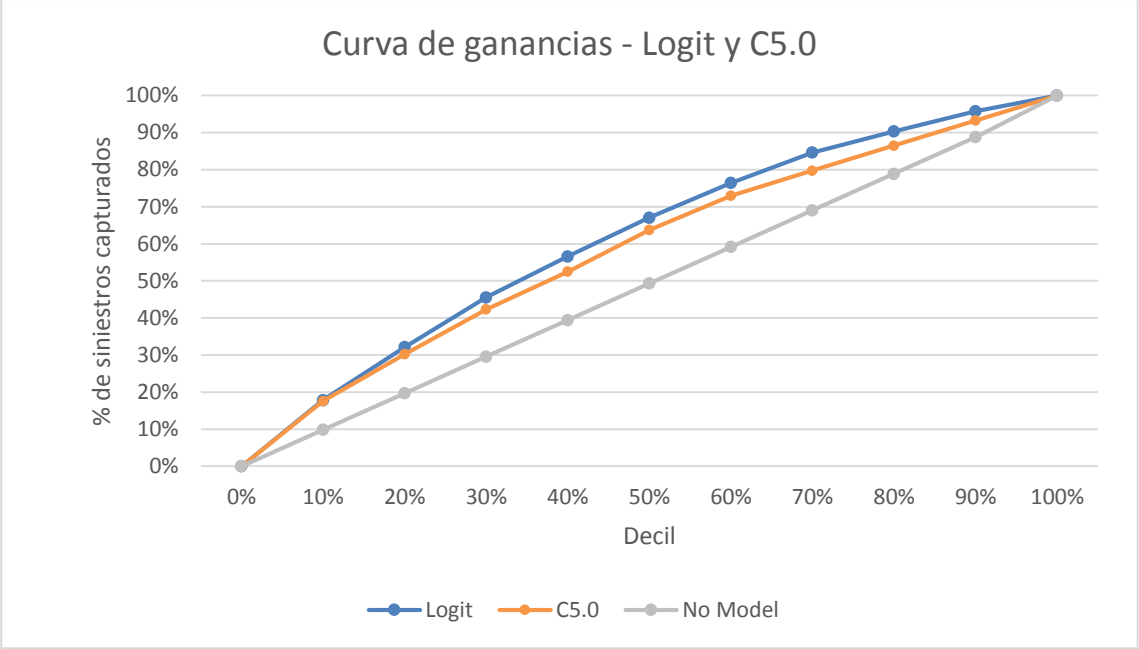

Figura 23. Comparación Curvas de Ganancias Modelo Logit y C5.0.

Matriz de Confusión

- Regresión Logística:

  Antes de desarrollar la Matriz de Confusión, es necesario escoger la probabilidad de corte para predecir los casos de siniestro. Para esto se dividen las instancias en deciles, y se escoge el punto de corte más cercano a cada decil. Los puntos de corte y la ganancia acumulada capturada en cada grupo se muestran en la Tabla 16.

% individuos	Corte	Total Acumulado	Siniestros Acumulados	% Siniestros Capturados
10%	≥ 0,334	461	181	18%
20%	≥ 0,281	928	327	32%
30%	≥ 0,247	1.386	460	45%
40%	≥ 0,221	1.859	576	57%
50%	≥ 0,198	2.318	680	67%
60%	≥ 0,177	2.790	777	77%
70%	≥ 0,159	3.245	859	85%
80%	≥ 0,140	3.709	917	90%
91%	≥ 0,121	4.189	976	96%
100%	≥ 0,000	4.626	1.014	100%

Tabla 16. Probabilidad de corte y siniestros capturados por decil – Regresión Logística.

Luego se estudian las distintas Matrices de Confusión y sus respectivas medidas de *True Positive*, *True Negative* y *Accuracy*. Se analizan los distintos casos y se escoge aquel que entrega un buen equilibrio entre ambas clases, considerando además el hecho de que para la empresa el error asociado a equivocarse en la clase negativa es considerablemente mayor que el caso contrario.

De esta forma, se escoge el punto de corte 0,247, el cual entrega la Matriz de Confusión que se muestra en la Tabla 17. Así, se obtiene un *Accuracy* de 68%, con un porcentaje de acierto de casos de no siniestro de un 74% y 45% de casos de siniestro.

corte=0,247	Real	Predicho		Total	% Acierto
		0	1		
	0	2.686	926	3.612	74%
	1	554	460	1.014	45%
	Total	3.240	1.386	4.626	68%

Tabla 17. Matriz de Confusión Regresión Logística – Corte=0,247.

- Árbol de decisión C5.0:

Siguiendo el mismo procedimiento que en el caso anterior se crean deciles, y como se puede observar en la Tabla 18, desde el sexto decil en adelante la probabilidad de siniestro es nula, por lo que se agregan todos los casos restantes en el último grupo.

% individuos	Corte	Total Acumulado	Siniestros Acumulados	% Siniestros Capturados
10%	≥ 0,367	462	177	17%
19%	≥ 0,250	927	316	31%
30%	≥ 0,165	1.402	465	46%
42%	≥ 0,080	1.940	601	59%
53%	≥ 0,070	2.431	703	69%
100%	≥ 0,000	4.626	1.014	100%

Tabla 18. Probabilidad de corte y siniestros capturados por grupo – Árbol de decisión C5.0.

Se analizan las distintas matrices de confusión que generan dichos cortes, y se escoge aquella con una probabilidad de corte de 0,25, la cual entrega un porcentaje de *True Positive* de un 83% y *True Negative* de 31%, además de un porcentaje de acierto general de un 72%, tal como se observa en la Tabla 19.

corte=0,250	Real	Predicho		Total	% Acierto
		0	1		
	0	3.001	611	3.612	83%
	1	698	316	1.014	31%
	Total	3.699	927	4.626	72%

Tabla 19. Matriz de Confusión Árbol de decisión C5.0 – Corte=0,250.

Curva ROC y AUC

Por último, se evalúa la curva ROC (Receiver Operating Characteristic), la cual compara el ratio de falsos positivos con el de falsos negativos según varía el umbral o corte de clasificación de la probabilidad de siniestro. En este caso se puede observar que el modelo de regresión logística también es superior al resultado entregado por el algoritmo C5.0, pero tampoco con una diferencia muy importante (ver Figura 24).

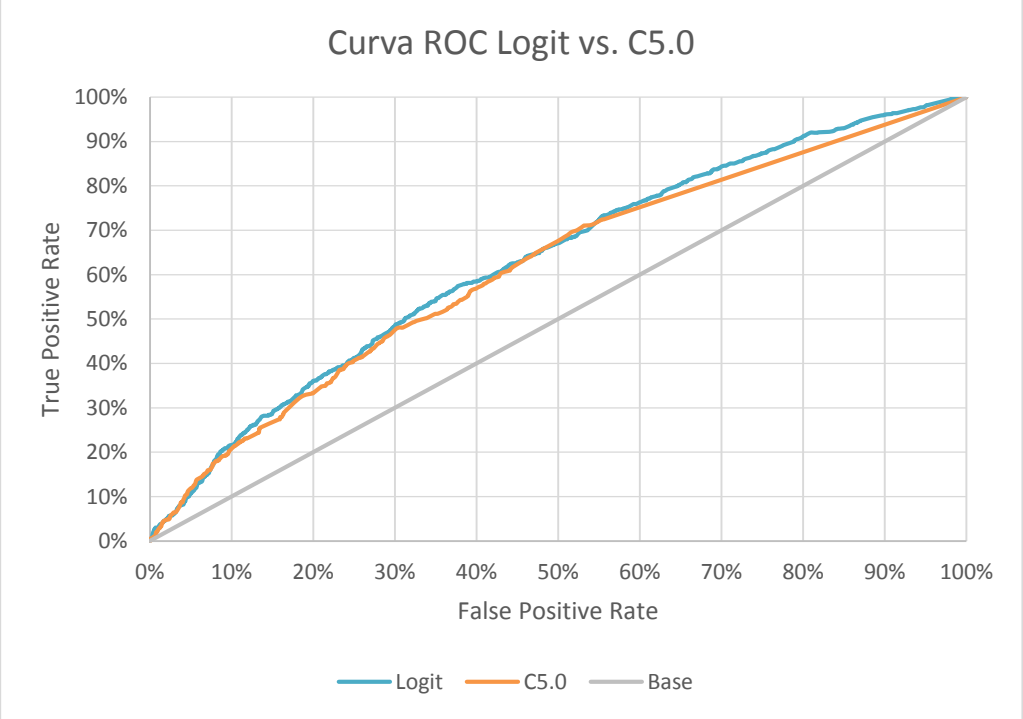

Figura 24. Curva ROC Regresión Logística vs. Árbol de Decisión C5.0.

Por último, el área bajo la curva ROC (AUC) para cada modelo se muestra en la Tabla 20. El modelo logit presenta un AUC de 0,649, es decir un 0,015 más que el algoritmo C5.0.

Modelo	AUC
Logit	0,649
C5.0	0,634

Tabla 20. AUC (Area Under ROC Curve).

A modo de resumen se muestra en la Tabla 21 las distintas métricas de evaluación y comparación de ambos modelos de estimación de la probabilidad de siniestro. Se puede observar que la regresión logística supera al modelo C5.0 en la mayoría de las métricas. Sin embargo este último presenta un mayor *Accuracy* y *True Negative Rate*, a pesar de que el umbral de corte utilizado es similar. Sin embargo, debido a que estas medidas cambian significativamente dependiendo del punto de corte, se le da mayor peso a las métricas relacionadas a la curva de ganancias y a la curva ROC.

Modelo	Ganancia 1° decil	True Positive	True Negative	Accuracy	AUC
Logit	18%	45%	74%	68%	0,649
C5.0	17%	31%	83%	72%	0,634

Tabla 21. Resumen métricas de evaluación y comparación de modelos de estimación de la probabilidad de siniestro.

Finalmente, en base a los resultados anteriores, se escoge el modelo de regresión logística para la estimación final de la probabilidad de siniestro de cada individuo.

Modelo Sin PSU

De forma adicional, se decide construir un modelo sin las variables asociadas a la PSU, ya que a la compañía le interesa saber si es necesario adquirir estas variables de forma permanente en el tiempo. De este modo, se construye un modelo de regresión logística (debido a que fue el mejor evaluado anteriormente) con las mismas variables anteriores, pero sin el atributo PSU de Matemáticas. Los resultados del modelo se muestran en la Tabla 22.

Variable	Beta	Error Std.	Sig.	Exp(Beta)
Siniestros por año	0,546	0,069	0,000	1,726
Años 1er siniestro			0,000	
2: $0 < X \leq 2$	0,344	0,060	0,000	1,411
3: $2 < X \leq 3$	0,002	0,087	0,984	1,002
4: $X \geq 4$	0,317	0,067	0,000	1,374
Años vehículo			0,000	
2: $1 < X \leq 4$	-0,215	0,053	0,000	0,807
3: $4 < X \leq 7$	-0,328	0,064	0,000	0,702
4 $X > 7$	-0,516	0,073	0,000	0,597
Tasación			0,000	
2: $3,3 < X \leq 4,3$	0,045	0,064	0,000	1,046
3: $4,3 < X \leq 6,7$	0,197	0,062	0,479	1,218

4 X > 6,7	0,345	0,069	0,002	1,412
Cantidad de Vehículos			0,000	
2: 2 vehículos	0,084	0,045	0,063	1,088
3: 3 o más vehículos	0,361	0,064	0,000	1,434
Edad			0,012	
2: 25 < X ≤ 29	-0,113	0,042	0,008	0,893
3: 29 < X ≤ 35	-0,180	0,079	0,022	0,835
Intercepto	-1,434	0,079	0,000	0,238

Tabla 22. Variables en la Regresión Logística – Modelo sin PSU.

