Machine Learning en modelos de predicción de la siniestralidad del asegurado en Seguros de Salud

Máster Universitario en Ciencias Actuariales y Financieras

Presentado por: Fernando Donado Gonzalez

Dirigido por: Lucio Fernández Juan José

Alcalá de Henares, a 22 de Junio de 2022

RESUMEN

Los seguros de salud se clasifican como el segundo ramo más importante en los seguros de no vida, el cual ha experimentado un crecimiento constante, especialmente en el último año debido a la pandemia del Covid-19. Este crecimiento comienza por cumplir los nuevos retos para las entidades aseguradoras, siendo los más importantes la digitalización, promocionar la salud del cliente y la creación de nuevos productos o coberturas.

En el presente trabajo, se utilizan métodos de Machine Learning enfocado al aprendizaje supervisado con datos reales del gasto realizado por parte de los asegurados con el fin de obtener un modelo que pueda predecir el gasto futuro y en consecuencia, tomar mejores decisiones.

La principal conclusión es que los modelos avanzados de Machine Learning son más eficaces y eficientes que los métodos tradicionales de regresión y clasificación, aunque estos últimos son más fáciles de entender. La posibilidad de personalización del producto o póliza con modelos en tiempo real será uno de los cambios mas importantes en el sector.

ABSTRACT

Health insurance ranks as the second most important branch in non-life insurance, which has experienced steady growth, especially in the last year due to the Covid-19 pandemic. This growth begins by meeting new challenges for insurers, the most important of which are digitalization, promoting customer health and the creation of new products or coverages.

In the present work, Machine Learning methods focused on supervised learning are used with real data of the expenditure made by the insured in order to obtain a model that can predict future spending and consequently make better decisions.

The main conclusion is that advanced Machine Learning models are more effective and efficient than traditional regression and classification methods, although the latter are easier to understand. The possibility of product or policy customization with real-time models will be one of the most important changes in the industry.

Contenido

Contenido de Tablas

Contenido de Ilustraciones

1. INTRODUCCIÓN

En España, así como en la mayoría de los países europeos, la sanidad está cubierta por el sistema público, aunque los agentes privados también tienen su mercado sobre una parte de la población. Se destaca la sanidad pública española, entre otros factores, como principal factor de que la esperanza de vida sea la mejor de Europa con 83,4 años frente a los 80,9 de la media europea(OECD/European Observatory on Health Systems and Policies 2019). Por tanto, la calidad del sistema sanitario español es, en cierta medida, una barrera de entrada para los seguros de salud privados.

En este trabajo se estudiará el seguro de salud de asistencia sanitaria. Se deberá distinguir entre dos modalidades de seguros de salud que se ofertan en el mercado. Los seguros de enfermedad que cubren las consecuencias económicas de la enfermedad del asegurado y los seguros de asistencia sanitaria que cubren los gastos médicos derivados de una enfermedad dentro del cuadro médico del seguro. Podría clasificarse un tercer seguro que se denomina de reembolso, que se diferencia de la asistencia sanitaria en que el asegurado puede elegir un profesional de su elección y el asegurador reembolsará un porcentaje de dicho pago. El reembolso y el seguro dental suele contratarse como coberturas adicionales del seguro de asistencia sanitaria, por tanto, se incluirán en el mismo apartado. Además de la cobertura del seguro, se distingue quien es el tomador del seguro. Los seguros colectivos son contratados por una empresa para sus empleados, es decir, el tomador del seguro es la propia empresa. Y los seguros individuales, una persona contrata como tomadora para su familia o para sí mismo el seguro de salud. También se debe destacar que dentro de los seguros colectivos nos encontramos a las Administraciones Públicas que son Muface, Isfas y Mugeju.

Los métodos estadísticos más utilizados en las Entidades aseguradoras para el estudio de perfiles para la fijación de precios (Pricing) y la suscripción se componen principalmente por los modelos lineales generalizados (GLMs), donde se intenta explicar el coste o los posibles siniestros con variables que describen al cliente como la edad, provincia, actividad que desarrolla, etc. Una vez que se crea el primer modelo y se empiezan a recopilar datos, este puede mejorarse para hacer

más eficiente el modelo y conseguir mejores resultados. La oportunidad surge ya que las compañías disponen de grandes bases de datos y las herramientas computacionales son cada vez más sencillas y accesibles.

Es necesario evolucionar y utilizar nuevos métodos estadísticos para mejorar el análisis de datos y predecir con mayor acierto. En palabras de Andrés Gonzalez (2016), socio y cofundador de clever data “Este nuevo enfoque disruptivo crea sistemas que aprenden de los datos y modelan el comportamiento y características de los clientes para establecer una prima de riesgo personalizada y adaptada a cada cliente”. Un ejemplo es la utilización de los árboles de decisión para la suscripción de las carteras, dando una respuesta inmediata y adaptada al cliente que vamos a suscribir.

El objetivo del presente trabajo es mostrar la utilidad de las herramientas de predicción en el proceso de Pricing y suscripción de una empresa. Se mostrará cómo los modelos de Machine Learning (en adelante, ML) nos pueden ayudar a crear un modelo predictivo para la toma de decisiones y mejorar el riesgo de la cartera. Para lograr este objetivo, nos apoyamos en los datos de una empresa de seguros española, de reciente creación y que ha cedido los datos de forma confidencial.

La muestra de datos se compone de 106.506 observaciones de asegurados, procesadas con la aplicación informática SAS, discretizando las variables seleccionadas para poder modelar los datos con el programa R. Con esta modelización de los datos se realizará un análisis exploratorio de los datos para detectar posibles datos erróneos o dependencias entre las variables. Posteriormente realizaremos un modelo lineal general con el cual compararemos los tres modelos de ML: modelo lineal múltiple, ramdom forest y support vector machine. Estos modelos se comparan con el Error residual estándar (RSE) y el Error cuadrático medio (RMSE), dando lugar a la comparación de los modelos para determinar qué modelo tiene mejor predicción.

2. MARCO TEÓRICO

En este capítulo se exponen los conceptos teóricos en los que se basa este trabajo. Por un lado, se describen aspectos fundamentales del seguro de salud y de los modelos predictivos aplicados a los seguros.

2.1 El seguro de salud

Antes de realizar el análisis de datos y desarrollar los modelos de ML es importante explicar el conceto de seguro de asistencia sanitaria y las particularidades que tiene frente al resto de seguros.

Un seguro de asistencia sanitaria cubre los gastos médicos de la asistencia médica, quirúrgica y hospitalaria, a cambio el asegurado se compromete a pagar una prima. Se debe destacar que un seguro de salud a diferencia de otros ramos se contrata para realizar siniestros. Es decir, el seguro de salud es un producto de servicios, es decir, se contrata para poder ir al médico y realizar ciertas pruebas periódicas. La asistencia se prestará en los centros concertados que se recogerán en el cuadro médico de la Entidad Aseguradora.

El cuadro médico recoge todos los profesionales, centros y hospitales que puede ir el asegurado y se definirá en las condiciones particulares del seguro. Las coberturas son las prestaciones que quedan cubiertas por el seguro y se clasifican generalmente de la siguiente forma:

- Medicina primaria
- Urgencias
- Especialidades médicas y quirúrgicas
- Hospitalización e intervención quirúrgica
- Medios de diagnostico
- Pruebas o tratamientos complejos

A su vez dentro del seguro de salud se puede contratar otras coberturas adicionales como puede ser la cobertura dental, reembolso, telemedicina, transporte

sanitario, etc. Estas coberturas adicionales dependerán de la compañía y se caracterizan por dar un valor adicional y característico al seguro, pudiendo distinguir entre varios productos por las coberturas incluidas y formando distintos productos.

En los seguros de salud, antes de contratar la póliza es muy importante la suscripción médica para determinar qué riesgo se incluye en la cartera. En la selección del riesgo se tiene en cuenta el historial médico y el estado de salud en el momento de la contratación del seguro, estos se definen como factores médicos y se pueden clasificar de la siguiente forma:

- Historial Clínico y Estado de Salud

El valorador medico se encarga de realizar un cuestionario médico para valorar si el cliente ha padecido o padece alguna enfermedad, tratamiento, dolencia u operación que pueda incrementar el riesgo de la póliza.

- Índice de Masa Corporal (IMC)

El índice de masa corporal es la relación entre el peso con la altura de una persona. Se calcula dividiendo el peso corporal entre la altura elevada al cuadrado. Es uno de los factores médicos determinantes en la contratación, cuando el cliente tiene un IMC superior a 35 tendría un impedimento para contratar la póliza de salud.

- Ejercicio Físico

La actividad física reduce los factores de riesgo de padecer una enfermedad. Una vida activa ayuda a mejorar nuestra salud y a reducir la posibilidad de padecer ciertas dolencias o enfermedades, por tanto, muchas compañías lo utilizan para fomentar la salud del asegurado y que se reduzca el riesgo de tener una enfermedad grave. En España, hay compañías que vinculan la actividad física del cliente con descuentos en la renovación de su póliza, con esta iniciativa el cliente genera descuentos y además tiene una mejor salud lo que implica menor gasto (futuro) para la compañía.

Si tras el cuestionario medico se detecta alguna patología o enfermedad, se pondrá una exclusión, esta se define como aquellas pruebas o garantías que no se

cubrirán por el seguro al reconocerse que el cliente ya las padecía antes de la contratación. Estas exclusiones pueden ser eliminadas a cambio de una sobreprima, para recoger el riesgo que conlleva asegurar las prexistencias del cliente se aplica un factor corrector que aumenta la prima.

Las carencias es el periodo de tiempo durante el cual algunas coberturas no puede ser utilizadas. Si los asegurados vienen de un seguro anterior activo se pueden eliminar ya que ha estado cubierto y es menos probable que tenga una dolencia, prueba o enfermedad anterior. Las carencias dependen de la garantía, a modo de ejemplo se suelen poner carencias de entre 8 a 10 meses en las garantías de hospitalización, pruebas o tratamientos complejos y de entre 3 a 6 meses en pruebas o tratamientos simples.

Tras explicar que es un seguro de salud y que procesos hay en la contratación, hay que tener en cuenta el marco legislativo que va a influir en todo el proceso de suscripción. El seguro de salud se rige por la Ley 50/1980, de 8 de octubre de Contrato de Seguro, por la Ley 20/2015, de 14 de julio, de ordenación, supervisión y solvencia de las Entidades Aseguradoras y Entidades Reaseguradoras, y su Reglamento de Desarrollo (Real Decreto 1060/2015, de 20 de noviembre, de ordenación, supervisión y solvencia de las entidades aseguradoras reaseguradoras). Entre las particularidades recogidas por ley de este seguro destacamos los siguientes puntos:

- No se puede discriminar por razón de VIH/SIDA u otras condiciones de salud. Esta condición implica que, en tu proceso de contratación, tarificación y de solvencia debes tener en cuenta que debes permitir y dar opción a contratar a una persona con patologías previas de mucho riesgo.
- Urgencia Vital, en el artículo 103 de LCS se recoge que ante una urgencia de vida o muerte la Entidad Aseguradora deberá obligatoriamente dar cobertura.
- Artículo 10: Este articulo recoge el compromiso del tomador del seguro, de declarar fehacientemente el cuestionario medico sometido por la Entidad Aseguradora. El articulo 10 exige contestar sobre lo que se

pregunta, la aseguradora debe realizar las preguntas oportunas para recoger todas las dolencias del asegurado y evitar un riesgo preexistente.

Las nuevas obligaciones establecidas en el Reglamento UE 2016/679, de 27 de abril de 2016, General de Protección de Datos (RGPD) y en la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPD-GDD) han obligado a las Entidades Aseguradoras a revisar el tratamiento de los datos del cliente y reforzar las medidas para cumplir con dicha ley.

El tratamiento de los datos solo debe ser necesario cuando el cliente da su consentimiento para tratar sus datos con el fin de ejecutar un contrato. En el caso de datos de salud se recogen en el artículo 35 de RGPD sonde se concreta que "Entre los datos personales relativos a la salud se deben incluir todos los datos relativos al estado de salud del interesado que dan información sobre su estado de salud física o mental pasado, presente o futuro. Se incluye la información sobre la persona física recogida con ocasión de su inscripción a efectos de asistencia sanitaria, o con ocasión de la prestación de tal asistencia, de conformidad con la Directiva 2011/24/UE del Parlamento Europeo y del Consejo".

2.2 Tarificación del seguro de salud

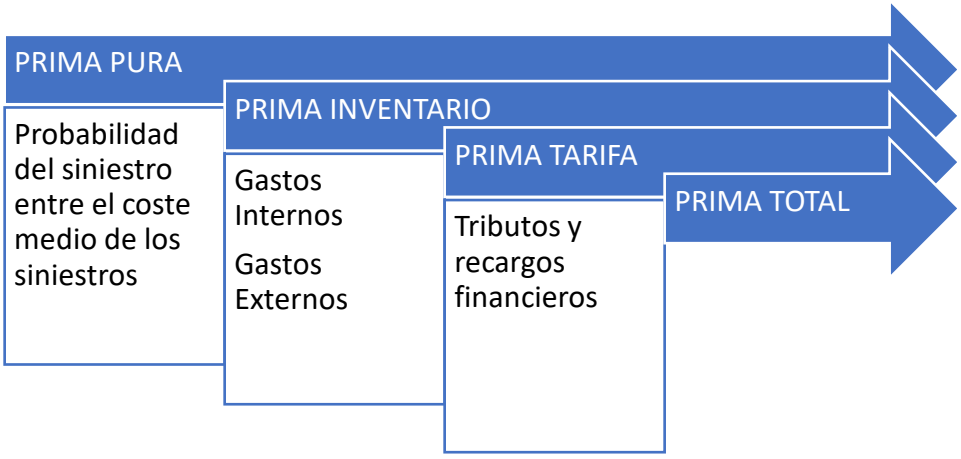
Para el Pricing del seguro debemos distinguir si el seguro es colectivo o individual. En el caso de los seguros de salud para colectivos se calcula la prima en base al método de grupo que calcula el riesgo del conjunto a asegurar. Se calcula la prima nivelada de la masa a asegurar obteniendo un precio único para todo el conjunto. En el caso individual la prima se calcula por el método de riesgo individual, se calcula las variables principales que se pueden obtener del individuo y se determina la prima sobre la base de estos factores.

En nuestro caso particular los datos son de seguros individuales y, por tanto, para obtener la prima nos basamos en el cálculo del riesgo individual. Para poder tarificar al cliente se necesitan recoger datos básicos del asegurado para poder dar un precio. En ese sentido es importante destacar que por protección de datos no se

pueden recoger datos personales del cliente, pero si se puede recopilar información anónima del mismo.

Para obtener la prima total, primero se debe calcular la prima pura, que se calcula como se muestra en la Ilustración 1:

Ilustración 1: Cálculo de la prima de salud



Fuente: Elaboración propia.

2.3 Evolución del ramo salud y competencia

Según datos de ICEA el seguro de Salud se sitúa en la segunda posición en importancia dentro del sector de los seguros no vida, como podemos ver en la Tabla 1 el volumen de primas es de 2.438 millones de euros a marzo 2021 y un crecimiento respecto al mismo periodo del año anterior del 3.9%.

Tabla 1. Número de Asegurados en el ramo de Salud.

	A 31/3/2021	Crecimiento desde Enero	Distribución
Asistencia Sanitaria sin AAPP	8.743.890	2,82%	66,78%
AAPP	1.781.223	-0,41%	13,60%
Asistencia Sanitaria	**10.525.113**	**2,26%**	**80,39%**
Reembolso de Gastos	760.745	-0,49%	5,81%
Prestación de Servicios	**11.285.858**	**2,07%**	**86,20%**
Subsidios e Indemnizaciones	1.807.317	3,52%	13,80%
Total Número de Asegurados	**13.093.175**	**2,27%**	**100,00%**

Fuente: ICEA (ICEA 2021)

En la Tabla 2 se recoge el número de asegurados, este crece en torno a la Asistencia Sanitaria en un 2.26%, la pandemia y saturación de la sanidad pública está provocando un aumento de los seguros privados.

Tabla 2. Volumen de primas imputadas en el ramo de Salud.

