universidad de león
Facultad de Ciencias Económicas y Empresariales

MUCAF
Máster Universitario en Ciencias Actuariales y Financieras

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

UNIVERSIDAD DE LEÓN

MÁSTER UNIVERSITARIO EN CIENCIAS ACTUARIALES Y FINANCIERAS (MUCAF)

TRABAJO FIN DE MÁSTER:

Análisis del Big Data en los Seguros: Modelos Predictivos

(Big Data Analytics in the Insurance Industry: Predictive Modeling)

AUTOR: *Samuel Álvarez Díaz*

TUTOR: *Rafael Santamaría Sánchez*

CURSO ACADÉMICO: 2017-2018

CONVOCATORIA: Junio

AGRADECIMIENTOS

Quisiera aprovechar este espacio para mostrar agradecimiento a aquellas personas que han hecho posible este trabajo.

En primer lugar agradecer a mi tutor, D. Rafael Santamaría Sánchez, por su dedicación y por animarme durante todo el proceso, ya que sin su guía y ayuda este trabajo no hubiera sido posible.

A Dña. Blanca Moreno Cuartas, tutora de mi Trabajo Fin de Grado, por haberme devuelto la confianza para afrontar este tipo de retos.

A mi familia y amigos, que son los responsables de cada meta que consigo, por su apoyo incondicional.

A mis compañeros de clase, por hacer de este Máster una experiencia inolvidable.

A todos los profesores que han compartido aula conmigo en estos dos años, por los conocimientos y valores que me han transmitido.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

RELACIÓN DE ABREVIATURAS UTILIZADAS

ACP: Análisis de Componentes Principales

AUC: *Area Under the Curve* (Área bajo la curva)

CTREE: *Conditional Inference Trees* (Árboles de decisión condicionales)

FIV: Factor de Inflación de la Varianza

GLM: *General Linearized Models* (Modelos Lineales Generalizados)

GPS: Global Positioning System (Sistema de posicionamiento global)

MCO: Mínimos Cuadrados Ordinarios

RGPD: Reglamento General de Protección de Datos

ROC: *Receiver Operating Characteristic* (Característica Operativa del Receptor)

SVM: *Support Vector Machine* (Máquina de vectores de soporte)

VICA: *Virtual International Conference of Actuaries* (Conferencia Virtual Internacional de Actuarios)

RESUMEN

Este trabajo realiza una exploración en torno al análisis del Big Data en el sector asegurador, tratando de crear un marco en el que se recojan los principales aspectos de éste, para posteriormente analizar los modelos predictivos que se están utilizando con más frecuencia en la actualidad. Se recuerdan brevemente los conceptos de aprendizaje supervisado y no supervisado, y se comentan los principales métodos predictivos con especial atención a la regresión logística y los árboles de decisión condicionales. Ambos métodos se explican mediante sendos ejemplos prácticos, al mismo tiempo que se definen y se tratan las principales características, con el objetivo de crear un marco en el que teoría y práctica estén alineadas con la intención de facilitar la comprensión de los métodos al lector.

Palabras clave: Big Data, análisis del Big Data, modelos predictivos, regresión logística, árboles de decisión condicionales.

ABSTRACT

This paper explores the analysis of Big Data in the insurance industry, trying to create a framework in which its main aspects are covered, to end with an analysis of the predictive models that are being used more frequently . The concepts of supervised and unsupervised learning are briefly recalled, and the main predictive methods are discussed, with special attention to logistic regression and conditional decision trees. Both methods are explained by two practical examples, the main characteristics are defined and explained, the intention is to create a framework in which theory and practice are aligned to facilitate the comprehension of the methods to the reader.

Key words: Big Data, Big Data analytics, predictive modeling, logistic regression, conditional inference trees.

1. INTRODUCCIÓN

Sin que el autor de este trabajo sea en absoluto un entendido del tema, siente gran curiosidad por el Big Data, y está decidido a **ampliar sus conocimientos** en la materia. Cuando comienza la etapa que ha abarcado la realización de este trabajo, el autor poseía conocimientos de cultura general respecto al Big Data. Nunca lo había tratado ni estudiado en curso alguno, sin embargo, entiende que es un tema de especial relevancia para su futuro profesional y quiere comenzar con el primer escalón de esta ascensión, realizando su *Trabajo Fin de Máster* en torno al Big Data. Concretamente, siente especial interés por el análisis del mismo y la creación de modelos predictivos entendiendo que en el sector asegurador cobra especial importancia. Durante el trabajo, se decide a estudiar a fondo la **regresión logística** y **los árboles de decisión condicionales**, al encontrarlos parte fundamental de los modelos predictivos en el análisis del Big Data. En este caso, respecto a la regresión logística el autor sí parte con conocimientos al haber recibido clases durante el Máster que está cursando, no siendo así en el caso de los árboles condicionales, los cuáles descubre realizando este proyecto.

El autor considera que **la literatura** en castellano que une **Big Data, sector asegurador, análisis y modelos predictivos** es aún escasa, y le entusiasma la idea de **poder contribuir** a la misma, realizando un trabajo que aborde todos estos temas. El autor ha aprovechado la oportunidad de tener acceso a la Virtual International Conference of Actuaries (*VICA 2018*), gracias al Máster en el que se desarrolla este trabajo, para inspirarse en la conferencia (Francis y Wolfstein, 2018), además de trabajos como (Padilla-Barreto, Guillén, y Bolancé, 2017), para elegir qué modelos predictivos se están usando más actualmente, y se decide por la regresión logística y los árboles condicionales como se ha mencionado. La intención en estas secciones es tratar de forma simultánea los conocimientos teóricos con los ejercicios prácticos que se realizan, en un intento de suavizar y mejorar la comprensión de la teoría. Es por esto que, los ejercicios prácticos tienen un carácter didáctico, siendo la intención del autor crear un marco que compagine la teoría sobre estos modelos predictivos con ejemplos prácticos. El primero, dentro del ámbito de la regresión logística, está inspirado en un ejemplo de (Pavía, 2016), tratando de completarlo y comentarlo de la forma más detallada posible, ya que el autor experimentó en primera persona la sencillez y capacidad de explicación que tiene. El

segundo, sobre los árboles de decisión condicionales usa datos de (De Jong y Heller, 2008).

Estos ejemplos prácticos se realizan a través de la herramienta R. La versión utilizada es: *R version 3.4.2* (2017-09-28). El código completo se encuentra adjunto en los Anexos 1 y 2. El autor tampoco es un experto en el tema al inicio del proyecto, si habiendo recibidos clases introductorias y no siendo su primer uso de técnicas estadísticas en ella. Sin embargo, nunca había realizado regresión logística o árboles condicionales con R. Al igual que con el Big Data y los modelos predictivos, el autor se encuentra interesado en mejorar sus competencias en esta herramienta, considerando que es de vital importancia en el entorno actuarial. También quiere aprovechar la ocasión para hacer una pequeña defensa del uso del software libre en este tipo de trabajos, siendo cierto que sí ha podido utilizarse en la parte práctica, no le ha sido posible realizar la redacción del proyecto en un programa de software libre como podía haber sido el sistema de composición de textos LaTeX. El autor espera que, herramientas de tanto valor como la mencionada, puedan ser utilizadas y aprovechadas en el futuro.

Resumiendo, el autor se marca como objetivos principales del trabajo comenzar a desarrollar habilidades en campos que considera de tremenda importancia en el mundo actuarial como son el Big Data, los modelos predictivos y la herramienta R, así como contribuir con su pequeño granito de arena a una literatura en castellano aún en desarrollo respecto al Big Data y los modelos predictivos. Otro objetivo del trabajo consiste en unir la parte teórica y práctica de los modelos predictivos para conseguir una lectura amena que ayude al que pueda encontrarse interesado en ampliar sus conocimientos en el tema, o a refrescarlos si ya los tuviere, así como realizar un repaso acerca de los principales aspectos del Big Data en el mundo asegurador, creando un marco en el que el lector pueda encontrar respuesta a la mayoría de cuestiones que puedan surgir, o en su defecto, dirigirle apropiadamente hacia lecturas más específicas sobre estos temas.

El autor considera que, una vez finalizado el trabajo, estos objetivos marcados se han conseguido.

El trabajo se divide en dos grandes bloques. El primero, el estudio y repaso de los principales aspectos del Big Data en el sector asegurador, comenzando con la definición y las características del Big Data (comienzo del Capítulo *2*), discutiendo su ámbito de aplicación en el sector (Sección *2.1*) y los límites que debe tener (Sección *2.2*). Además,

se remarca la importancia del análisis del Big Data y se justifica la elección de los métodos a utilizar (Sección *2.3*) en el segundo gran bloque del trabajo: los modelos predictivos. En este bloque el autor hace un breve repaso de los principales modelos predictivos (comienzo del Capítulo *3*), comentando el aprendizaje supervisado y no supervisado (Secciones *3.1* y *3.2*), y finalmente abordando los dos métodos seleccionados para el estudio en profundidad: la regresión logística (Sección *3.2.1*) y los árboles de decisión (Sección *3.2.2*). Seguidamente, se encuentra un apartado con las principales conclusiones (Capítulo *4*) y, por último, las referencias bibliográficas utilizadas (Capítulo *5*).

2. ¿QUÉ ES EL BIG DATA?

En pleno año 2018, ya podemos confirmar que el Big Data no es una tendencia o una revolución por llegar, sino que el Big Data es una realidad. No hace falta mucho más que unos segundos de reflexión pensando en el día a día de una persona cualquiera, para darse cuenta de que la cantidad de información que genera y queda registrada constantemente ha aumentado de forma exponencial.

Esto queda remarcado en el hecho de que se cumplen 21 años de la que es reconocida como una de las primeras referencias al término Big Data (Cox y Ellsworth, 1997). Éste es un artículo de dos investigadores de la NASA, en el que se defendía que la velocidad a la que estaban creciendo los datos, empezaba a ser un quebradero de cabeza para los métodos y sistemas que se utilizaban entonces. El concepto de Big Data original se definía como aquel volumen de datos que no podía ser procesado de forma eficiente por los métodos y las herramientas tradicionales (Kaisler, Armour, Espinosa, y Money, 2013). Esta definición podría utilizarse de la misma manera hoy en día, pero no tendría el mismo sentido: lo que hace diez años se conocía como “métodos y herramientas tradicionales” ahora tiene un significado completamente distinto. Ahora hay métodos y herramientas que hace diez años no existían. Por ponernos en contexto, hace poco más de una década no existían: iPhone, Uber, Android, Facebook, YouTube, Spotify, Instagram o WhatsApp. El lector se puede hacer una idea en este momento de que no existe una definición globalmente aceptada del término, y que lo principal es que el concepto de Big Data es dinámico.

Sí se ha llegado a un consenso sin embargo en cuanto a las características que presenta el Big Data, siendo muy conocida la clasificación de las tres “V’s” del Big Data. En los

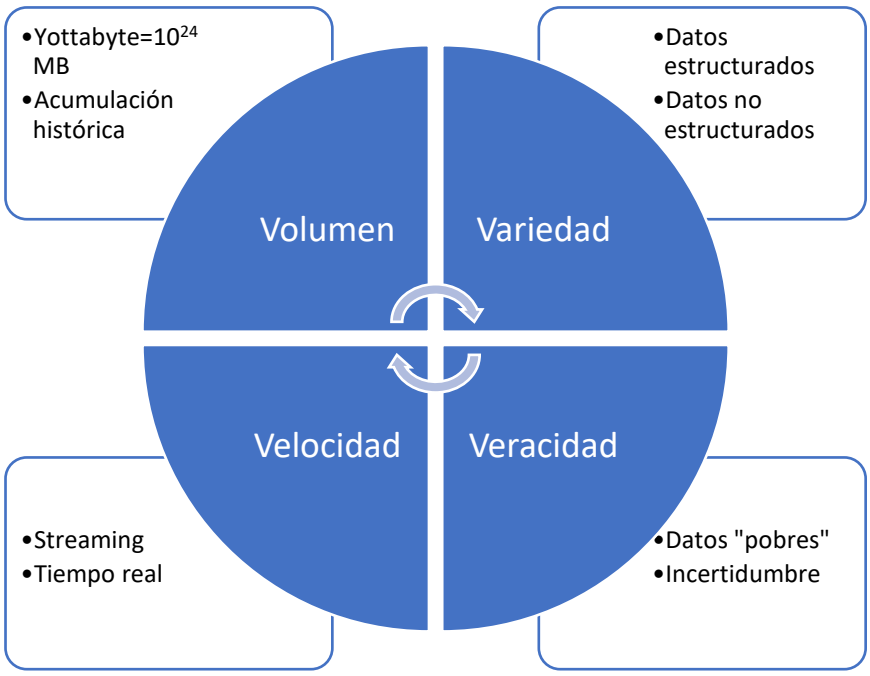
últimos tiempos, esta clasificación ha visto incorporada una cuarta "V", pasando a ser las "cuatro V's del Big Data".

Estas cuatro características son: volumen, variedad, velocidad y veracidad. Basándose en estas características, IDC (*International Data Corporation*) uno de los principales proveedores globales de servicios de consultoría y tecnología de la información, publicó una definición más actual de Big Data:

"El Big Data describe una nueva generación de tecnologías y arquitecturas diseñadas para extraer valor de manera económica de **volúmenes** muy grandes de una amplia **variedad** de datos, permitiendo la captura, el descubrimiento o el análisis a alta **velocidad**." (Gantz y Reinsel, 2011).

En la *Figura 2.1* se presenta un esquema de las cuatro características que se van a tratar seguidamente.

Figura 2. 1 Resumen 4 V's de Big Data



Fuente: Elaboración propia

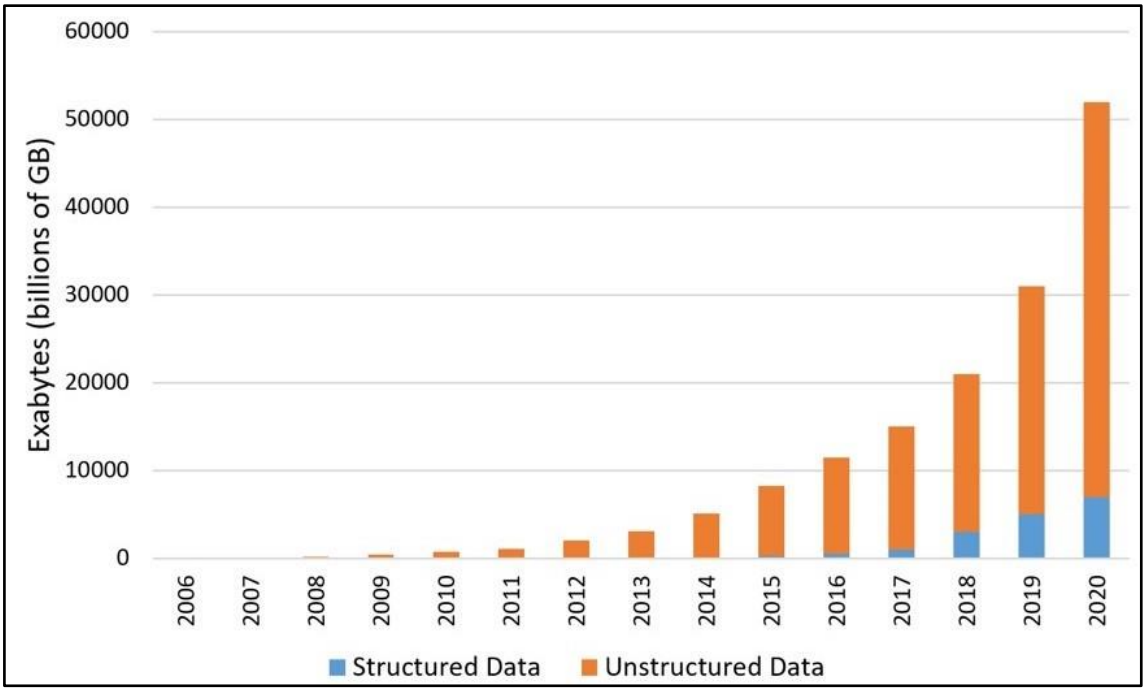
Volumen es quizás, la característica más obvia del Big Data. Lo primero que nos pasa por la cabeza a la hora de pensar en Big Data son enormes cantidades de datos. No nos equivocamos, es definitivamente algo muy característico del Big Data. Tanto el volumen de los datos que se generan cada día, como el volumen de los datos que se van acumulando

a lo largo del tiempo, crecen de forma exponencial. Se presentan algunos datos que ilustran esta característica: cuarenta Zettabytes (10^{12} GB) de datos serán creados en 2020, lo que son unos cuarenta y tres billones de Gigabytes, trescientas veces lo que se creaba en 2005. Existen aproximadamente seis mil millones de teléfonos móviles en propiedad. La gran mayoría de las empresas en Estados Unidos tienen como mínimo cien mil Gigabytes de datos almacenados (IBM, 2018). En el caso del sector al que está enfocado este trabajo, la cantidad de datos disponibles para las aseguradoras ha aumentado de forma exponencial en los últimos años. Las redes sociales han introducido datos de nuevas fuentes, muy diferentes entre sí, lo que podría tener impacto en la marca, los productos o la percepción del cliente entre otros aspectos (Mills y Forder, 2012). Además, esta característica trae consigo un problema añadido, la creación de una nueva necesidad en materia de almacenamiento escalable de datos. Es aquí donde entra en acción la arquitectura de datos y las tecnologías del Big Data, no sólo para esta característica, sino que será de vital importancia también para las demás. La arquitectura de datos y las tecnologías del Big Data son materias de gran interés para cualquier persona interesada en este tema, lamentablemente, se escapan del campo de conocimiento del autor y del enfoque de este trabajo, por tanto, no se tratarán. Se recomienda (Gökalp et al., 2017) para aquellas personas interesadas en la arquitectura de datos en Big Data.