Este modelo entrega un R^2 de Nagelkerke de 0,069 (tan sólo 0,2% menos que en el modelo con las variables PSU). En cuanto al ajuste del modelo, se realizó un test de Hosmer y Lemeshow, el cual entrega un p-valor de 0,38, por lo tanto no se puede rechazar la hipótesis nula de ajuste a los datos. De este modo, el modelo sigue ajustándose a lo observado en la realidad.

Además, al comparar la ganancia obtenida en cada decil de la Curva de Ganancia (Tabla 23) se observa que no presenta una gran diferencia, de alrededor de un 1% menos en algunos deciles.

Modelo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Con PSU	18%	32%	45%	57%	67%	77%	85%	90%	96%	100%
Sin PSU	17%	32%	45%	56%	66%	76%	83%	91%	96%	100%

Tabla 23. Ganancia acumulada por decil Modelo con PSU vs. Modelo sin PSU.

En la Matriz de Confusión (Tabla 24) también se observa que no existe gran diferencia entre las distintas métricas, de hecho, en ambos modelos se tiene el mismo *Accuracy*, de un 68%. En este caso se obtiene un 44% de verdaderos positivos (frente a un 45% del modelo con PSU), y un 75% de verdaderos negativos (frente a un 74%). Adicionalmente se compara el área bajo la curva (AUC) de ambos modelos, y se obtiene una pequeña diferencia de 0,003 (el AUC del modelo sin PSU es de 0,646, mientras que el con PSU es de 0,649).

corte=0,247	**Real**	**Predicho**		**Total**	**% Acierto**
		0	1		
	0	2.696	917	3.612	75%
	1	567	447	1.014	44%
	Total	3.262	1.364	4.626	68%

Tabla 24. Matriz de Confusión Regresión Logística Modelo Sin PSU – Corte=0,247.

De esta forma, se decide no incorporar las variables de la PSU, ya que no entregan mejoras sustanciales en el desempeño del modelo, y se continúa utilizando este último modelo desarrollado (regresión logística sin PSU) tanto para la integración con la estimación del costo medio, como para la evaluación de los resultados del negocio.

6.4.2 Costo Medio de Siniestros

Para la construcción del modelo de estimación del costo medio asociado a cada individuo se llevan a cabo distintas medidas:

1. Primero se seleccionan sólo aquellos casos que tienen costo medio 2014 mayor a 0. Esto se debe a que el modelo anterior (regresión logística binaria) entrega la estimación de aquellas personas que no presentan siniestros y, por lo tanto, tampoco costo medio. De esta manera, sería redundante considerar el primer grupo como aquellos individuos con costo medio nulo.

2. Se categoriza la variable objetivo en tres grupos, para lo que se consultó a expertos del negocio de la compañía. Las tres categorías para el costo medio son:

    - Leve (entre 0 y 10 UF): Esto ya que la prima promedio que cobra la compañía es de 15 UF anuales, ya que para marginar se necesita un máximo de 60% de siniestralidad; lo que, considerando esta prima media, significan 9 UF de costo, por lo tanto, 6 UF de margen.

    - Medio (mayor a 10 UF y menor o igual a 26 UF): Debido a que el costo medio por siniestro[18] es de 26 UF se utiliza este corte para diferenciar entre los siniestros medios y graves. Es importante destacar que este grupo de clientes no necesariamente significan pérdidas para la empresa.

    - Grave (mayor a 26 UF): Este grupo de clientes sobrepasa el límite del costo medio de siniestro de la compañía, y además, representa el grupo de individuos que generan pérdidas para la empresa.

3. Debido a que, a pesar de haber eliminado los casos con costo medio nulo, las clases no están totalmente balanceadas, se procede a equilibrar las categorías. Para esto se aumentan las clases con menor número de instancias de la forma que se muestra en la Tabla 25. Se utilizan estos factores debido a que son los que logran alcanzar igual distribución ($^1/_3$ de probabilidades en cada una) utilizando un método sencillo de *over-sampling* [23].

[18] Considerando sólo aquellas personas que efectivamente presentan siniestros.

Grupos Costo Medio 2014	Factor
1	1,415
2	1,277
3	1,000

Tabla 25. Factores de balanceo de clases.

Es importante mencionar que este balanceo de clases se realiza únicamente en la partición de entrenamiento, y luego se evalúan los resultados del modelo en la partición de prueba, la que sigue teniendo la misma distribución de clases que en un principio.

4. Al igual que en el modelo de estimación de la probabilidad de siniestro se divide la data en partición de entrenamiento y prueba, en un 80% y 20% respectivamente. La cantidad de registros en cada partición se muestra en la Tabla 26.

Partición	Número de Registros	Proporción de Registros
Entrenamiento	3.860	80%
Prueba	964	20%
Total	4.824	100%

Tabla 26. Especificaciones Partición de Entrenamiento y Prueba – Estimación de Costo Medio.

Al igual que en el modelo anterior se fueron introduciendo las variables una a una y evaluando en cada paso el porcentaje de acierto del modelo, así como también la complejidad del árbol. De esta forma, los atributos utilizados en el modelo son los que se muestran en la Tabla 27.

Rol	Variable	Tipo
Objetivo	Costo medio 2014	Nominal
Predictor	Siniestros por año	Continua
	Años 1er siniestro	Continua
	Tasación	Continua
	Años vehículo	Continua
	Cantidad de vehículos	Continua
	Marca	Nominal
	NEM	Continua
	NEM sobre media	Binaria

Tabla 27. Variables (dependientes e independientes) – Modelo Estimación Costo Medio.

El árbol resultante se encuentra en el Anexo J, en el cual se puede observar que las variables más importantes son las relacionadas al costo medio que ha tenido la persona en el pasado, y las asociadas al valor del vehículo, como la marca o la tasación.

También, en general se muestra que las variables relacionadas a la PSU (en este caso el NEM) influyen positivamente en el costo medio del individuo, es decir, que a mayor NEM (o sobre la media) existe una mayor proporción de casos con Costo Medio 2014 en los grupos 2 o 3.

También es importante mencionar que, al contrario del caso de la probabilidad de siniestro, el puntaje PSU de Matemáticas o Lenguaje parece no ser importante, ya que aunque estas variables se probaron en el modelo, estas no fueron seleccionadas.

Evaluación del Modelo

Para evaluar este modelo se utilizó la Matriz de Confusión generada, en conjunto con sus métricas de acierto (ver Tabla 28).

		Predicho			Total	% Aciertos
		1	**2**	**3**		
Real	**1**	137	59	71	267	51%
	2	115	76	120	311	24%
	3	162	74	150	386	39%
	Total	414	209	341	964	38%

Tabla 28. Matriz de Confusión – Árbol de Decisión C5.0.

Se puede observar que el modelo presenta una baja tasa de acierto, tanto general (38% de *Accuracy*) como en cada una de las clases. La primera categoría, es decir, las personas con un menor costo medio, son las que se logran predecir con mayor precisión, sin embargo el porcentaje de acierto sigue siendo bastante bajo, de un 51%.

De esta forma, se piensa que al integrar los resultados de los dos modelos (probabilidad y costo medio) para generar un *score final*, la precisión obtenida por el primer modelo se puede ver afectada negativamente. De todos modos se realiza el cálculo de este puntaje y se evalúa la Curva de Ganancia según éste, para de esta forma analizar si es que realmente se ve afectado el resultado.

Para esto se calcula el promedio del costo asociado a cada una de las categorías, lo que se muestran en la Tabla 29. Luego a cada individuo se le asocia un costo medio según su clase estimada.

Grupos Costo Medio 2014	Promedio de Costo Medio 2014
1	5,84
2	17,602
3	77,079

Tabla 29. Promedio de Costo Medio 2014 por categoría.

Integración de los modelos

Se calcula el puntaje resultante de multiplicar ambas predicciones, es decir, la probabilidad de siniestro de cada individuo por su costo medio estimado.

Para evaluar si es que este modelo *Score* afecta el desempeño de la regresión logística se toman los casos de entrenamiento del primer modelo (es decir 4.626 casos) y se estima tanto la probabilidad como el costo medio de siniestro, para así calcular el puntaje asociado. A continuación se ordenan los casos según este *score* de forma descendente y se grafica la Curva de Ganancia obtenida en este caso. La Figura 25 muestra la comparación de ambas curvas, donde se puede ver que los resultados son bastante similares, a excepción de los casos entre los deciles 2 y 6 aproximadamente.

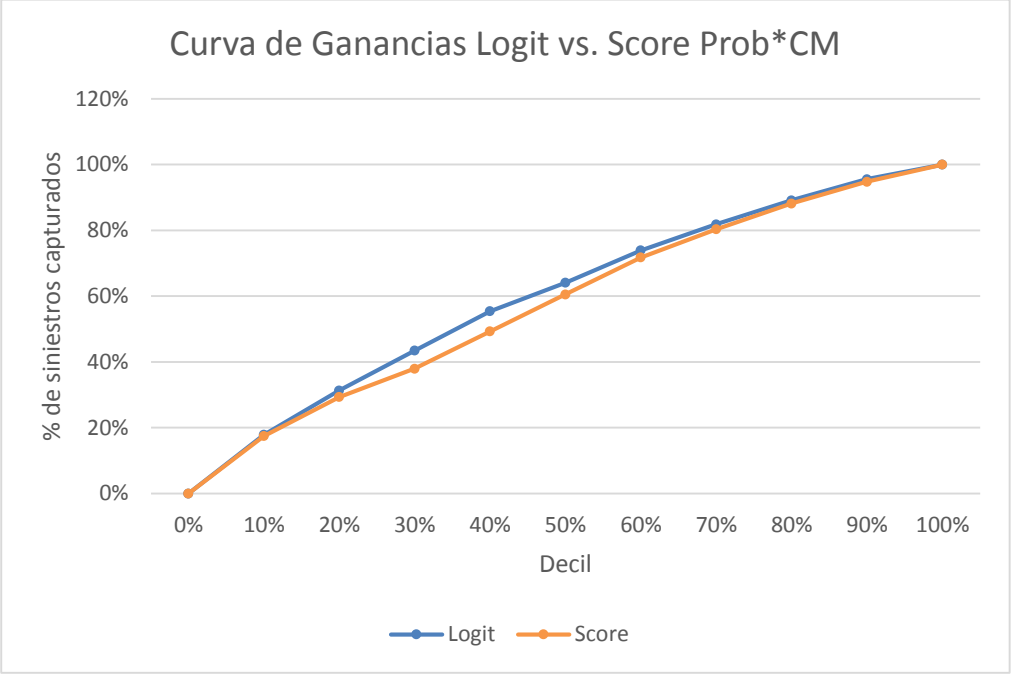

Figura 25. Comparación curvas de ganancia Logit vs *Score ProbCosto Medio.**

Debido a que la estimación del costo medio no genera valor adicional al entregado por la probabilidad de siniestro, se decide utilizar tan solo el modelo logit para estimar el riesgo de cada individuo.