	Enero - Marzo 2021	Crecimiento Interanual	Distribución
Asistencia Sanitaria sin AAPP	1.777.677.592,90	5,53%	72,92%
AAPP	406.562.780,59	-0,88%	16,68%
Asistencia Sanitaria	**2.184.240.373,49**	**4,27%**	**89,60%**
Reembolso de Gastos	190.391.225,18	1,31%	7,81%
Prestación de Servicios	**2.374.631.598,66**	**4,03%**	**97,41%**
Subsidios e Indemnizaciones	63.232.166,13	-2,17%	2,59%
Total Volumen de Primas Imputadas	**2.437.863.764,79**	**3,86%**	**100,00%**

Fuente: ICEA (ICEA 2021)

En la Tabla 3 se indica el volumen de primas imputadas. La prima se ha incrementado en un 4.27%, la demanda de productos con primas más altas ha supuesto un incremento importante en las primas imputadas en el ramo de Salud.

Tabla 3. Importe de prestaciones pagadas.

	Enero - Marzo 2021	Crecimiento Interanual	Distribución
Asistencia Sanitaria sin AAPP	1.435.203.648,43	6,13%	70,66%
AAPP	402.628.937,06	-1,62%	19,82%
Asistencia Sanitaria	**1.837.832.585,49**	**4,33%**	**90,48%**
Reembolso de Gastos	157.371.163,84	6,64%	7,75%
Prestación de Servicios	**1.995.203.749,33**	**4,51%**	**98,23%**
Subsidios e Indemnizaciones	35.923.006,38	13,47%	1,77%
Total Importe de Prestaciones Pagadas	**2.031.126.755,71**	**4,65%**	**100,00%**

Fuente: ICEA (ICEA 2021)

A su vez, este incremento de productos con garantías más completas ha provocado que el coste por acto se incremente como se puede apreciar en la Tabla 4. Además, durante la pandemia se llegó a un acuerdo entre profesionales de la sanidad privada y los asegurados para hacer un recargo por acto para compensar las pérdidas producidas por el confinamiento en 2020.

En la Tabla 4 se representa el Loss Ratio de la operación de Salud, este se sitúa en 84.14% en las Asistencia sanitaria, un % maduro que aporta solidez en las carteras de riesgos. El crecimiento de la prima de estos últimos crece en menor medida (3.7%) frente a los individuales (3.9%).

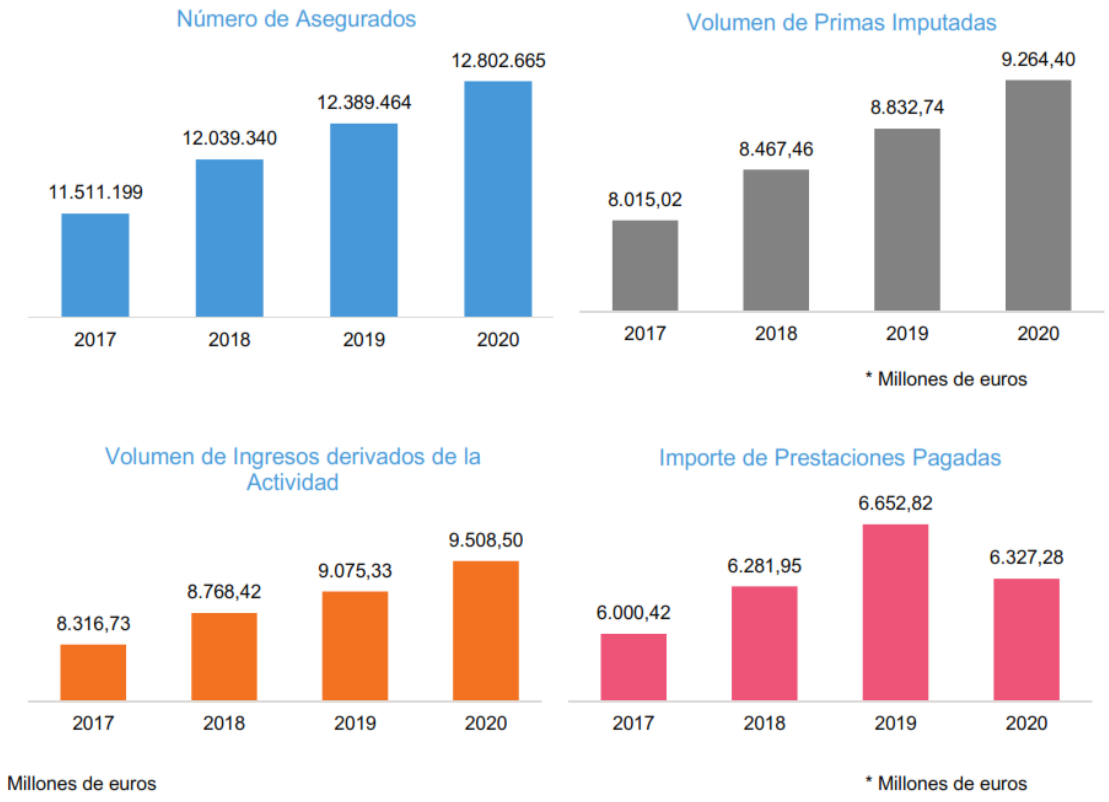
Tabla 4. Ratio de prestaciones pagadas/Volúmenes de Primas Imputadas

	Enero - Marzo 2021	Crecimiento Interanual
Asistencia Sanitaria sin AAPP	80,74%	0,57%
AAPP	99,03%	-0,75%
Asistencia Sanitaria	84,14%	0,05%
Reembolso de Gastos	82,66%	5,26%
Prestación de Servicios	84,02%	0,46%
Subsidios e Indemnizaciones	56,98%	15,95%
Ratio Importe de Prestaciones Pagadas/Primas Imputadas	83,32%	0,76%

Fuente: ICEA (ICEA 2021)

Si observamos la Tabla 5 que contiene datos anuales, vemos el gran potencial que tiene el ramo de salud.

Tabla 5. Principales volúmenes en el ramo de salud



Fuente: ICEA (ICEA 2021)

Este incremento de volúmenes generales crea nuevas oportunidades en las Entidades Aseguradoras con ramo de Salud y, por tanto, el aumento de la competencia entre las mismas para ganar la mayor cuota de mercado y conseguir los mejores perfiles para su cartera.

La competencia en el seguro de salud ha crecido con los años con la creación de nuevas compañías o ramos dentro de las compañías. Se puede observar en la Tabla 6 las Entidades Aseguradoras con mayor mercado, que son Adeslas y Sanitas, ambas representan casi el 50% del mercado. Este dato se explica por dos cuestiones, la primera es por el peso de los colectivos de empresa y la segunda por las exclusiones que se pueden llegar a poner cuando cambias de seguro y sufres de una patología previa.

Tabla 6.Volúmenes de prima por compañía sin colectivos de AAPP

Nº	Entidad (*)	Volumen de Primas Imputadas de Asistencia Sanitaria sin AAPP		
		Euros	Crecimiento	Cuota de Mercado
1	SEGURCAIXA ADESLAS (1)	505.182.817,11	5,15%	28,42%
2	SANITAS (1)	343.438.820,35	4,57%	19,32%
3	ASISA (1)	174.677.317,46	6,69%	9,83%
4	MAPFRE ESPAÑA	106.311.173,95	13,01%	5,98%
5	DKV SEGUROS	86.718.645,76	12,78%	4,88%
6	IMQ (1)	56.513.114,79	1,01%	3,18%
7	ASISTENCIA SANITARIA COLEGIAL	50.756.563,47	0,50%	2,86%
8	FIATC	39.814.605,94	-3,12%	2,24%
9	AXA SEGUROS GENERALES	36.558.247,14	7,48%	2,06%
10	AGRUPACIO AMCI	30.741.541,00	2,50%	1,73%
11	BBVA SEGUROS	23.726.497,26	6,82%	1,33%
12	MGC INSURANCE (1)	20.887.636,12	1,53%	1,17%
13	AEGON ESPAÑA	20.167.018,10	-6,93%	1,13%
14	CIGNA LIFE (1)	15.562.155,46	4,90%	0,88%
15	ACUNSA	15.388.881,22	13,41%	0,87%
16	PLUS ULTRA SEGUROS	14.687.150,75	-1,37%	0,83%
17	GENERALI SEGUROS	11.763.082,61	20,91%	0,66%
18	NUEVA MUTUA SANITARIA	8.650.709,05	6,16%	0,49%
19	ASEFA	6.174.764,88	3,83%	0,35%
20	SANTALUCIA	5.909.560,59	3,51%	0,33%
21	LINEA DIRECTA	5.708.548,00	28,49%	0,32%
22	ALLIANZ	5.392.187,10	-6,20%	0,30%
23	SEGUROS CATALANA OCCIDENTE	3.885.827,26	7,01%	0,22%
24	SANTANDER GENERALES	3.564.787,57	56,31%	0,20%
25	HNA	2.807.707,69	14,54%	0,16%
26	MUTUALIDAD DE LA ABOGACIA	2.244.617,90	5,79%	0,13%
27	UNION MEDICA LA FUENCISLA	1.540.673,34	0,76%	0,09%
28	SEGUROS BILBAO	1.066.918,72	26,55%	0,06%
29	HELVETIA SEGUROS	963.380,66	45,40%	0,05%
30	GES SEGUROS	472.155,75	5,31%	0,03%
31	MURIMAR	263.975,99	-2,66%	0,01%
32	AMIC SEGUROS GENERALES	25.076,31	-13,23%	0,00%

Fuente: ICEA (ICEA 2021)

Para poder destacar frente a la competencia es necesario innovar continuamente y adaptar los productos a los asegurados. Entre las grandes innovaciones de los últimos años nos encontramos con la digitalización del seguro. Por ejemplo, poder realizar la contratación y firma de las condiciones del seguro o pedir cita médica desde una aplicación. Junto con la digitalización destacamos el impulso que ha recibido la telemedicina, esta cobertura ya existía, pero durante el confinamiento por la pandemia de covid-19 se impulsó y mejoró para poder ofrecer servicio a todos los clientes que no podían desplazarse al médico.

En un estudio de competencia realizado por Elena Alfaro, CEO y socia de EMO Insigt (2021) se destacan dos motivos por el cual un cliente elige una compañía: la recomendación de un tercero y el precio. Es importante ser competitivo dando un buen servicio y a la vez dar un precio competitivo para llegar al máximo número de clientes.

Como hemos visto la competencia entre las entidades aseguradoras se traduce en la búsqueda del mejor servicio y de cómo destacar para atraer al cliente. Este punto nos lleva a mejorar los procesos y ser más eficientes. Un cliente que debe llamar en varias ocasiones al servicio de la entidad aseguradora crea un coste de servicio para la misma, como un cliente insatisfecho. Estos procesos están evolucionando a la automatización mediante ML para dar una respuesta más rápida al cliente u ofrece un precio más competitivo.

2.4 Machine Learning en los seguros

El análisis predictivo se debe considerar como una de las áreas de las ciencias actuariales que a través de la recolección de los datos permite extraer información y analizar patrones y tendencias de comportamiento, lo cuales se aplican a eventos del pasado, presente o futuro. También permite la tipificación de relaciones entre variables de sucesos pasados, con la finalidad de predecir posibles resultados en sucesos futuros.

Un actuario es una persona que analiza datos estadísticos y ayuda en la gestión del riesgo para minimizar y prevenir perdidas. Para ello, debe tener habilidades

estadísticas y conocimiento general sobre finanzas, matemáticas e informática. Para ello la disposición de datos, en cantidad y calidad es objeto prioritario para la ciencia actuarial, ya que permite la generación de modelos que servirán en la toma de decisiones. Los avances en herramientas tecnológicas permiten mejorar la interpretación y utilización de los datos por parte del actuario, pero surge un nuevo problema que es la cuantía de datos que cada día crece y su procesamiento se hace más difícil.

John McCarthy, conocido como el padre de la Inteligencia Artificial (en adelante IA), dijo que la inteligencia artificial es "la ciencia e ingeniería para fabricar maquinas inteligentes, especialmente programas informáticos inteligentes. Podemos relacionarlo como una tarea similar de usar computadoras para comprender la inteligencia humana, pero la IA no tiene que limitarse a métodos que sean biológicamente observables" (McCarthy, 2007). La IA esta implementada en muchos aspectos de nuestra vida, por ejemplo, los asistentes virtuales como Siri de Apple o Alexa de Amazon o los coches autónomos de Tesla. La utilización de estas herramientas todavía es limitada en el sector seguro, el 55% de las empresas afirman haber comenzado a usarlos o planean hacerlo en los próximos tres años, solo el 30% de las empresa ya están aplicando activamente herramientas de IA y ML. (Eiopa)

El Machine Learning es una rama de la IA que se encarga de aprender de eventos pasados sin ser expresamente programado para ello. Lo realmente valioso es que puede analizar grandes cantidades de datos para deducir cual es el resultado óptimo para una serie de condiciones, reconocer estos patrones da poder a la máquina de crear reglas sin la necesidad de la intervención humana. Algunos ejemplos de ML en el campo actuarial son "el ajuste de los modelos lineales generalizados a conjuntos de datos de reclamaciones para predecir la frecuencia y gravedad de las reclamaciones, o la predicción de las bajas de una póliza "(Richman, 2018). Los primeros modelos predictivos en seguros se utilizaron para identificar los perfiles más rentables y para elaborar las tarifas. En la actualidad, las aplicaciones de los modelos predictivos al sector de seguros son mucho más amplias, como se resume en la Tabla 7.

Tabla 7. Modelos predictivos del pasado, presente y futuro en el ámbito actuarial.

Modelos predictivos del pasado	**Modelos predictivos del presente y futuro**

Suscripción de nuevos riesgos	Fidelización de Siniestros
Gestión de reclamaciones	Detección de fraude
Reservas	Prevención de Bajas
	Fijación de precios y selección de riesgos
	Clasificación de Siniestros
	prevención de consultas, reclamaciones y aceleración de resoluciones
	Ofrecimiento de nuevos productos a los clientes
	Identificación de clientes en riesgo de cancelación o de impago
	Mejora en los cálculos actuariales
	monitorización de procesos de venta externalizada

Fuente:Elaboración propia a partir de datos del articulo Analítica Avanzada en el Sector Asegurador: Machine Learning – Cleverdata https://cleverdata.io/sector-asegurador-machine-learning/

Entre las áreas claves en las cuales dichos modelos pueden resultar muy útiles para las Entidades Aseguradoras se encuentra la identificación de riesgos, la detección de fraude, y la gestión del cliente (Boodhun 2017).

En el caso de la suscripción y Pricing, no supone una novedad el uso de los modelos predictivos, pero si lo supone el potencial de mejora en su predicción gracias a la cantidad de datos que se reciben de los dispositivos, redes sociales, etc., a diferencia de las fuentes externas tradicionales. La prima que pagan los consumidores por su póliza de seguro depende de una serie de características individuales que evalúa la compañía tanto en la cotización como en la renovación. Al disponer de mayores datos se dispone de más factores para calificar al cliente y realizar una mejor suscripción. En los seguros de salud los factores con relación directa para determinar el precio y el riesgo del asegurado se sitúa en el 67% frente al 80% de los seguros de automóvil(Eiopa, 2021).

3. METODOLOGÍA

En este capítulo se exponen los conceptos en los que se basa este trabajo. Se describen los aspectos fundamentales de los modelos predictivos aplicados en los seguros y se detalla los cuatro modelos a utilizar en este trabajo. Se desarrollan las descripciones técnicas y matemáticas de los modelos utilizados. Finalmente, se presentan las validaciones de los modelos.

3.1 El aprendizaje predictivo

Los modelos de Machine Learning sigue tres corrientes principales:

- El **aprendizaje por refuerzo** se basa en que la máquina aprenda por medio de prueba y error hasta ser capaz de poder completar la tarea. Estos modelos aprenden por sí mismos y mejoran realizando procesos de prueba y error hasta conseguir un objetivo o recompensa. Un ejemplo de aprendizaje supervisado es la IA “AlphaGo”, que tras entrenar durante meses y analizar miles de partidas del juego “Go” ha sido capaz de ganar al campeón del mundo de este juego.
- El **aprendizaje supervisado** se basa en el entrenamiento de la maquina con datos etiquetados, normalmente se estudia el pasado y se intenta dar una respuesta a un comportamiento futuro anticipando su comportamiento. Por ejemplo, si nosotros etiquetamos 2 grupos de fotos de perro y gatos, la maquina aprenderá a identificar en nuevas imágenes si se trata de un perro o un gato.
- El **aprendizaje no supervisado** busca similitudes en los datos, a diferencia del supervisado que busca patrones en los datos etiquetados. Los algoritmos buscan datos que compartan rasgos y que puedan agrupar. Por ejemplo, el reconocimiento facial busca patrones comunes e identifican el rosto con estas similitudes.

Los modelos que utilizaremos en nuestro estudio son de aprendizaje supervisado, ya que los datos están etiquetados. Los modelos que se utilizaran en el

presente trabajo son: Modelo Lineal múltiple, Random Forest y Support Vector Machine.