La segunda característica que tratar es la **variedad**, y aquí es necesario explicar dos conceptos antes de continuar. Estamos hablando de los **datos estructurados** y los **datos no estructurados** (*structured data vs. unstructured data*). Los datos estructurados, son aquellos que se encuentran organizados y se pueden ordenar y procesar de forma sencilla. Este tipo de datos son los que estamos acostumbrados a utilizar de forma habitual. Son ejemplos de datos estructurados las bases de datos organizadas en filas y columnas con títulos. Los datos no estructurados son aquellos que no tienen una estructura identificable, ni están organizados de alguna forma. La mayoría de este tipo de datos se espera que sean en formato de texto, pero aludiendo a la característica que estamos definiendo, la variedad, podemos decir que estos datos pueden ser muy variados, en forma de imágenes, fechas, números… Se presenta la *Figura 2.2* donde se puede ver las expectativas de crecimiento de este tipo de datos. Remarcar que en el gráfico se habla de *billions*, es decir, miles de millones. Observamos que los datos no estructurados son mayores, lo que nos lleva a comentar la "regla del ochenta por ciento". Una regla comúnmente aceptada en el

entorno Big Data, que señala que la proporción actual de datos no estructurados rondaría esa cifra y tan sólo un veinte por ciento serían datos estructurados (Malone, 2007).

Figura 2. 2 Datos Estructurados y No Estructurados



Fuente: Patrick Cheesman

La variedad de datos es una medida de la riqueza de los datos: texto, imágenes, video, audio, etc. Desde una perspectiva analítica, es probablemente el mayor obstáculo para usar de manera efectiva grandes volúmenes de datos. Los formatos de datos que sean incompatibles, las estructuras de datos no alineadas y la semántica de datos inconsistente representan desafíos significativos que pueden conducir a la expansión analítica (Kaisler et al., 2013). En cuanto a las compañías de seguros, históricamente se ha confiado en los datos estructurados para tomar decisiones. La aparición de las redes sociales, y la nueva gran variedad y cantidad de datos no estructurados significa que las empresas tienen una nueva oportunidad de conseguir información extra acerca de aspectos tan importantes como sus productos, sus empleados o sus clientes entre otros (Mills y Forder, 2012).

La siguiente característica es la **velocidad.** Las empresas obtienen datos de forma cada vez más rápida. De hecho, muchas empresas obtienen datos en tiempo real, como por ejemplo datos *GPS* (*Global Positioning System*). Esto afecta directamente al mundo de los seguros, ya que el uso del *GPS* genera datos telemáticos de la conducción de vehículos, que pueden ser utilizados por las aseguradoras. La velocidad de los datos hace referencia a la velocidad de creación, de transmisión y agregación de estos. El comercio

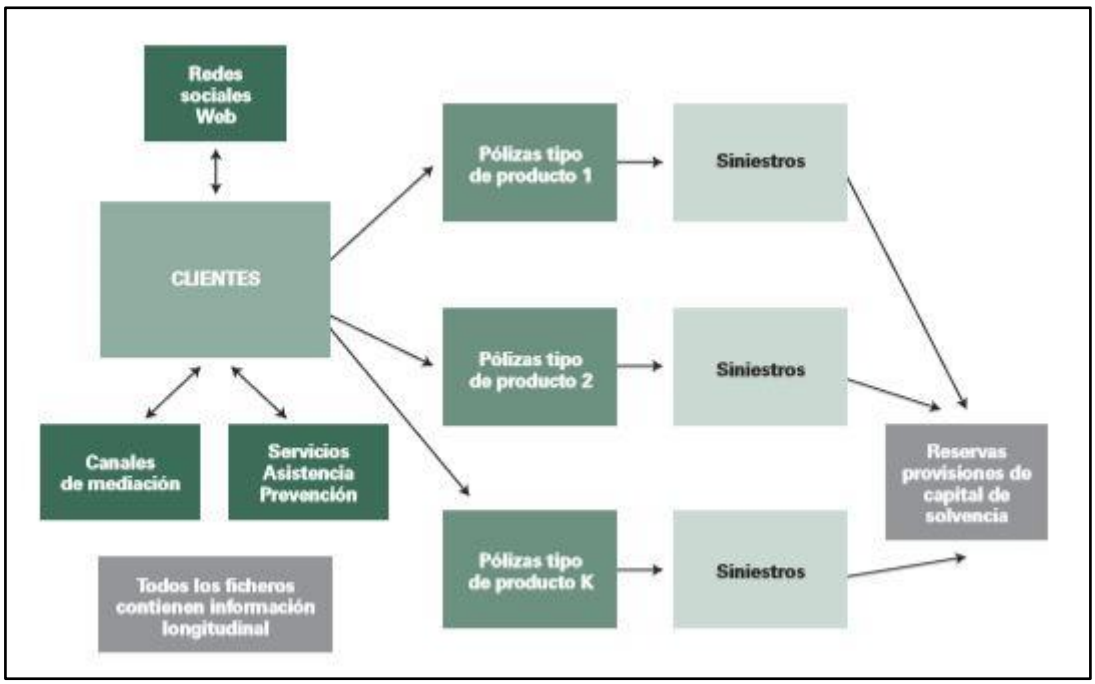
electrónico ha aumentado rápidamente la velocidad de los datos utilizados para diferentes transacciones (por ejemplo, clics en el sitio web) (Kaisler et al., 2013).

La última característica incorporada es la **veracidad**, a veces excluida de la clasificación, pero en determinados sectores, uno de ellos el asegurador, puede tratarse de una característica de vital relevancia. Afecta a la incertidumbre sobre los datos, y se refiere al sesgo, el ruido y la falta de normalidad en los mismos. La veracidad en el análisis de datos es un gran desafío y el coste que tienen los "datos pobres" (*poor data*) puede ser trascendental. Se decide seguir el enfoque de las cuatro Vs por su relevancia para el sector asegurador. Según *IBM*, el coste de "datos pobres", es decir, de mala calidad, en la economía americana es de 3.1 billones de dólares (IBM, 2018). Teniendo en cuenta que la detección del fraude en las aseguradoras es uno de los aspectos relevantes del negocio, nos indica la importancia de que la información de la empresa sobre el cliente sea veraz.

En la *Figura 2.3* se muestra un esquema de base de datos en una entidad aseguradora, donde podemos apreciar la complejidad del mismo. Se observa como los clientes tienen la posibilidad de contratar una o varias pólizas, a través de distintos medios como canales de mediación, agentes o corredores. Los siniestros se encuentran vinculados a los contratos y a los clientes. Además las pérdidas esperadas de estos siniestros han de ser dotadas como reservas para garantizar la cobertura de las mismas (Guillén, 2016). La información que se tiene de los clientes llega por diferentes medios, y ha de ser incorporada a la póliza para concretar una prima. Todo este conglomerado de información es parte del Big Data de las empresas de seguros.

En las empresas de seguros el uso de métodos estadísticos, de la ciencia actuarial y la recolección masiva de datos siempre han formado parte del negocio. Es por esto por lo que, al mismo tiempo que es un gran desafío, también es una gran oportunidad, al partir de un punto más avanzado que muchos otros sectores. En la siguiente sección veremos diferentes casos en los que el uso del Big Data puede aportar gran valor al sector asegurador.

Figura 2. 3 Esquema de base de datos en una entidad aseguradora



Fuente: (Guillén, 2016)

2.1 ÁMBITO DE APLICACIÓN DEL BIG DATA EN LOS SEGUROS

En la sección anterior se ha argumentado por qué es importante el Big Data en el sector asegurador. En ésta, veremos algunos de los usos que pueden darle las compañías aseguradoras.

El sector asegurador se encuentra en un período en el que ya no es una opción adaptarse o no al Big Data. Las empresas que mejor se adapten a esta nueva herramienta, obtendrán una ventaja competitiva muy importante para el futuro. Como veremos a lo largo de esta sección, hay muchos y muy variados usos del Big Data, y será clave para las empresas detectar cuáles son más importantes para su negocio. *IBM* recomienda tres pasos que cualquier aseguradora debería tener en cuenta a la hora de incorporar Big Data a su negocio (Mills y Forder, 2012). El primer paso sería identificar los posibles casos de uso del Big Data en su negocio. El segundo, clasificarlos por la importancia y el valor que tendrían para la empresa, y, por último, la creación de un "laboratorio de datos" (*data labs*), donde estudiar estos casos y validarlos antes de incorporarlos al negocio.

Pasamos a discutir los casos en los que el uso de Big Data puede agregar valor al negocio asegurador. Para ello, utilizamos datos extraídos de dos encuestas de la conocida empresa Willis Tower Watson, que tratan sobre las áreas del negocio en las que las empresas ven mayor impacto del Big Data (Willis Towers Watson, 2015, 2016). Se presenta un cuadro resumen en la *Tabla 2.1* de estas dos encuestas a aseguradoras americanas. En estas encuestas, las aseguradoras respondían a la pregunta: ¿En qué áreas considera que el Big Data ayuda más a su negocio? Los valores de la tabla representan el porcentaje sobre el total de respuestas. Vemos que de la encuesta de 2015 a la de 2016, hay un cambio importante. En la encuesta de 2015, el 77% de las aseguradoras consideraba que la tarificación, la suscripción y la selección de riesgos, eran áreas donde el Big Data iba a ayudar más a su negocio en el corto plazo (dos años). En la encuesta del 2016, este dato ya subía al 92%. Otro dato que destacar es el cambio en la percepción sobre las decisiones de gestión. En 2015, un 19% respondía que el Big Data ayudaba en ese momento en las decisiones de gestión. En 2016 este dato crecía hasta un 41% y se esperaba que en la actualidad el dato subiera hasta el 84%.

Tabla 2. 1 Encuestas Big Data 2015/2016

	Encuesta 2015		Encuesta 2016	
	Actualidad (2015)	Dentro de dos años (2017)	Actualidad (2016)	Dentro de dos años (2018)
Tarificación, suscripción y selección de riesgos	42%	77%	57%	92%
Decisiones de gestión (*management*)	19%	60%	41%	84%
Control de pérdidas y reclamaciones	17%	58%	27%	76%

Fuente: Elaboración propia a partir de datos de (Willis Towers Watson, 2015, 2016)

Procedemos ahora a comentar algunos de los casos específicos en los que la industria aseguradora puede aprovechar el Big Data.

Detección de fraude

Las aseguradoras pueden utilizar el Big Data para mejorar la detección de fraude a través del análisis de datos y modelos predictivos. Una de las opciones es cruzar constantemente bases de datos de reclamaciones actuales con bases de datos de reclamaciones históricas fraudulentas, datos de los clientes y datos de las redes sociales, y estudiar los casos en los que existan coincidencias. Cuantas más variables tengamos y podamos cruzar de manera eficiente, más opciones de éxito tiene este sistema.

Las variables en estas bases de datos no provienen únicamente de las pólizas y las reclamaciones, sino que también de los datos de los clientes, su comportamiento, su entorno… Este proceso es tremendamente complicado para ser realizado por una persona y cada caso consumiría muchos recursos, en cambio utilizando análisis de Big Data es algo mucho más factible.

Especialmente interesante es el caso de las redes sociales y la información telemática en este aspecto. Una empresa aseguradora podría constatar si una reclamación es válida o no, utilizando datos no estructurados de información reciente en las redes sociales. Otros ejemplos son: ¿Están dos personas que presentan una reclamación conectadas por las redes sociales? ¿Confirman los datos GPS que la persona que reclama estaba en lugar del suceso en ese momento? Éstas son algunas de las cuestiones que se plantearían en un modelo de detección de fraude (Mills y Forder, 2012).

Gestión de riesgos

La gestión de riesgos no sólo se considera uno de los ámbitos donde el Big Data podría ser de mayor valor, sino que es uno de los ámbitos más importantes de una empresa aseguradora. Desde la entrada en efecto en 2016 de la normativa Solvencia II, un nuevo marco normativo que afecta a la aseguradoras (ver (Pwc, 2012) para más información sobre la influencia de Solvencia II en la gestión de riesgos), la gestión de riesgos ha ganado mucho peso en las aseguradoras.

Uno de los aspectos más importantes para las aseguradoras es determinar las primas de las pólizas. Siendo las primas una expresión del riesgo transferido por parte de los clientes a las aseguradoras, entendemos el porqué de la importancia de la gestión de riesgos.

Algunas compañías de seguros, principalmente en los ramos de automóvil, hogar y salud, están empezando a sacar provecho al uso de los datos telemáticos, tecnología portátil (*smartwatch*, pulseras de actividad diaria, etc.) o del internet de las cosas (*IoT* o *Internet of Things*), para rastrear a sus clientes a fin de predecir y calcular riesgos (siempre con la aprobación de estos, por supuesto).

Mediante el uso de modelos predictivos, las aseguradoras pueden identificar si hay mayor probabilidad de que sus clientes participen en un accidente o su coche sea robado, combinando sus datos de comportamiento con datos de factores exógenos, como las condiciones de la carretera o previsiones meteorológicas.

Un uso similar se puede observar en el ramo de salud y el seguro de vida debido al creciente uso de tecnología portátil. Los rastreadores de actividad pueden controlar los comportamientos y hábitos de los usuarios y proporcionar evaluaciones continuas de sus niveles de actividad. Muchas aseguradoras ahora ofrecen servicios extra y descuentos basados en el uso de estos dispositivos, para saber más sobre tecnología portátil (*wearables*) y su uso en los seguros (ver (Verma, van Deel, Nadimpalli, Sahoo, y Vervuurt, 2016) o (Young, 2017)).

Relación con los clientes (*Customer Relations Management*)

El Big Data no sólo es una oportunidad de crear valor para el negocio de las aseguradoras, también es una oportunidad de crear valor para el cliente. El Big Data permite a las aseguradoras personalizar al máximo el trato con el cliente. No sólo eso, los clientes que han tenido una mala experiencia y que no informen a la empresa sobre ello, es muy posible que sí expresen su descontento a través de las redes sociales. Las empresas pueden aprovechar toda esa información que se publica en las redes sociales para detectar debilidades en su trato con los clientes y actuar en consecuencia.

El uso de *wearables* como relojes o pulseras de actividad, tiene más de una utilidad para las empresas de seguro. Además de servir para prevenir riesgos, mejora el trato al cliente. Una práctica que está comenzando a extenderse en las aseguradoras es incluir con sus seguros de salud este tipo de dispositivos, ofreciendo servicios a los clientes a través de ellos, como por ejemplo la prevención de la diabetes. También existen aplicaciones para el móvil propiedad de aseguradoras que evalúan la conducción del cliente y le aconsejan sobre malos hábitos (Eyastax, 2017).

Finanzas/Decisiones de negocio

"Imagine poder realizar ajustes automáticos diarios a estrategias de reaseguro, tarificación y límites de suscripción mediante la combinación de datos internos estructurados (por ejemplo, actuariales, financieros y políticos), con datos externos no estructurados, como comentarios de prensa y analistas de Twitter, blogs y sitios web" (Mills y Forder, 2012).

Éste es sin duda uno de los usos más importantes, y ya no es una cuestión de si puede hacerse, es una cuestión de cuánto tardará en hacerse. Las opciones son muchas y muy variadas. Tenemos el ejemplo de *pay as you drive*, un seguro de auto basado en el uso, en el que la prima se personaliza hasta tal punto que literalmente los conductores pagan por el uso que le den al automóvil. De esta manera, los usuarios menos frecuentes pagarán menos (Guillén, 2016).

En (McAfee y Brynjolfsson, 2012) se considera al Big Data directamente una revolución del *management*, por todas las oportunidades que ofrece. Se recomienda esta referencia para una mejor comprensión de cómo afecta el Big Data en las decisiones de negocio.

En definitiva, existen diversas y numerosas aplicaciones en los seguros para el Big Data. Este trabajo se centra en el análisis de estos datos, concretamente en los modelos predictivos que nos permiten hacer previsiones a futuros sobre muy variados aspectos del negocio.

2.2 LOS LÍMITES DEL BIG DATA

Es un tema de gran trascendencia en la actualidad. Los medios que se hacen eco de escándalos acerca del uso de información personal de forma ilegal, y la entrada en vigor de forma reciente del Reglamento General de Protección de Datos (RGPD) recogido en (R (UE) nº679/2016 del Parlamento Europeo, de 27 de abril de 2016), han puesto el foco sobre este asunto. Puede parecer que es un tema puramente de ética en los negocios, pero es mucho más complejo. Cuando se está hablando de utilizar literalmente toda la información disponible y que el uso de esa información va a crear valor para las empresas, es necesario un marco de legalidad muy claro para proteger a los individuos de los que se tiene la información. En el sector seguros es especialmente sensible es el caso de la información respecto a la salud de las personas.