Más adelante se realiza una evaluación del resultado de negocio utilizando este modelo, donde se realiza un análisis de escenarios según distintos puntos de corte de esta probabilidad.

6.4.3 Comparación con el Modelo Actual

Actualmente la compañía tiene un modelo de estimación del riesgo asociado a cada individuo, el cual entrega un puntaje derivado de otros dos puntajes (de frecuencia y de monto de siniestro).

Al igual que con el *score* anterior, se ordenaron los casos de mayor a menor puntaje para luego construir la Curva de Ganancia que entrega el modelo actual. La comparación de ambas curvas se muestra en la Figura 26, donde se puede observar que las ganancias por decil que entrega el modelo actual (modelo K-0) son bastante menores que las ganancias entregadas por la regresión logística.

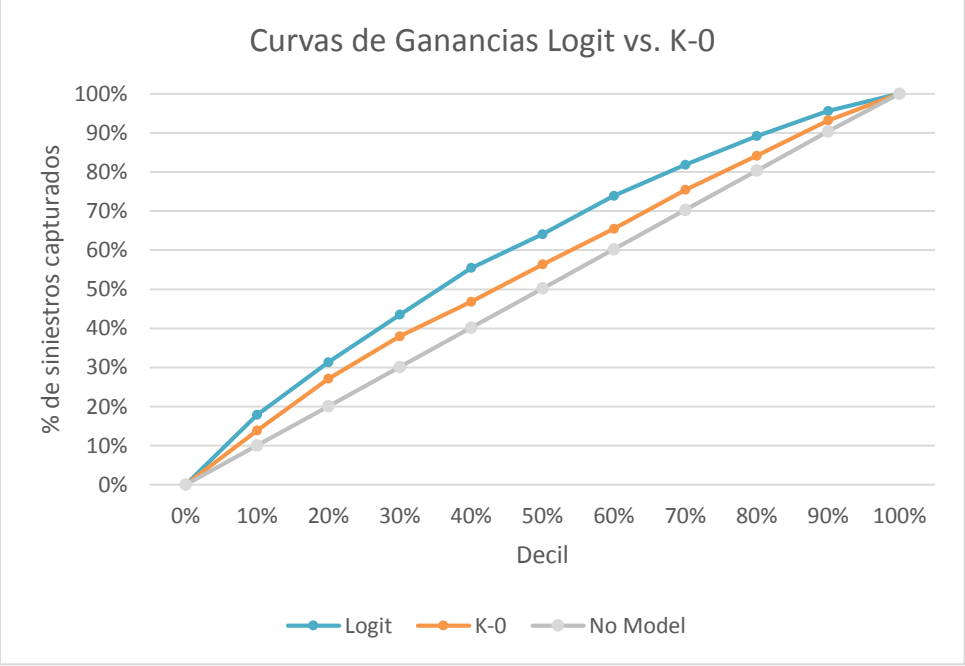

Figura 26. Comparación Curvas de Ganancias Modelo Logit vs. Modelo K-0.

Más específicamente, en la Tabla 30 se puede ver en más detalle la ganancia acumulada por decil en cada uno de los modelos.

Modelo	1	2	3	4	5	6	7	8	9	10
Logit	17%	32%	44%	56%	66%	76%	84%	91%	96%	100%
K-0	13%	27%	38%	47%	56%	65%	75%	84%	93%	100%

Tabla 30. Ganancia acumulada por decil Regresión Logística vs. Modelo K-0.

Así, se puede observar que el modelo planteado entrega valor adicional al negocio, aunque no con una diferencia muy grande, entre un 5% y un 10% dependiendo de cada decil.

6.5 Evaluación

Para evaluar el impacto del nuevo modelo en los objetivos del negocio se realiza un análisis de escenarios, donde se comparan las utilidades obtenidas en cada uno de éstos. Se consideran cinco escenarios:

- Escenario Actual: No ofrecerle seguro a prácticamente ningún individuo dentro del segmento, es decir, cobrar primas que los clientes no estén dispuestos a aceptar.

- Escenario No Diferenciado: Ofrecerle seguro a todos los clientes pertenecientes al segmento. Para esto se asumió una prima asequible y plana para todos, la prima base.

- Escenario Blando: Si bien existe diferenciación según el riesgo de los clientes, se ofrecen tarifas asequibles (incluyendo descuentos) para muchos de ellos.

- Escenario Exigente: También existe diferenciación, pero en general se ofrecen tarifas elevadas para muchos individuos. En este escenario no se ofrecen descuentos sobre la prima base para ningún grupo de riesgo.

- Escenario Propuesto: Se propone un escenario donde existe un mayor grado de diferenciación entre los grupos de riesgo establecidos. De esta forma se logra capturar de mejor manera a aquellos clientes menos riesgosos, y evitar a los con una mayor probabilidad de siniestro.

Para la construcción de los últimos tres escenarios se realizan distintas pruebas, donde se implementan diferentes puntos de corte según la probabilidad de siniestro, y luego, a cada uno de los grupos generados se les asigna una política de precio. De esta forma, a aquellas personas dentro de los grupos más riesgosos se les asigna un recargo sobre la prima base, y a aquellas menos riesgosas se les asigna un descuento, o bien, la misma prima base.

Para escoger el monto del recargo o descuento se consultó con expertos de la compañía, específicamente con el área de Tarificación de la empresa.

Para realizar la evaluación de los distintos umbrales y sus políticas de precio asociadas, se utilizó la misma base de datos con la que se desarrolló el modelo de regresión logística, es decir, con 22.756 instancias.

Seguidamente, se comparan las utilidades obtenidas en el año 2014 utilizando la política actual, con las que hubiesen tenido de haber utilizado las políticas de precio en cada uno de los escenarios. Es importante mencionar que el escenario actual también es ficticio, debido a que los clientes con que se realiza

el análisis pertenecen a otra compañía[19], ya que Penta prácticamente no tiene clientes pertenecientes al segmento joven.

Para la creación de los escenarios se siguen los siguientes pasos:

1. Se crean cuantiles, donde el mayor cuantil representa los casos más riesgosos, es decir, con mayor probabilidad de siniestro.

2. A cada cuantil se le asigna una política de precio (nivel de recargo, nivel de descuento o mantención de la prima base).

3. Se calculan las utilidades tanto para el caso real como para los escenarios. Para el cálculo de los ingresos se considera que la prima base de la compañía es de 15 UF[20], y para los costos se considera el monto real de cada individuo en el año 2014.

Es importante señalar que las políticas de precio propuestas, es decir, el nivel de recargo o descuento, varía entre 0,85 (15% de descuento) y 2,4 (140% de recargo)[21].

También es importante mencionar que estos escenarios están sujetos a ciertos supuestos. Primero, se considera que el individuo contrata el seguro ofrecido sólo si la tarifa no supera en más de 1,25 veces la prima base, es decir, los clientes toleran un máximo de 25% de recargo. Este supuesto intenta hacer más real el análisis, ya que la decisión de contratar un seguro o no, depende en gran medida de la prima establecida. Además, se intenta capturar en alguna medida el efecto de la competencia, el cual es muy importante en este tipo de negocios, ya que al ser los seguros de automóviles muy similares en todas las compañías, el precio es uno de los principales factores al momento de seleccionar entre las ofertas del mercado. De todas formas, se considera que si se ofrece una tarifa a un precio menor a 1,25 veces la prima base, el cliente siempre contrata el seguro; por lo tanto, no se está considerando el efecto de la competencia en este caso.

De esta forma de crean los cinco escenarios mencionados anteriormente, cuyas utilidades se muestran en la Figura 27.

[19] Seguros Falabella.
[20] Se considera una prima constante ya que no se cuenta con los datos de la prima base real asociada a cada individuo.
[21] Actualmente se las políticas de precio varían entre 1 y 2,4 veces la prima base.

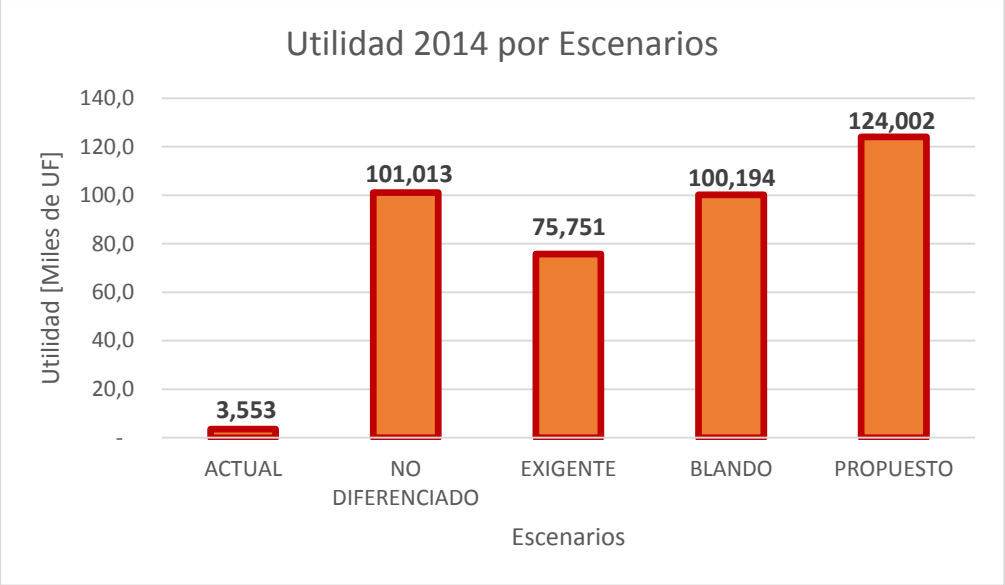

Figura 27. Comparación de Utilidad 2014 por Escenarios.

Se puede observar que la utilidad obtenida en el escenario actual es mucho menor que en los demás escenarios, lo que se debe a que con las primas impuestas actualmente sólo se está capturando clientes con 34 y 35 años de edad (son los únicos que luego del proceso de tarificación tienen primas menores a 1,25 veces la prima base). Esto se debe a que las tarifas impuestas para aquellos individuos entre 18 y 33 años varían entre 1,7 y 2,4 veces la prima base.

El escenario No Diferenciado, es decir, aquel en que se le cobra la prima base a todos los individuos obtiene 101 mil UF de utilidad. Este valor es bastante alto comparado con el escenario actual. Sin embargo, hay que recordar que bajo los supuestos establecidos, al ofrecer la prima base, todos los clientes contratan el seguro automotriz. Además, debido a que en la data utilizada la mayoría de los clientes no presentan siniestros, se obtienen altas ganancias de aquellos individuos que no presentan costo de siniestro para la compañía.

Como se mencionó anteriormente, para crear los tres últimos escenarios se implementan distintos puntos de corte según la probabilidad de siniestro, y luego se asignan diferentes políticas de precio basadas en estos umbrales. En la Figura 28 se puede observar la distribución de los clientes en cada política de precio de estos escenarios.