3.2 Modelo Lineal Múltiple

Los modelos lineales (Montero,2016) expresan de forma cuantitativa relaciones entre un conjunto de variables. La variable que vamos a explicar es llamada variable respuesta o dependiente y el resto son llamadas variables explicativas o independientes.

La muestra aleatoria se cuantifica como n $\{(y_i, x_{i1}, \dots., x_{ip})\}$: $i = 1,2, \dots, n\}$, donde Y es la variable respuesta, y $X_1, X_2, \dots, X_i$ las variables independientes que se relacionan linealmente, la observación de la muestra se puede expresar como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{21} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad i = 1, \dots, n$$

Siendo ε_i el error aleatorio con esperanza cero y varianza σ^2 constante; y dos errores cualesquiera ε_i y $\varepsilon_i, \forall \neq i'$ son incorrelacionados entre sí.

Hay cinco condiciones que debe cumplir una regresión lineal múltiple:

- Multicolinealidad: Los regresores deben ser independientes, y no deben estar relacionados entre sí mismos:

$$E(u_i * u_p) = 0, \forall i \neq p$$

- Linealidad: Los parámetros deben estar relacionados linealmente entre la variable a explicar y las variables explicativas cuando estas se mantienen constantes:

$$Y = X * B + U$$

- Homocedasticidad: La varianza de los residuos debe ser constante a lo largo de las observaciones:

$$V(u_i) = \sigma^2$$

- Normalidad: la distribución de la perturbación aleatoria tiene una distribución normal:

$$U \approx N(0, \sigma^2)$$

- Independencia: Las observaciones no deben estar relacionadas entre sí, estadísticamente serán independientes si la probabilidad conjunta es el producto de las marginales.

El Coeficiente de determinación se utiliza para comprobar en que porcentaje la variabilidad de la variable respuesta es explicada por la regresión, se expresa como:

$$R^2 = \frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

Este coeficiente aumenta siempre que se aumentan el número de variables regresoras. Por ello, se utiliza también el coeficiente de determinación corregido (ajustado) por el número de grados de libertad:

$$R^2\ ajustado = 1 - \frac{VNE/(n-(p+1)}{\mathrm{VT}/(n-1)}$$

3.2.1 Modelo Lineal Generalizado

Un Modelo lineal Generalizado (McCullagh, and Nelder, 1990) (o GLM) es una extensión de los modelos lineales que expresan de forma cuantitativa las relaciones entre un conjunto de variables.

La principal diferencia con una regresión lineal es que la variable respuesta no necesita seguir una normal. Este modelo se utiliza frecuentemente para distribuciones de la familia exponencial. Además, existe una relación lineal entre las variables explicativas y una transformación de la media de la variable respuesta.

Se denotará como Y la variable respuesta aleatoria y como X las variables explicativas, las observaciones de Y pueden seguir una distribución no normal de los errores.

Los componentes principales del Modelo Lineal Generalizado son:

- Componente aleatoria: Determina la variable respuesta y la distribución a la familia exponencial.
- Componente sistemática: Determina las variables explicativas que se utilizaran en la función predictora lineal.

- Función Link: Es la función que crea una relación no lineal a una lineal para poder ajustar un modelo lineal.

Se denota la componente aleatoria Y_i al valor de la respuesta para el sujeto $i, y\ x_{i1}, \ldots\ldots, x_{ip}$ del valor de las predicciones para el mismo sujeto. El objetivo es ver cómo se comporta la variable respuesta en función de las variables predictoras, es decir:

$$E\left(Y_i \middle| x_{i1}, \ldots\ldots, x_{ip}\right) = \mu_i$$

Los supuestos del GLM son los siguientes:

- Y_i Los datos deben ser independientes y aleatorios; $E(Y) = \mu_i$.
- La componente sistemática especifica las variables explicativas en el modelo que determinan el conjunto de datos predictores lineales η_i:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{21} + \cdots + \beta_p x_{ip}$$

- La función $g()$ denominada función link establece que:

$$g_{(\mu_i)} = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{21} + \cdots + \beta_p x_{ip}$$

La función Link transforma las probabilidades de la variable dependiente para linealizar la relación con las variables independientes. Las funciones Link más utilizadas se recogen en la Tabla 8:

Tabla 8. Funciones Link más comunes.

FUNCIÓN DE VÍNCULO	FÓRMULA	USO
IDENTIDAD	$\boldsymbol{\mu}$	Continua con errores normales.
LOGARÍTMICA	$\boldsymbol{\log(\mu)}$	Función tipo Poisson para la suma de errores.
LOGIT	$\boldsymbol{\log\left(\frac{\mu}{n-\mu}\right)}$	Proporciones con errores binomiales
RECIPROCA	$\boldsymbol{\frac{1}{\mu}}$	Datos continuos con errores gamma
RAÍZ CUADRADA	$\boldsymbol{\sqrt{\mu}}$	Computo datos.
EXPONENCIAL	$\boldsymbol{\mu^n}$	Funciones de potencia

Fuente: Elaboración propia a partir de McCullagh, Peter & Nelder, John A. (1983, 1989) Generalized Linear Models, Chapman & Hall

La estimación de los parámetros $\beta_1, \dots, \beta_p$ se realiza por el método de máxima verosimilitud. Los parámetros se maximizan por el logaritmo de máxima verosimilitud estimando los valores que maximizan la probabilidad de los datos observados. Este método también se utiliza para comparar dos modelos cuando estos se aproximan a una chi-cuadrado

Los ajustes de $\hat{\mu}_i$ se calculan como $g^{-1}\left(\sum_{j=1}^{P} \hat{\beta} j\, x_{ji}\right)$, una vez estimados los parámetros del vector β. Para el modelo lineal generalizado necesitamos extender la noción de residuos para todas las distribuciones que pueden reemplazar a la normal. Los residuos de Pearson son los más utilizados y mide la discrepancia entre el valor observado y el pronóstico por el modelo:

$$r_\rho = \frac{\left(y_{i-\gamma i}^{a}\right)}{\sqrt{V(\hat{\mu} i)}}$$

3.2.1 Multivariable Adaptative Regression Splines

Multivariable Adaptative Regression Splines (o MARS) es un modelo algorítmico utilizado para predecir variables continuas. Este modelo se desarrolla para resolver datos no lineales, con un conjunto de funciones lineales unidas o varias funciones bisagra.

El modelo genera funciones para capturar las relaciones no lineales de los datos mediante puntos de corte similares a las funciones escalonadas. El proceso termina cuando el modelo encuentra los suficientes puntos de corte para generar una ecuación de predicción. En este punto el modelo comenzara a eliminar los puntos de corte que no contribuyan significativamente a la precisión de la predicción, lo que se conoce como “poda”. Una vez realizada la poda podemos utilizar la validación cruzada para seleccionar el numero óptimo de puntos de corte y ajustaremos el modelo

MARS es una generalización del modelo Recursive Partitioning Regression (RPR). Se puede representar como (Vanegas and Vásquez, 2017):

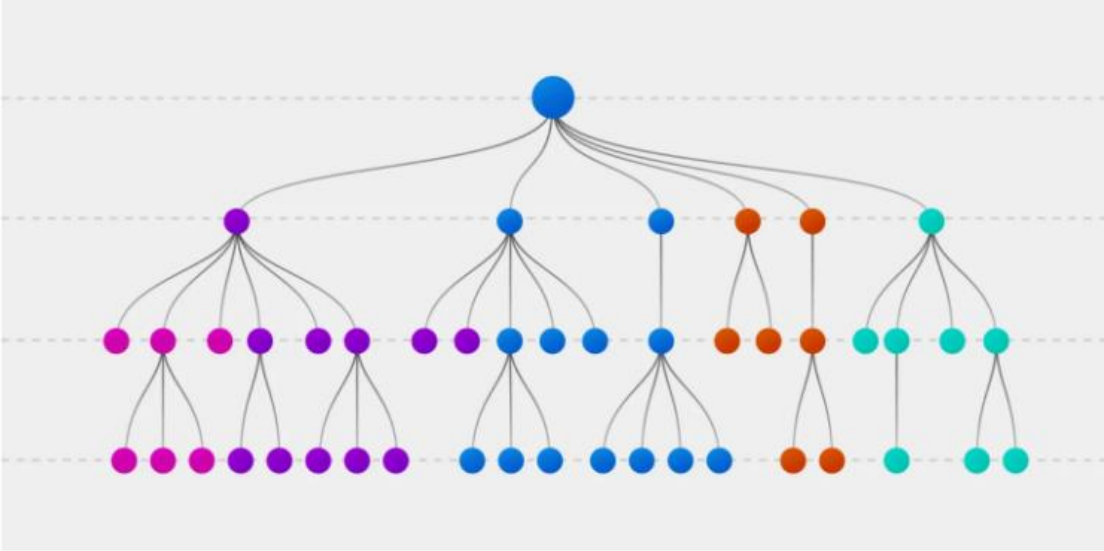
$$y_t = f(x_t) = \beta_0 + \Sigma_{i=1}^{k} \beta_i B(x_{it})$$

donde y_t representa la variable dependiente predicha por el modelo, β_0 es una constante, $B(x_{it})$ es la función x-ésima que puede representar la función base de un punto de corte y $\beta_i$que es el coeficiente de la función de base x-ésima.

3.3 Arboles de decisión

Una de las técnicas más utilizadas en Machine Learning es el árbol de decisión. Es un método de aprendizaje supervisado no paramétrico que genera una solución grafica que incluya todas las posibles respuestas a un problema. Se representa mediante una estructura jerárquica en forma de árbol. Se puede observar en la Ilustración 3, parte de un punto llamado raíz, se dividen en nodos llamados ramas y termina en los nodos finales llamados hojas, que contiene la solución al problema.

Ilustración 2. Estructura de un Árbol de decisión



Fuente: https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/

El algoritmo de un árbol de decisión detecta qué combinación es preferible para tomar las decisiones. Esto sucede al segmentar de manera recursiva la región de interés, de modo que cada partición genera un subgrupo que resulte lo más homogéneo posible. La gran ventaja de este método es que son fáciles de interpretar gráficamente y las relaciones entre las variables se pueden ver fácilmente.

En nuestro caso vamos a utilizar un Random Forest (o Bosques aleatorios) que surge como una combinación entre la técnica de *Classification And Regression Tree*

(CART) y *Booststrap Aggregation* (Bagging). Esta técnica genera diferentes combinaciones de árboles predictores en la que cada árbol se forma por los valores de un vector aleatorio probado independientemente y mantiene la distribución para cada árbol.

3.3.1 El árbol de decisión y regresión (CART)

Para entender cómo funciona el algoritmo de Ramdom Forest se estudiarán las dos técnicas que utiliza. La primera de ellas es el árbol de decisión y regresión (CART), desarrollado por (Breiman 1984), que utiliza datos históricos para construir arboles de clasificación o de regresión con fines de clasificación o predicción de nuevos datos. Puede utilizar fácilmente tanto variables cuantitativas como variables cualitativas.

Este método es muy robusto ante outliers, ante una baja variabilidad en la estructura de los árboles de clasificación o regresión, además de la gran facilidad de interpretabilidad.

La construcción del árbol empieza con el nodo principal que contiene todos los datos de la base de entrenamiento. El algoritmo comienza a separar los datos según el criterio de participación que determinara la media de impureza ya que establece el grado de homogeneidad entre grupos.

El nodo N_m es el número de instancias de entrenamiento que alcanza el nodo m. Para el nodo raíz, esto es N. N_m^i de N_m , que pertenece a las clases C_i, donde $\sum_i N_m^i = Nm$.

Dado que una instancia alcanza el nodo m, la estimacion de la probabilidad de la clase Ci es (Medina; Merino and Ñique; Chacón 2017):

$$\hat{P}(C_i|x,m) \equiv P_m^i = \frac{N_m^i}{N_m}$$

Por tanto, para el nodo m es puro si P_m^i para todo i son 0 o 1. Obtendremos un valor 0 cuando ninguna de las observaciones del nodo m son de clase C_i, y obtendremos 1 cuando todos los casos son de clase C_i. Cuando la división de los nodos es pura, no es necesario dividir más y se añadirá el nodo final P_m^i es 1.

3.3.2 Bagging

El *Bootstrap aggregation* (Bagging), desarrollado por (Breiman 1996) es un método para reducir la varianza evitando el sobreajuste los hiperparametros. Se aplica un muestreo repetido para reducir la varianza de las estimaciones mediante el promedio de distintos modelos de ML.

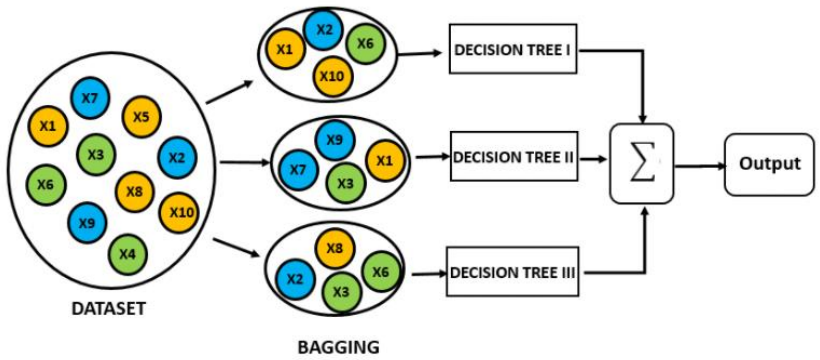
Si tenemos un conjunto de n observaciones independientes $x_{1,}x_{2,}, \dots, x_{1,n}$, y cada observación con varianza σ^2 / n; entonces, la varianza de la media $\bar{X}$ de las observaciones está dada por σ^2 / n; con esto se puede concluir que utilizando un conjunto de observaciones se reduce la varianza. Se selecciona una gran cantidad de observaciones que permitan crear conjuntos de entrenamiento de la población y para cada conjunto de observaciones se crea un modelo de predicción independiente.

La forma de implementar la técnica de bagging tiene los siguientes pasos:

- Se divide los datos de entrenamiento en varios subgrupos, obteniendo como resultado arboles como subgrupos se hayan generado.
- Se crea un modelo predictivo con cada subgrupo, obteniendo modelos diferentes.
- Por último, obtenemos un único modelo predictivo que será el promedio de los modelos anteriormente calculados.

Se puede visualizar en la Ilustración 4 cómo funciona el algoritmo del proceso bagging:

Ilustración 3. Funcionamiento del proceso bagging



Fuente: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html

3.3.3 Random Forest

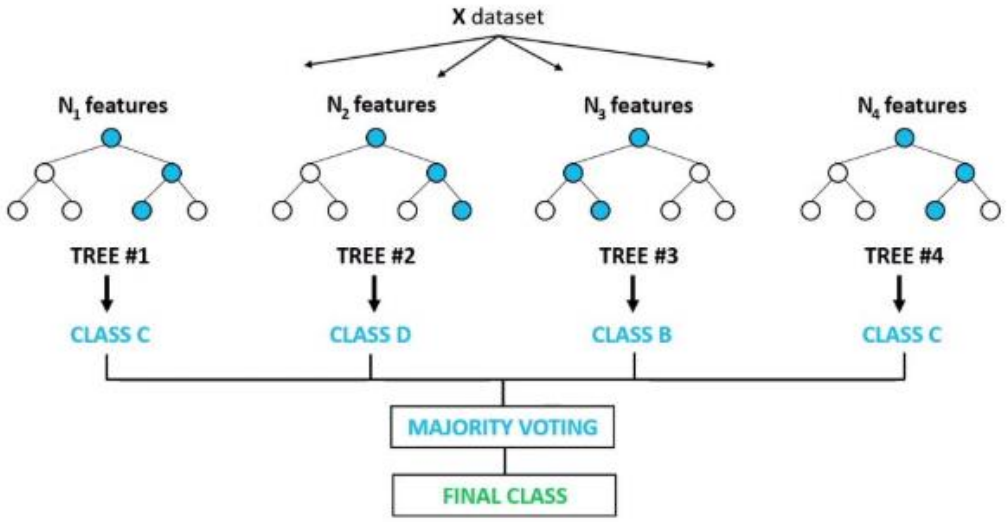
Como hemos visto en los dos puntos anteriores, el método de Random forest combina la técnica de CART y Bagging. Mediante este método mejoramos la precisión en la clasificación con la incorporación de aleatoriedad en la construcción de cada clasificador individual.