El tema es tan amplio y complejo que merece investigación específica y un trabajo con el Big Data y la protección de datos como tema principal. Existen organismos oficiales como el Supervisor Europeo de Protección de Datos, autoridad supervisora europea nombrada por el Parlamento Europeo o la Information Commissioner's Office que responde directamente al Parlamento Británico, (las páginas web de ambos organismos se encuentran en la bibliografía).

Los principales puntos por los que estas leyes son tan importantes en relación al Big Data son: la protección de los individuos del efecto intrusivo que pueda llegar a tener el análisis del Big Data, remarcar que las empresas deben de usar los datos personales en el análisis del Big Data en un marco de lo que los individuos consideren expectativas razonables y que las empresas sean reticentes a la transparencia debido poseer innovadores métodos de análisis que no quieran compartir por temor a perder una ventaja competitiva. Para una lectura más completa y extensa sobre el tema ver (Information Commissioner's Office, 2017; Louveaux, 2016).

2.3 ANÁLISIS DEL BIG DATA Y SU IMPORTANCIA

¿Por qué es tan importante el análisis del Big Data? "No puedes controlar aquello que no puedes medir" es una frase que se atribuye a W. Edwards Deming, uno de los estadísticos más exitosos del siglo XX. Gracias al Big Data podemos medir, y por tanto conocer profundamente mucho más sobre cualquier aspecto de la sociedad, los negocios y por supuesto, de los seguros (McAfee y Brynjolfsson, 2012).

La importancia del Big Data queda destacada en la anterior sección. El análisis del Big Data implica conocer más sobre los clientes, las tendencias de mercado, las campañas de marketing, el rendimiento de los activos y mucho más. Es por esto que, muchas empresas de seguros necesitan de un departamento de Big Data, departamento de *Analaytics* o un *data lab*.

El concepto de los *data labs* es un tema a tratar de forma independiente. Existen multitud de artículos que explican la importancia de departamentos o grupos de trabajo cuya tarea es el análisis de Big Data, y concretamente (Padilla-Barreto et al., 2017) defiende durante gran parte de su trabajo la necesidad para las empresas de seguros de contar con un departamento de *Analytics* o de Big Data.

Dentro del análisis del Big Data, el autor de este trabajo se interesa especialmente por los modelos predictivos. Actualmente, los modelos predictivos engloban diferentes áreas del sector asegurador, como pueden ser modelos sobre la detección del fraude, la probabilidad de reclamación, las reservas, la tarifación… La aparición del Big Data hace que sea una tarea cada vez más interesante. Dentro de los modelos predictivos, en este trabajo se dedica especial atención a la regresión logística y a los árboles condicionales. La decisión se ha tomado teniendo en cuenta la información proporcionada por encuestas como (Willis Towers Watson, 2017), donde las aseguradoras confirman que los Modelos Lineales Generalizados, en adelante *GLM*, y los árboles de decisión son, con un amplio margen, las herramientas más utilizadas en cuanto a modelos predictivos. Se presenta la *Tabla 2.2* que resume esos datos:

Tabla 2. 2 ¿Para qué aplicaciones planean las aseguradoras utilizar los siguientes métodos?

Modelo Predictivo	Modelos sobre pérdidas	Modelos sobre reclamaciones	Modelos sobre Marketing
GLM	100%	42%	37%
Árboles de decisión	48%	31%	30%
Modelos combinados	35%	23%	19%
Métodos Gradient boosting	30%	17%	22%
Otros métodos de Machine-Learning	26%	33%	19%

Fuente: Elaboración propia a partir de datos de (Willis Towers Watson, 2016)

3. MODELOS PREDICTIVOS PARA EL ANÁLISIS DEL BIG DATA

A la hora de realizar modelos predictivos para analizar casos de Big Data, lo primero es matizar qué es **aprendizaje supervisado** (*supervised learning*) y **aprendizaje no supervisado** (*unsupervised learning*). Estos términos nacen en el contexto de *Machine Learning*, que puede ser definido brevemente como los métodos de computación que utilizan la experiencia para realizar mejores predicciones. Se hace referencia a la experiencia como la información pasada disponible, que habitualmente toma forma de datos en un aparato electrónico y son utilizados para el análisis. Pertenece al campo de la Ciencia de la Computación, campo que se escapa de los conocimientos del autor de este trabajo y que recomienda la lectura (Mohri, Rostamizadeh, y Talwalkar, 2012), considerada una de las referencias principales en este tema.

En el aprendizaje supervisado se usan datos históricos que están etiquetados (*labeled*) para hacer predicciones. Para ello, se suele trabajar con dos muestras, una denominada **muestra de entrenamiento** o *training set*, que se usa para realizar el ajuste del modelo/análisis/algoritmo pertinente. Con los resultados obtenidos de ese *training set*, se realizan predicciones en otra muestra de la información, denominada **muestra de prueba**, que habitualmente estará formada por la parte restante que no sea utilizada por el *training set*. El objetivo es predecir en la muestra de prueba los patrones o comportamiento que se han visto en la muestra de entrenamiento. Para realizar el aprendizaje supervisado de forma correcta, la muestra de entrenamiento y la muestra de prueba han de tener la misma distribución de probabilidad. Son métodos de aprendizaje supervisado la regresión logística, los árboles de decisión condicionales, las máquinas de vectores de soporte o las redes neuronales.

En el aprendizaje no supervisado, en cambio, se usan datos históricos que no se encuentran etiquetados, con el objetivo de dar con una estructura o forma en la que se organicen. Toda la información a incorporar en el análisis se utiliza como *training set*, ya que no se realizan predicciones. Son ejemplos de aprendizaje no supervisado el análisis de componentes principales (en adelante *ACP*) o el análisis *Cluster*.

En las siguientes secciones, se definen algunos de los métodos existentes para modelos predictivos (*predictive modeling*). El grueso del trabajo está centrado en la regresión logística y los árboles de decisión condicionales, donde se explican dos ejemplos prácticos con ayuda de la herramienta R para ayudar a su comprensión. El resto de métodos serán brevemente comentados. A la hora de decidir qué métodos estudiar, y qué clasificación de los mismos realizar, el autor se ha fijado en los métodos utilizados en (Padilla-Barreto et al., 2017) y en (Francis y Wolfstein, 2018).

3.1 MÉTODOS DE APRENDIZAJE NO SUPERVISADO

Como hemos explicado en la introducción del capítulo, los métodos de aprendizaje no supervisado son utilizados para agrupar o clasificar elementos de unos datos históricos. Se suelen utilizar en situaciones en las que se dispone de mucha información, con gran cantidad de variables que están correlacionadas entre sí en mayor o menor grado (Aldás y Uriel, 2017).

El **análisis de componentes principales** o *ACP* es una técnica estadística utilizada para simplificar una base de datos. El origen del *ACP* suele atribuirse a (Pearson, 1901),

aunque el primer estudio que usa este método tal y como se conoce hoy en día pertenece a (Hotelling, 1933). El objetivo del *ACP* es reducir la dimensión de datos multivariantes al mismo tiempo que se preserva la mayor cantidad posible de información relevante (Jolliffe, 2002). Es una técnica descriptiva y factorial, que parte de un conjunto de variable independientes altamente correlacionadas, y cuyo objetivo es agruparlas, para conseguir un nuevo conjunto de variables incorreladas entre sí. Dentro del análisis multivariante, se trata de uno de los métodos más comúnmente utilizados, y suele utilizarse como técnica previa a otros análisis o métodos (ver (Jolliffe, 2002) para obtener más información acerca del *ACP)*.

El análisis **clúster** o de **conglomerados** es una técnica estadística multivariante que, sobre una serie de variables de un conjunto de individuos, reorganiza la información original de forma que queda agrupada en grupos que se denominan clúster o agrupaciones. Tiene su origen en el campo de la Biología (Sokal y Sneath, 1963) y actualmente se utiliza en todos los campos. Es una técnica de clasificación de individuos, aunque también puede aplicarse para clasificar variables. El objetivo del análisis clúster es clasificar el conjunto de individuos en función de las variables analizadas y caracterizar los grupos obtenidos. Los individuos han de quedar clasificados en un único grupo, y los individuos del mismo grupo han de ser lo más parecidos (homogéneos) entre sí y ser lo más diferentes (heterogéneos) respecto a los otros que sea posible (Mures y Vallejo, 2018). Para una lectura más amplia sobre el análisis clúster o de conglomerados consultar por ejemplo (Ferrán, 1996), (Hait, Anderson, Tatham, y Black, 1995) o (Johnson y Wichern, 1998).

3.2 MÉTODOS DE APRENDIZAJE SUPERVISADO

Dentro de los métodos de aprendizaje supervisado, este trabajo se centra en la regresión logística y los árboles condicionales, ya que son las más metodologías más utilizadas para modelos predictivos en las aseguradoras (Willis Towers Watson, 2017).

Otros de los métodos de este tipo son las **máquinas de vectores de soporte (***SVM-Support Vector Machine***)**, "un método de predicción alternativo y comúnmente utilizado en problemas de clasificación, análisis de regresión y detección de valores extremos" (Padilla-Barreto et al., 2017). Las *SVM* aplican un método lineal simple a los datos, pero en un espacio de características de alta dimensión que no está linealmente relacionado con el espacio de entrada. Además, aunque podemos pensar en *SVM* como un algoritmo

lineal en un espacio de alta dimensión, en la práctica, no implica ningún cálculo en ese espacio de alta dimensión. Esta simplicidad, combinada con el rendimiento otros métodos de aprendizaje (clasificación, regresión…), ha contribuido a la popularidad de las *SVM*. Sus principales limitaciones son el tiempo y recursos que consume entre la fase de entrenamiento y la de prueba (Karatzoglou, Meyer, y Hornik, 2006).

Las **redes neuronales artificiales** también pertenecen a este tipo de aprendizaje. Están inspiradas en el funcionamiento de las redes neuronales del cerebro humano, y tratan de aproximar sus diferentes funciones a la computación. Las redes neuronales artificiales en el contexto del aprendizaje supervisado consisten en que, a partir de un conjunto de datos de entrenamiento compuesto por patrones de entrada y de salida, la red construya un modelo a partir del proceso desconocido generado entre entradas y salidas. Estos modelos pueden ser de tres tipos: por corrección de error, por refuerzo y estocástico (Casanovas Ramón, Merigó Lindahl, y Torres Martínez, 2014). Los resultados de las redes neuronales son poco intuitivos y generan poca información sobre la influencia de las variables explicativas en la variable objetivo. Destacan por su capacidad para encontrar relaciones complejas no lineales y su capacidad de aprender de dichas relaciones (Padilla-Barreto et al., 2017). Se recomiendan las lecturas (Casanovas Ramón et al., 2014) y (Livingstone, Manallack, y Tetko, 1997) para conocer más información respecto a las redes neuronales.

3.2.1 Regresión logística

Muchas veces, la variable objetivo que queremos estudiar no es un valor numérico sino la ocurrencia o no de un evento. Este tipo de variables, se denominan dicotómicas o de respuesta binaria. En esta sección se desarrollará un ejemplo práctico en el que se da esta situación. El ejemplo práctico presenta una base de datos de una reaseguradora, en la que los individuos son un conjunto de pequeñas mutuas, las cuales pueden hacer reclamaciones a la reaseguradora. Por ejemplo, en nuestro ejercicio práctico, la variable objetivo es la existencia o no de una reclamación de una mutua de seguros. Esta variable dicotómica puede por tanto tomar únicamente dos valores: 0 y 1. El valor 1 indicaría la ocurrencia del evento en cuestión, en este caso, la existencia de una reclamación, mientras que el valor 0 indica la no ocurrencia de este evento, es decir, no hay reclamación.

Para estudiar este tipo de variables, una de las técnicas más adecuadas es la regresión logística. Se trata de un método de regresión múltiple que se emplea cuando la variable

dependiente es no métrica. Puede tomar dos formas: cuando esta variable tenga más de dos niveles, regresión logística multinomial, y cuando tenga sólo dos niveles, regresión logística binomial, (Aldás y Uriel, 2017). Este tipo es el más utilizado y el más común, se trata de un modelo lineal generalizado de respuesta binaria, muy utilizado a nivel actuarial. También es una herramienta muy utilizada en el sector asegurador, teniendo muchas aplicaciones en estudios de morosidad, fraude, reclamaciones, predicción de catástrofes naturales… Una de las ventajas de este modelo es que es bastante conocido y sus resultados sencillos de comprender y explicar. En cambio, uno de sus principales inconvenientes es la complejidad de extrapolar los resultados si hay sobreajuste (*overfitting*) (Padilla-Barreto et al., 2017). Por estos motivos, la regresión logística es uno de los modelos más utilizados en entornos Big Data.

Un modelo de regresión logística se construye a partir de una base de datos con observaciones para las cuales el evento ya es conocido. Como hemos dicho, la regresión logística es una técnica de dependencia y de elección discreta (la variable objetivo muestra decisiones individuales de respuesta cerrada). Esta variable objetivo o variable dependiente se expresará como la probabilidad de pertenencia al grupo de interés. Las variables independientes introducidas podrán ser métricas o categóricas (Mures y Vallejo, 2018). Es decir, el objetivo del análisis de regresión logística es predecir la probabilidad de ocurrencia de un evento o suceso de interés. Como hemos mencionado previamente, en el caso de nuestro ejemplo práctico, este evento es la existencia de reclamación en una póliza de seguros. Hallar esta probabilidad de ocurrencia, nos permitirá también estudiar el grado de relación que existe entre la variable dependiente y las variables independientes.

Para explicar el funcionamiento de la regresión logística nos apoyaremos en un ejemplo práctico. Como hemos comentado en la introducción, uno de los objetivos de este trabajo es combinar en un mismo estudio la base teórica del análisis de Big Data y ejemplos prácticos sencillos de entender. Se pretende que el lector interesado pueda comprobar por sí mismo cómo se ponen en práctica estas técnicas, y cómo se interpretarían los resultados, a la vez que pueda comprobar las bases teóricas que soportan este tipo de análisis. Para este ejercicio utilizaremos una base de datos publicada por el Catedrático de Universidad de Métodos Cuantitativos José Manuel Pavía Miralles, de la Universitat de Valencia, y continuaremos un ejemplo práctico publicado por él mismo (Pavía, 2016). El enlace a la base de datos se encuentra en la bibliografía. Siguiendo esta

base de datos, nos pondremos en la piel de una empresa reaseguradora. La empresa reaseguradora tiene franquiciadas con un conjunto de mutuas de automóvil las reclamaciones por responsabilidad civil, estas **mutuas** serán los **individuos *i*** de nuestro modelo. En la base de datos, disponemos de datos de 173 mutuas diferentes, con 5 variables de cada una:

Variable **Zona**: Variable categórica, en R un factor con 4 niveles, dependiendo de la zona geográfica de cada mutua. Area1: Nada lluviosa, Area2: Algo lluviosa, Area3: Bastante lluviosa, Area4: Muy lluviosa.

Variable **Tipo**: Variable categórica, en R un factor con rango 1-3 indicando la franja de negocio (además de particulares) donde opera la mutua. 1: Profesionales y Empresas; 2: Profesionales o Empresas; 3: Sólo particulares.

Variable **Volumen**: variable numérica que recoge el volumen de primas percibidas (datos en miles de euros).

Variable **Reclamaciones**: variable numérica que recoge el número de reclamaciones.

Variable **Pólizas**: variable numérica que recoge el número de pólizas suscritas.

Estas variables representan de forma general y resumida, el tipo de variables del que podríamos disponer en un modelo real en una compañía de seguros de automóvil, que podría llegar a ser una lista tremendamente amplia como puede ser: edad, sexo, otras pólizas contratadas, antigüedad de la póliza, tipo de pago, tipo de vehículo, segundo conductor… y así hasta utilizar prácticamente toda la información de la que dispone la compañía.

Se ha mencionado en la introducción de este trabajo, que uno de los objetivos es contribuir a la bibliografía de los modelos predictivos en Big Data. Es por esto, que el ejercicio práctico que se va a realizar se irá explicando al mismo tiempo que la teoría sobre la regresión logística. La intención es que el ejercicio práctico ayude a una mejor comprensión de la teoría, para que el lector pueda ir comprobando según lee las hipótesis y la estimación del modelo, cómo se traducen éstas directamente a casos reales en el sector asegurador y, además, cómo se realizan a través de la herramienta R.

Especificación del modelo de regresión logística binomial

Partimos de la variable dependiente, a partir de ahora Y_i, para comenzar a explicar el modelo. Como hemos mencionado al principio de esta sección, Y_i indicará si ha habido reclamación. Esta variable dependiente será expresada a través de las variables explicativas anteriormente definidas, ahora X_i, que en nuestro ejemplo son 4 variables generales, que comprimen gran cantidad de información cada una (Zona geográfica, Tipo de clientes de la mutua, Volumen de primas recibidas y Pólizas suscritas).

Para modelizar este escenario en *GLM*, la distribución de la variable dependiente se establece por la distribución binomial (ver (McCullagh y Nelder, 1989) para un mejor entendimiento de la relación entre la distribución binomial y los Modelos Lineales Generalizados). La predicción generada por el modelo será la probabilidad de que el evento en cuestión ocurra (valor de Y_i igual a 1). En otras palabras, estimaremos la probabilidad de que haya reclamación, por tanto:

$$Y_i \sim B(\pi_i)$$

donde π_i es la probabilidad de que para la mutua i haya reclamación.