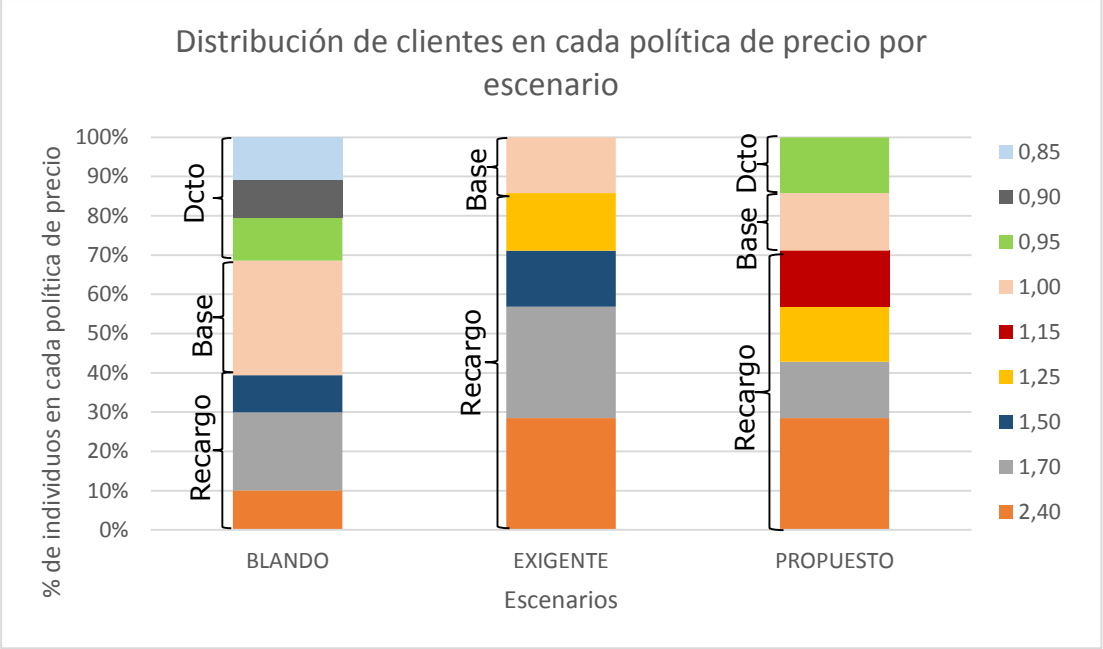

Figura 28. Distribución de clientes en cada política de precio por escenario.

En el escenario Exigente se ofrecen políticas muy estrictas, en donde se les cobra un alto recargo a aquellas personas con alta probabilidad de siniestro. Estos recargos varían entre 1,25 y 2,4 veces la prima base. Además, este escenario no considera descuentos para ningún cliente, aunque sí ofrece a algunos la base de 15 UF. Se puede observar que al imponer primas tan estrictas la utilidad disminuye considerablemente con respecto al escenario anterior (25 mil UF de diferencia aproximadamente). Esto se debe a que al definir estas altas tarifas se pierde una gran cantidad de clientes, ya que éstos optan por ofertas de la competencia.

El escenario Blando genera utilidades similares al No Diferenciado, pero en este caso se utilizan distintas políticas de precio según el riesgo de los distintos individuos. Además, este es el escenario que ofrece menos cantidad de primas con recargos (excluyendo el No Diferenciado). Se puede observar que al crear tarifas diferenciadas se logra capturar aquellos clientes con una menor probabilidad de siniestro y evitar a los con una mayor probabilidad, ya que se obtienen utilidades similares al escenario No Diferenciado, pero imponiendo diferencias en las tarifas.

Para crear el escenario Propuesto se probaron distintas combinaciones de corte de probabilidad y política de precio, y luego se escogió la que se muestra en la Figura 28 debido a que ésta es la que entregaba mejores resultados con un mayor grado de diferenciación. Se puede observar que en este escenario se logra una mejor distribución de las tarifas ofrecidas, con lo que se generan mayores utilidades para la compañía, cerca de 24 mil UF más que en el escenario Blando (y en el No Diferenciado), y 50 mil UF más que en el Exigente.

Esto se logra gracias a la alta diferenciación en las primas ofrecidas y a la distribución de éstas dentro del grupo de clientes con que se está trabajando. De este modo, sería posible encontrar combinaciones que generaran mayores utilidades para la empresa, lo que se convierte en un problema de optimización, el cual está fuera del alcance del proyecto.

Luego de obtener las políticas propuestas para los diferentes puntos de corte de la probabilidad de siniestro, se analiza cómo se distribuyen actualmente las políticas de precio según los distintos grupos de riesgo creados.

En la Figura 29 se muestra dicha distribución[22]. Se puede observar que actualmente, en todos los grupos, la política más utilizada es imponer un 70% de recargo. Además, a muy pocos clientes se les ofrece la prima base o un recargo de tan sólo el 10%, sólo a aquellos clientes entre 34 y 35 años de edad. Además se puede observar que la cantidad de individuos a los que se le ofrece una prima sobre 2 veces la tarifa base aumenta de 4.496 a 6.479, por lo que en la política propuesta se está siendo más exigente con una mayor cantidad de clientes altamente riesgosos.

Las tarifas ofrecidas con un 70% de recargo disminuyen considerablemente, ya que actualmente esta es la mínima prima que se le cobra a alguien menor de 33 años. En este caso se disminuye de 16.706 a 3.264 personas. De este modo, se está siendo menos severo con los individuos medianamente riesgosos, ya que muchos de los que hoy en día se encuentran en este rango de precio, se propone ofrecerles menores tarifas, como un 25% o 15% de recargo, por ejemplo.

Otro punto importante a destacar es la definición de tarifas con descuento dentro del grupo joven, que actualmente no existen. Se puede observar que dentro de este grupo no existe una gran cantidad de individuos con prima actual sobre 1,87, sólo 497 personas. De esta forma, se observa que este grupo también es catalogado como uno menos riesgoso en la actualidad.

[22] En el Anexo L se muestra esta distribución en forma más detallada.

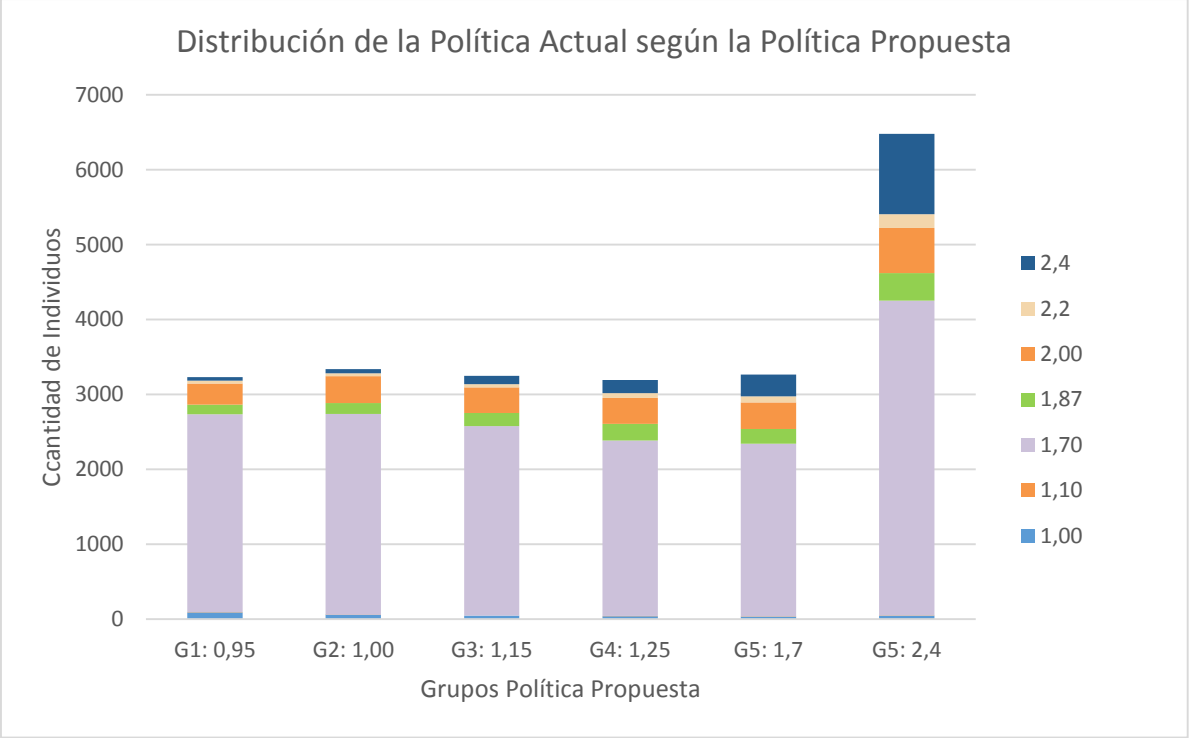

Figura 29. Distribución de la Política Actual según la Política Propuesta.

En la Tabla 31 se muestra la prima promedio actual y la propuesta según cada grupo de riesgo creado. Se puede observar que las tarifas actuales son muy similares entre sí, y muy cercanas al 70% de recargo (política más utilizada hoy en día). De todas formas, las tarifas actuales aumentan según los grupos de riesgo definidos, donde ya en los grupos 5 y 6, ésta supera el 80% de recargo en promedio.

Grupo	Prima Promedio Actual	Prima Propuesta
Grupo 1	1.72	0.95
Grupo 2	1.74	1.00
Grupo 3	1.76	1.15
Grupo 4	1.78	1.25
Grupo 5	1.81	1.70
Grupo 6	1.86	2.40

Tabla 31. Comparación de Políticas Actual y Propuesta, por Grupo de Riesgo.

Como en el escenario actual prácticamente no genera ingresos para la compañía proveniente del segmento joven, y según las utilidades calculadas para el escenario propuesto, se justifica ofrecer descuento a un grupo de individuos, mientras se aseguren de evitar a los clientes más riesgosos. Por este motivo, se aumenta la cantidad de clientes con una prima sobre 2 veces la prima base.

7. Conclusiones

7.1 Conclusiones del trabajo realizado

El objetivo inicial del presente proyecto fue estimar el nivel de riesgo de los potenciales clientes jóvenes en el negocio de seguros de automóviles, para de esta forma poder ofrecerles una prima acorde a éste, y poder así también capturar a aquellos clientes que presentan un menor nivel de riesgo de siniestro.

Para hacer esto se pretendía estimar dos factores, la probabilidad de siniestro y el costo medio de siniestros de cada individuo y, con esto poder generar un puntaje que lograra medir la gravedad de los siniestros, además de la propensión a éstos. Sin embargo, los resultados entregados por la estimación del costo medio fueron bastante deficientes, a tal punto que el considerar este factor disminuía el valor entregado por el modelo de estimación de la probabilidad. Debido a esto se decidió no considerar el segundo factor en la implementación del modelo final.

Es importante mencionar que para la estimación de la probabilidad de siniestro de cada individuo se utilizaron dos modelos, una regresión logística y un árbol de decisión C5.0. Los resultados de ambos métodos fueron bastante similares, sin embargo, el primer modelo superó levemente al segundo al entregar una mejor curva de ganancias y una mayor área bajo la curva (AUC). Por otro lado, se prefiere la regresión logística, ya que es un método sencillo de entender e implementar en la compañía; y que además, al ser un modelo paramétrico se logra un mejor entendimiento de los coeficientes dentro de éste, y por lo tanto de su efecto en el resultado.

En cuanto a los resultados entregados por la regresión logística, se obtuvo mejoras en la curva de ganancias con respecto al modelo actual de la compañía, por lo que el trabajo realizado entrega valor adicional al estado presente de la empresa.