La aleatoriedad del proceso permite la construcción del árbol, en todas las etapas que desarrollara el algoritmo: Entrenamiento, convalidación y prueba. En términos generales el algoritmo del modelo Random Forest de desarrolla en los siguientes pasos (Montillo and Ling 2009):

1. Una etapa de entrenamiento, donde el algoritmo intenta optimizar los parámetros de las funciones divisorias a partir de la muestra de entrenamiento. Estos datos seleccionan aleatoriamente m predictores, si $m = p$, siendo p las variables de la muestra, los resultados de Random forest y Bagging serán los mismos. Hay datos que no se utilizaran para formar el árbol. Los datos que no formen parte del conjunto de entrenamiento son denominados *Out of bag data*, que permitirán calcular el error de clasificación del modelo. Para escoger el valor óptimo de m se compara el valor del error del *Out of Data* para los distintos valores de m.
2. Se busca alcanzar un nodo de tamaño n_{min} con la construcción de forma recursiva de T_b (Random forest). El algoritmo seleccionara m variables al azar de las p variables y escoge la mejor variable ente las m variables, dando como resultado un conjunto de arboles $\{T_b\}_1^B$.
3. Se realizará la votación para cada resultado previsto T_b. En un problema de clasificación utilizara la moda, y para un problema de regresión, utilizara la media aritmética. En este punto utilizara las muestras OOB que funciona como un conjunto de test, de las cuales se puede obtener una predicción válida para cada elemento data original.
4. Finalmente, el algoritmo selecciona el resultado de la predicción con más votos dando lugar a la predicción final.

En la Ilustración 4 podemos ver gráficamente cómo funciona el algoritmo del modelo Random forest:

Ilustración 4. Algoritmo de Random forest



Fuente: https://www.freecodecamp.org/espanol/news/random-forest-classifier-tutorial-how-to-use-tree-based-algorithms-for-machine-learning

El valor de este modelo es que genera un gran número de árboles, a lo que llamamos bosque. Estos árboles son generados a partir de Bootstrap (o remuestreo) de la muestra original. Además, se introduce aleatoriedad por la división de los nodos. El modelo seleccionara al azar el subconjunto de las p variables y restringe la selección de la variable a este subconjunto. En la Tabla 9 podemos ver qué ventajas y desventajas presenta el modelo RF:

Tabla 9. Ventajas y desventajas de Random Forest

Ventajas	Desventajas
La preparación de los datos es mínima y es más preciso que un árbol de decisión.	Se pierde interpretación.
Permite la entrada de muchas variables identificando las más significativas.	Es uno de los mejores modelos para la clasificación, pero no es tan bueno estimando en regresión.
Determina la importancia de las variables del modelo.	En los modelos de regresión no es posible predecir fuera de los rangos de

	valores del conjunto de entrenamiento.
Se prestan métodos efectivos para estimar valores restantes.	Se tiene poco control en lo que hace el modelo.
Es posible utilizarlo como metido no supervisado (clustering) y detección de outliers.	

Fuente: Elaboración propia

Para conocer la importancia de cada variable en el modelo RF tenemos que considerar las siguientes medidas (Medina-Merino and Ñique-Chacón 2017):

I. **Mean Decrease Accuracy** (MDA): Cada árbol genera errores en la clasificación y selecciona aquellos excluidos en la submuestra para la construcción del árbol, generada por Bootstrap. Este proceso no utiliza al menos un tercio de los datos de la muestra, generando el error del *Out of bag data* R_{OOB}. Una vez obtenido el R_{OOB} seleccionaremos aleatoriamente una de las variables permutando sus valores entre los datos de entrenamiento, descorrelacionando dicha variable con el modelo. Se repite el proceso calculando de nuevo el OOB R_{PERM}, para luego compararlo con el error calculado inicialmente. Para conocer la importancia del modelo recurrimos a la diferencia entre el número de clasificaciones correctas antes de R_{OOB} y después de la permutación R_{PERM} resultante. Este proceso se repetirá para todas las variables y se ordenaran estás de acuerdo con los cambios en los errores OOB. Como resultado obtenemos la medida MDA, siendo esta la media de las diferencias de los b-enésimos árboles, $b = 1, \dots, B$ en donde interviene la variable.

$$MDA = \frac{1}{B}\, \Sigma_b^B (R_{OOB} - R_{PERM})$$

II. **Mean decrease Gini** (MDG): Esta medida calcula la probabilidad de error de una variable cuando es seleccionada al azar para cada árbol. Para obtener la importancia de cada variable que utilizamos en el árbol se mide como la

suma de los decrementos atribuidos a esa variable y la importancia final, como la media en todos los árboles.

$$MG = 1 - \sum_{j=1}^{c} (p_j)^2$$

Donde $j = 1, \dots, c$ es el número de clases de la variable respuesta categórica y p_j es la frecuencia relativa de la clase C en el modelo.

3.4 Support Vector Machine

Support Vector Machine (o Maquinas de Vectores Soporte) es un algoritmo del grupo de aprendizaje supervisado utilizado para clasificación, regresión o detección de valores atípicos. El método SVM es considerado uno de los modelos más precisos y robustos posibles dentro de los algoritmos de clasificación binaria. Para la clasificación utiliza un hiperplano en $\mathbb{R}^M$, donde M es la cantidad de atributos que separa lo mejor posible una clase de otra. El objetivo es maximizar la distancia entre el hiperplano óptimo y los hiperplanos canónicos que representan los bordes de cada clase. Para lograr esto se minimiza la norma Euclidiana de w, que corresponde a los coeficientes que definen el hiperplano, dando origen al siguiente problema de minimización que se formula como (Flores, Maldonado, and Weber 2015):

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

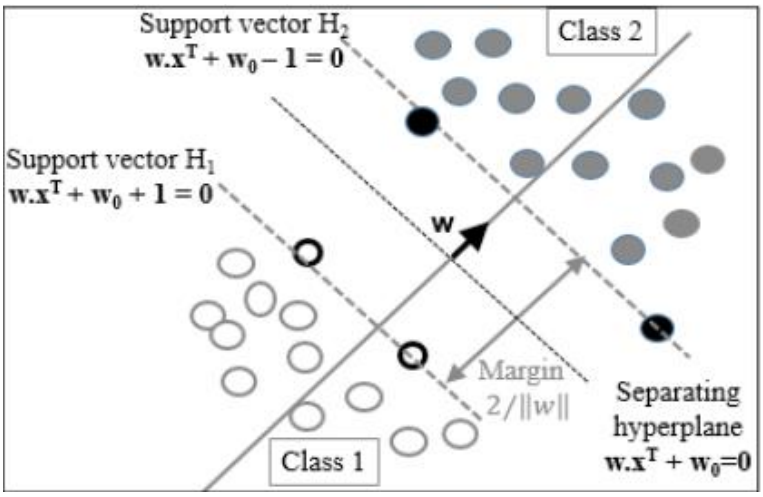
$$s.a \quad y_i \cdot (w^T \cdot x_i + b) \geq 1 - \xi_i, \qquad i = 1, \dots, N,$$

$$\xi_i \geq 0, \qquad i = 1, \dots, N,$$

Para entender cómo funciona el concepto de separar un conjunto de observaciones de entrenamiento con un hiperplano podemos representarlo en dos dimensiones (x, y) y con dos clases C_1, C_2. En la Ilustración 5 se representa los datos separados en el hiperplano de una ecuación lineal $y = w.x^T + w_0$, que se encuentra en el punto medio entre los puntos de datos del límite para la clase $C_1(H_1: w.x^T + w_0 + 1 = 0$ y la clase $C_2(H_2: w.x^T + w_0 - 1$. Como resultado se

generan los planos H_1 y H_2como soporte del vector con el mínimo margen entre las clases C_1 y C_2.

Ilustración 5. Support vector machine



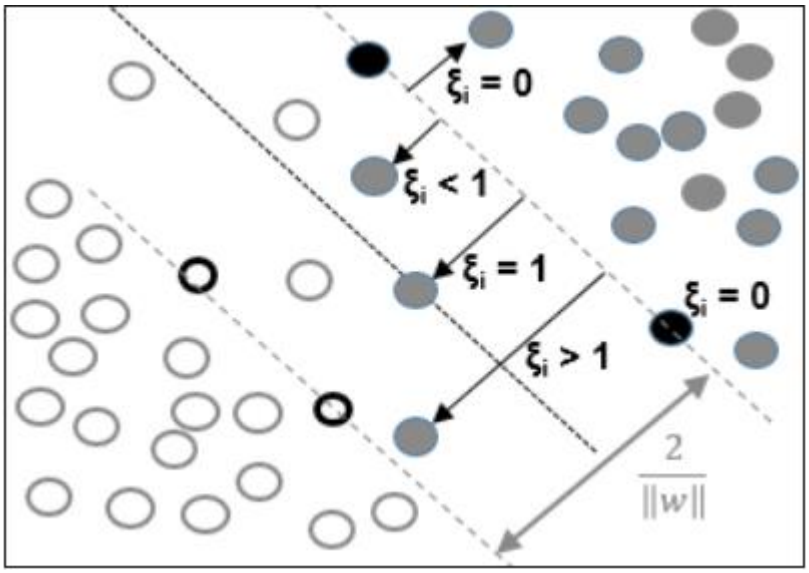
Fuente: Kernel Models and Support Vector Machines (Nicolas 2014:282)

Cuando los vectores de soporte no pueden separar fácilmente las observaciones de los datos de entrenamiento, generan variables de holgura ξ donde (Nicolas 2014):

$$x_i \epsilon \mathbb{R}^M, y_i \in \{-1,1\}, y\ \xi_i\ (i = 1, \dots, N)\ ,$$

Las variables de holgura generan errores para relajar las restricciones, y estas se controlarán con el parámetro C que es el coste. En la ilustración 6 podemos ver cómo la variable de holgura ξ mide la desviación desde el borde del margen de la clase respectiva, cuando $0 < \xi < 1$ corresponde a datos no separables, pero bien clasificados, y $\xi > 1$ corresponde a datos no separables y además mal clasificados.

Ilustración 6. Variables de holgura en Support Vector Machine



Fuente: Kernel Models and Support Vector Machines

Si nuestro problema es de regresión, podemos utilizar un SVM que comúnmente se llama Support Vector Regression (SVR). El objetivo es maximizar el margen entre los vectores de soporte con restricciones para separarlas, este método es muy similar a las variables de holgura para los datos no separables. Los parámetros del modelo $\{w_i\}$ se vuelven a escalar durante la optimización para garantizar que el margen es menor a 1. El problema de ajustar un modelo no lineal en las observaciones no es tarea fácil, para resolver este problema se utiliza una transformación no lineal del hiperplano a un hiperplano lineal en un nuevo espacio. La formulación no lineal es muy similar al caso lineal, se diferencia por la restricción del vector soporte:

$$min \ \ \frac{1}{2} < w, w > + C \sum_{i=1}^{N} (\xi_i^+ + \xi_i^-)$$

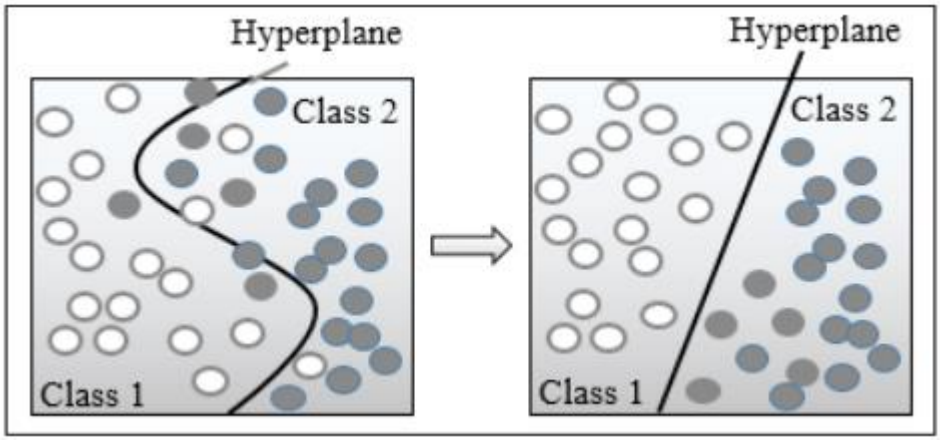
$$s.a \quad (< w, w > + b) - y_i - \epsilon - \xi_i^+ \leq 0$$

$$y_i - (< w, x_i > + b) - \epsilon - \xi_i^{\pm} \leq 0$$

$$\xi_i^+, \xi_i^- \geq 0, \quad i = 1, \dots, n$$

Podemos observar en la Ilustración 7 la formulación para los casos de SVR que transforma los datos de un hiperplano no lineal al plano lineal:

Ilustración 7. Support Vector Regression, transformación del hiperplano



Fuente: Kernel Models and Support Vector Machines (Nicolas 2014:285)

Si mediante esta transformación de datos no lineales a lineales de los vectores de soporte no se puede ajustar, se debe recurrir a una transformación de los parámetros a un nuevo espacio, en el que si sea posible ajustar los parámetros como un regresor lineal (Carmona, 2016).

3.5 Medidas de comparación de los modelos

Para poder comparar los modelos estudiados se ven tres medidas estadísticas que van a medir el desempeño del modelo.

- Root Mean Squared Error (RMSE): Es la métrica más popular para medir la tasa de error de un modelo de regresión.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{r}(y_j - \hat{y}_j)^2}$$

donde n es el número de muestras, $\hat{y}_i$ el valor predicho de la variable objetivo y y_j el valor real de la variable objetivo.

- Coeficiente de determinación (R^2): Resume el poder explicativo del modelo de regresión y se calcula a partir de los términos de las sumas de

cuadrados. El coeficiente R^2 toma valores entre 0 y 1, si $R^2 = 1$ la regresión es perfecta.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y} - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y - \bar{y})^2}$$

donde n es el número de muestras, $\bar{y}$ es la media de la variable objetivo, $\bar{\hat{y}}$ el valor predicho de la variable objetivo y y_i el valor real de la variable objetivo.

4. ANALISIS EXPLORATORIO DE LOS DATOS

En este capítulo se presenta en detalle el análisis previo de los datos para ejecutar los modelos de Machine Learning. Se detallan los datos y las variables utilizadas y por último, se realizará un análisis exploratorio de los datos para poder desarrollar los modelos. Finalmente, se detallan el software y paquetes utilizados.

4.1 Datos

Los datos utilizados en este trabajo fueron facilitados por una Entidad Aseguradora del ramo de salud. Se trata, por tanto, de datos reales que abarca un periodo cronológico de 12 meses desde el momento de extracción de los datos, en concreto hasta el 20/07/2021.

4.1.1 Variables

La base de datos contiene de partida 106.510 registros de asegurados que tienen actualmente el seguro de salud en la compañía. Cada registro representa un asegurado que ha estado activo con una póliza de salud en los últimos 12 meses. El producto de salud es nuevo en la compañía y comparte negocio con otros ramos de auto, moto y hogar.

Para cada asegurado se han identificado como variables independientes un total de 14 características de los clientes del seguro, entre las cuales se incluyen aquellas que el cliente debe aportar en el momento de la contratación y tarifican para determinar la prima final del cliente.

En las Tablas 10, 11 y 12 se muestran dichas variables, que se agrupan sobre la base de diferentes criterios: las variables que son propias del cliente, es decir, es información que aporta el cliente en el momento de la contratación; variables de producto, que son relativas a las características del producto que ha contratado el cliente; por ultimo las variables médicas, que son relacionadas al cuestionario médico y uso. También se hace referencia al tipo de variable original que obtenemos de la BBDD de la compañía y el tipo de variable modelo que vamos a utilizar en el presente trabajo.

Tabla 10. Variables Descriptivas del cliente

VARIABLES CLIENTE	DESCRIPCION	TIPO VARIABLE ORIGINAL	TIPO VARIABLE MODELO
SEXO	Sexo del asegurado: Hombre, Mujer	Factor	Factor
EDAD_ASEG	Edad del asegurado.	Numérico	Numérico
PROVINCIA	Provincia de residencia del asegurado.	Factor	Factor
CLIENTE	Determina si el cliente en el momento de contratación de salud ya es cliente de la compañía: SI/NO	Numérico	Factor
ESTADO_CIVIL	Estado civil: Soltero y casado (en casado se incluye pareja de hecho, divorciado, viudo)	Factor	Factor
SEGURO_ANT	Indica si ha tenido seguro de salud anterior: SI/NO	Numérico	Factor

Fuente: Elaboración propia

Tabla 11. Variables del producto

VARIABLES PRODUCTO	DESCRIPCION	TIPO VARIABLE ORIGINAL	TIPO VARIABLE MODELO
PRIMA_PAGADA	Prima que ha pago el asegurado anualizado	Numérico	Numérico
PRODUCTO_CONTRATADO	Producto contratado. Básico/Completo	Factor	Factor
TIPO_COPAGO	Si el cliente tiene copago o no: SI/NO	Factor	Factor
CANAL_ENTRADA	Canal por el cual entra el cliente: Teléfono, Técnica, emisión, Internet, Otro.	Factor	Factor
ASEGURADOS	Número de asegurados que tiene la póliza	Numérico	Numérico

Fuente: Elaboración propia

Tabla 12. Variables Medicas

VARIABLES MEDICAS	DESCRIPCION	TIPO VARIABLE ORIGINAL	TIPO VARIABLE MODELO

EXCLUSION	Si se marca una exclusión al asegurado; SI/NO	Factor	Factor
VALORADO	Si es valorado por el departamento medico: SI/NO	Factor	Factor
ACTO_TOT	Numero de actos totales que realiza el asegurado	Numérico	Numérico

Fuente. Elaboración propia.