Resumiendo, para cada mutua i, $i = 1, \ldots, n$, el comportamiento de Y_i (hay reclamación, valor de Y_i igual a 1, o no hay reclamación, valor de Y_i igual a 0) sigue una distribución de Bernoulli de parámetro π_i. Este parámetro es distinto para cada individuo y depende de sus características individuales (Pavía, 2016).

Sabemos por la teoría clásica que en los *GLM* la ecuación a seguir sería la siguiente:

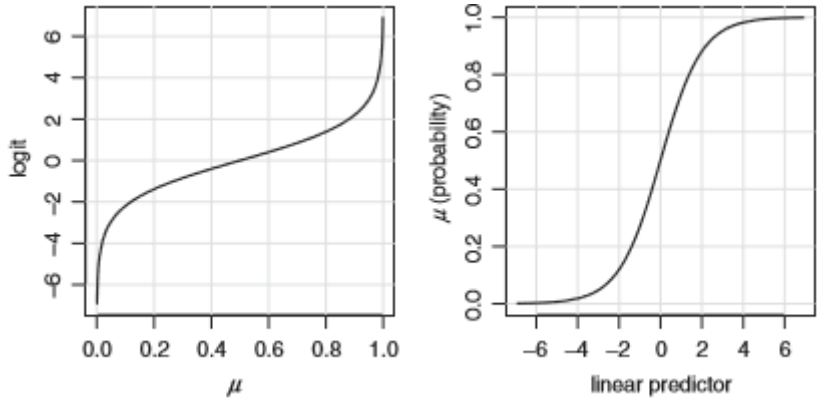
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon \qquad (1)$$

Para modelizar una variable dicotómica mediante distribución binomial, se ha de usar una función enlace (*link function*). Esto es debido a que una de las propiedades básicas de los *GLM* es que el predictor lineal, el segundo miembro en la ecuación (1), puede tomar cualquier valor entre [-∞,∞] (Goldburd, Khare, y Tevet, 2016). La media de la distribución binomial en los casos de regresión logística será una medida de probabilidad, por tanto, ha de encontrarse en el intervalo [0,1]. Necesitamos una función enlace que nos permita satisfacer estas condiciones.

Hay diferentes funciones enlace disponibles para realizar esta terea, la más conocida es la función enlace **logit:**

$$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

Figura 3. 1 Gráficas Logit-Función Logística



Fuente: (Goldburd et al., 2016)

Sobre estas líneas se puede observar la *Figura 3.1*. La gráfica de la parte izquierda representa la función logit. Como se puede ver, la función logit se aproxima a -∞ cuando μ, para nosotros π (el parámetro de la distribución), se aproxima a 0, y toma valores arbitrariamente grandes cuando π se aproxima a 1. La gráfica de la parte de la derecha representa la inversa de la función logit, que llamaremos la función logística, definida como:

$$\pi = 1/(1 + e^{-x}),$$

siendo π la probabilidad de ocurrencia de *Y*.

En *GLM*, está función traslada el valor del predictor lineal a una predicción de probabilidad. Un número negativo grande indicaría una baja probabilidad de ocurrencia del evento, en el caso contrario, indicaría una alta probabilidad de ocurrencia del evento (Goldburd et al., 2016).

Dependiendo de la distribución de probabilidad para el término de perturbación ε existirán diferentes modelos. Si la función utilizada es la logística, como en este caso, obtendremos el modelo **logit**. Otra opción sería tomar como función de distribución una

normal estándar, en ese caso, tendríamos el modelo **probit** (Levy y Varela, 2003). Por tanto, la especificación final del modelo quedaría de la siguiente forma:

$$Y_i \sim B(\pi_i)$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \qquad (2)$$

De aquí extraemos la probabilidad de que el suceso ocurra:

$$\pi_i = P(Y_i = 1) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}} + \varepsilon = \frac{e^{(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}}{1 + e^{(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}} + \varepsilon$$

También, de que no ocurra:

$$P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \pi_i$$

$$1 - \pi_i = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}} + \varepsilon$$

Es importante destacar la expresión *odds* o razón de probabilidades (verosimilitudes):

$$Odds = {}^{\pi_i}/_{1-\pi_i}.$$

Expresa el número de veces que es más probable la ocurrencia del evento (en nuestro caso, que haya reclamación) frente a la no ocurrencia. Sólo presenta valores positivos, sin límite superior. Cuando el valor de las *odds* es superior a 1, nos está indicando que hay una mayor probabilidad de que ocurra el suceso (frente a que no ocurra). Cuando las *odds* son próximas a 1, indica que las dos probabilidades (que haya reclamación y que no la haya) son similares. La ecuación (2) se puede interpretar como el logaritmo de las *odds*. Si aplicamos la forma exponencial a ambos miembros de la ecuación (2), la función logística *GLM* se convierte en una serie multiplicativa de términos que produce las *odds*. Esto nos lleva a poder aplicar una interpretación natural de los coeficientes del GLM, describiendo el efecto del predictor en las *odds*. Por ejemplo, un coeficiente de *0.24* estimado para un predictor continuo *x*, incrementa las odds en

$$e^{0.24} - 1 \cong 27\% \, .$$

Un coeficiente estimado de *0.24* para un nivel dado de una variable categórica indica que las *odds* para ese nivel son un *27%* mayores (Goldburd et al., 2016).

Estimación del modelo y contraste de hipótesis

Antes de comenzar con los resultados del ejercicio, es importante repasar algunas de las hipótesis del modelo de regresión logística (Mures y Vallejo, 2018)

- Ausencia de multicolinealidad. Que exista multicolinealidad puede dar lugar a tener coeficientes y errores elevados, de forma que pueda afectar a la significatividad de los coeficientes. Para su estudio se pueden utilizar medidas como matrices de correlaciones, FIV (factor de inflación de la varianza) o tolerancia (1/FIV).
- Análisis de los residuos y de los residuos logit. Para cada individuo, será la diferencia entre el valor real y el pronosticado. En un modelo logit, se modelizan a través de una distribución binomial. La varianza del error es considerada en función de la media. Los residuos han de ser estandarizados por sus errores típicos con la siguiente expresión:

$$Z_i = \frac{E_i}{\sqrt{\pi(1-\pi)}}$$

- En el modelo logit, a diferencia de otros *GLM*, no se asume relación lineal entre la variable objetivo y las variables independientes, sino que la relación lineal es entre la función enlace y las variables independientes.
- Para estimar los parámetros, no se usan *MCO* (Mínimos Cuadrados Ordinarios), se aplica el método de máxima verosimilitud (ver (Levy y Varela, 2003) (Aldás y Uriel, 2017)).

  La función a seguir es:

$$LL = \sum_{i=1}^{N} [Y_i \ln(P(Y_i)) + (1 - Y_i)\ln(1 - P(Y_i))] \qquad (3)$$

Esta función es una suma de valores que miden aciertos (que serán valores pequeños) y desaciertos (que serán valores grandes). Por tanto, un mayor valor

de la función significará menor capacidad de estimación para replicar los valores reales. Los parámetros β_j de la ecuación (2) serán los que minimicen la función de máxima verosimilitud (Aldás y Uriel, 2017).

El desarrollo de este ejercicio práctico se hace a través de la herramienta R. El código entero sin interrupciones se publica en el Anexo 1.

Para comenzar este ejercicio práctico, abrimos la base de datos mencionada y observamos la información disponible en ella:

```
> DatosMutuas<-read.csv(file="http://www.uv.es/pavia/seguros_no_vida/mutuas.csv",sep=";")
> View(DatosMutuas)
```

El primer paso consiste en hacer una pequeña modificación en la base de datos. Como podemos observar con el comando *levels* la variable *Tipo* viene por defecto en formato numérico, sin niveles. Como hemos explicado al comienzo del capítulo la variable *Tipo* es una variable de tres niveles que categoriza las mutuas según el tipo de clientes que traten.

```
> levels(DatosMutuas$Tipo)
> DatosMutuas$Tipo<-factor(DatosMutuas$Tipo)
> levels(DatosMutuas$Tipo)
```

Ahora ya disponemos de todas las variables en el correcto formato para la realización del ejercicio.

```
> summary(DatosMutuas)
```

.

Tabla 3. 1 Resumen variable regresión logística

Zona	Tipo	Volumen	Claims	Polizas
Area1: 12	Tipo1: 37	Min: 21.0	Min: 0.000	Min: 1200
Area2: 95	1er Cuartil: 15	1er Cuartil: 24.9	1er Cuartil: 0.000	1er Cuartil: 2000
Area3: 44	Media: 121	Media: 26.1	Media: 2.000	Media: 2350
Area4: 22		Mediana: 26.3	Mediana: 2.919	Mediana: 2437
		3er Cuartil.:27.7	3er Cuartil.: 5.000	3er Cuartil.:2850
		Max. :33.5	Max. :15.000	Max. :5200

Fuente: Elaboración propia a partir datos de R

Al introducir la orden *summary* nos encontraremos con una tabla similar a la *Tabla 3.1*. En ella observamos las cinco variables antes mencionadas y los rangos en los que se mueven los valores de éstas.

Comenzaremos con el ejemplo más sencillo. Estudiaremos la probabilidad de que exista reclamación dependiendo del volumen de primas. Es una buena variable para iniciar el proceso, ya que el volumen de primas nos indica la parte del riesgo que ha sido transferido a la reaseguradora por parte de las mutuas. Realizaremos el ejemplo más sencillo y posteriormente explicaremos cómo añadir y realizar el ejercicio con más variables, lo cual no añade ninguna dificultad extra, simplemente un mayor cuidado a la hora de respetar las hipótesis y controlar la significatividad del modelo.

Como queremos estudiar el impacto de si hay reclamación o no, necesitamos crear una nueva variable a partir de la variable *Claims*, que nos indique simplemente si hay reclamación o no, la cual denominaremos **Reclama**. Además, construiremos un *data frame* que contiene todas las observaciones sobre volumen de primas con reclamación o sin ella, eliminando los que tengan frecuencia 0 (es decir, que no estén representados en nuestra base de datos). El objeto *data frame* no es sencillo de traducir al castellano. La traducción literal sería "marco de datos", algo así como una hoja de datos que creamos dentro de R, para agrupar la información con la que queremos trabajar. Tiene una estructura similar a una matriz. Introduciendo los siguientes comandos, habremos conseguido lo escrito en este párrafo.

```
> DatosMutuas$Reclama <- as.numeric(DatosMutuas$Claims > 0)
> DataFrame <- as.data.frame(table(DatosMutuas$Reclama, DatosMutuas$Volumen))
> DataFrame <- as.data.frame(apply((DataFrame[Datos1$Freq != 0, ]), 2, as.numeric))
```

Tabla 3. 2 Data frame Reclama-Volumen

	Var1	Var2	Freq
1	0	21.0	1
2	0	22.0	1
.	.	.	.
.	.	.	.
.	.	.	.
92	1	33.5	1

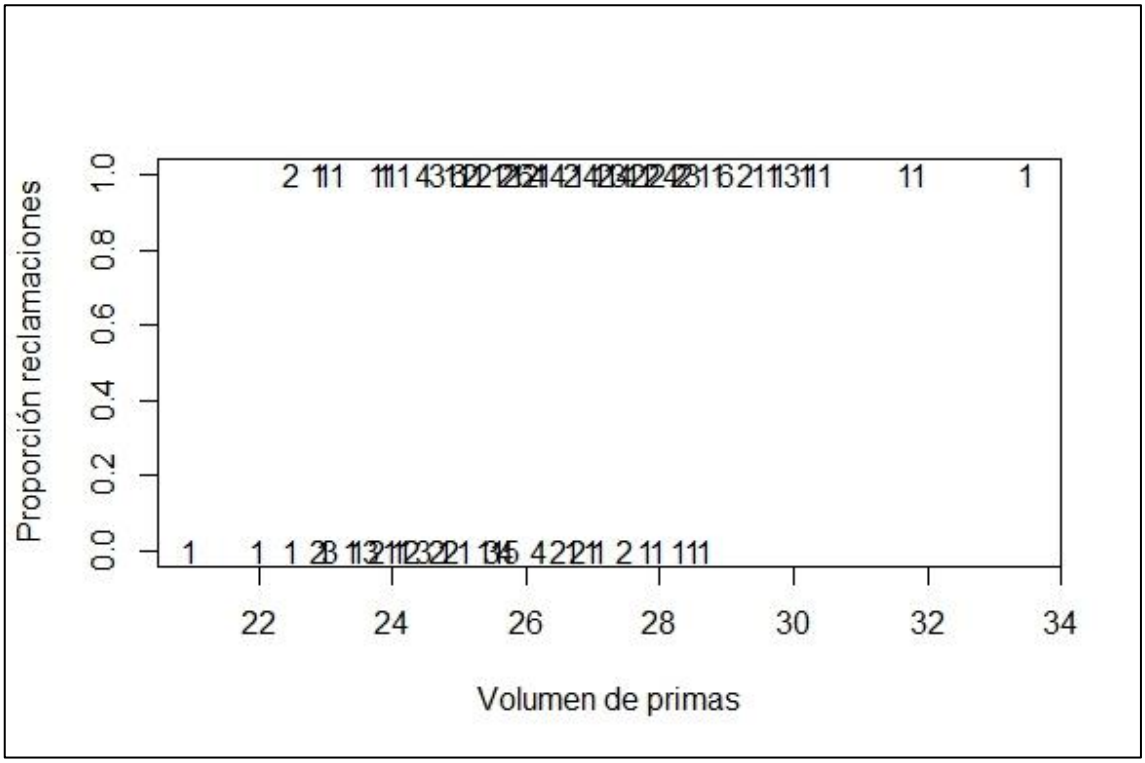
Fuente: Elaboración propia a partir datos de R

Sobre estas líneas podemos observar la estructura del *data frame* que hemos construido. Var1 sería la variable *Reclama* que hemos creado, indicando si existe reclamación. Var2 indicaría el *Volumen* de primas. *Freq* es la frecuencia con la que se repite esta situación. Por tanto, en la primera fila, tendríamos valor 0 para la variable *Reclama*, es decir no hay reclamación, para el volumen de primas valor de 21.0 (miles de

euros) y se repite una vez en toda la base de datos (frecuencia es igual a 1). Para visualizar mejor esta información, vamos a representarla en un gráfico:

```
> with(DataFrame,plot(Var2,Var1,type="n",xlab="Volumen de primas",ylab="Proporción
reclamaciones"))
> points(DataFrame$Var2,DataFrame$Var1,pch=as.character(Datos1$Freq))
```

Gráfico 3. 1 Frecuencia en el data frame Reclama-Volumen



Fuente: Elaboración propia a partir datos de R

Estas instrucciones indican lo siguiente: con el *data frame* creado, realizamos un gráfico donde el Volumen de primas esté en el eje X, la variable Reclama (hay o no reclamación) en el eje Y, y los puntos del gráfico será la existencia o no de reclamaciones, a la altura del volumen que a cada valor del *data frame* corresponda, y el caracter de cada punto en el gráfico será la frecuencia de esa combinación de valores.

Como observamos, todos los puntos del *Gráfico 3.1* se distribuyen en dos líneas, una para el valor 0 en la proporción de reclamaciones (no hay reclamación) y otra en la parte superior para el valor *Reclama* igual a 1. Por ejemplo, el primer "*2*" de la parte superior del gráfico indica que para el volumen *21.0* (miles de euros) de primas, hay dos observaciones en las que hay reclamación. En cambio, perpendicularmente en la parte inferior del gráfico observamos que, para ese mismo volumen de primas (*21000 euros*)

hay una observación que no reclama ("*1*" en el gráfico). Sobre este mismo gráfico, representaremos las líneas del ajuste Logit y también del ajuste Probit (se usa Logit/Probit, con la primera letra mayúscula para referirnos a los ajustes específicos que realizamos en el ejercicio) con la ayuda de R.

Pero, por supuesto, primero hemos de realizar dichos ajustes:

```
> Logit<-glm(Reclama~Volumen,data = DatosMutuas,family=binomial)
> Probit<-glm(Reclama~Volumen,data=DatosMutuas,family=binomial("probit"))
```

A la hora de realizar este paso, recomendamos encarecidamente una lectura de la ayuda que proporciona R sobre estos tipos de ajustes. En "*help*" buscar "*glm*" (R Core Team, 2017).