Un punto importante dentro del proyecto fue la utilización de variables relacionadas al rendimiento académico de los individuos al momento de dejar su etapa escolar. El modelo realizado demostró que dichos atributos parecen tener influencia en el nivel de riesgo de las personas; sin embargo, este efecto no es exactamente el esperado, ya que en un principio se pensaba que estas variables eran un indicador del grado de responsabilidad y buen comportamiento de conducción de los individuos, pero los resultados del modelo muestran que esto no es necesariamente cierto. De hecho, las variables de rendimiento escolar indican que a mayor puntaje en la PSU (especialmente en la de Matemáticas) existe una mayor propensión a incurrir en un accidente automovilístico. Esto se debe probablemente a la alta relación

entre los puntajes de la prueba y el nivel socioeconómico de los individuos, quienes, según los resultados del estudio, tienen una mayor probabilidad de siniestro. Aunque se intentó compensar este efecto socioeconómico en las variables PSU, al estandarizarlas según el puntaje promedio de la comuna de cada persona, no se logró vislumbrar ninguna relación que muestre que la responsabilidad de los individuos conlleve a cierto comportamiento de conducción. Esto se puede deber a que no existe tal relación, o bien, no se logró retirar el efecto de los grupos socioeconómicos en los atributos utilizados.

De todas formas, al realizar un modelo de regresión logística sin considerar las variables PSU se obtienen resultados muy similares en cuanto al desempeño de los modelos, por lo que se decide no utilizar dichos atributos en el desarrollo del modelo final de la compañía. Esto ya que el aporte de éstas variables no es mucho, y el adquirir estos datos año a año supone un costo para a empresa.

Dentro de las variables utilizadas las más importantes son las relacionadas al comportamiento de conducción de los clientes, donde el siniestro medio (o siniestros por año) es el atributo más relevante de todos. Luego, las demás variables que influyen en la propensión a incurrir en un accidente, están relacionadas a las características del vehículo, donde la antigüedad y la tasación de éste son las más importantes. La cantidad de vehículos que posee el individuo también influye en la probabilidad de siniestro, donde alguien que posee tres o más vehículos presenta una mayor propensión a incurrir en un accidente.

Luego de desarrollar el modelo y obtener las probabilidades de siniestro de cada individuo, se crearon distintos escenarios y se calculó la utilidad que generaría cada uno de éstos para la compañía. Específicamente, se compararon cinco escenario: el escenario actual, uno no diferenciado, uno diferenciado pero con políticas blandas, uno con políticas exigentes, y el propuesto (el cual presenta un mayor grado de diferenciación y una mejor distribución de las políticas).

El escenario propuesto define 6 grupos de riesgo en base a distintos umbrales de corte de la probabilidad de siniestro, donde a cada uno de ellos se le asigna una política (un nivel de recargo, uno de descuento, o bien, la prima base). Este escenario logra generar 124 mil UF al año[23], es decir, 120 mil UF más que las actuales; esto se debe a que hoy en día la compañía no asegura a prácticamente ningún individuo del segmento joven. Por otro lado, este escenario genera 24 mil UF más que en el escenario blando, y cerca de 50 mil UF adicionales al escenario exigente.

Luego, se analizaron las distribuciones de las primas actuales según las propuestas, donde se puede observar que, a pesar de que actualmente se

[23] Bajo ciertos supuestos.

definen tarifas muy altas y homogéneas para todos los clientes, existe una relación entre los grupos de riesgo creados y las primas actuales. Es decir, los grupos menos riesgosos presentan una menor prima, mientras que los más riesgosos presentan una mayor.

Luego de este análisis se propone ofrecer descuentos al primer grupo de riesgo, ofrecer la prima base al segundo, y diferenciar a los individuos riesgosos en cuatro grupos con recargo (entre un 15% a un 140% de recargo). De esta forma, se lograría aumentar la cantidad de clientes en los primeros grupos y disminuir o evitar a los clientes más riesgosos mediante la definición de altas primas.

7.2 Limitaciones del trabajo

El trabajo presenta distintas limitantes. Primero, las variables disponibles a utilizar en el modelo no son muy representativas del problema a resolver, ya que los atributos más explicativos de todos son los relacionados al historial de conducción. Este hecho influye negativamente al momento de estimar el riesgo de los individuos jóvenes, ya que en general quienes pertenecen a este segmento no poseen mucho historial de conducción. En consecuencia, para aquellas personas se debe basar la predicción en características que no tienen un alto grado de asociación con la variable objetivo.

Una segunda limitación del trabajo está relacionada a la calidad de los datos que tiene actualmente la empresa, ya que existen muchos atributos con una gran cantidad de datos faltantes o erróneos. El principal ejemplo de esto es la comuna, la que a pesar de que se intentó limpiar lo mejor posible, continuó teniendo muchos datos sin información.

Por último, otra limitante del trabajo es el nivel de sesgo de la base de datos, que si bien se intentó reducir en la mayor medida posible, sigue existiendo en los datos finalmente utilizados para construir el modelo de predicción. Esto se debe a que no se tiene información de la tenencia (o no tenencia) de seguro de los individuos en cada uno de los años utilizados en el trabajo, ni tampoco de los accidentes que no son declarados por los clientes. Esto podría conducir a considerar casos que realmente tuvieron siniestros como un caso de no siniestro.

7.3 Recomendaciones y trabajos futuros

Como primera recomendación se sugiere la incorporación de nuevas variables a sus sistemas de información. Desde el punto de vista de los segmentos jóvenes podría ser útil contar con información relacionada a la carrera que estudia la persona (o estudió), el número de años que dura dicha carrera, para de este modo poder conocer el estado laboral en que se encuentra actualmente

la persona, o si tiene alguna profesión o no. Los datos relacionados al estado laboral también podrían ser de ayuda, así como también el estado civil. Otro atributo que se cree que sería de gran aporte es el grupo socioeconómico asociado a cada individuo y su ingreso monetario estimado, de esta forma, quizás se podría calcular mejores variables que eliminaran el efecto de estos estratos en las variables de rendimiento académico. Una última idea de información a incorporar son las variables relacionadas a las multas de tránsito de cada individuo, tanto la cantidad de multas, como el monto de éstas, e idealmente una clasificación que pueda dar una idea de cuál fue el motivo de dicha penalización.

La segunda recomendación también está ligada a la información que tiene la compañía actualmente, ya que como se menciona en el punto anterior, la calidad de la información actual es deficiente. De esta forma, sería recomendable tener un sistema de depuración y actualización de los mismos, constante en el tiempo. Con estos dos conjuntos de mejoras en las variables, sería posible encontrar algoritmos de predicción del riesgo de cada persona que se ajusten aún más a la realidad.

También, con respecto a los resultados obtenidos del trabajo, se sugiere implementar distintas políticas de precios a las actuales, aplicando un modelo como el indicado en este trabajo, principalmente porque hoy en día, para los clientes con bajo nivel de riesgo, no se está ofreciendo ningún tipo de incentivo a escoger la oferta de la compañía frente a otras. También podría ser beneficioso para la empresa penalizar al segmento de individuos que, según los resultados obtenidos, representan un riesgo mayor.

Por último, se recomienda como trabajo futuro desarrollar un proyecto de optimización del punto de corte de la probabilidad de siniestro, para de esta forma encontrar los umbrales y las políticas de precio que logran la mayor rentabilidad para la compañía.

8. Bibliografía

[1] Méndez,M. (2013). "Desarrollo de un modelo de recomendación de compra para clientes de una empresa de seguros", memoria para optar al título de ingeniero civil industrial, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, Santiago de Chile, marzo 2013.

[2] Pereira, N. (2014). "Identificación de clientes con patrones de consumo eléctrico fraudulento", memoria para optar al título de ingeniero civil industrial, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, Santiago de Chile, agosto 2014.

[3] Maldonado, S. (2007). "Utilización de support vector machines no lineal y selección de atributos para credit scoring", memoria para optar al título de ingeniero civil industrial, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, Santiago de Chile, julio 2007.

[4] Dionne, G., & Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. Astin Bulletin, 19(2), 199-212.

[5] Lindskog, F., & McNeil, A. J. (2003). Common Poisson shock models: applications to insurance and credit risk modelling. Astin Bulletin, 33(2), 209-238.

[6] Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847-856.

[7] West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27(11), 1131-1152.

[8] Dolgui, A., & Proth, J. M. (2010). Pricing strategies and models. Annual Reviews in Control, 34(1), 101-110.

[9] Ridout, M., Hinde, J., & DemeAtrio, C. G. (2001). A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics*, *57*(1), 219-223.

[10] Mitchell, T. (1997). Machine Learning, *Chapter 3: Decision Tree Learning*. Ed. McGraw Hill.

[11] Rulequest Research (2012). "Is See5/C5.0 Better Than C4.5?" [Disponible en] http://rulequest.com/see5-comparison.html.

[12] Biggs, D., De Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics, 18(1), 49-62.

[13] Rokach, L. (2007). Data mining with decision trees: theory and applications. World scientific.

[14] Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine, 16(9), 965-980.

[15] Balakrishnan, N. (2013). Handbook of the logistic distribution. CRC Press.

[16] Hosmer Jr, D. W., & Lemeshow, S. (2004). Applied logistic regression. John Wiley & Sons.

[17] Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*(1), 35-46.

[18] Wenzel, T. P., & Ross, M. (2005). The effects of vehicle model and driver behavior on risk. *Accident Analysis & Prevention*, *37*(3), 479-494.

[19] Evans, L., & Wasielewski, P. (1983). Risky driving related to driver and vehicle characteristics. *Accident Analysis & Prevention*, *15*(2), 121-136.

[20] D. Riano. "Árboles de Decisión ID3-C4.5" [Disponible en] http://banzai-deim.urv.net/~riano/teaching/id3-m5.pdf

[21] Demre, "Compendio Estadístico Proceso de Admisión Año Académico 2014" [Disponible en] http://www.demre.cl/text/pdf/p2014/compendio_p2014.pdf

[22] Nacional, "Municipales superan hasta en 61 puntos a subvencionados en el puntaje promedio PSU". [Disponible en] http://papeldigital.info/lt/2014/12/30/01/paginas/018.pdf

[23] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on, 21(9), 1263-1284.

9. Anexos

Anexo A: Variables Iniciales en las Bases de Datos

Variables en cada Base de Datos		
RNVM	**SISGEN**	**PSU**
RUT	RUT	RUT
DV	DV	DV
Nombre	Cantidad Siniestros 0-12 meses	Año PSU
Género	Cantidad Siniestros 12-24 meses	NEM
Año de Nacimiento	Cantidad Siniestros 24-36 meses	PSU Lenguaje
Comuna	Cantidad Siniestros 36-48 meses	PSU Matemáticas
Ciudad	Monto Siniestros 0-12 meses	PSU Historia y Cs. Sociales
Región	Monto Siniestros 12-24 meses	PSU Ciencias
Patente	Monto Siniestros 24-36 meses	
Tipo de Vehículo	Monto Siniestros 36-48 meses	
Marca del Vehículo		
Modelo		
Año del Vehículo		
Color del Vehículo		
Tasación del Vehículo		
Total Vehículos		

Tabla 32. Variables Iniciales en las distintas Bases de Datos.

Anexo B: Análisis de Probabilidades Condicionales

Género	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
M	11091	3034	14125	0.21479646	1.003677258	0.003670514	0.003670514	62%
F	6795	1836	8631	0.212721585	0.99398201	-0.006036171	0.006036171	38%
Total General	17886	4870	22756	0.214009492	1			**0.5%**

Tabla 33. Análisis de Probabilidades Condicionales – Variable Género.