Se observa en la Tabla 13 los principales estadísticos de las variables numéricas. De las cinco variables numéricas vamos a explicar las más relevantes. En el gasto obtenemos una media de 166€ con un gasto máximo de 74.445,54€ que deriva de una estancia hospitalaria larga. La media de los actos por asegurado es de 5,03 con una máximo de 398.

Tabla 13. Variables numéricas

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
gasto	0	1	166.18	627.74	0.04	9.46	37.64	155.21	74445.64	▇▁▁▁▁
Prima_pagada	0	1	209.44	137.22	0.32	107.05	187.32	283.27	1516.03	▇▂▁▁▁
acto_tot	0	1	5.03	9.28	0.00	0.00	2.00	6.00	398.00	▇▁▁▁▁
Asegurados	0	1	2.67	1.28	1.00	2.00	3.00	4.00	8.00	▇▂▅▂▁
EDAD_ASEG	0	1	32.84	17.45	0.00	17.00	36.00	47.00	88.00	▆▆▇▃▁

Fuente: Elaboración propia.

En la Tabla 14 se recogen las frecuencias para las variables categóricas que se utilizaran. Las variables categóricas se han reprocesado antes para construir categorías lógicas que puedan dar más peso a la variable para poder utilizarla en el modelo.

Tabla 14. Distribución de las variables categóricas

variable	levels	count	percentage
Canal_Entrada	TECNICA	11885	13.6%
Canal_Entrada	OTRO	15518	17.8%
Canal_Entrada	INTERNET	16934	19.4%
Canal_Entrada	EMISION	17905	20.5%
Canal_Entrada	TELEFONO	25090	28.7%
CLIENTE	NO	29230	33%
CLIENTE	SI	58102	67%
Estado_civil	SOLTERO	43502	49.81%
Estado_civil	CASADO	43830	50.19%
Exclusion	Si	22826	26%
Exclusion	No	64506	74%
Producto_contratado	BS	40646	46.5%
Producto_contratado	CM	46686	53.5%
Provincia	CADIZ	3121	3.57%
Provincia	ALICANTE	3842	4.40%
Provincia	VALENCIA	6078	6.96%
Provincia	SEVILLA	6438	7.37%
Provincia	MALAGA	6997	8.01%
Provincia	MADRID	11688	13.38%
Provincia	BARCELONA	13514	15.47%
Provincia	OTRO	35654	40.83%
Seguro_Ant	Si	29125	33%
Seguro_Ant	No	58207	67%
Sexo	H	41327	47.3%
Sexo	M	46005	52.7%
Tipo_Copago	CON COPAGO	14637	17%
Tipo_Copago	SIN COPAGO	72695	83%
valorado	No	36423	42%
valorado	Si	50909	58%

Fuente: Elaboración propia

4.1.1 Variable Dependiente

La variable dependiente de nuestro estudio es el gasto, que es la suma del importe de los actos médicos realizados por el asegurado. Las características básicas de la variable gasto son:

Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo
0,04	9,46	37,63	166,19	155,21	74.445,64

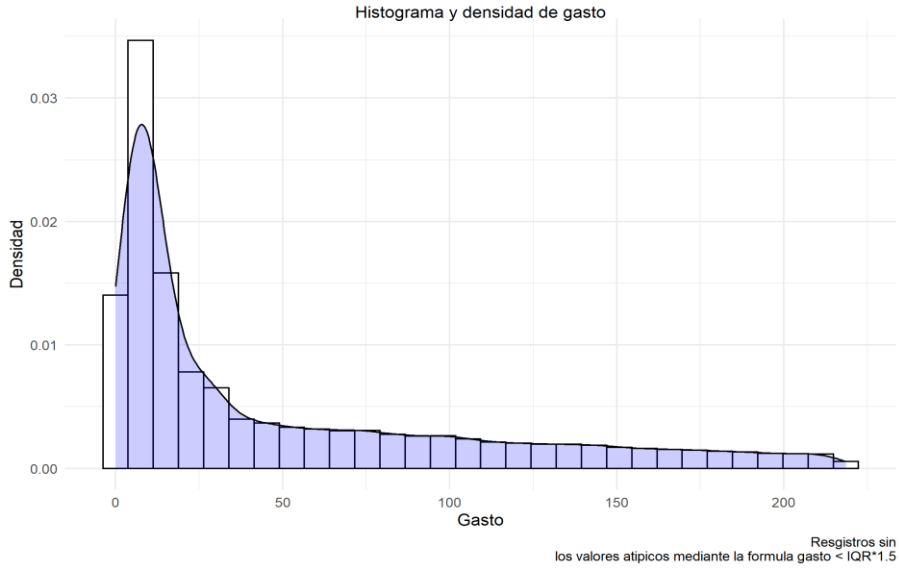
Como vemos en las medidas básicas estadísticas hay una gran diferencia entre la media y mediana, esto se debe por los outliers o valores atípicos y nos indica que hay sesgo (asimétrica) hacia la derecha, nuestra distribución se acumula en los primeros valores. Este tipo de outliers se llaman siniestros punta, y en muchas ocasiones están reasegurados. En la Ilustración 8 podemos ver el histograma de la variable gasto quitándonos valores atípicos según la fórmula:

$$gasto < (IQR) * 1.5$$

siendo IQR el rango intercuartílico que es la diferencia del intercuartil tercero $Q3$ y el intercuartil primero $Q1$.

Se debe atender que el valor mínimo es 0.04, esto es debido a las capitas que parte de su coste se distribuye entre todos los asegurados. Las capitas son paquetes de actos, por ejemplo, una cápita de análisis de sangre incluye 1000 prueba, y parte del coste de estas se distribuye entre todos los asegurados.

Ilustración 8. Histograma y densidad de gasto



Fuente: Elaboración propia

Podemos observar que nuestra variable es asimétrica positiva (o sesgada a la derecha), y solo obtenemos una moda (distribución unimodal). Se han eliminado 19.173 datos que contenían un gasto superior a 218,62€, se queda la BBDD en 87.333.

Se observa en la Tabla 15 los principales estadísticos de las variables numéricas sin los valores outliers. En este caso vemos que las variables independientes no han variado mucho respecto a los datos con outliers.

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
gasto	0	1	50.07	57.34	0.04	8.14	21.07	80.29	218.62	
Prima_pagada	0	1	188.45	124.80	0.32	94.67	168.26	257.75	1486.87	
acto_tot	0	1	2.12	3.27	0.00	0.00	1.00	3.00	47.00	
Asegurados	0	1	2.73	1.27	1.00	2.00	3.00	4.00	8.00	
EDAD_ASEG	0	1	31.66	17.76	0.00	15.00	34.00	46.00	88.00	

Fuente: Elaboración propia

Se observa en la Tabla 17 las frecuencias para las variables categóricas que se utilizarán sin los valores outliers. En este caso no ha variado mucho las frecuencias.

Tabla 17. Variables cualitativas sin outliers.

variable	levels	count	percentage
Canal_Entrada	TECNICA	11885	13.6%
Canal_Entrada	OTRO	15518	17.8%
Canal_Entrada	INTERNET	16934	19.4%
Canal_Entrada	EMISION	17905	20.5%
Canal_Entrada	TELEFONO	25090	28.7%
CLIENTE	NO	29230	33%
CLIENTE	SI	58102	67%
Estado_civil	SOLTERO	43502	49.81%
Estado_civil	CASADO	43830	50.19%
Exclusion	Si	22826	26%
Exclusion	No	64506	74%
Producto_contratado	BS	40646	46.5%
Producto_contratado	CM	46686	53.5%
Provincia	CADIZ	3121	3.57%
Provincia	ALICANTE	3842	4.40%
Provincia	VALENCIA	6078	6.96%
Provincia	SEVILLA	6438	7.37%
Provincia	MALAGA	6997	8.01%
Provincia	MADRID	11688	13.38%
Provincia	BARCELONA	13514	15.47%
Provincia	OTRO	35654	40.83%
Seguro_Ant	Si	29125	33%
Seguro_Ant	No	58207	67%
Sexo	H	41327	47.3%
Sexo	M	46005	52.7%
Tipo_Copago	CON COPAGO	14637	17%
Tipo_Copago	SIN COPAGO	72695	83%
valorado	No	36423	42%
valorado	Si	50909	58%

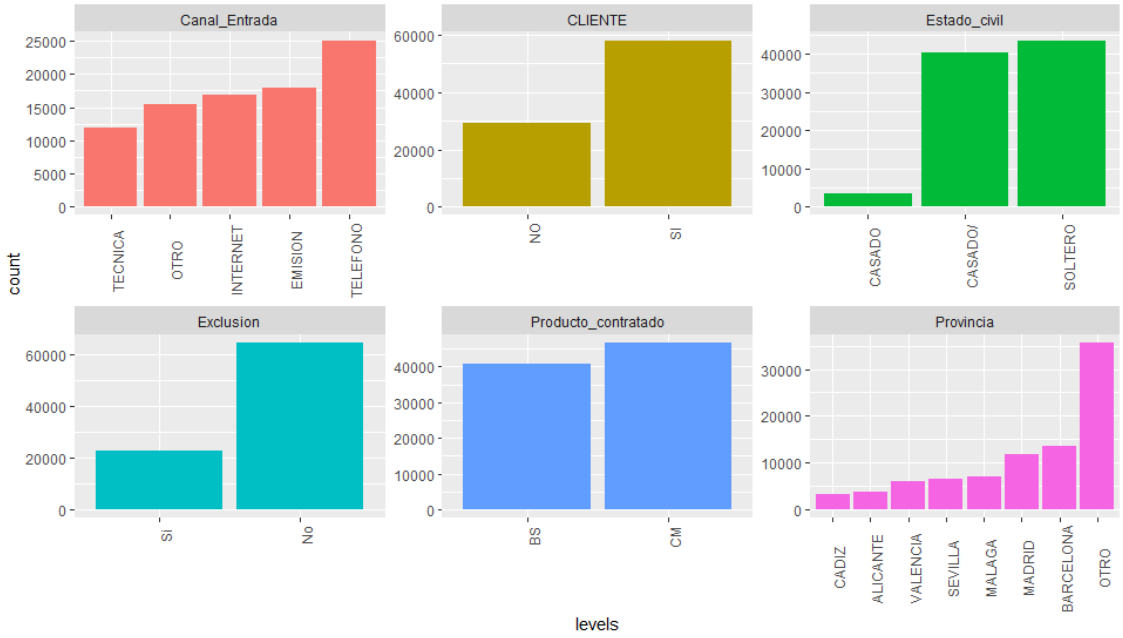
Fuente: Elaboracion propia

4.1.2 Variables Independientes

En este apartado se estudiará las variables independientes que contiene la base de datos. Se realizará mediante la ayuda de Diagramas y Boxplot.

Se observa en la Ilustración 9 la distribución de las variables cualitativas.

Ilustración 9. Diagrama de Barras de Canal entrada, Cliente, Estado civil, exclusión, Producto contratado y provincia.



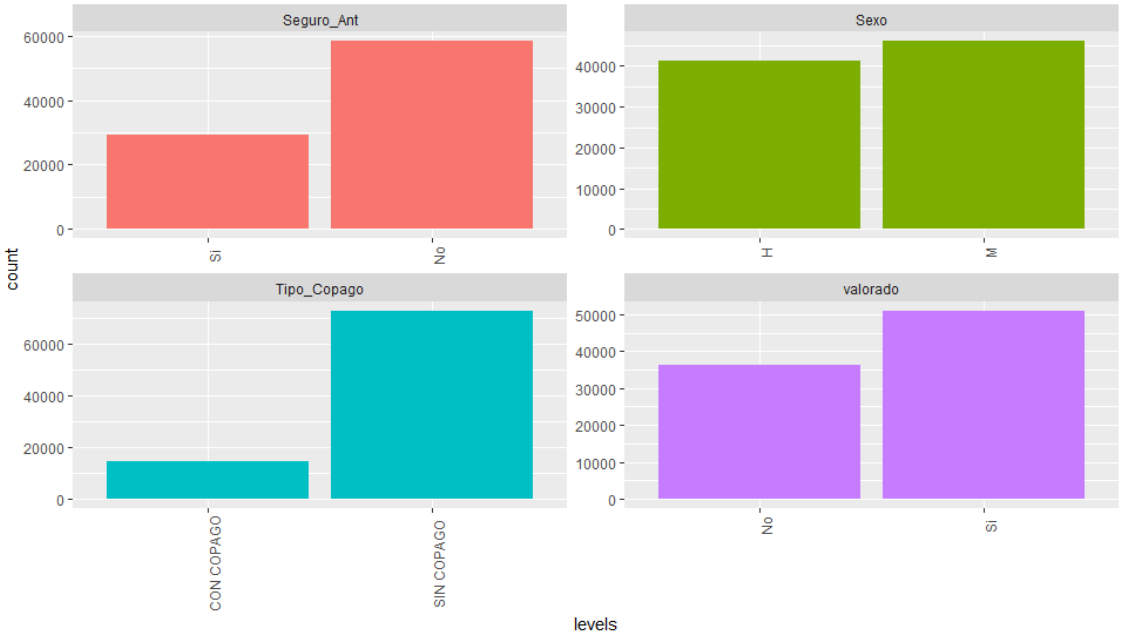
Fuente: Elaboración propia

El 44% de los asegurados tienen el producto básico. Se observa que hay un 50% de asegurados con estado civil Casado/Otro y un 50% con estado Soltero. El canal de entrada del asegurado en un 28% es por telefono, esta categoria recoge aquellos asegurados que llaman por teléfono a la compañía. En un 20% el canal de emision, que son las llamadas emitidas por los teleoperadores a los asegurados. En un 20% el canal de internet, que representa cuando el cliente cotiza y se informa en la web del seguro o algun comparador de seguros. En un 15% se realizan por Técnica, que se categoriza de esta forma cuando se debe realizar una nueva poliza para realizar algun cambio respecto la anterior. Y por ultimo un 17% otros canales, que son categorizaciones internas de la compañía. Por provincia destacando a Barcelona

que representa el 17% de los asegurados. Solo hemos categorizado aquellas provincias con una representatividad mayor al 4%. De los que fueron valorados un 20% ha resultado tener alguna exclusión en el seguro de salud.

En la Ilustración 10 se puede observar la distribución de las variables cualitativas.

Ilustración 10. Diagrama de Barras de Seguro anterior, sexo, tipo de copago y valorado.