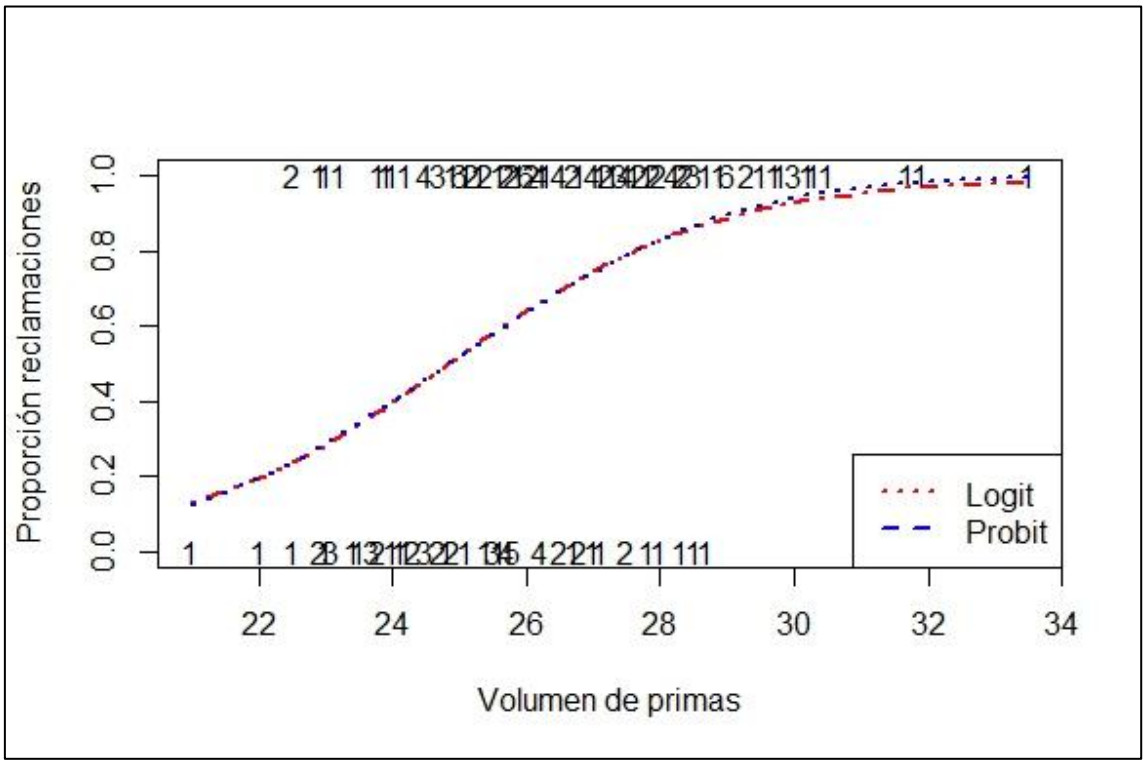
Con las órdenes que se encuentran sobre estas líneas, introducimos la variable dependiente *Reclama*, la variable independiente *Volumen*, la base de datos a utilizar y los dos ajustes que queremos realizar: el ajuste logit, y el ajuste probit.

Y ahora, los representamos encima del *Gráfico 3.1*:

```
> indice<-order(DatosMutuas$Volumen)
> points(DatosMutuas$Volumen[indice],y=Logit$fitted.values[indice],type="l",col="red",lty=4,lwd=2)
> points(DatosMutuas$Volumen[indice],y=Probit$fitted.values[indice],type="l",col="blue",lty=3,lwd=2)
> legend("bottomright",c("Logit","Probit"),col=c("red","blue"),lty=c(3,2),lwd=c(2,2),bty="o")
```

Primero hemos creado el objeto índice, que servirá para representar de forma ordenada (de menor a mayor volumen) los valores ajustados tanto del modelo logit como del modelo probit. Así, la línea de puntos roja representará el ajuste Logit y la línea azul el ajuste Probit.

Gráfico 3. 2 Ajustes Logit-Probit



Fuente: Elaboración propia a partir datos de R

Resumimos ahora los resultados de ambos ajustes.

Tabla 3.3 Resumen resultados Logit

Variable	Coeficiente	Error Std	Z valor	P-Valor
(Intercepto)	-12.3508	2.6287	-4.698	2.62e-06 ***
Volumen	0.4972	0.1017	4.887	1.02e-06 ***

Fuente: Elaboración propia a partir datos de R

Tabla 3.4 Resumen resultados Probit

Variable	Coeficiente	Error Std	Z valor	P-Valor
(Intercepto)	-7.5020	1.5071	-4.978	6.44e-07***
Volumen	0.3020	0.0580	5.204	1.95e-07 ***

Fuente: Elaboración propia a partir datos de R

Nuestro modelo Logit, siguiendo la ecuación (2), quedaría de la siguiente manera:

$$\ln\frac{\pi_i}{1-\pi_i} = -12.3508 + 0.4972X_i$$

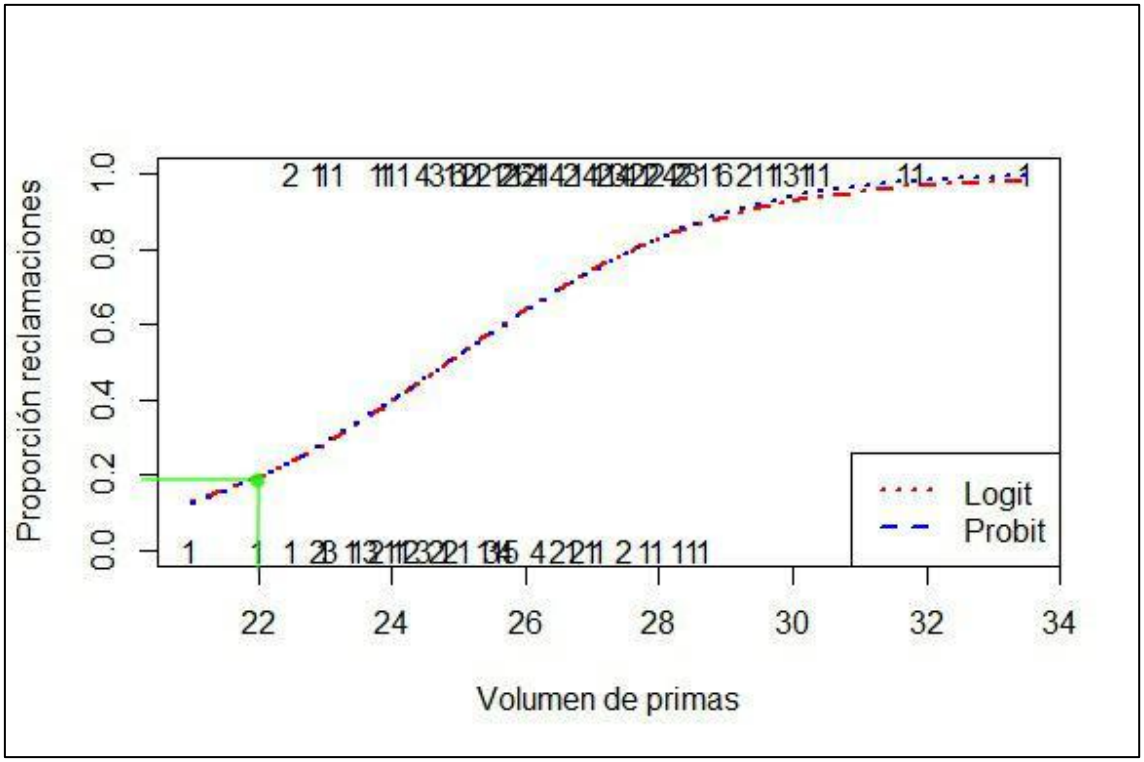
Introduciendo un valor para la variable X_i, véase 22.0 (miles de euros), de la siguiente manera obtendremos la probabilidad de reclamación π_i:

$$-12.3508 + 0.4972(22.0) = -1.4124$$

$$\frac{e^{-1.4124}}{1+e^{-1.4124}} = 0.1958558\,.$$

Esto significa que, para una mutua con un volumen de primas de *22000 euros*, la probabilidad de que haya reclamación es de *0.195856*. Estaríamos hablando del siguiente punto en el *Gráfico 3.2*:

Gráfico 3. 3Ajustes Logit-Probit con marca



Fuente: Elaboración propia a partir datos de R

El siguiente paso es realizar un **contraste de significatividad global del modelo**. Para ello, primero contrastaremos la hipótesis nula de que todos los coeficientes de regresión son nulos. Para ello, se calcula la máxima verosimilitud si la/s variable/s no jugaran ningún papel (*LL(0)*) y la máxima verosimilitud del modelo (*LL(M)*). Siguiendo la ecuación (3). Si el segundo valor es significativamente más pequeño que el primero, querrá decir que al menos alguna de las variables que se introducen está influyendo de

forma significativa en la predicción de nuestra variable dependiente. Debido a que para saber si esta diferencia es significativa, se precisa que siga una distribución conocida, y la función *LL* no cumple esta condición, hemos de realizar el cálculo con -2 veces esta función (*-2LL*), que se distribuye como una χ^2 con ($k_M - k_0$) grados de libertad (Aldás y Uriel, 2017), donde k_M representa los parámetros a estimar en el modelo y k_0 los parámetros a estudiar en el modelo si sólo existiera la constante.

La información que necesitamos para realizar esto la obtenemos a través de los siguientes comandos:

```
> Deviance.Modelo<-Logit$deviance
> Deviance.Base<-Logit$null.deviance
> chi<-Deviance.Base-Deviance.Modelo
> chi.df<-Logit$df.null-Logit$df.residual
> sig.chi<-1-pchisq(chi,df=chi.df)
```

En R es común utilizar el término ***deviance***. Hace referencia a la razón de máxima verosimilitud *-2LL*, explicada en el párrafo anterior. Por tanto, lo que hemos hecho es, con los datos del modelo que antes hemos realizado, obtener la razón de máxima verosimilitud para el modelo (*-2LL*(*M*)) y la base (*-2LL*(*0*)). La diferencia será nuestra variable χ^2. Obtenemos también los grados de libertad, que como sabemos en nuestro ejemplo serán igual a 1 (sólo hemos introducido una variable) y ya podemos analizar la significatividad global de nuestro modelo:

Deviance.Modelo	Deviance.Base	χ^2	Grados de libertad	Significatividad
194.4527	225.7585	31.30586	1	2.2041e-08***

Con estos datos, podríamos decir que rechazamos la hipótesis nula de que las dos razones de máxima verosimilitud (*-2LL*) sean iguales. Como la del modelo es más pequeña, es significativamente mejor. De esta forma, descartamos que todos los coeficientes sean nulos.

Pasamos ahora a estudiar la significatividad de los coeficientes individuales. Para comparar la significatividad de los coeficientes se usa el estadístico de Wald (W_k). La hipótesis nula es H_0: $\beta_k = 0$ y el estadístico de Wald:

$$W_k = \left(\frac{\beta_k}{S_{\beta_k}}\right)^2 \approx \chi^2$$

Para ello, no necesitamos realizar ninguna acción añadida, nos fijamos en la *Tabla 3.3*, donde tenemos el resumen de los resultados obtenidos en nuestro modelo. En la última fila tenemos la estimación del coeficiente no estandarizado β para el modelo,

vemos que es significativo a un nivel de α igual a *0.01*, (β*=0.4972,* W_x *=4.887* $p<0.01$). En consecuencia, rechazamos la hipótesis nula y concluimos que la variable independiente influye en la predicción de la probabilidad de la variable dependiente.

A la hora de interpretar los coeficientes de la regresión es importante hacer una matización. La poca frecuencia con la que se usa el término *odds* en castellano, puede llevarnos a realizar una interpretación errónea de los coeficientes. Hemos definido previamente el concepto de *odd*s y ahora definiremos el término *odds* ratio. Las *odds* ratio juegan el papel de los coeficiente estandarizados β_i en la regresión lineal, su utilidad es ver la contribución relativa de cada variable independiente (Aldás y Uriel, 2017). Se conoce como *odd* ratio al término $e^{\beta i}$. Para nuestro ejemplo, $e^{0.4972} = 1.644$. Esto **no quiere** decir que un incremento de una unidad en la variable independiente provoque un aumento del *64.4%* en la variable dependiente, provoca un incremento del *64.4%* en las *odds*. Trasladar este dato a términos de probabilidad de ocurrencia del suceso es sencillo: si en nuestro ejemplo nos encontramos con un valor de la variable *X* (volumen) de 22.0 (miles de euros), tenemos una probabilidad de ocurrencia de *0.195856* (ejemplo que hemos puesto en la página 29). Veamos lo que pasa ahora si incrementamos el valor de la variable independiente de *22.0* a *23.0*, es decir, un aumento de una unidad. Las *odds* iniciales son *0.195856, odds ratio* es *1.644*, por tanto, las *odds* nuevas son *0.195856*1.644=0.3220*. La nueva probabilidad de ocurrencia, será:

$$0.3220/(1+0.3220)= 0.2435753$$

Resumiendo, al pasar *X* del valor 22.0 al valor 23.0, la probabilidad de que ocurra el evento ha subido un

$$((0.2435753-0.195856)/0.195856)= 0.2436448,$$

esto es, en un 24.36%.

Para comprobar el ajuste del modelo global hay diferentes coeficientes y técnicas que utilizar. Algunas son: G razón de verosimilitud o χ^2, Coeficiente R^2 de Cox y Snell, Pseudo-R^2… Calculamos dos de ellas de forma breve

El Pseudo-R^2 es la proporción de la reducción de la *deviance* que incorpora el modelo, respecto al modelo sin predictor línea (Aldás y Uriel, 2017):

$$R^2_{MF} = \frac{-2LL(0) - (-2LL(M))}{-2LL(0)},$$

$$R^2_{MF} = \frac{225.7585 - (194.4527)}{225.7585} = 0.1387.$$

Con el R^2 de Cox y Snell, queremos estimar qué parte de la varianza de la variable dependiente explica el modelo, comparando los logaritmos de la función de máxima verosimilitud base con el del modelo. Este resultado nos indica que el *16.55%* de la varianza de la variable dependiente es explicada por el modelo:

$$R^2_{CS} = 1 - e^{\left[\frac{1}{N}(2LL(M) - 2LL(0)\right]},$$

$$R^2_{CS} = 1 - e^{\frac{1}{173}(194.4527 - 225.7585)} = 0.1655299\,.$$

Ambos valores obtenidos para evaluar el ajuste del modelo, son resultados muy bajos para un modelo fiable. Esto es debido a que únicamente hemos introducido una variable para explicar el modelo, y hay una gran parte de la varianza de la variable dependiente que queda por explicar.

Otra opción para evaluar la precisión de las estimaciones son las matrices de confusión. A continuación, comprobamos cuántos casos hemos acertado y cuántos hemos errado en nuestro modelo. Es una forma mucho más visual, que encaja mejor con este tipo de regresión que los coeficientes calculados anteriormente. Para ello, introducimos los siguientes comandos:

```
> Predict.Ajuste<-Logit$fitted.values
> Predict.Ajuste[Predict.Ajuste>=.50]<-1
> Predict.Ajuste[Predict.Ajuste<.50]<-0
> CrossTable(DatosMutuas$Reclama,Predict.Ajuste,prop.chisq=FALSE,prop.c = FALSE,prop.r=FALSE)
```

Obtendremos una tabla similar a la que se encuentra bajo estas líneas (*Tabla 3.5*). En ella, tendremos por un lado los valores predichos por el modelo, y por el otro los valores reales de la base de datos. Lo que hacemos en comparar las veces que acierta y las que falla el modelo. A simple vista vemos que acierta en *(27+95) =122* casos, que representan un *70.52%* de los casos totales. Pero no nos podemos quedar aquí, ya que en este tipo de modelos no todos los errores son iguales.

Tabla 3.5 Matriz de confusión

		Predicciones		
		0(No Reclama)	1(Reclama)	Total Fila
Valores Reales	0(No Reclama)	27	35	62
	1(Reclama)	16	95	111
	Columna total	43	130	173

Fuente: Elaboración propia a partir datos de R

De las matrices de confusión podemos extraer mucha información. Pueden ser incluso indicadores de la capacidad discriminante del modelo. La capacidad discriminante de un modelo se considera en muchas ocasiones incluso más importante que el ajuste que puedan indicarnos los distintos R^2 (Aldás y Uriel, 2017). Definimos los siguientes términos:

Falsos positivos: % de negativos clasificados como positivos *(35/62): 56.54%*.

Falsos negativos: % de positivos clasificados como negativos *(16/111):14.41%*.

Sensibilidad: % de positivos que se clasifican como positivos (95/111): *85.59%*.

Especificidad: % de negativos que son clasificados como negativos (27/62): *43.46%*.