Edad	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[18,25]	4554	1332	5886	0.226299694	1.057428304	0.055839832	0.055839832	26%
[26,30]	11963	3209	15172	0.211508041	0.988311496	-0.011757352	0.011757352	67%
[31,35]	1369	329	1698	0.193757362	0.905368074	-0.099413706	0.099413706	7%
Total General	17886	4870	22756	0.214009492	1			**3%**

Tabla 34. Análisis de Probabilidades Condicionales – Variable Edad.

Sector	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
Norte	416	101	517	0.195357834	0.912846584	-0.091187448	0.091187448	2%
Centro - Grupo 1	1457	335	1792	0.186941964	0.873521836	-0.135222152	0.135222152	8%
Centro - Grupo 2	1621	427	2048	0.208496094	0.974237599	-0.026100063	0.026100063	9%
Centro - Grupo 3	550	173	723	0.239280775	1.118084868	0.111617282	0.111617282	3%
Centro - Grupo 4	518	190	708	0.268361582	1.253970464	0.226314888	0.226314888	3%
Centro - Grupo 5	1424	422	1846	0.228602384	1.068188057	0.065963809	0.065963809	8%
Sur	1274	284	1558	0.182284981	0.851761195	-0.160449078	0.160449078	7%
Sin Sector	10626	2938	13564	0.216602772	1.012117594	0.012044763	0.012044763	60%
Total General	17886	4870	22756	0.214009492	1			**5%**

Tabla 35. Análisis de Probabilidades Condicionales – Variable Sector.

Años Vehículo	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[1,3]	7030	2439	9469	0.257577358	1.203579128	0.185299724	0.185299724	42%
[4,7]	5489	1460	6949	0.210102173	0.98174231	-0.018426419	0.018426419	31%
[8,75]	5367	971	6338	0.153202903	0.715869664	-0.334257162	0.334257162	28%
Total General	17886	4870	22756	0.214009492	1			**18%**

Tabla 36. Análisis de Probabilidades Condicionales – Variable Años del Vehículo.

Tipo Vehículo	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
Automóvil	13960	3806	17766	0.214229427	1.001027688	0.00102716	0.00102716	78%
Motocicleta	1109	298	1407	0.211798152	0.989667094	-0.010386661	0.010386661	6%
Camioneta	190	35	225	0.155555556	0.726862879	-0.319017431	0.319017431	1%
S. Wagon	2148	609	2757	0.220892274	1.032161107	0.031654766	0.031654766	12%
Otros	479	122	601	0.202995008	0.948532733	-0.05283898	0.05283898	3%
Total General	17886	4870	22756	0.214009492	1			**1%**

Tabla 37. Análisis de Probabilidades Condicionales – Variable Tipo de Vehículo.

Marca	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
Precio alto	449	201	650	0.309230769	1.444939504	0.368067455	0.368067455	3%
Precio medio	3309	902	4211	0.214200902	1.000894401	0.000894002	0.000894002	19%
Precio bajo	10960	2836	13796	0.205566831	0.960550063	-0.040249177	0.040249177	61%
Otras marcas	3168	931	4099	0.227128568	1.061301374	0.059495867	0.059495867	18%
Total General	17886	4870	22756	0.214009492	1			**5%**

Tabla 38. Análisis de Probabilidades Condicionales – Variable Marca del Vehículo.

Ítems	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
1	12343	3220	15563	0.206900983	0.966784142	-0.033780033	0.033780033	68%
2	4069	1128	5197	0.217048297	1.014199394	0.014099527	0.014099527	23%
3 o más	1474	522	1996	0.261523046	1.222016106	0.200502041	0.200502041	9%
Total General	17886	4870	22756	0.214009492	1			**4%**

Tabla 39. Análisis de Probabilidades Condicionales – Variable Ítems (o Cantidad de Vehículos).

NEM	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[236, 458)	3433	845	4278	0.197522207	0.922960028	-0.080169352	0.080169352	19%
[458, 538)	3766	984	4750	0.207157895	0.96798461	-0.03253909	0.03253909	21%
[538, 599)	3475	902	4377	0.206077222	0.962934961	-0.037769407	0.037769407	19%
[599, 661)	3188	849	4037	0.210304682	0.98268857	-0.017463025	0.017463025	18%
[661, 826)	4024	1290	5314	0.242754987	1.134318784	0.126032281	0.126032281	23%
Total General	17886	4870	22756	0.214009492	1			**6.2%**

Tabla 40. Análisis de Probabilidades Condicionales – Variable Puntaje NEM.

Lenguaje	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[159, 418)	3651	873	4524	0.192970822	0.90169282	-0.103481372	0.103481372	20%
[418, 492)	3621	819	4440	0.184459459	0.86192186	-0.148590662	0.148590662	20%
[492, 561)	3635	961	4596	0.209094865	0.977035472	-0.02323232	0.02323232	20%
[561, 636)	3584	1057	4641	0.22775264	1.064217467	0.062239756	0.062239756	20%
[636, 850)	3395	1160	4555	0.254665203	1.189971532	0.173929384	0.173929384	20%
Total General	17886	4870	22756	0.214009492	1			**10.2%**

Tabla 41. Análisis de Probabilidades Condicionales – Variable Puntaje PSU Lenguaje.

Matemáticas	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[163, 441)	3732	814	4546	0.179058513	0.836684912	-0.178307729	0.178307729	20%
[441, 509)	3705	845	4550	0.185714286	0.867785274	-0.141810975	0.141810975	20%
[509, 570)	3570	981	4551	0.21555702	1.007231121	0.007205101	0.007205101	20%
[570, 640)	3412	1009	4421	0.228228907	1.06644292	0.064328736	0.064328736	19%
[640, 850]	3467	1221	4688	0.260452218	1.21701246	0.196399053	0.196399053	21%
Total General	17886	4870	22756	0.214009492	1			**11.8%**

Tabla 42. Análisis de Probabilidades Condicionales – Variable Puntaje PSU Matemáticas.

NEM Estand	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[0.44, 0.83)	3639	898	4537	0.197928146	0.924856858	-0.078116302	0.078116302	20%
[0.83, 0.95)	3594	943	4537	0.207846595	0.971202692	-0.029220087	0.029220087	20%
[0.95, 1.06)	3586	924	4510	0.204878049	0.957331597	-0.043605451	0.043605451	20%
[1.06, 1.18)	3534	1009	4543	0.222099934	1.037804127	0.037107064	0.037107064	20%
[1.18, 1.52]	3533	1096	4629	0.2367682	1.106344388	0.101061236	0.101061236	20%
Total General	17886	4870	22756	0.214009492	1			**5.8%**

Tabla 43. Análisis de Probabilidades Condicionales – Variable Puntaje NEM Estandarizado.

Leng Estand	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[0.30, 0.80)	3638	885	4523	0.195666593	0.91428932	-0.089608215	0.089608215	20%
[0.80, 0.94)	3687	876	4563	0.191978961	0.89705816	-0.10863458	0.10863458	20%
[0.94, 1.06)	3621	929	4550	0.204175824	0.954050319	-0.047038863	0.047038863	20%
[1.06, 1.20)	3508	1033	4541	0.227482933	1.062957213	0.061054848	0.061054848	20%
[1.20, 1.69]	3432	1147	4579	0.250491374	1.170468521	0.157404114	0.157404114	20%
Total General	17886	4870	22756	0.214009492	1			**9.3%**

Tabla 44. Análisis de Probabilidades Condicionales – Variable Puntaje PSU Lenguaje Estandarizado.

Mat Estand	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[0.33, 0.82)	3722	829	4551	0.182157768	0.851166768	-0.161147203	0.161147203	20%
[0.82, 0.94)	3678	858	4536	0.189153439	0.883855372	-0.123461836	0.123461836	20%
[0.94, 1.05)	3578	965	4543	0.212414704	0.99254805	-0.007479855	0.007479855	20%
[1.05, 1.18)	3535	1017	4552	0.223418278	1.043964338	0.04302533	0.04302533	20%
[1.18, 1.68]	3373	1201	4574	0.262571054	1.226913121	0.204501357	0.204501357	20%
Total General	17886	4870	22756	0.214009492	1			**10.8%**

Tabla 45. Análisis de Probabilidades Condicionales – Variable Puntaje PSU Matemáticas Estandarizado.

NEM sup	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
0	9206	2365	11571	0.204390286	0.955052433	-0.045989036	0.045989036	51%
1	8680	2505	11185	0.223960662	1.04649873	0.045450049	0.045450049	49%
Total General	17886	4870	22756	0.214009492	1			**5%**

Tabla 46. Análisis de Probabilidades Condicionales – Variable NEM Sup.

Leng sup	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
0	8948	2152	11100	0.193873874	0.9059125	-0.098812556	0.098812556	49%
1	8938	2718	11656	0.233184626	1.089599455	0.085810157	0.085810157	51%
Total General	17886	4870	22756	0.214009492	1			**9%**

Tabla 47. Análisis de Probabilidades Condicionales – Variable Lenguaje Sup.

Mat sup	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
0	9101	2126	11227	0.189364924	0.884843574	-0.122344402	0.122344402	49%
1	8785	2744	11529	0.2380085	1.112139925	0.106286019	0.106286019	51%
Total General	17886	4870	22756	0.214009492	1			**11%**

Tabla 48. Análisis de Probabilidades Condicionales – Variable Matemáticas Sup.

Cant PSU	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
1	15085	4102	19187	0.213790587	0.998977127	-0.001023397	0.001023397	84%
Más de 1	2801	768	3569	0.215186327	1.005498984	0.005483919	0.005483919	16%
Total General	17886	4870	22756	0.214009492	1			**0.2%**

Tabla 49. Análisis de Probabilidades Condicionales – Variable Cantidad de PSU rendidas.

Sin por año	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
[0, 0.33)	10887	2109	12996	0.162280702	0.758287402	-0.276692807	0.276692807	57%
[0.33, 0.66)	4032	1374	5406	0.254162042	1.187620417	0.171951656	0.171951656	24%
[0.66, 1)	1795	744	2539	0.293028751	1.369232499	0.314250363	0.314250363	11%
[1, 2.33]	1172	643	1815	0.354269972	1.655393736	0.504038887	0.504038887	8%
Total General	17886	4870	22756	0.214009492	1	0	0	**27%**

Tabla 50. Análisis de Probabilidades Condicionales – Variable Siniestros por año.

Costo Medio	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
0	10939	2132	13071	0.163109173	0.762158591	-0.27160062	0.27160062	57%
(0, 5]	948	344	1292	0.26625387	1.244121779	0.218429883	0.218429883	6%
(5, 22]	2928	1183	4111	0.287764534	1.344634443	0.296122187	0.296122187	18%
(22, 50]	1606	641	2247	0.285269248	1.332974744	0.287413095	0.287413095	10%
(50, 365]	1465	570	2035	0.28009828	1.308812415	0.269120173	0.269120173	9%
Total General	17886	4870	22756	0.214009492	1	0	0	**27%**

Tabla 51. Análisis de Probabilidades Condicionales – Variable Costo Medio de Siniestros.