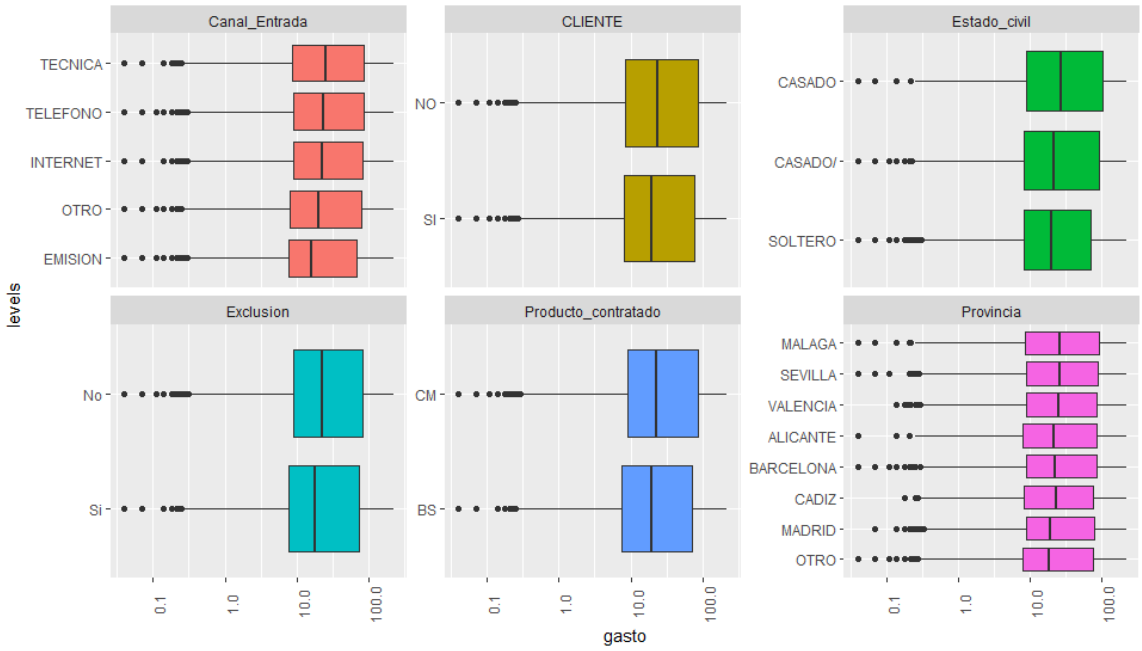


Fuente: Elaboración propia

Se puede observar la distribución de la cartera por tipo de copago siendo este el 85% con copago. El sexo se distribuye en un 55% de mujeres y 45% por hombres. Los asegurados que se han valorado suponen el 63%. En un 66% los asegurados han tenido un seguro de salud anterior.

En la Ilustración 11 se representan los Boxplot de las variables categóricas:

Ilustración 11. Boxplot Variables categóricas por gasto

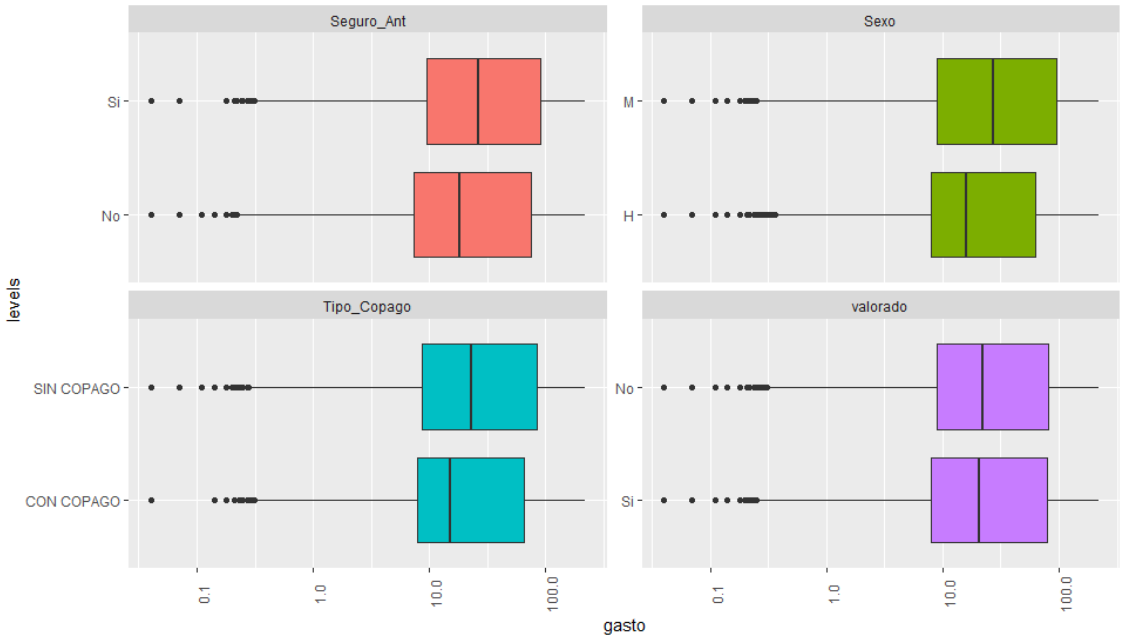


Fuente: Elaboración propia.

Se observa que los asegurados con alguna exclusión tienen menos gasto que aquellos que la tienen. También se observa una pequeña diferencia en el gasto de los que no son clientes, superior a los que son clientes de la compañía.

En la Ilustración 12 se representan los Boxplot de las variables categóricas:

Ilustración 12. Boxplot Variables categóricas

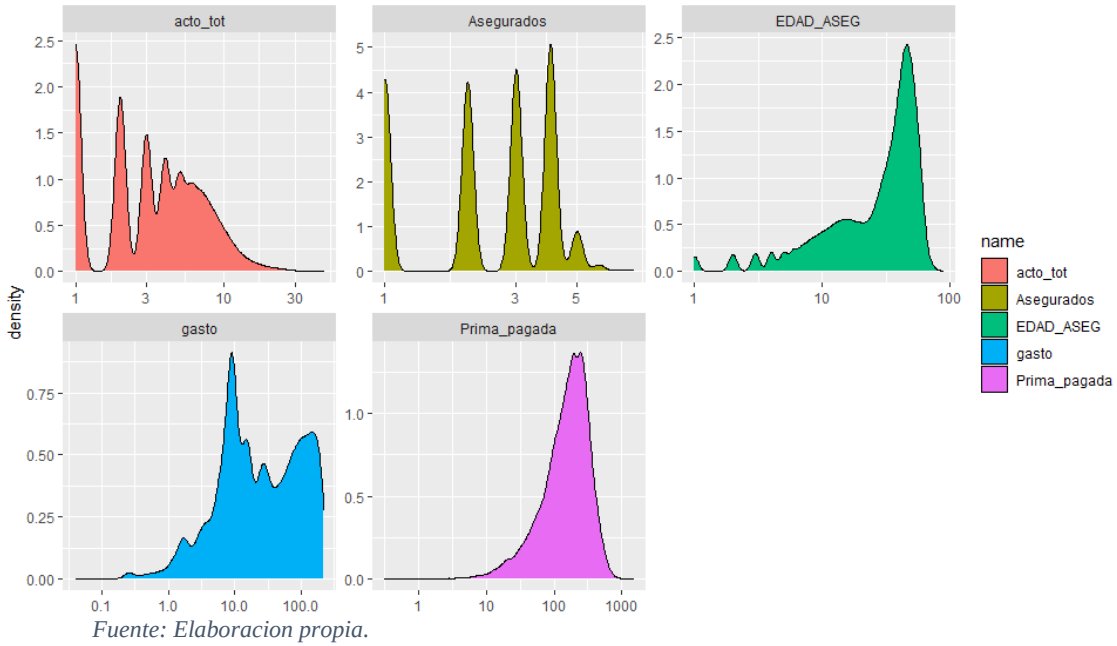


Fuente: Elaboración propia.

Se observa que los asegurados con estado civil casado tienen más gasto que los solteros. Los que tienen el producto sin copago tienen más gasto que un producto con copago. Nos encontramos la mayor diferencia en el sexo, las mujeres tienen mucho más gasto que los hombres.

En la Ilustración 13 se representan las distribuciones de las variables numéricas. La prima media por asegurado es de 209.44€, se utiliza en media 5 veces al año por asegurado y la media de asegurados por póliza es de 2.6, lo que determina que en la cartera tienen un gran peso las pólizas familiares. Además, la media de edad es de 36 años, lo que representa una cartera joven.

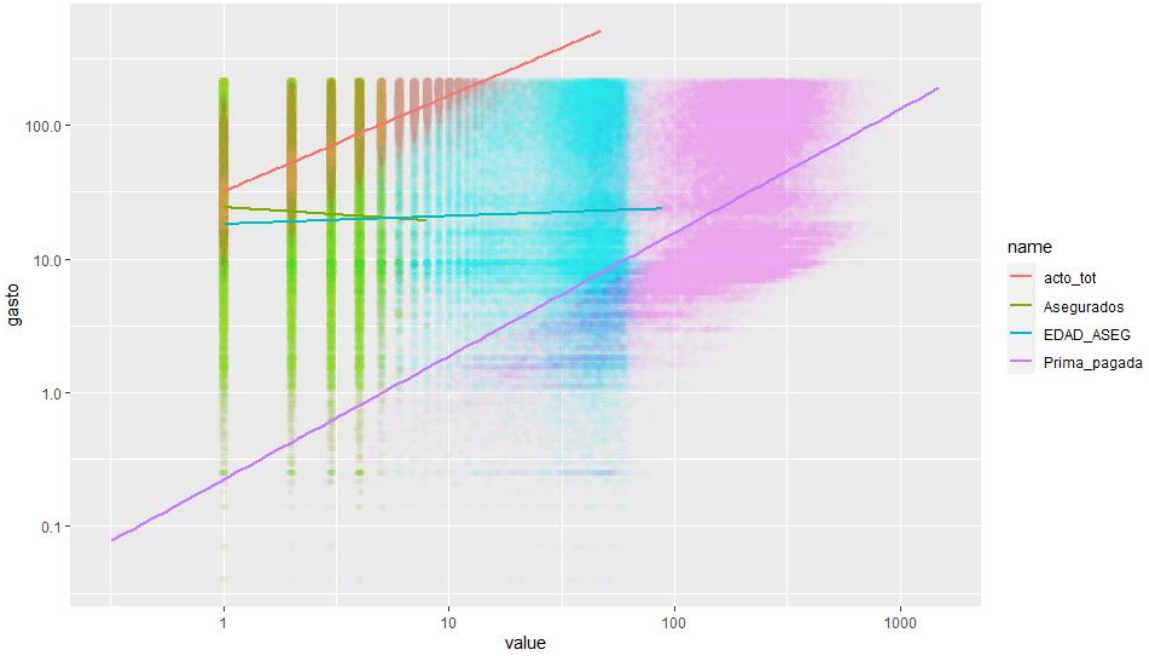
Ilustración 13. Histograma de las variables continuas con escala logarítmica.



Fuente: Elaboracion propia.

En la Ilustración 14 se representan las variables continuas respecto al gasto por el método de independencia en media, cada punto encima de la recta se interpreta en términos de esperanzas condicionales $E[y|x] = E[y]$.

Ilustración 14. Scatterplot de las variables numéricas.



Fuente: Elaboracion propia.

Podemos observar que hay una tendencia lineal con la variable prima pagada, y se concentran en el extremo derecho de la variable gasto. También se observa que hay una correlación alta del gasto con los actos totales.

4.1.3 Correlaciones

Realizamos el estudio de las correlaciones. En la Tabla 16 se presentas las correlaciones de las variables numéricas concluyendo que no hay correlaciones altas entre las variables independientes.

Tabla 16. Tabla de correlaciones numéricas

	gasto	Prima_pagada	acto_tot	Asegurados	EDAD_ASEG
gasto	1.00000000	0.2926003	0.82478106	-0.06472379	0.06712463
Prima_pagada	0.29260033	1.0000000	0.19248087	-0.12927146	0.20190980
acto_tot	0.82478106	0.1924809	1.00000000	-0.07585520	0.07335052
Asegurados	-0.06472379	-0.1292715	-0.07585520	1.00000000	-0.25301609
EDAD_ASEG	0.06712463	0.2019098	0.07335052	-0.25301609	1.00000000

Fuente: Elaboración propia

En la Tabla 17 se puede observar las correlaciones de las variables categóricas. Este estudio se ha realizado con la V de Cramer para comprobar la independencia entre las variables categóricas. Se puede observar que no hay una correlación alta entre las variables.

Tabla 17. Tabla de correlaciones de variables categóricas.

	Seguro_Ant	Sexo	Provincia	CLIENTE	Estado_civil
Seguro_Ant	1.000000000	0.0049349534	0.09603435	0.03932878	0.0121829126
Sexo	0.004934953	1.0000000000	0.02690078	0.04258103	0.0003620112
Provincia	0.096034346	0.0269007833	1.00000000	0.13175677	0.0295253714
CLIENTE	0.039328785	0.0425810281	0.13175677	1.00000000	0.0231441918
Estado_civil	0.012182913	0.0003620112	0.02952537	0.02314419	1.0000000000
Producto_contratado	0.355705751	0.0401566103	0.15162023	0.01105306	0.0279757992
Tipo_Copago	0.011418087	0.0207241117	0.05313329	0.06431616	0.0269786716
Canal_Entrada	0.068643858	0.0221945513	0.05202383	0.32129785	0.0155863325
Exclusion	0.059610529	0.0123363331	0.04639664	0.02640623	0.0494183671
valorado	0.055759978	0.0102321803	0.04259001	0.01873875	0.0414715225

	Producto_contratado	Tipo_Copago	Canal_Entrada	Exclusion
Seguro_Ant	0.35570575	0.01141809	0.06864386	0.05961053
Sexo	0.04015661	0.02072411	0.02219455	0.01233633
Provincia	0.15162023	0.05313329	0.05202383	0.04639664
CLIENTE	0.01105306	0.06431616	0.32129785	0.02640623
Estado_civil	0.02797580	0.02697867	0.01558633	0.04941837
Producto_contratado	1.00000000	0.21980217	0.09033286	0.02431226
Tipo_Copago	0.21980217	1.00000000	0.10710249	0.11186668
Canal_Entrada	0.09033286	0.10710249	1.00000000	0.04236550
Exclusion	0.02431226	0.11186668	0.04236550	1.00000000
valorado	0.02448754	0.07016729	0.05111454	0.49348703

	valorado
Seguro_Ant	0.05575998
Sexo	0.01023218
Provincia	0.04259001
CLIENTE	0.01873875
Estado_civil	0.04147152
Producto_contratado	0.02448754
Tipo_Copago	0.07016729
Canal_Entrada	0.05111454
Exclusion	0.49348703
valorado	1.00000000

Fuente: Elaboración propia.

4.2 Software

En este trabajo se emplearán los programas de SAS Interprise guide 7.1 Y y R Studio.

SAS es una herramienta para el análisis estadístico y la generación de informes, incluye parte de su propio lenguaje SAS y SQL, y es un software con licencia privada.

R Es una herramienta de análisis estadístico que utiliza su propio lenguaje, es gratuito y consta de innumerables paquetes que proporcionan formulación o

algoritmos creados por usuarios de este entorno. El paquete que se utilizará para desarrollar los modelos es Tidymodels, es una colección de paquetes para el modelado y aprendizaje automático. Esta colección de paquetes integra todo lo necesario para desarrollar todas las fases de creación del modelo.

El paquete tidymodels contine cientos de funciones, pero los más importantes son las siguientes funciones:

- parsnip: Paquete para definir de los modelos.
- recipes: Paquete para el tratamiento de datos y su reprocesado.
- rsample: Paquete para la validación de los modelos.
- dials: Paquete para modificar y manejar los valores de los hiperparametros.
- tune: Paquete para realizar el "tunning" de los modelos.
- yardstick: Paquete que calcula las métricas de los modelos.
- workflows: Paquete que unifica todo el proceso del modelo.

Los datos se han extraído de tablas en SAS, y se han preparado para poder tratarlas directamente con R. Se podrá ver el código utilizado en el ANEXO.

Para la ejecución de los modelos se ha utilizado Amazon SageMaker. Es una plataforma que permite crear, entrenar e implementar modelos de ML. Se ha realizado en esta plataforma para poder ejecutar los modelos ya que necesitan una potencia que un ordenador personal no puede aportar. SageMaker permite utilizar el lenguaje de R con el cual se han construido los modelos. Este servicio se ofrece en la nube y permite elegir la potencia de computación utilizada para desarrollar los modelos finales.

5. MODELOS

En este capítulo se preparan los datos de entrenamiento y testeo para los modelos. Se explica la configuración de los parámetros de los modelos y demás elementos técnicos y estadísticos para la modelización.

5.1 Datos de entrenamiento y test.

Los modelos de predicción tienen por objetivo el estudio de datos conocidos para obtener datos desconocidos. Para realizar dicha comparación es necesario dividir los datos en dos conjuntos, uno para entrenar el modelo y otro para validar el modelo.

Creamos un grupo de entrenamiento (trainset) sobre el cual se generará el modelo, y otro de prueba (o Hold-out) sobre el cual se probará el mismo. La parte de entrenamiento abarca un 90% de nuestra base de datos, y el 10% restante será nuestro Hold-out. Los datos de entrenamiento son 78.597 y los datos de Hold-out son 8.735.