Es importante distinguir entre la diferencia de un tipo de error u otro. En el caso de una reclamación puede ser menos claro, pero en el caso de detección de una enfermedad, por ejemplo, no es el mismo error predecir un positivo (existe enfermedad) que luego no lo sea, que predecir un negativo (no hay enfermedad) y que luego sea un positivo. El primer error es mucho más grave. Ahora que conocemos estos términos, podemos hablar de la curva ROC. La curva ROC compara, para distintos puntos de corte de la probabilidad, cual es la tasa de clasificaciones correctas (sensibilidad) y la de falsos positivos (1-especifidad) (Aldás y Uriel, 2017). El ejemplo que hemos calculado sobre estas tasas ha sido suponiendo que con valor de probabilidad igual o superior a 0.5 se consideraba predicción de reclamación (valor de Y=1) y con menos de un 0.5 no había reclamación. La curva ROC hace esta predicción para multitud de diferentes puntos de corte. Con esta curva, más el estadístico c asociado a las curvas ROC (el área por debajo de la curva), tendremos suficiente información para comprender la capacidad discriminante de nuestro modelo. Para ello, tenemos los siguientes comandos, siguiendo las pautas de (Grzasko, 2016):

```
> Positivos.Ajuste <- Logit$fitted.values[DatosMutuas$Reclama== 1]
> Negativos.Ajuste <- Probit$fitted.values[DatosMutuas$Reclama== 0]
```

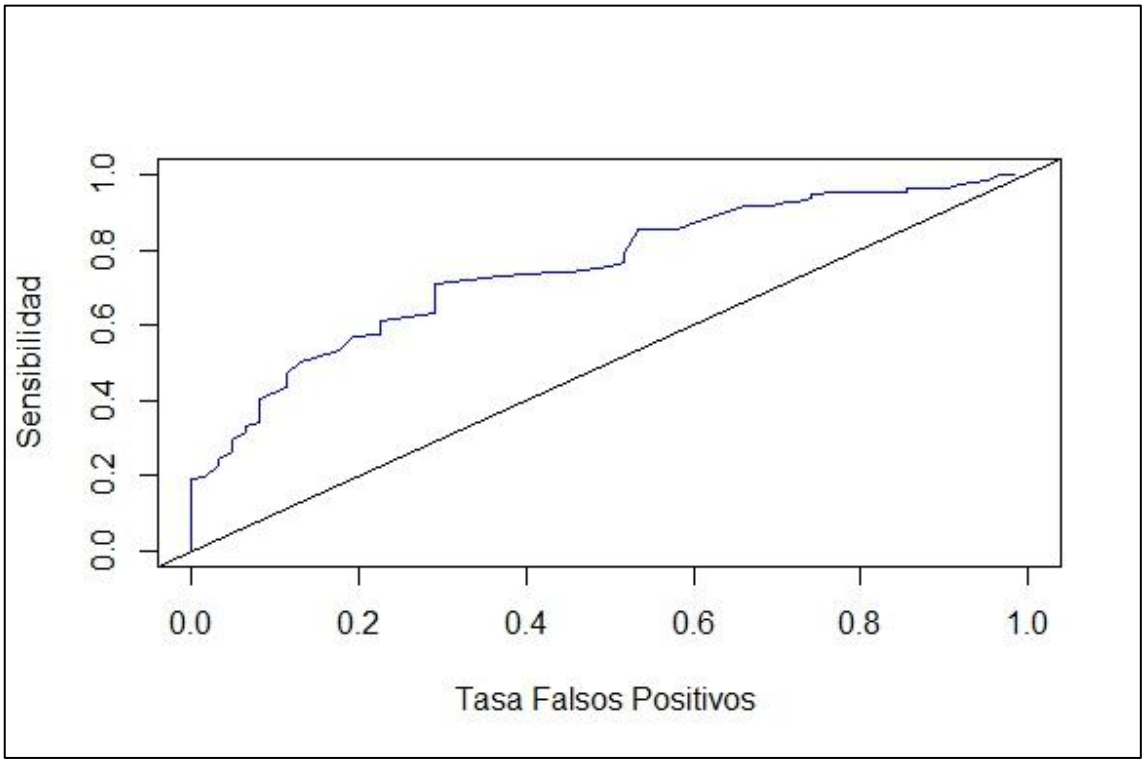
```
> Orden.Ajuste <- sort(Logit$fitted.values)
> Sensibilidad <- 0
> TasaFalsosPositivos <- 0
> for (i in length(Orden.Ajuste):1){Sensibilidad <- c(Sensibilidad, mean(Positivos.Ajuste >= Orden.Ajuste
[i]))
> TasaFalsosPositivos <- c(TasaFalsosPositivos, mean(Negativos.Ajuste >= Orden.Ajuste[i]))}
> plot(TasaFalsosPositivos, Sensibilidad, xlim = c(0, 1), ylim = c(0, 1), type = "l",
    xlab = " TasaFalsosPositivos ", ylab = "Sensibilidad", col = 'blue')
> abline(0, 1, col= "black")
```

Existen diferentes opciones mucho más sencillas para representar la curva ROC, por ejemplo, con la ayuda de los paquetes Epi, o pRoc (Carstensen, Plummer, Laara, y Hills, 2018; Robin et al., 2011), los cuales la representan de forma automática, y nos indican el valor del área debajo de la curva. Sin embargo, esta manera es muy visual, ya que nos permite entender perfectamente cómo se realiza esta curva. Con estos comandos estamos indicando que para cada valor de probabilidad nos represente el valor sensibilidad/tasa de falsos positivos que tendríamos. Una opción mucho más simple de representar la curva ROC, si los comandos anteriores dieran problemas sería:

```
> Prediccion = prediction(Logit$fitted.values, DatosMutuas$Reclama)
> roc = performance(Prediccion, measure="tpr", x.measure="fpr")
> plot(roc, col="blue",xlab = "Tasa Falsos Positivos ", ylab = "Sensibilidad", lwd=2)
> lines(x=c(0, 1), y=c(0, 1), col="black", lwd=2)
```

Obtendremos de cualquiera de las dos formas, el siguiente gráfico:

Gráfico 3.4 Curva ROC



Fuente: Elaboración propia a partir datos de R

Para calcular el área debajo de la curva (estadístico c asociado a las curvas ROC) usaremos lo siguiente:

```
> library(pROC)
> roc_obj<-roc(DatosMutuas$Reclama,Logit$fitted.values)
> auc(roc_obj)
```

Nos indicará finalmente que el área debajo de la curva es igual a *0.7424*. Este estadístico varía entre 0.5 y 1. Cuando el valor es 0.5 significa que el poder de predicción del modelo es nulo, ya que está realizando predicciones aleatorias. Por el contrario, cuando el valor es 1, estaríamos ante el clasificador perfecto, la tasa de falsos positivos sería igual a 0 y la verdaderos positivos 1, no fallaría nunca. Cuanto mayor sea el área debajo de la curva, mejor clasificador es el modelo (Aldás y Uriel, 2017).

Si se tuviera un tamaño muestral muy elevado, deberíamos calibrar el modelo al respecto. Se puede seguir el procedimiento propuesto por (Hosmer y Lemeshow, 2000) y la explicación en (Aldás y Uriel, 2017) sobre cómo realizarlo en R.

Este modelo que hemos realizado tiene un único predictor, para mayor sencillez a la hora de explicarlo. Por supuesto, estos modelos pueden realizarse utilizando varios predictores. El siguiente es un ejemplo incluyendo la variable Zona:

Tabla 3.6 Resumen resultados Logit varios predictores

Variable	Coeficiente	Error	P-Valor
(Intercepto)	-12.3508	2.6287	-4.698 2.62e-06 ***
Volumen	0.4972	0.1017	1.95e-07 ***
ZonaArea2	0.0724	0.7399	0.922
ZonaArea3	-0.2238	0.7770	0.773
ZonaArea4	-1.330	0.8525	0.119

Fuente: Elaboración propia a partir datos de R

A simple vista, con el cuadro resumen nos damos cuenta de que, para esta base de datos, no tiene sentido incluir la variable zona a la variable volumen para el ajuste Logit, debido a la significatividad de los coeficientes. Destacar que, al introducir una variable categórica de cuatro niveles, R lo que hace es descomponerla en tres variables *dummies* por lo que hemos de interpretar las variables introducidas en el modelo fijándonos en este aspecto. Por ejemplo, en este caso tenemos signos negativos en los coeficientes de Zona 3 y Zona 4, lo que nos indica que pertenecer a estas dos zonas reduce la probabilidad de que haya reclamación respecto a la Zona 1 (la zona omitida en el modelo). En cambio, Zona 2 tendría un signo del coeficiente positivo, lo que se traduce en menor probabilidad de que haya reclamación que en la Zona 1.

Hemos visto la simplicidad de añadir predictores a la ecuación, la siguiente pregunta sería ¿Cómo elegimos qué variables incluir? ¿Qué modelo seleccionamos?

La respuesta se obtiene comparando las distintas combinaciones posibles de modelos. Volvemos a hacer el ajuste Logit para las combinaciones que deseemos:

```
> Logit3<-glm(Reclama~Volumen+Zona+Tipo, data=DatosMutuas, family=binomial)
> Logit4<-glm(Reclama~Volumen+Zona*Tipo, data=DatosMutuas, family=binomial)
```

Tabla 3.7 Comparación de modelos

Modelo	Predictores	Desviación Residual	Grados de libertad	AIC
Logit1	Volumen	194.45	171	198.50
Logit2	Volumen+Zona	187.46	168	197.50
Logit3	Volumen+Zona+Tipo	186.61	166	200.60
Logit4	Volumen+Zona*Tipo	177.61	160	203.60

Fuente: Elaboración propia a partir datos de R

A la hora de decidir con qué modelo quedarnos, debemos ver la *deviance* y el criterio de información de Akaike (*AIC*), que en Regresión Logística sería el equivalente al

coeficiente de determinación R^2 ajustado en regresión lineal simple. Cuanto menor sea este valor, mejor será para el modelo. Este criterio no tiene un rango de valores sobre el que valorar un modelo, sino que se utilizara para comparar los valores del criterio entre diferentes modelos, con el objetivo de ver qué modelos incorporan más información al modelo, teniendo en cuenta los parámetros introducidos. Para mayor información al respecto, ver (Akaike, 1974) .

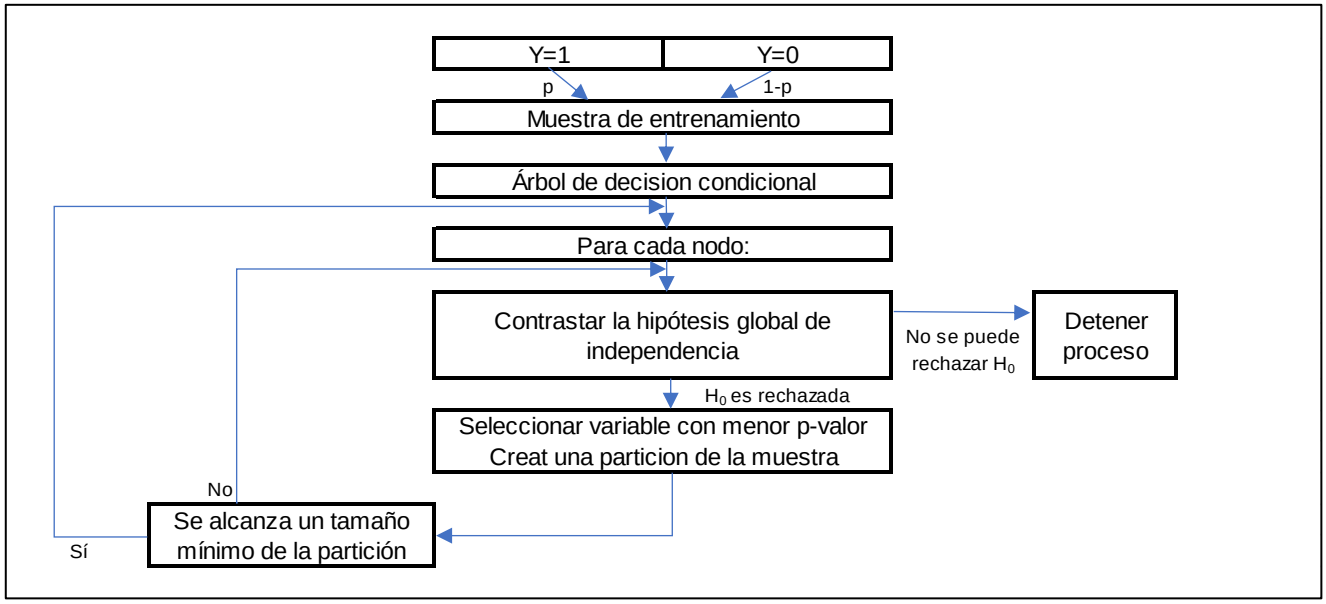
3.2.2 Árboles de decisión condicionales

Los Árboles de decisión condicionales, denominados en inglés *Conditional Tree*, (en adelante *CTREE),* son una clase no paramétrica de árboles de regresión que integran modelos de regresión con estructura de árbol de decisión en una teoría bien definida de procesos de inferencia condicionales. Son aplicables a todo tipo de problemas de regresión incluidas las variables de respuesta nominal, ordinal, numérica así como variables de respuesta multivariante (Hothorn y Hornik, 2006).

Los árboles de decisión nacen por la necesidad de incorporar la información de las distribuciones de probabilidad de las variables al particionamiento recursivo en Estadística. Como hemos mencionado en la introducción del apartado, estas técnicas son potentes herramientas para clasificar, agrupar y visualizar gran cantidad de datos. Se trata de una herramienta que sirve para tratar dos grandes problemas en la investigación: la predicción y la explicación. Los *CTREE* son un tipo especial de árbol de decisión donde se seleccionan variables en dos fases. Primero se estudia la dependencia entre la variable objetivo y las variables independientes planteando una hipótesis global de independencia en términos de *m* hipótesis parciales (Padilla-Barreto et al., 2017). Si la hipótesis nula de independencia no se puede rechazar, se detiene el particionamiento recursivo (observar *Figura 3.3*). En cambio, si es rechazada, la segunda fase sería medir cómo se asocian y en qué grado la variable dependiente y cada una de las variables independientes.

Bajo estas líneas se presenta la *Figura 3.3,* como ayuda al lector a la hora de visualizar el proceso que sigue un árbol de decisión condicional.

Figura 3. 3 Algoritmo para árbol de decisión condicional



Fuente: Elaboración propia a partir de (Guelman, Guillén, y Pérez Marín, 2014)

Para el ejemplo práctico de esta sección se utiliza la base de datos "car.csv" original de (De Jong y Heller, 2008), a la que se puede acceder de forma libre. La base de datos contiene *67856* pólizas de auto de una aseguradora australiana correspondientes al año 2005. Las diez variables en la base de datos son las siguientes:

Variable **ValorAuto**: variable numérica, que representa el valor del vehículo de la póliza en cuestión, expresada en $10,000 (australianos).

Variable **Exposicion**: variable numérica entre 0 y 1, que mide la exposición al riesgo adjudicada a la póliza

Variable **Reclama**: variable categórica, de valor igual a 0 si no hay reclamación o valor igual a 1 si la hay.

Variable **NumReclamaciones**: variable numérica que recoge el número de reclamaciones.

Variable **ValorReclamacion**: variable numérica que recoge el valor total de las reclamaciones, si existieran (si no hay reclamación, valor igual a 0).

Variable **TipoAuto**: variable categórica, en R un factor con 13 niveles, que recoge las categorías de autos que se consideran: *BUS*= autobús, *CONVT* = convertible, *COUPE*, *HBACK* = hatchback, *HDTOP* = capota rígida, *MCARA* = caravana, *MIBUS* = minibús, *PANVN* = furgoneta, *RDSTR* = descapotable biplaza, *SEDAN*, *STNWG* = familiar, *TRUCK*= camioneta, *UTE* = utilitario.

Variable **AntiguedadAuto**: variable categórica, en R un factor de 4 niveles, que agrupa los automóviles por antigüedad, siendo el valor igual a 1 para los automóviles nuevos y el valor igual a 4 los automóviles más antiguos.

Variable **Sexo**: variable categórica, en R un factor de 2 niveles, que divide a los conductores por género, siendo F las mujeres y M los hombres.

Variable **Zona**: Variable categórica, en R un factor con 6 niveles, dependiendo del área de residencia del conductor. Los valores son *A,B,C,D,E,F*.

Variable **GrupoEdad**: Variable categórica, en R un factor con 6 niveles, que clasifica a los conductores según su edad, siendo el valor igual a 1 el grupo más joven y el valor igual a 6 el grupo de mayor edad.

Para disponer de las variables con estos nombres, primero debemos modificar la base de datos. Usamos los siguientes comandos para abrir la base de datos y nombrar las variables según se propone. Como en el ejemplo de la regresión logística (Sección *3.2.2*), debemos cambiar el formato a "*factor*" para algunas variables, que por defecto vienen en formato numérico.

```
> DatosPolizasAuto = read.csv("car.csv")
> colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="veh_value"] <- "ValorAuto"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="exposure"] <- "Exposicion"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="clm"] <- "Reclama"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="numclaims"] <- "NumReclamaciones"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="claimcst0"] <- "ValorReclamacion"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="veh_body"] <- "TipoAuto"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="veh_age"] <- "AntiguedadAuto"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="area"] <- "Zona"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="gender"] <- "Sexo"
>colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="agecat"] <- "GrupoEdad"
>View(DatosPolizasAuto)
```

Ahora que ya disponemos de los datos que necesitamos y los tenemos bien identificados y en su correcto formato, podemos comenzar con el ejercicio. El primer paso será dividir la muestra en dos partes. Como se ha explicado en la Sección *3.2* los *CTREE* se emplean bajo el supuesto de aprendizaje supervisado, por este motivo dividiremos la base de datos en dos muestras, utilizaremos el 70% para entrenamiento (*training set*) y el 30% para prueba (*test set*). Construiremos el árbol de decisión condicional con el 70% de la muestra y comprobaremos su capacidad de predicción con el 30% restante.