Anos 1° Sin	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
0	10887	2109	12996	0.162280702	0.758287402	-0.276692807	0.276692807	57%
[1,2]	2652	1042	3694	0.282079047	1.318067925	0.276166971	0.276166971	16%
[3,5]	2559	902	3461	0.260618318	1.217788594	0.197036586	0.197036586	15%
[6,15]	1788	817	2605	0.313627639	1.465484714	0.38218605	0.38218605	11%
Total General	17886	4870	22756	0.214009492	1	0	0	**28%**

Tabla 52. Análisis de Probabilidades Condicionales – Años desde el primer siniestro.

Tasación [MM]	0	1	Total Casos	Pr(sin=1/total)	Ratio	ln(Ratio)	|ln(Ratio)|	Pr(total casos)
<2.7	3844	704	4548	0.154793316	0.723301169	-0.323929589	0.323929589	20%
[2.7, 3.8)	3755	777	4532	0.171447485	0.801120936	-0.221743362	0.221743362	20%
[3.8, 4.9)	3616	955	4571	0.208925837	0.976245655	-0.024041028	0.024041028	20%
[4.9, 6.8)	3444	1096	4540	0.241409692	1.128032637	0.120475086	0.120475086	20%
>6.8	3227	1338	4565	0.293099671	1.369563886	0.314492357	0.314492357	20%
Total General	17886	4870	22756	0.214009492	1			**20%**

Tabla 53. Análisis de Probabilidades Condicionales – Tasación del Vehículo (en millones de pesos).

Anexo C: Test de Comparación de Medias y Test de Proporciones

Test de Comparación de Medias

Un Test de comparación de medias sirve para comprobar si los valores de una característica que es posible cuantificar difieren al agruparlas en dos o más grupos (en este caso, si presenta o no presenta siniestros). Una de las pruebas más utilizadas para comparar medias, y a la que se recurre en este estudio, es la prueba t de Student para datos independientes.

Antes de realizar la prueba de comparación de medias es necesario comprobar que las varianzas de ambos grupos sean distintas. Para esto se utiliza el test de la razón de varianzas o test de Levene. Bajo el supuesto de que ambas siguen una distribución normal y tienen igual varianza ($H_o: \sigma_1 = \sigma_2$) se espera que la razón de varianzas siga una distribución F de Snedecor con parámetros $(n-1)$ y $(m-1)$:

$$F = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2} = \frac{\widehat{S_1^2}}{\widehat{S_2^2}}$$

Si el p-valor es menor a un α (generalmente 0,05) se rechaza la hipótesis nula, por lo tanto la varianza no es homogénea.

Luego, la hipótesis nula para la comparación de medias es $H_o: \mu_1 = \mu_2$. El estadístico a utilizar es:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n-1)\widehat{S_1^2} + (m-1)\widehat{S_2^2}}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Del mismo modo, si el p-valor es menor a α se rechaza la hipótesis nula, y por lo tanto las medias de ambos grupos son distintas.

Test de Proporciones

Para comparar una respuesta que se mide como una proporción entre dos o más niveles, se utilizan pruebas que indican si hay diferencias entre estas proporciones, es decir, si se distribuyen homogéneamente entre los niveles de la variable. En el caso de comparar una variable ordinal o nominal en función de dos categorías, se busca comparar una variable de dos categorías con otra de dos categorías.

La prueba χ^2 es una de las más frecuentemente utilizadas para el contraste de variables cualitativas. Para su cálculo se computan las frecuencias esperadas para compararlas con las observadas en la realidad. Se calcula el valor del estadístico χ^2 como:

$$\chi^2 = \sum \frac{\left|O_{ij} - E_{ij}\right|^2}{E_{ij}} \sim \chi^2_{(f-1)\cdot(c-1)}$$

Donde O_{ij} corresponden a las frecuencias observadas dentro de la casilla de la fila i y columna j, E_{ij} corresponden a las frecuencias esperadas o teóricas, f es el número de filas y c el número de columnas.

De esta forma, si se obtiene un p-valor menor a α se rechaza la hipótesis nula de igualdad de proporciones.

Anexo D: Test de Multicolinealidad

La multicolinealidad surge cuando en un modelo de regresión se incluyen variables explicativas con una alta correlación entre ellas, o algunas de ellas. El problema de la multicolinealidad hace referencia, en concreto, a la existencia de relaciones aproximadamente lineales entre los regresores del modelo.

Para analizar la existencia de multicolinealidad se realizan regresiones lineales de cada variable independiente con todas las demás. De este modo se calcula el coeficiente de determinación de cada una de estas regresiones (R^2), con lo que se puede calcular el factor de inflación de la varianza (o VIF) y la tolerancia (T), los cuales se definen como:

$$VIF_i = \frac{1}{1 - R_i^2} \qquad T_i = \frac{1}{VIF_i} = 1 - R^2$$

De este modo, se puede observar que si se tiene una tolerancia baja, cercana a 0, significa que las variables están altamente correlacionadas, por lo que es redundante utilizarlas todas en el modelo a desarrollar. Si se tiene una tolerancia cercana a 1, las variables no están fuertemente correlacionadas, y por lo tanto no existe el problema de multicolinealidad.

Una regla empírica, citada por Kleinbaum, consiste en considerar que existen problemas de colinealidad si algún VIF es superior a 10, lo que corresponde a un R^2 de 0,9 y a una tolerancia de 0,1. Además, si el problema de todos los VIF es superior a 10 también se está en presencia de multicolinealidad.

Anexo E: Extracto Matriz de Correlación

	genero	edad	anos_veh	tasacion	anossin1	nem	leng	mat	s1_bin	cm	sinporano	nem_est	mat_est	leng_est	nem_sup	mat_sup	leng_sup
genero	1	-0.001	0	0	-0.001	0	-0.002	0.004	-0.003	0.005	-0.001	0.001	0.006	0	0.004	0.013	0
edad	-0.001	1	0.03	-0.004	0.052	-0.064	-0.039	-0.082	-0.021	-0.03	-0.021	-0.054	-0.073	-0.027	-0.025	-0.047	-0.018
anos_veh	0	0.03	1	-0.499	-0.069	-0.12	-0.106	-0.089	-0.087	-0.063	-0.101	-0.108	-0.079	-0.093	-0.086	-0.08	-0.085
tasacion	0	-0.004	-0.499	1	0.162	0.114	0.125	0.146	0.109	0.109	0.154	0.104	0.131	0.11	0.085	0.119	0.1
anossin1	-0.001	0.052	-0.069	0.162	1	0.093	0.18	0.199	0.118	0.356	0.654	0.076	0.157	0.141	0.062	0.146	0.129
nem	0	-0.064	-0.12	0.114	0.093	1	0.581	0.591	0.039	0.028	0.07	0.992	0.578	0.567	0.809	0.504	0.484
leng	-0.002	-0.039	-0.106	0.125	0.18	0.581	1	0.767	0.057	0.062	0.121	0.561	0.738	0.976	0.476	0.642	0.805
mat	0.004	-0.082	-0.089	0.146	0.199	0.591	0.767	1	0.07	0.086	0.138	0.568	0.971	0.733	0.48	0.773	0.63
s1_bin	-0.003	-0.021	-0.087	0.109	0.118	0.039	0.057	0.07	1	0.065	0.155	0.036	0.065	0.051	0.024	0.059	0.048
cm	0.005	-0.03	-0.063	0.109	0.356	0.028	0.062	0.086	0.065	1	0.368	0.024	0.077	0.052	0.023	0.067	0.042
sinporano	-0.001	-0.021	-0.101	0.154	0.654	0.07	0.121	0.138	0.155	0.368	1	0.063	0.122	0.106	0.051	0.115	0.091
nem_est	0.001	-0.054	-0.108	0.104	0.076	0.992	0.561	0.568	0.036	0.024	0.063	1	0.583	0.572	0.815	0.506	0.484
mat_est	0.006	-0.073	-0.079	0.131	0.157	0.578	0.738	0.971	0.065	0.077	0.122	0.583	1	0.755	0.486	0.791	0.641
leng_est	0	-0.027	-0.093	0.11	0.141	0.567	0.976	0.733	0.051	0.052	0.106	0.572	0.755	1	0.48	0.652	0.816
nem_sup	0.004	-0.025	-0.086	0.085	0.062	0.809	0.476	0.48	0.024	0.023	0.051	0.815	0.486	0.48	1	0.445	0.421
mat_sup	0.013	-0.047	-0.08	0.119	0.146	0.504	0.642	0.773	0.059	0.067	0.115	0.506	0.791	0.652	0.445	1	0.615

Figura 30. Extracto Matriz de Correlación.

Anexo F: Codificación de Variables Categóricas – Modelo Logit

Variable		Frecuencia	Codificación de parámetro		
			1	2	3
Años Vehículo	1	3.002	0	0	0
	2	6.108	1	0	0
	3	3.967	0	1	0
	4	5.053	0	0	1
Años 1er Siniestro	1	10.417	0	0	0
	2	2.937	1	0	0
	3	1.289	0	1	0
	4	3.487	0	0	1
Tasación	1	5.458	0	0	0
	2	3.620	1	0	0
	3	5.434	0	1	0
	4	3.618	0	0	1
Cantidad Vehículos	1	12.392	0	0	
	2	4.187	1	0	
	3	1.551	0	1	
Edad	1	4.691	0	0	
	2	12.090	1	0	
	3	1.349	0	1	

Tabla 54. Codificación de las variables categóricas – Modelo de Regresión Logística.