5.1 Modelado en R.

Los modelos se ejecutarán en el entorno de Tidyverse, que se compone de las siguientes características(Pimpler, 2017):

- División de los datos en un conjunto de entrenamiento del modelo y un conjunto de testeo.
- Reprocesamiento de los datos, con el fin de facilitar el trabajo al algoritmo de ML, se procesan los datos eliminando NAs, centrando y escalando los datos numéricos y creando nuevas variables dummy para las variables cualitativas.
- Creación del modelo, se utilizará el paquete tidymodels para crear un modelo optimizado, los pasos para la optimización de los modelos son los siguientes(Wickhman, Hadley ; Grolemund, 2016):

    I. Entrenamiento: Se elige el modelo que queremos ejecutar del paquete parsnip, se define qué implementación del modelo se quiere emplear y se ajusta para conseguir el modelo ajustado.
    II. Evaluación del modelo: Para mejorar la estimación del modelo, se utilizan estrategias de validación de los errores para obtener el mejor modelo.
    III. Optimización de hiperparametros: los algoritmos de ML contienen en su ecuación uno o varios parámetros que no se

aprenden con los datos, a estos se les conoce como hiperparametros. Estos parámetros se pueden modificar para mejorar la predicción del modelo.

IV. Predicción: una vez creado el modelo, este se emplea para predecir nuevas observaciones.

Se programan los pasos que se han comentado en un 'Workflows', que combinan en un solo objeto todos los elementos que se encargan del reprocesamiento, modelado y post procesado del modelo. Para la elección del mejor modelo se ha utilizado la medida de RMSE, que seleccionará el modelo con menor RMSE.

En la evaluación y validación del modelo, para estimar el error que comete el modelo utilizamos la estrategia de validación $Repeated\ K-Fold\ Cross-Validation$, que repite n veces la estrategia de validación $K-Fold\ Cross-Validation$. Este método consiste en dividir los datos aleatoriamente en k grupos del mismo tamaño, para entrenar el modelo se utilizarán $k-1$ grupos y el restante se utilizará para evaluar el rendimiento del modelo. Una vez que se ha repetido k veces para todos los subconjuntos, genera un error de predicción tomando el promedio de los errores de predicción en cada caso.

En la búsqueda de los mejores hiperparametros hemos utilizado la búsqueda por muestreo de hipercubo latino (o LHS)[1], esta búsqueda es una estrategia de optimización de búsqueda local. El LHS examina más valores para cada hiperparametros y garantiza que cada valor aparezca solo una vez en combinaciones combinadas aleatorias. Al identificar buenos valores para cada hiperparametro, se puede hacer combinaciones más fuertes a partir de estos buenos valores. Esta estrategia permite una búsqueda más eficiente utilizando el mismo número de puntos.

5.1.1 Modelo lineal generalizado

[1] Para la búsqueda de los hiperparametros se han tenido en cuenta las ideas del libro: Sarker, Ruhul Amin, and Charles S. Newton. 2007. Optimization Modelling.

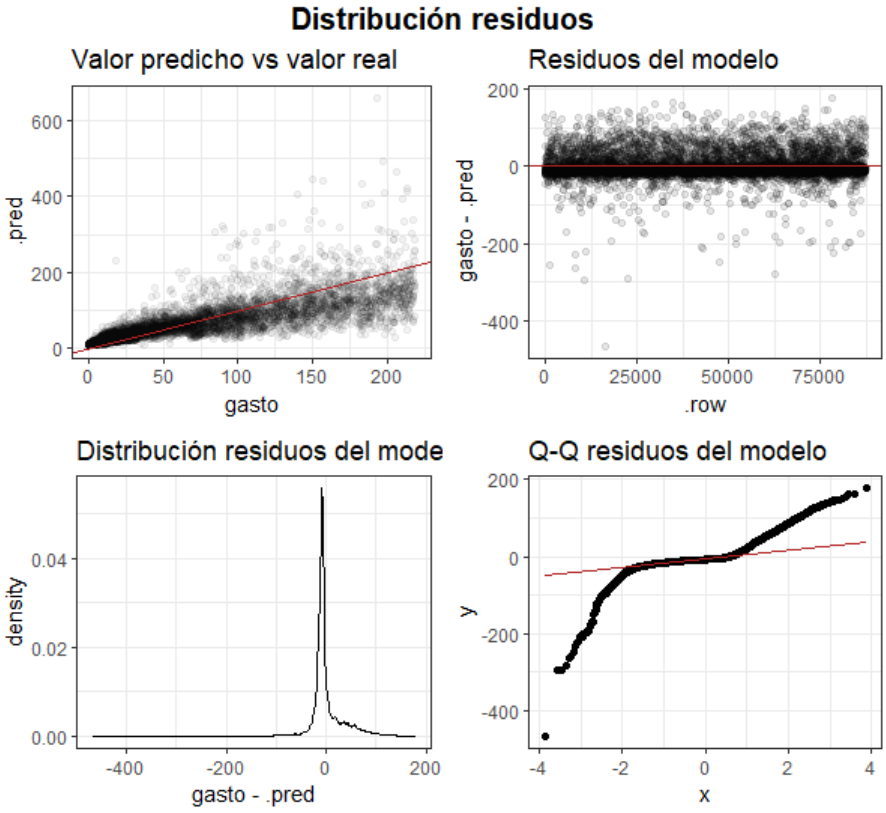
Se muestra el modelo final:

$$gasto = 50.03 + 8.068\ Prima\ Pagada + 54.61\ actos\ Tot + 0.6472\ Asegurados - 0.9086\ Edad\ Asegurado$$

term	estimate	penalty
(Intercept)	50.03	2.022e-10
Prima_pagada	8.068	2.022e-10
acto_tot	45.61	2.022e-10
Asegurados	0.4672	2.022e-10
EDAD_ASEG	-0.9086	2.022e-10

Se puede observar en la ilustración 15 el análisis grafico de los residuos:

Ilustración 15. Distribución de los residuos del modelo GLM.



Fuente: Elaboracion propia.

Se pueden extraer las siguientes conclusiones:

- Valor predicho vs Valor real: Se visualiza el rendimiento del modelo, se representa en el eje de las x que muestran los valores calculados por el modelo y en el eje Y el valor real. Cuando más próximos a la línea roja, mejor será el modelo. En nuestro caso vemos incluso valores atípicos cuando crece el gasto y la mayor concentración se visualiza en los valores de gasto más bajo.
- Residuos del modelo: Se puede observar un patrón de residuos normal, sin autocorrelación.
- Distribución residuos del modelo: Sigue una distribución apuntada con cola larga.
- Q-Q residuos del modelo: Los datos no siguen una distribución normal, tenemos datos extremos con cola larga.

Medidas de bondad del modelo:

$RMSE$	32.9
R^2	0.683

Tras el análisis de las medidas de desempeño, se puede concluir que obtenemos unos resultados medios, y que debemos utilizar otro modelo para mejorar las predicciones.

5.1.2 MARS

Se muestra las variables más importantes de nuestro modelo y los hiperparametros del modelo final:

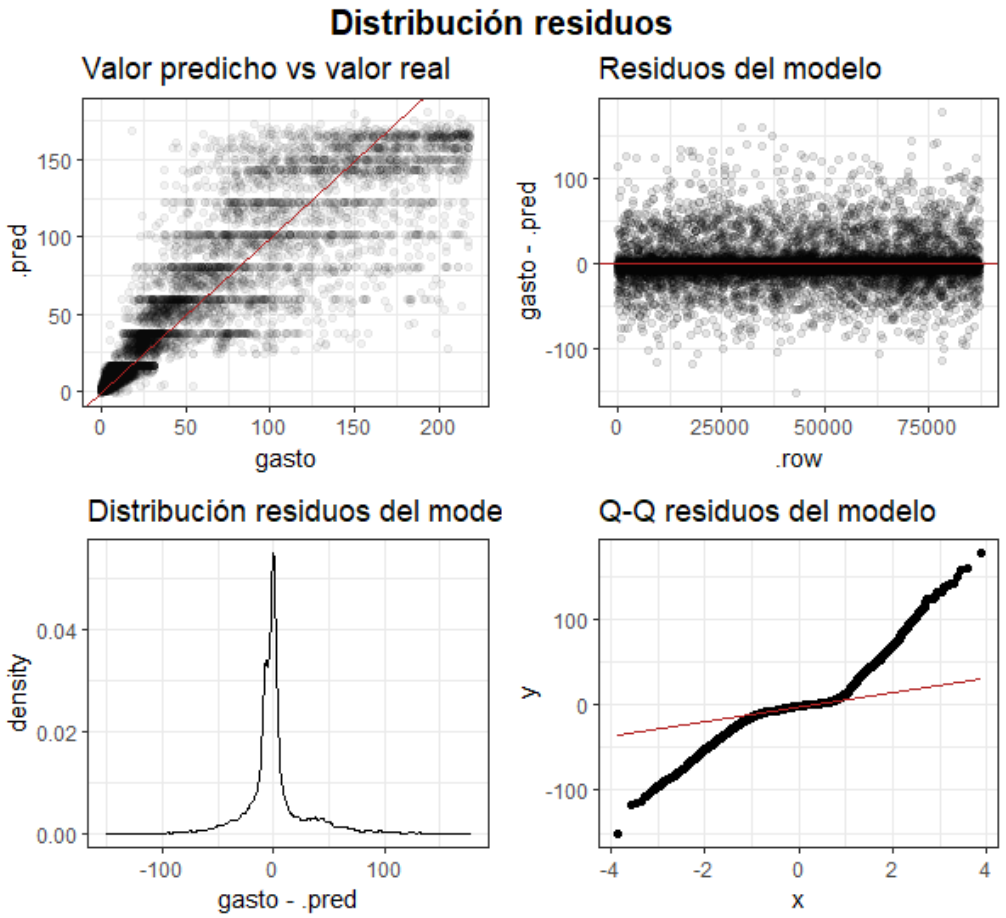
```
model_fit$`MARS` %>%
   extract_fit_parsnip()

## parsnip model object
##
## Fit time:  321ms
## Selected 5 of 7 terms, and 2 of 4 predictors (nprune=5)
## Termination condition: RSq changed by less than 0.001 at 7 terms
## Importance: acto_tot, Prima_pagada, Asegurados-unused, EDAD_ASEG-
unused
## Number of terms at each degree of interaction: 1 4 (additive model)
## GCV 630.286    RSS 49524722    GRSq 0.8076291    RSq 0.807678
```

El modelo selecciona las variables más relevantes, en nuestro caso actos totales, prima ganada, número de asegurados y edad del asegurado. El número máximo de términos para el modelo es de 5, de los cuales obtenemos el mejor modelo con las variables de actos totales y Prima ganada.

Se puede observar en la ilustración 16 el análisis grafico de los residuos:

Ilustración 16. Distribución de los residuos del modelo MARS.



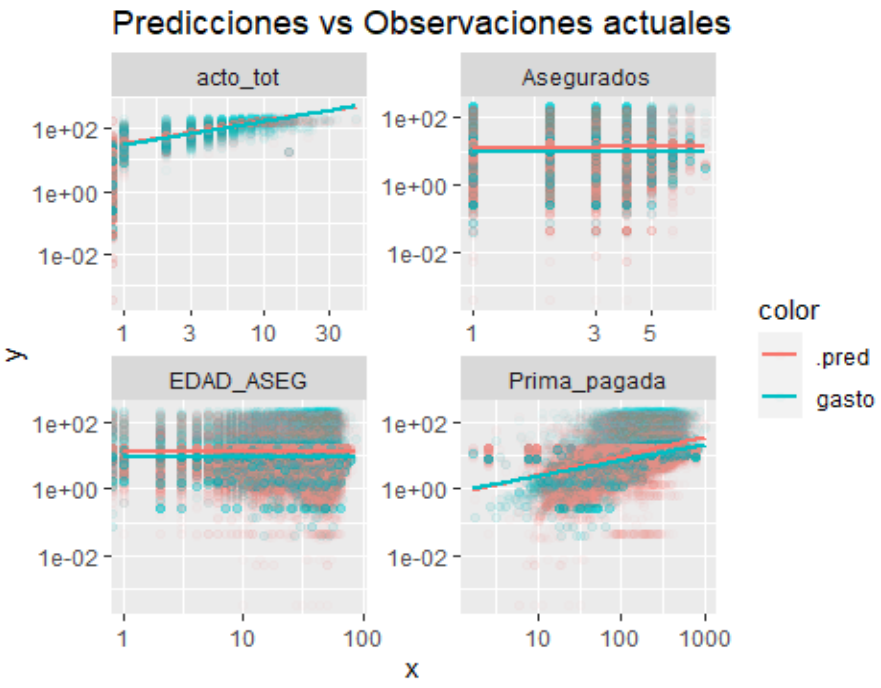
Fuente: Elaboracion propia.

Se pueden extraer las siguientes conclusiones:

- Valor predicho vs Valor real: Para visualizar mejor, realizamos un análisis por variable en la siguiente ilustración.
- Residuos del modelo: Se puede observar un patrón de residuos normal, sin autocorrelación.
- Distribución residuos del modelo: Sigue una distribución apuntada con cola larga.
- Q-Q residuos del modelo: Los datos no siguen una distribución normal, tenemos datos extremos con cola larga.

Se puede observar en la Ilustración 17, las predicciones de las variables más importantes sobre el gasto. Se observa que predicen bien las variables salvo la prima ganada que ante valores altos hay un error mayor.

Ilustración 17. Predicciones vs Observaciones actuales del modelo MARS.



Fuente: Elaboracion propia.

Medidas de bondad del modelo:

$RMSE$	25.91
R^2	.8023

Tras el análisis de las medidas de desempeño, se puede concluir que obtenemos unos resultados buenos, pero no excelentes. Mejoramos la capacidad predictiva respecto al modelo GLM.

5.1.4 SVM

Se puede extraer los hiperparametros del modelo SVM y las medidas más importantes de nuestro modelo final:

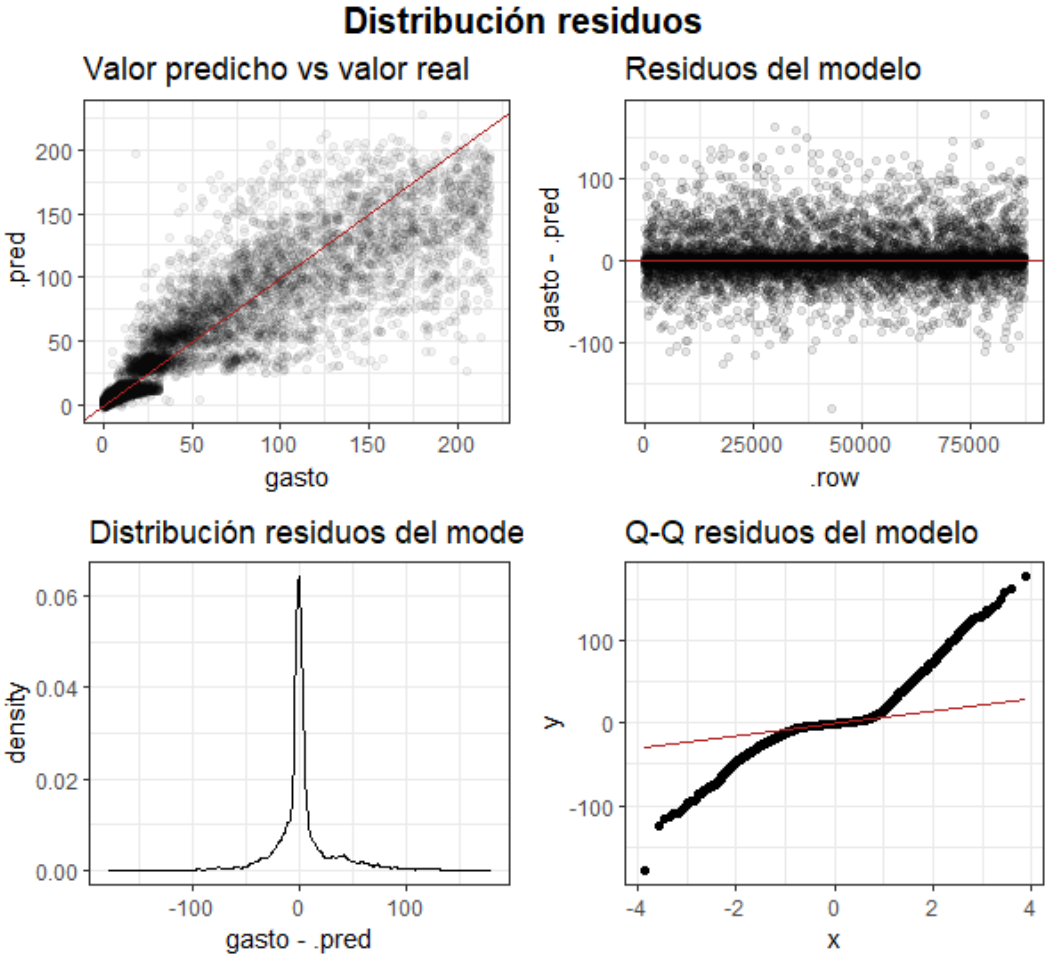
```
model_fit$`SVM_poly` %>%
   extract_fit_parsnip()

## parsnip model object
##
## Fit time:  34m 53.5s
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 0.0847297434863019
##
## Polynomial kernel function.
##  Hyperparameters : degree =  3  scale =  1  offset =  1
##
## Number of Support Vectors : 36914
##
## Objective Function Value : -1218.683
## Training error : 0.196354
```

En los casos de Support vector de regresión se debe utilizar el hiperparametro épsilon que establece un margen de tolerancia para las predicciones, que en nuestro caso será de 0.1 y el coste que determina la penalización aplicada por violar el margen, en nuestro modelo es de 0.08472. La función que utilizamos para transformar y procesar nuestros datos se llama ‘Kernel’ y se utilizara en este caso el polinómico de tres grados con escala (‘scale’) a 1 y compensación (‘offset’) a 1.