Para ello, utilizamos el paquete "caret" (Kuhn, 2018) y con los siguientes comandos realizamos la partición de la muestra:

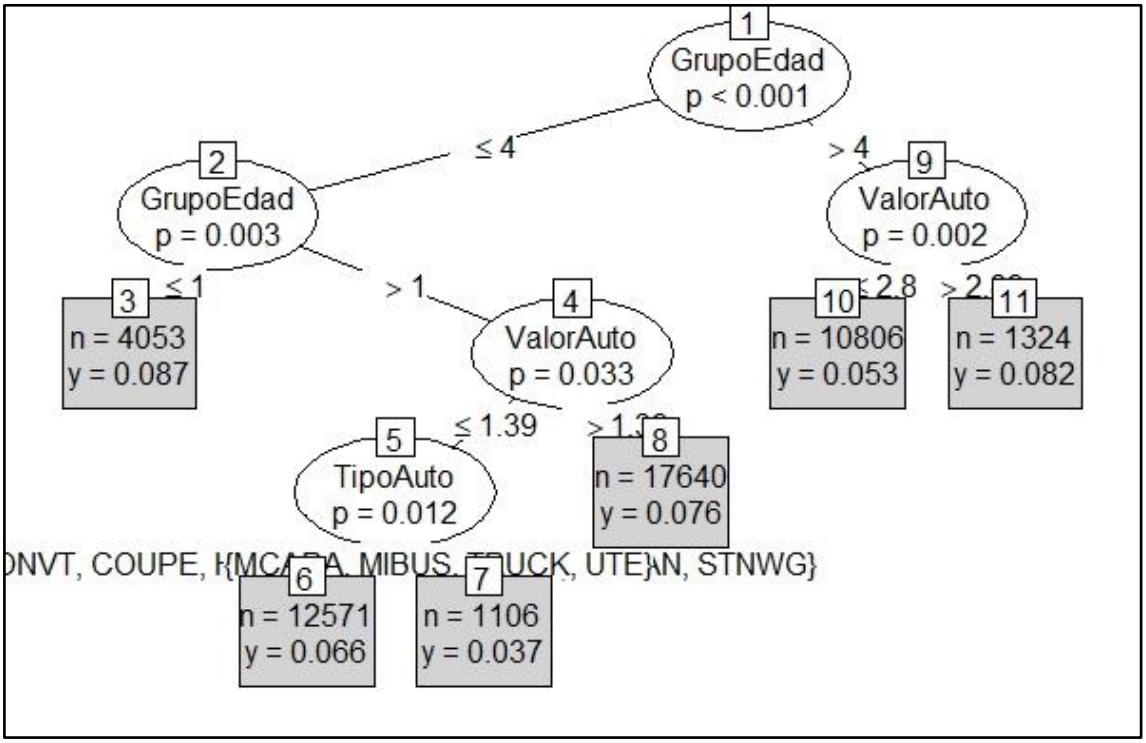
```
>IndiceValidez <- createDataPartition(DatosPolizasAuto$Reclama, p=0.70, list=FALSE)
>MuestraPrueba <- DatosPolizasAuto[-IndiceValidez,]
>MuestraEntrenamiento <- DatosPolizasAuto[IndiceValidez,]
```

En *MuestraEntrenamiento* ya disponemos de nuestro *training set*, ahora necesitamos especificarle al programa la fórmula que queremos que siga y ejecutar el árbol de decisión condicional. Con el comando *Fmla*, estamos indicando que queremos que la variable Reclama sea la variable dependiente (también variable respuesta) y el resto de variables como *inputs* siendo las variables independientes

```
>Fmla = Reclama ~ ValorAuto + TipoAuto + AntiguedadAuto + Sexo + Zona + GrupoEdad
>ModeloArbol= ctree(Fmla, data = MuestraEntrenamiento)
>ModeloArbol
>summary(ModeloArbol)
```

Nos encontraremos con un gráfico similar al que se presenta bajo estas líneas (*Gráfico 3.5*). La interpretación es similar a los árboles de decisión tradicionales. Para aquellas personas que no estén familiarizadas con el formato de árbol de decisión, se presentan dos tablas (*Tabla 3.8 y Tabla 3.9*) donde se resume la información obtenida en el modelo.

Gráfico 3. 5 Árbol de decisión condicional 1



Fuente: Elaboración propia a partir de datos de R

Denominaremos **nodos** a los números del gráfico enmarcados en un cuadro blanco. Los nodos son los puntos en los que el árbol divide la base de datos. Se denomina nodo raíz al primer nodo que divide la base de datos en dos subgrupos, en el *Gráfico 3.5* el número 1. Los nodos definitivos son los que tienen un cuadro gris debajo, y son los que confeccionan los últimos subgrupos del árbol y en los que nos hemos de fijar. Se recomienda comprobar con la *Tabla 3.9* en caso de duda sobre el subgrupo que genera un nodo en concreto. Los árboles de decisión condicionales, al igual que los árboles de decisión tradicionales se leen de arriba abajo. De esta forma, observamos que efectivamente, es el número 1 el primer nodo o nodo raíz del árbol. Es la primera clasificación que se realiza en la muestra de entrenamiento, y como vemos, comienza dividiendo la muestra con la variable GrupoEdad, que recordamos era una variable categórica que agrupaba a los conductores en seis grupos ordenados de menor a mayor edad. Por tanto, la primera **clasificación** que realiza el árbol, es entre los conductores que pertenezcan a los grupos 1,2,3 y 4 de la variable GrupoEdad y los que, al contrario, pertenezcan a los grupos 5 y 6. Del nodo raíz, salen otros dos nodos, los números 2 y 9. El nodo número 9 divide su población según el valor del automóvil, diferenciando entre los que poseen un automóvil con valor mayor o igual a 2.89 y los que no. Seguidamente llegamos a los dos primeros nodos finales o definitivos, el número 10 y el número 11. Los nodos finales están en la *Tabla 3.8* entre paréntesis en color amarillo. Son importantes ya que son dos de los subgrupos finales que nos propociona el modelo. El nodo número 10 cuenta con 10806 pólizas, y una probabilidad de reclamación de *0.053 (y=0.053)* en la *Tabla 3.8*. De la misma manera se interpretan el resto de subgrupos. Se denominará como subgrupo de Tipo de Auto 1: BUS, CONVT, COUPE, HBACK, HDTOP, PANVN, SEDAN, STNWG, y subgrupo de Tipo de Auto 2: MCARA, MIBUS, TRUCK, UTE.

Tabla 3. 8 Árbol de decision Condicional 2

```
Variable respuesta:  Reclama
Inputs:  ValorAuto, TipoAuto, AntiguedadAuto, Sexo, Zona, GrupoEdad
Numero de observaciones:  47500

1. GrupoEdad <= 4
     2. GrupoEdad <= 1
                    (3)*  casos = 4053, y=0.087
     2. GrupoEdad > 1
           4. ValorAuto <= 1.39
                5. TipoAuto == {BUS, CONVT, COUPE, HBACK, HDTOP, PANVN,
                                                                SEDAN, STNWG}
                    (6)*  casos = 12571, y=0.066
                5. TipoAuto == {MCARA, MIBUS, TRUCK, UTE}
                    (7)*  casos = 1106, y=0.037
           4. ValorAuto > 1.39
                    (8)*  casos = 17640, y=0.076
1. GrupoEdad > 4
     9. ValorAuto <= 2.89;
                    (10)*  casos = 10806, y=0.053
     9. ValorAuto > 2.89
                    (11)*  casos = 1324, y=0.082
```

Fuente: Elaboración propia a partir de datos de R

Como vemos, el modelo adjudica una probabilidad a cada subgrupo. Como en la fórmula que hemos introducido queremos que sea la variable *Reclama* nuestra variable objetivo (*response* en las salidas de R), la probabilidad que adjudica el árbol a cada subgrupo es la probabilidad de reclamación para ese subgrupo. Como vemos, lo que hace el árbol es clasificar las observaciones en subgrupos y adjudica una probabilidad a cada subgrupo, por lo que, si escogemos una observación al azar, simplemente sabiendo el subgrupo al que pertenece, conoceremos si hay una mayor o menor probabilidad de que reclame. Por este motivo se considera a los árboles de decisión condicionales poderosas herramientas de clasificación (Guelman et al., 2014).

Tabla 3.9 Resumen Nodos CTREE

Subgrupo	Nodo	Casos	Probabilidad de reclamación
Conductores del GrupoEdad = 1	(3)	4053	y=0.087
Conductores del GrupoEdad= 2,3,4 + Valor de Auto ≤ 1.39 + Grupo Tipo de Auto 1	(6)	12571	y =0.066
Conductores del GrupoEdad = 2,3,4 + Valor de Auto ≤ 1.39 + Grupo Tipo de Auto 2	(7)	1106	y =0.037
Conductores del GrupoEdad = 2,3,4 + Valor de Auto > 1.39	(8)	17640	y =0.076
Conductores del GrupoEdad > 4 + valor de Auto ≤ 2.89	(10)	10806	y =0.053
Conductores del GrupoEdad > 4 + valor de Auto > 2.89	(11)	1324	y =0.082

Fuente: Elaboración propia a partir de datos de R

Ahora que tenemos los resultados del modelo, al igual que con la regresión logística, debemos comprobar la capacidad de predicción del mismo. Para ello, utilizaremos los datos de prueba (el *test set*), o como los hemos denominado en el ejercicio *MuestraPrueba*. Lo que haremos será utilizar los valores predichos por el modelo con la muestra de entrenamiento y comprobar su acierto en la muestra de prueba. Lo conseguimos con la siguiente lista de comandos:

```
>pred <- predict(ModeloArbol, newdata=MuestraPrueba)
>pred[pred>=.06]<-1
>pred[pred<.06]<-0
>CrossTable(MuestraPrueba$Reclama,pred,prop.chisq=FALSE,prop.c = FALSE,prop.r=FALSE)
```

Primero generamos un objeto que recoja los valores predichos por el modelo. Al igual que en la regresión logística, establecemos un valor, en este caso de 0.06, a partir del cual se clasificará el valor predicho como reclamación (valor igual a 1). Si está por debajo de este nivel se tomará el valor igual a 0, es decir, no hay reclamación. Se fija en 0.06 el valor tras probar diferentes opciones, teniendo en cuenta sobre todo la proporción de reclamaciones en la muestra inicial. Se recomienda probar diferentes valores para observar cómo cambia la matriz de confusión y la curva ROC. Por último, creamos la

matriz de confusión de la misma forma que en el ejemplo de regresión logística (Sección *3.2.2*), comparando valores reales de la muestra de prueba con los valores predichos por el modelo en la muestra de entrenamiento. Como podemos apreciar en la esquina inferior derecha de la *Tabla 3.10*, las pólizas totales de la muestra de prueba son *20356*, es decir, la diferencia entre las totales iniciales (*67856 pólizas)* y las observaciones de la muestra de entrenamiento (*47500 pólizas*).

Tabla 3.10 Matriz de confusión CTREE

		Predicciones Muestra Entrenamiento		Total Fila
		0(No Reclama)	1(Reclama)	
Valores Reales Muestra Prueba	0 (No Reclama)	7691	11316	19007
	1 (Reclama)	224	1125	1349
Columna total		7915	12441	20356

Fuente: Elaboración propia a partir de datos de R

Falsos positivos: % de negativos clasificados como positivos *(11316/19007): 59.54%*

Falsos negativos: % de positivos clasificados como negativos *(224/1349): 16.60%*

Sensibilidad: % de positivos que se clasifican como positivos *(1125/1349): 83.40%*

Especificidad: % de negativos clasificados como negativos *(7691/19007): 40.46%*

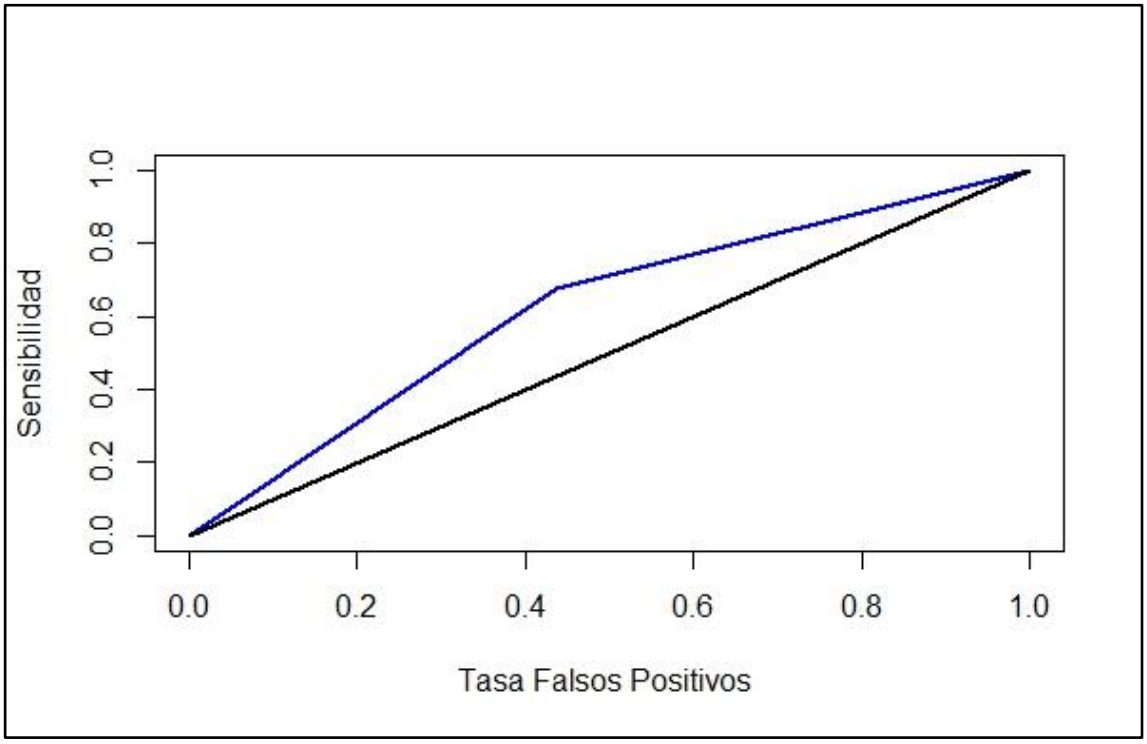
Obtenemos resultados similares, aunque algo peores, que en el ejemplo de la regresión logística. Pero esto es algo que no debe preocuparnos en exceso. La mayor utilidad de los árboles de decisión condicionales es la de clasificación y agrupación. También es una herramienta con la que hacer predicciones, pero con mayor dificultad que en la regresión logística. Como observamos, el modelo predice bien los casos en los que hay reclamación (*83.40%* de sensibilidad), sin embargo, tiene un alto valor de falsos positivos. Debemos estar atentos a esta situación, ya que tener demasiadas predicciones de reclamaciones puede ser un duro trabajo para la empresa, pero es un error menos grave que las predicciones de no reclamación que fallan, los falsos negativos. Una empresa con capacidad para estudiar gran cantidad de posibles reclamaciones, puede permitirse una tasa alta de falsos positivos, ya que acertará una gran parte de las reclamaciones. En cambio, ninguna empresa puede permitirse una alta tasa de falsos negativos, un modelo así no tendría mucho valor.

Ahora que tenemos las tasas de sensibilidad y falsos positivos podemos representar la curva ROC. En este caso se usa el paquete *ROCR* (Sing, Sander, Beerenwinkel, y Lengauer, 2005). Con la predicción que utilizamos para la matriz de confusión dibujamos

la curva ROC comparando estos valores, de nuevo, con la muestra de prueba. De la misma forma que en el ejemplo de regresión logística calculamos el área debajo de la curva con las órdenes siguientes:

```
>pred = prediction(pred, MuestraPrueba$Reclama)
>roc = performance(pred, measure="tpr", x.measure="fpr")
>plot(roc, col="blue",xlab = "Tasa Falsos Positivos ", ylab = "Sensibilidad", lwd=2)
>lines(x=c(0, 1), y=c(0, 1), col="black", lwd=2)
>auc = performance(pred, 'auc')
>slot(auc, 'y.values')
```

Gráfico 3.6 Curva ROC CTREE



Fuente: Elaboración propia a partir de datos de R

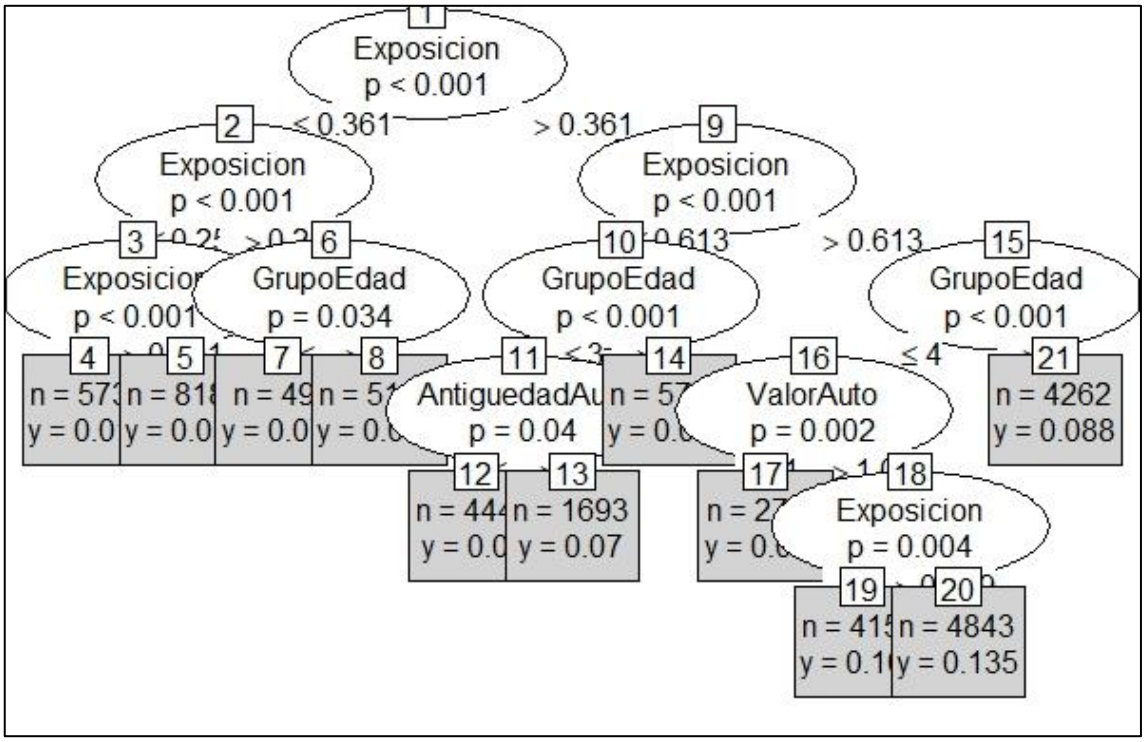
Apreciamos que la curva sobre el gráfico no parece tan buena como la obtenida en el ejemplo de la sección anterior (*Gráfico 3.4*), y cuando observamos el valor del área bajo la curva (*AUC*) confirmamos esta sospecha *auc=0.6192957*. Como apreciábamos en la matriz de confusión, el árbol de decisión de condicional no tiene un gran poder de predicción sobre esta muestra.