Anexo G: Árbol de Decisión C5.0 – Estimación de probabilidad

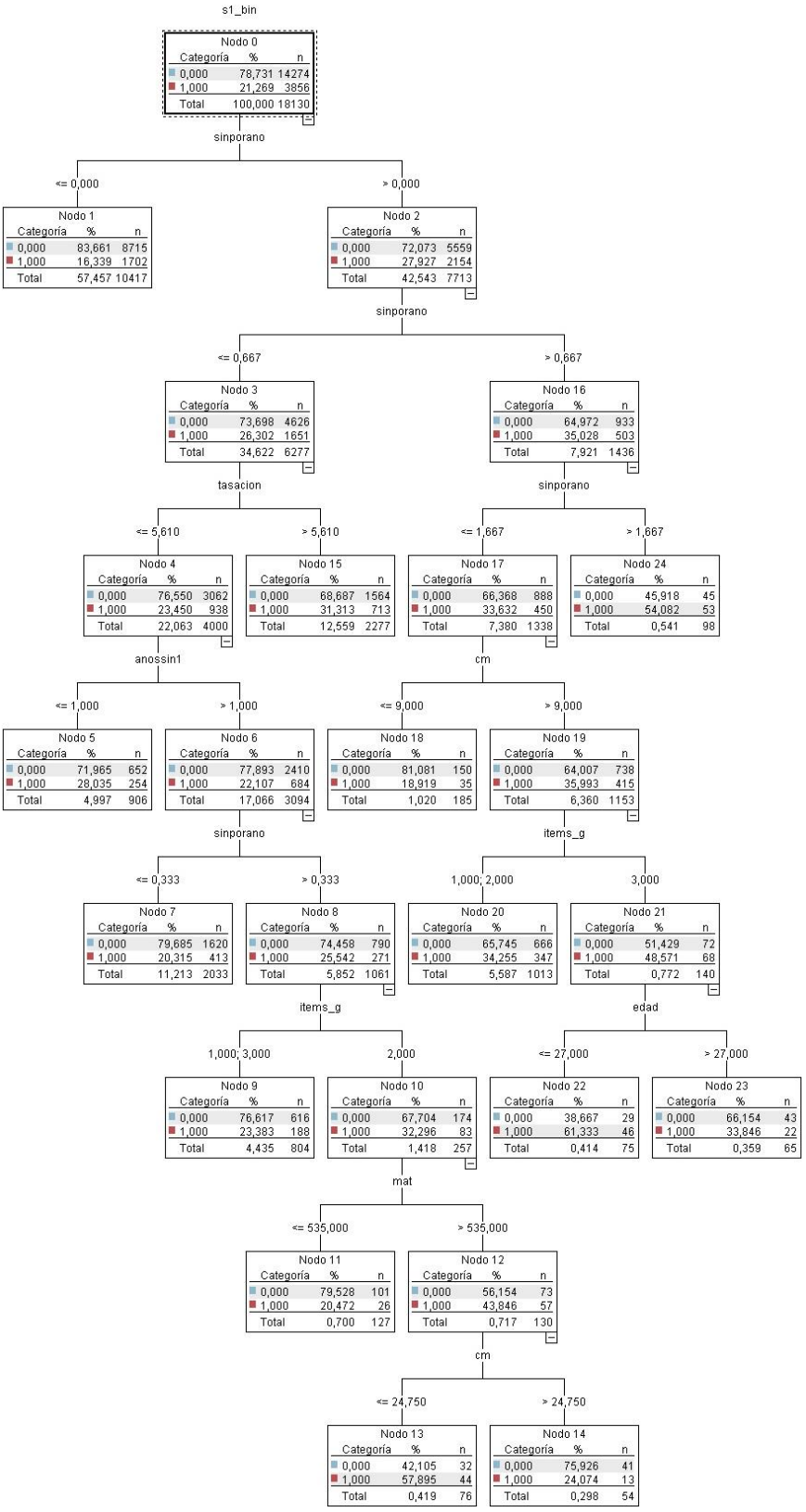

Figura 31. Árbol de Decisión C5.0 – Estimación de probabilidad.

Anexo H: Matrices de Confusión – Modelo Logit

corte=0.121	Real	Predicho		Total	% Acierto
		0	1		
	0	399	3213	3612	11%
	1	38	976	1014	96%
	Total	437	4189	4626	30%

Tabla 55. Matriz de Confusión Modelo Logit – Corte=0.121.

corte=0.140	Real	Predicho		Total	% Acierto
		0	1		
	0	820	2792	3612	23%
	1	97	917	1014	90%
	Total	917	3709	4626	38%

Tabla 56. Matriz de Confusión Modelo Logit – Corte=0.140.

corte=0.159	Real	Predicho		Total	% Acierto
		0	1		
	0	1226	2386	3612	34%
	1	155	859	1014	85%
	Total	1381	3245	4626	45%

Tabla 57. Matriz de Confusión Modelo Logit – Corte=0.159.

corte=0.177	Real	Predicho		Total	% Acierto
		0	1		
	0	1599	2013	3612	44%
	1	237	777	1014	77%
	Total	1836	2790	4626	51%

Tabla 58. Matriz de Confusión Modelo Logit – Corte=0.177.

corte=0.198	Real	Predicho		Total	% Acierto
		0	1		
	0	1974	1638	3612	55%
	1	334	680	1014	67%
	Total	2308	2318	4626	57%

Tabla 59. Matriz de Confusión Modelo Logit – Corte=0.198.

corte=0.221	Real	Predicho		Total	% Acierto
		0	1		
	0	2329	1283	3612	64%
	1	438	576	1014	57%
	Total	2767	1859	4626	63%

Tabla 60. Matriz de Confusión Modelo Logit – Corte=0.221.

corte=0.247	Real	Predicho		Total	% Acierto
		0	1		
	0	2686	926	3612	74%
	1	554	460	1014	45%
	Total	3240	1386	4626	68%

Tabla 61. Matriz de Confusión Modelo Logit – Corte=0.247.

corte=0.281	Real	Predicho		Total	% Acierto
		0	1		
	0	3011	601	3612	83%
	1	687	327	1014	32%
	Total	3698	928	4626	72%

Tabla 62. Matriz de Confusión Modelo Logit – Corte=0.281.

corte=0.334	Real	Predicho		Total	% Acierto
		0	1		
	0	3332	280	3612	92%
	1	833	181	1014	18%
	Total	4165	461	4626	76%

Tabla 63. Matriz de Confusión Modelo Logit – Corte=0.334.

Anexo I: Matrices de Confusión – Árbol de Decisión C5.0

corte=0.070	Real	Predicho		Total	% Acierto
		0	1		
	0	1884	1728	3612	52%
	1	311	703	1014	69%
	Total	2195	2431	4626	56%

Tabla 64. Matriz de Confusión Modelo C5.0 – Corte=0.070.

corte=0.080	Real	Predicho		Total	% Acierto
		0	1		
	0	2273	1339	3612	63%
	1	413	601	1014	59%
	Total	2686	1940	4626	62%

Tabla 65. Matriz de Confusión Modelo C5.0 – Corte=0.080.

corte=0.165	Real	Predicho		Total	% Acierto
		0	1		
	0	2675	937	3612	74%
	1	549	465	1014	46%
	Total	3224	1402	4626	68%

Tabla 66. Matriz de Confusión Modelo C5.0 – Corte=0.165.

corte=0.25	Real	Predicho		Total	% Acierto
		0	1		
	0	3001	611	3612	83%
	1	698	316	1014	31%
	Total	3699	927	4626	71%

Tabla 67. Matriz de Confusión Modelo C5.0 – Corte=0.250.

corte=0.367	Real	Predicho		Total	% Acierto
		0	1		
	0	3327	285	3612	92%
	1	837	177	1014	17%
	Total	4164	462	4626	76%

Tabla 68. Matriz de Confusión Modelo C5.0 – Corte=0.367.

Anexo J: Árbol de Decisión C5.0 – Estimación del Costo Medio

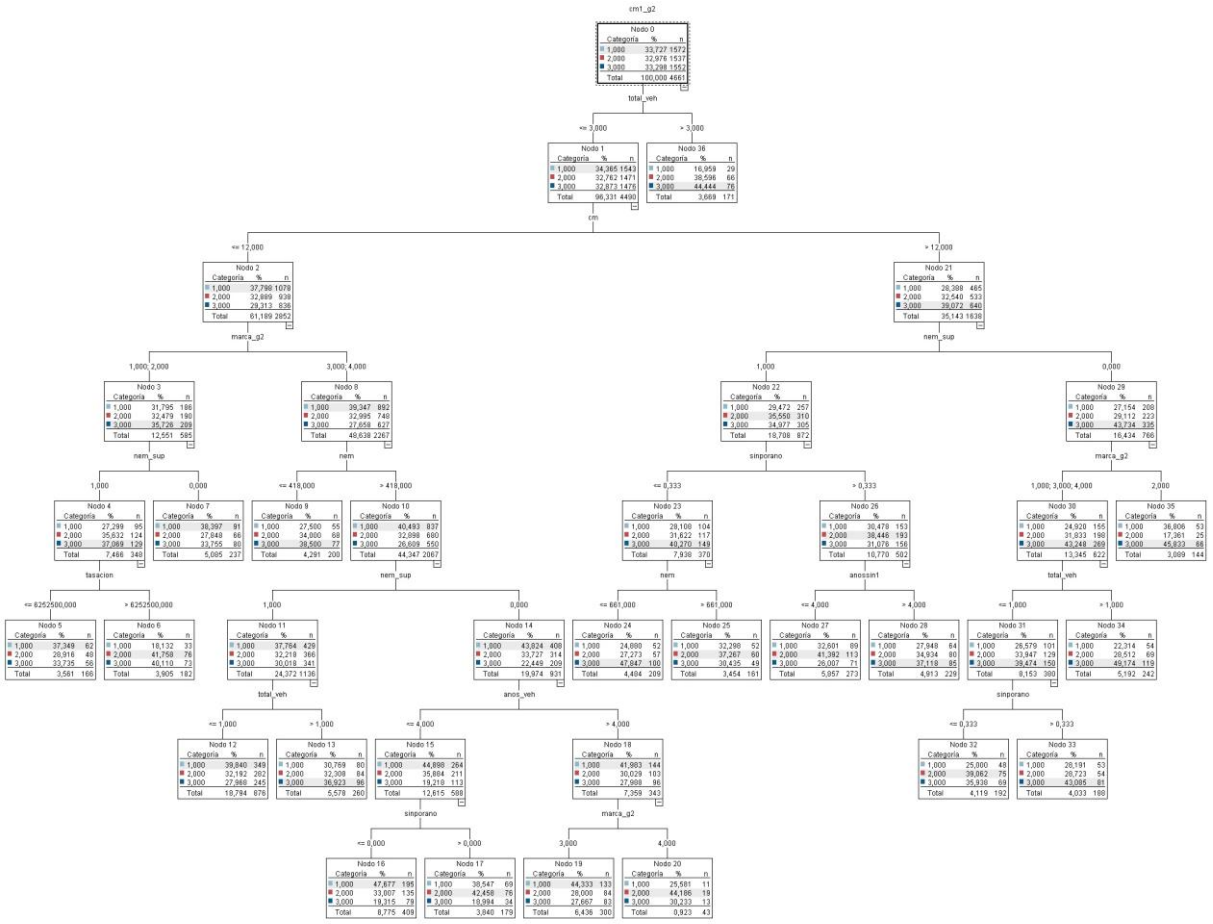

Figura 32. Árbol de Decisión C5.0 – Estimación del Costo Medio.

Anexo K: Análisis de Escenarios.

Escenario Blando		Evaluación	
Corte	Política	Métrica	Valor Métrica
0.330	2.40	Utilidad [UF]	100,193.75
0.245	1.70	Clientes con Dcto	7145
0.221	1.50	Clientes con Recargo	8975
0.158	1.00	Clientes con Prima Base	6636
0.139	0.95		
0.125	0.90		
0.000	0.85		

Tabla 69. Evaluación Económica – Escenario Blando.

Escenario Exigente		Evaluación	
Corte	Política	Métrica	Valor Métrica
0.250	2.40	Utilidad [UF]	75,751.25
0.181	1.70	Clientes con Dcto	0
0.155	1.50	Clientes con Recargo	19525
0.130	1.25	Clientes con Prima Base	3231
0.000	1.00		

Tabla 70. Evaluación Económica – Escenario Exigente.

Escenario Propuesto		Evaluación	
Corte	Política	Métrica	Valor Métrica
0.250	2.40	Utilidad [UF]	124,002.00
0.210	1.70	Clientes con Dcto	3231
0.181	1.25	Clientes con Recargo	16186
0.155	1.15	Clientes con Prima Base	3339
0.130	1.00		
0.000	0.95		

Tabla 71. Evaluación Económica – Escenario Propuesto.

Anexo L: Distribución de la Política Actual según la Política Propuesta

		Política Actual						
		1,00	1,10	1,70	1,87	2,00	2,2	2,4
Política Propuesta	G1: 0,95	89	2	2643	128	284	37	48
	G2: 1,00	58	1	2680	145	360	38	57
	G3: 1,15	47	0	2529	177	343	41	113
	G4: 1,25	38	3	2344	220	349	65	174
	G5: 1,7	34	2	2305	196	358	80	289
	G5: 2,4	44	5	4205	365	603	187	1070

Tabla 72. Tabla de Distribución de la Política Actual según la Política Propuesta.