Queremos destacar que, aunque se reproduzcan los comandos de la misma forma que se presenta en el código en el Anexo 2, es difícil que los resultados sean exactamente iguales a los presentados en este trabajo. Esta circunstancia se debe a que estamos seleccionando al azar una muestra para entrenamiento del 70%, por lo que los valores

seleccionados cada vez que se ejecute el código serán diferentes. Aun así, los resultados no deberían diferir mucho de los presentados en este capítulo.

Para finalizar la sección, se presenta en el *Gráfico 3.7* el árbol de decisión que se obtendría al utilizar las máximas variables posibles de la muestra para su confección. Este árbol no es el más apropiado para explicar el funcionamiento de los *CTREE*, ya que la presencia de gran cantidad de nodos dificulta su comprensión. Aun así, los árboles de decisión condicionales son muy utilizados por la sencillez de interpretación de sus resultados y su gran capacidad para clasificar y agrupar.

Gráfico 3.7 CTREE complejo



Fuente: Elaboración propia a partir de datos de R

4. CONCLUSIONES

Para finalizar redactamos este capítulo de conclusiones.

El autor espera que el lector haya podido seguir el desarrollo del trabajo de forma agradable, y que haya encontrado las explicaciones prácticas comprensibles y alineadas con la teoría que se iba mostrando. Eso querrá decir que uno de los objetivos principales destacados en la introducción se ha cumplido.

En el primer bloque se ha discutido el significado, las características, la importancia, los límites y el análisis del Big Data, consiguiendo crear un marco común para todos estos aspectos en torno al Big Data en el sector asegurador. En el segundo bloque, se han presentado diferentes modelos predictivos y se han explicado en detalle dos de ellos, la regresión logística y los árboles condicionales. Estos dos eran otros de los objetivos marcados al comienzo del proyecto.

Por último, el autor ha conseguido satisfacer su deseo de ampliar conocimientos en la materia, disfrutando y descubriendo a partes iguales gran cantidad de valiosas publicaciones. Espera haber conseguido realizar una pequeña aportación en la literatura en castellano del Big Data, más concretamente a la aplicación de modelos predictivos en el sector asegurador.

5. BIBLIOGRAFÍA

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Aldás, J., y Uriel, E. (2017). *Análisis multivariante aplicado con R* (2º edición). Paraninfo.

Carstensen, B., Plummer, M., Laara, E., y Hills, M. (2018). Epi: A Package for Statistical Analysis in Epidemiology.

Casanovas Ramón, M., Merigó Lindahl, J. M., y Torres Martínez, A. (2014). *Inteligencia computacional en la gestión del riesgo asegurador: operadores de agregación OWA en procesos de tarificación*. Madrid: Fundación MAPFRE.

Cox, M., y Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, (July), 235-244,. https://doi.org/10.1109/VISUAL.1997.663888

De Jong, P., y Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data (International Series on Actuarial Science)*. Cambdrige University Press. https://doi.org/10.1017/CBO9780511755408

Eyastax. (2017). Top 7 Big Data Use Cases in Insurance Industry. Recuperado a partir de https://www.exastax.com/big-data/top-7-big-data-use-cases-in-insurance-industry/

Ferrán, M. (1996). *SPSS para Windows. Programación y análisis estadístico*. Madrid: McGraw-Hill.

Francis, L., y Wolfstein, A. (2018). ASTIN big data working party phase II: Predictive modeling. En *VICA 2018*. Berlin, Germany: IAA.

Gantz, B. J., y Reinsel, D. (2011). Extracting Value from Chaos State of the Universe : An Executive Summary, (June), 1-12.

Gökalp, M. O., Kayabay, K., Zaki, M., Koçyiğit, A., Eren, P. E., y Neely, A. (2017). Big-Data Analytics Architecture for Businesses : a comprehensive review on new open-source big-data tools, (October).

Goldburd, M., Khare, A., y Tevet, D. (2016). *Generalized Linear Models for Insurance Rating*. *Casual Actuarial Society*.

Grzasko, A. (2016). ROC Curve Example Using Base R.

Guelman, L., Guillén, M., y Pérez Marín, A. M. (2014). A survey of personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics*. https://doi.org/10.1016/j.insmatheco.2014.06.009

Guillén, M. (2016). Big data en seguros. *Revista de Estadística y Sociedad, 2016, vol. 67*, 28-30.

Hait, J. F., Anderson, R. E., Tatham, R. L., y Black, W. (1995). *Multivariate Data Analysis.* (4th ed.). Englewood Cliffs: Prentice Hall.

Hosmer, D. W., y Lemeshow, S. (2000). Applied Logistic Regression. *Wiley Series in Probability and Sattistics*. https://doi.org/10.2307/2074954

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, *24*, 417-441.

Hothorn, T., y Hornik, K. (2006). ctree : Conditional Inference Trees. *Journal of Computational and Graphical Statistics*.

IBM. (2018). The Four Vs of Big Data. Recuperado 10 de junio de 2018, a partir de http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Information Commissioner's Office. (2017). Big data , artificial intelligence , machine learning and data protection.

Johnson, R. A., y Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis* (4th ed.). Englewood Cliffs: Prentice Hall.

Jolliffe, I. T. (2002). Journal of Educational Psycholog. *Springer Series in Statistics*, *2*.

Kaisler, S., Armour, F., Espinosa, J. A., y Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995-1004. https://doi.org/10.1109/HICSS.2013.645

Karatzoglou, A., Meyer, D., y Hornik, K. (2006). Support Vector Algorithm in R. *Journal of Statistical Software*, *15*(9), 1-28. https://doi.org/10.18637/jss.v081.b02

Kuhn, M. (2018). caret: Classification and Regression Training. Recuperado a partir de https://cran.r-project.org/package=caret

Levy, J.-P., y Varela, J. (2003). *Análisis Multivariable para las Ciencias Sociales*. Pearson Educación.

Livingstone, D. J., Manallack, D. T., y Tetko, I. V. (1997). Data modelling with neural networks: advantages and limitations. *Journal of computeraided molecular design*, *11(2)*, 135-142.

Louveaux, S. (2016). Big data and the new EU data protection Regulation The role of Big Data in Healthcare Big Data Means Opportunities, (November).

Malone, R. (2007). Structuring Unstructured Data. *Forbes*.

McAfee, A., y Brynjolfsson, E. (2012). Big Data. The management revolution. *Harvard Buiness Review*, *90*(10), 61-68. https://doi.org/10.1007/s12599-013-0249-5

McCullagh, P., y Nelder, J. A. (1989). Generalized Linear Models, Second Edition. Chapman and Hall.

Mills, S., y Forder, S. (2012). IBM Smarter Analytics: Big Data and Insurance. *IBM Smarter Analytics*.

Mohri, M., Rostamizadeh, A., y Talwalkar, A. (2012). *Foundations of Machine Learning*. (T. Dietterich, Ed.). The MIT Press.

Mures, M. J., y Vallejo, E. (2018). Análisis de Regresión Logística. *Transparencias de Estadística Actuarial II*, (Ult. Vez visitado 11/06), Departamento de Estadística. Universidad de León.

Padilla-Barreto, A., Guillén, M., y Bolancé, C. (2017). Big-Data Analytics en Seguros. *Revista Anales del Instituto de Actuarios Españoles*, *23*, 1-19.

Pavía, J. (2016). Modelos Lineales Generalizados. *Transparencias de Modelos Lineales Generalizados*, (Ult.Vez visitado 11/06/2018), Dpto de Economía Aplicada. Universidad de Valencia.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *6*(2), 559-572.

PWC. (2012). Pillar 2 - Operational issues of risk management.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena, Austria.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., y Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-12-77

Sing, T., Sander, O., Beerenwinkel, N., y Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 7881.

Sokal, R. R., y Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco and London. Freeman.

Verma, S., van Deel, L., Nadimpalli, R., Sahoo, D., y Vervuurt, M. (2016). *Wearable Devices and their Applicability in the Life Insurance Industry Table of Contents*.

Willis Towers Watson. (2015). How U . S . P y C insurers are using or plan to use predictive analytics and big data, 2015-2016.

Willis Towers Watson. (2016). How U . S . P y C insurers use predictive models and big data.

Willis Towers Watson. (2017). Predictive modeling: new applications, new questions 2016, (March).

Young, K. (2017). The integration of wearables and insurance. Recuperado 20 de agosto de 2006, a partir de http://institute.swissre.com/research/library/Medical_Wearables_Kelvyn_Young.html

ANEXO 1 CÓDIGO REGRESIÓN LOGÍSTICA

```
#paquetes que podemos necesitar durante el proceso
library(odfWeave)
library(party)
library(MASS)
#Abrimos y nombramos la base de datos a utilizar
DatosMutuas<-read.csv(file="http://www.uv.es/pavia/seguros_no_vida/mutuas.csv",sep=";")
View(DatosMutuas)
#Comprabamos que toda las variables están en el formato correcto
levels(DatosMutuas$Tipo)
DatosMutuas$Tipo<-factor(DatosMutuas$Tipo)
levels(DatosMutuas$Tipo)
summary(DatosMutuas)
#Creamos la variable Reclama a partir de la variable claims, si es mayor que 0, hay reclamación
DatosMutuas$Reclama<-as.numeric(DatosMutuas$Claims>0)
#creamos un dataframe con los datos de las variables reclama y volumen
DataFrame<-as.data.frame(table(DatosMutuas$Reclama,DatosMutuas$Volumen))
DataFrame<-as.data.frame(apply((Datos1[Datos1$Freq !=0,]),2,as.numeric))
View(DataFrame)
#Creamos un gráfico con los datos del data frame con la frecuencia de las reclamaciones como puntos en
el gráfico
with(DataFrame,plot(Var2,Var1,type="n",xlab="Volumen de primas",ylab="Proporción reclamaciones"))
points(DataFrame$Var2,DataFrame$Var1,pch=as.character(Datos1$Freq))
#Realizamos los ajustes Logit y Probit
Logit<-glm(Reclama~Volumen,data = DatosMutuas,family=binomial)
Probit<-glm(Reclama~Volumen,data=DatosMutuas,family=binomial("probit"))
#Representamos los datos de los ajustes en el gráfico anterior
indice<-order(DatosMutuas$Volumen)
points(DatosMutuas$Volumen[indice],y=Logit$fitted.values[indice],type="l",col="red",lty=4,lwd=2)
points(DatosMutuas$Volumen[indice],y=Probit$fitted.values[indice],type="l",col="blue",lty=3,lwd=2)
legend("bottomright",c("Logit","Probit"),col=c("red","blue"),lty=c(3,2),lwd=c(2,2),bty="o")
#Observamos los principales datos de los ajustes
Logit
summary(Logit)
Probit
summary(Probit)
#Calculamos un ejemplo
exp(-1.4124)/(1+exp(-1.4124))
#Otro gráfico para visualizar el ajuste
plot(DatosMutuas$Volumen,DatosMutuas$Reclama,xlab="Volumen",ylab="Probabilidad de
Reclamacion")
g=glm(Reclama~Volumen,family=binomial,DatosMutuas)
curve(predict(g,data.frame(Volumen=x),type="resp"),add=TRUE)
points(DatosMutuas$Volumen,fitted(g),pch=20)
#calculo de las deviance y su grado de significación
Deviance.Modelo<-Logit$deviance
Deviance.Base<-Logit$null.deviance
chi<-Deviance.Base-Deviance.Modelo
chi.df<-Logit$df.null-Logit$df.residual
sig.chi<-1-pchisq(chi,df=chi.df)
#valores predichos y cálculo de la curva ROC
Prediccion = prediction(Logit$fitted.values, DatosMutuas$Reclama)
roc = performance(Prediccion, measure="tpr", x.measure="fpr")
#representación de la curva roc
plot(roc, col="blue",xlab = "Tasa Falsos Positivos ", ylab = "Sensibilidad", lwd=2)
```

```
lines(x=c(0, 1), y=c(0, 1), col="black", lwd=2)
#Comprobamos algunos datos del modelo
Deviance.Modelo
Deviance.Base
chi
chi.df
sig.chi
#realizamos la matriz de confusión
Predict.Ajuste<-Logit$fitted.values
Predict.Ajuste[Predict.Ajuste>=.50]<-1
Predict.Ajuste[Predict.Ajuste<.50]<-0
CrossTable(DatosMutuas$Reclama,Predict.Ajuste,prop.chisq=FALSE,prop.c = FALSE,prop.r=FALSE)
#Se presenta una forma alternativa de calcular la curva ROC
Positivos.Ajuste <- Logit$fitted.values[DatosMutuas$Reclama== 1]
Negativos.Ajuste <- Logit$fitted.values[DatosMutuas$Reclama== 0]
Orden.Ajuste <- sort(Logit$fitted.values)
Sensibilidad <- 0
TasaFalsosPositivos  <- 0
for (i in length(Orden.Ajuste):1){Sensibilidad <- c(Sensibilidad, mean(Positivos.Ajuste >= Orden.Ajuste[i]))
TasaFalsosPositivos  <- c(TasaFalsosPositivos , mean(Negativos.Ajuste >= Orden.Ajuste[i]))}
plot(TasaFalsosPositivos , Sensibilidad, xlim = c(0, 1), ylim = c(0, 1), type = "l",
    xlab = "Tasa Falsos Positivos ", ylab = "Sensibilidad", col = 'blue')
abline(0, 1, col= "black")
#cálculo del AUC
library(pROC)
roc_obj<-roc(DatosMutuas$Reclama,Logit$fitted.values)
auc(roc_obj)
#Ejemplos de otros modelos
Logit2<-glm(Reclama~Volumen+Zona, data=DatosMutuas,family=binomial)
summary(Logit2)
Logit3<-glm(Reclama~Volumen+Zona+Tipo, data=DatosMutuas, family=binomial)
Logit4<-glm(Reclama~Volumen+Zona*Tipo, data=DatosMutuas, family=binomial)
```

ANEXO 2 CÓDIGO ÁRBOLES DE DECISION CONDICIONALES

```
#paquetes que podemos necesitar durante el proceso
library(party)
library(caTools)
library(caret)
library(party)
library(MASS)
library(ROCR)
#Abrimos la base de datos y la nombramos
DatosPolizasAuto = read.csv("car.csv")
#Renombramos las variables
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="veh_value"] <- "ValorAuto"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="exposure"] <- "Exposicion"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="clm"] <- "Reclama"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="numclaims"] <- "NumReclamaciones"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="claimcst0"] <- "ValorReclamacion"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="veh_body"] <- "TipoAuto"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="veh_age"] <- "AntiguedadAuto"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="area"] <- "Zona"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="gender"] <- "Sexo"
colnames(DatosPolizasAuto)[colnames(DatosPolizasAuto)=="agecat"] <- "GrupoEdad"
View(DatosPolizasAuto)
#resumen de los datos
summary(DatosPolizasAuto)
#Creamos el training set y el test set
IndiceValidez <- createDataPartition(DatosPolizasAuto$Reclama, p=0.70, list=FALSE)
# seleccionamos el 30% de los datos para la prueba
MuestraPrueba <- DatosPolizasAuto[-IndiceValidez,]
# Usamos el 80% restante para realizar el modelo
MuestraEntrenamiento <- DatosPolizasAuto[IndiceValidez,]
#Indicamos la formula a seguir para el arbol y creamos el modelo con ctree
Fmla = Reclama ~ ValorAuto + TipoAuto + AntiguedadAuto + Sexo + Zona + GrupoEdad+Exposicion
ModeloArbol = ctree(Fmla, data = MuestraEntrenamiento)
print(ModeloArbol)
summary(ModeloArbol)
plot(ModeloArbol, type="simple")
#Creamos un objeto donde se encuentran las predicciones, y creamos la matriz de confusión
pred <- predict(ModeloArbol, newdata=MuestraPrueba)
pred[pred>=.06]<-1
pred[pred<.06]<-0
CrossTable(MuestraPrueba$Reclama,pred,prop.chisq=FALSE,prop.c = FALSE,prop.r=FALSE)
#utilizamos las predicciones para crear la curva roc y calcular el estaístico c de la curva ROC
pred = prediction(pred, MuestraPrueba$Reclama)
roc = performance(pred, measure="tpr", x.measure="fpr")
plot(roc, col="blue",xlab = "Tasa Falsos Positivos ", ylab = "Sensibilidad", lwd=2)
lines(x=c(0, 1), y=c(0, 1), col="black", lwd=2)
auc = performance(pred, 'auc')
slot(auc, 'y.values')
```