Se puede observar en la ilustración 18 el análisis grafico de los residuos:

Ilustración 18. Distribución de residuos del modelo SVM



Fuente: Elaboracion propia.

Se pueden extraer las siguientes conclusiones:

- Valor predicho vs Valor real: el modelo predice bien para valores de gasto bajos y observamos bastantes outliers cuando esta crece con una dispersión alta en estos valores.
- Residuos del modelo: Se puede observar un patrón de residuos normal, sin autocorrelación.
- Distribución residuos del modelo: Sigue una distribución apuntada.
- Q-Q residuos del modelo: Los datos no siguen una distribución normal, tenemos datos extremos con cola larga.

Medidas de bondad del modelo:

$RMSE$	26.24
R^2	.7991

Tras el análisis de las medidas de desempeño, se puede concluir que obtenemos unos resultados buenos, pero no excelentes. Mejoramos la capacidad predictiva respecto al modelo GLM.

5.1.5 Random Forest

Se muestra las variables más importantes de nuestro modelo y los hiperparametros del modelo final:

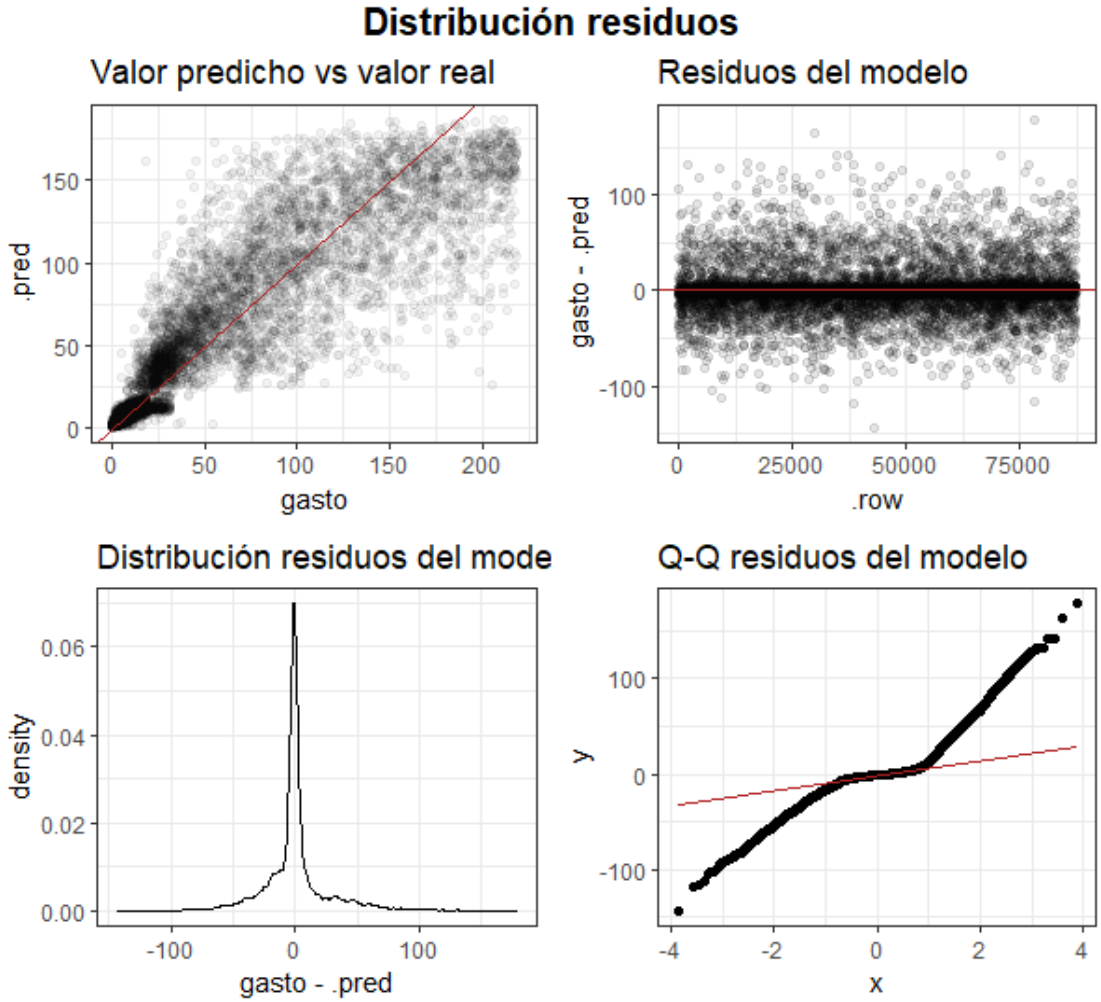
```
model_fit$`Random forest` %>%
   extract_fit_parsnip()

## parsnip model object
##
## Fit time:  20.8s
## Ranger result
##
## Call:
##  ranger::ranger(x = maybe_data_frame(x), y = y, num.trees = ~300,
min.node.size = min_rows(~39L, x), num.threads = 1, verbose = FALSE,
seed = sample.int(10^5, 1))
##
## Type:                             Regression
## Number of trees:                  300
## Sample size:                      78597
## Number of independent variables:  4
## Mtry:                             2
## Target node size:                 39
## Variable importance mode:         none
## Splitrule:                        variance
## OOB prediction error (MSE):       626.6128
## R squared (OOB):                  0.8087477
```

En nuestro modelo final de Random Forest se han utilizado 300 árboles para su simulación, se han seleccionado 4 variables independientes y el modelo se ha optimizado con 2 ramificaciones.

Se puede observar en la ilustración 19 el análisis grafico de los residuos:

Ilustración 19. Distribución de residuos del modelo Random Forest.



Fuente: Elaboracion propia.

Se pueden extraer las siguientes conclusiones:

- Valor predicho vs Valor real: el modelo predice bastante bien para valores de gasto bajos y observamos bastantes outliers cuando esta crece con una alta dispersión en estos valores.
- Residuos del modelo: Se puede observar un patrón de residuos normal, sin autocorrelación.
- Distribución residuos del modelo: Sigue una distribución apuntada con cola larga hacia la derecha.
- Q-Q residuos del modelo: Los datos no siguen una distribución normal, tenemos datos extremos con cola larga.

Medidas de bondad del modelo:

$RMSE$	25.76
R^2	0.8046

Tras el análisis de las medidas de desempeño, se puede concluir que obtenemos unos resultados buenos, pero no excelentes. Mejoramos la capacidad predictiva respecto al modelo GLM.

5.2 Comparación y resultados de los modelos

Se ha incluido el modelo de árbol de decisión para conocer si obteníamos alguna diferencia respecto a un algoritmo más complejo como el de Random forest. En este caso vemos que tras ajustar ambos modelos tenemos los mismos resultados. En la Tabla 18 se recogen las medidas propuestas para comparar los modelos utilizados:

Tabla 18. Tabla con las medidas de cada modelo.

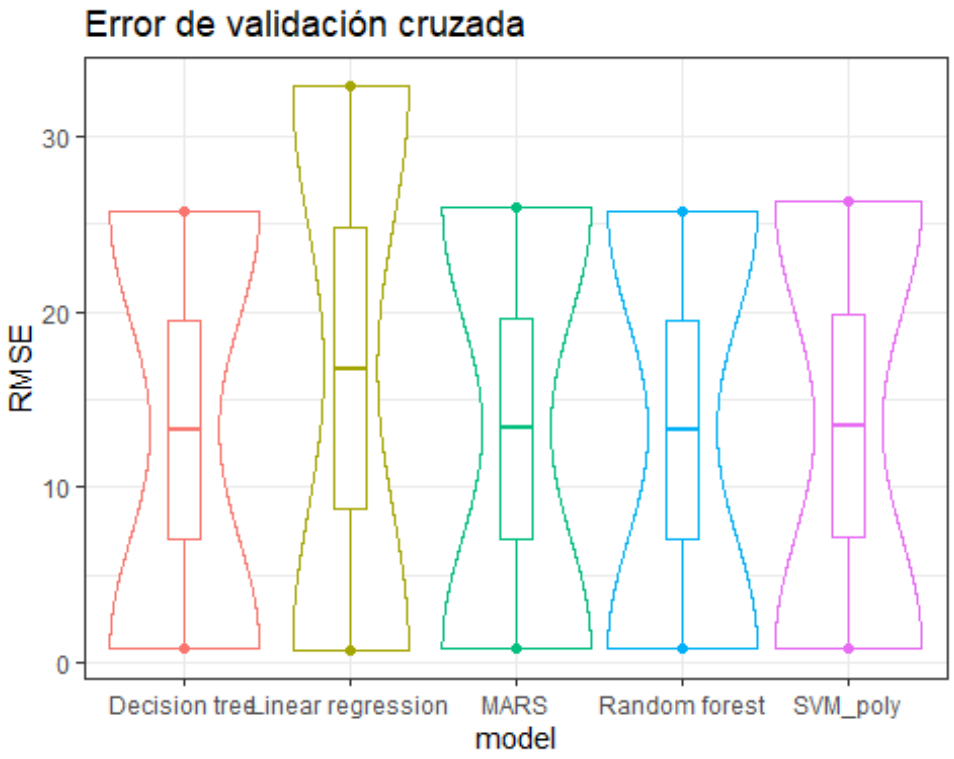
model	.metric	.estimate
Linear regression	rsq	0.6826
SVM_poly	rsq	0.7991
MARS	rsq	0.8023
Random forest	rsq	0.8046
Decision tree	rsq	0.8046
Decision tree	rmse	25.76
Random forest	rmse	25.76
MARS	rmse	25.91
SVM_poly	rmse	26.24
Linear regression	rmse	32.86

Fuente: Elaboración propia.

Revisando los resultados nos decantamos por el modelo MARS con un R^2 de 0.8023 y una mejor distribución de los errores. Tanto el modelo de Random Forest y SVM obtienen buenos resultados.

Se observa en la Ilustración 20 el error obtenido por la validación cruzada y se destaca que todos los modelos salvo el GLM obtenemos un RMSE parecido.

Ilustración 20. Ilustración del error de validación cruzada de los modelos.



Fuente: Elaboracion propia.

5.3 Discusión de los resultados

Los resultamos muestran que tanto el modelo Random Forest como MARS obtienen los mejores resultados, seguido del modelo SVM. El peor resultado lo obtenemos del GLM.

Se han observado como las variables que nos predicen el modelo se limitan a las continuas, por tanto, se debe buscar nuevas variables que aporten más poder predictivo de los modelos. La ventaja de Machine Learning es poder incluir tantas variables como se dispongan y además tratar gran cantidad de datos para mejorar la predicción de los modelos.

Meyers y Hoyweghen (2020) relatan con detalle un reciente experimento realizado en Bélgica con el objetivo de relacionar la mejora de la conducción con la disminución de la siniestralidad utilizando grandes bases de datos con información de los asegurados. En dicho estudio se recoge la dificultad que presenta obtener datos de calidad a través de un dispositivo móvil, y como pueden ser utilizados estas nuevas variables como son la calidad de conducción del asegurado para una tarificación personalizada. Se puede extrapolar a los seguros de salud, en la recogida de datos del asegurado con nuevas variables que contengan el comportamiento de la actividad física del asegurado y de su alimentación. Estas nuevas variables pueden nutrir el modelo y proporcionar nuevas estimaciones que ayuden en la tarificación del seguro.

En el estudio de Porrini (2017) se plantea como puede afectar el nuevo marco regulatorio europeo en la obtención y manejo de los datos masivos que manejan las Entidades Aseguradoras. En dicho estudio se enumeran las principales preocupaciones que deberá tener la Entidad Aseguradora para mantener la privacidad de los datos asegurados, utilizar estos de forma responsables y no crear comportamientos o patrones de discriminación que puedan perjudicar a ciertos clientes. Se puede esperar que en el seguro de salud los datos médicos deban ser protegidos y solo utilizarlos antes el consentimiento expreso del cliente. Como se comentó ya hay regulaciones para que ciertas enfermedades, como el caso de VIH, no perjudiquen en la contratación de un seguro de salud. En la recolección de los datos, actualmente se plantea como se debe regular nuevos datos sobre el asegurado, en el punto de mira están los datos genéticos que proporcionan la probabilidad de tener cierta enfermedad futura y si es ético utilizarlo para decidir si asegurar o no a ciertos clientes.

6. Conclusiones

Mediante este estudio se ha proporcionado conocimiento sobre como las nuevas herramientas proporcionan mayor facilidad para la creación de algoritmos de

Machine Learning y poder comparar diferente técnica para Un estudio en el sector asegurador.

Este trabajo se ha centrado en comparar cuatro modelos predictivos de aprendizaje automático para predecir el gasto de los asegurados y así poder conocer cómo se comportarán los nuevos clientes de la compañía. Para ello, primero se ha realizado el análisis de los datos que ha permitido presentar las variables más más explicativas y cómo filtrar los datos para eliminar outliers y que nuestro modelo pueda predecir mejor. Seguidamente, se han entrenado varios modelos, obteniendo los modelos finales que mejor nos predicen la variable de estudio. Además, se ha obtenido que las variables de actos y prima pagada son las variables que explican mejor nuestro modelo, anotando que las técnicas más sencillas como el modelo lineal generalizado han obtenido peores resultados que las técnicas más avanzadas como SVM o MARS.

En conclusión, a este trabajo, podemos determinar que el mejor modelo que se adapta a nuestros datos es el algoritmo MARS. Una de las fortalezas de este estudio es haber utilizado datos reales tanto para el training como para la validación. Sin embargo, el hecho de que la muestra proceda de una empresa específica, y que esta sea de recién creación, determina que se debe ser cautos en extrapolar el modelo a otras compañías.

Este trabajo abre nuevos caminos en la tarificación y valoración de riesgo mediante modelos que utilicen la entrada de grandes bases de datos y que puedan utilizar modelos complejos para personalizar la prima al potencial cliente. Uno de los retos podría consistir en obtener nuevas variables de datos o la utilización de novedosos algoritmos como es el caso de lightGBM, un nuevo algoritmo con buenos resultados en regresión.

Una aplicación de nuestro modelo para el conocimiento de este sector seria la aplicación de la predicción del gasto como una variable factor para la determinación de la prima del asegurado. Siendo el cálculo de la tarifa para el cliente en tiempo real, con el aprendizaje del modelo se podría obtener la mejor prima para un nuevo cliente.

Para finalizar, se debe añadir que en el entorno del Big Data se debe seguir investigando en tres aspectos: los algoritmos de aprendizaje y su optimización. En la

reducción de los errores en los resultados del modelo y su estabilidad en el funcionamiento del modelo y cómo proceder ante la alimentación constante de datos. Y finalmente el uso de indicadores sintéticos que permitan “comprender” las variables que se están recogiendo.

BIBLIOGRAFIA

Boodhun, Noorhannah. 2017. "A Review of Data Analytical Approaches in the Insurance Industry." *Journal of Applied Technology and Innovation* 1(1):58–73.

Breiman, Leo; Jerome H. Friedman ;. Richard A. Olshen ;Charles J. Stone. 1984. *Classification and Regression Trees*. CHAPMAN & HALUCRC.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(2):123–40. doi: 10.1007/bf00058655.

Carmona, E. .. 2016. "Tutorial Sobre Máquinas de Vectores Soporte (SVM)." (November):1–27.

Cheek, P. J., P. McCullagh, and J. A. Nelder. 1990. "Generalized Linear Models, 2nd Edn." *Applied Statistics* 39(3):385.

Eiopa. 2019. "Big Data Analytics in Motor and Health Insurance: A Thematic Review." 68. doi: 10.2854/54208.

Europea, Unión. 2002. "Tratado Constitutivo de La Comunidad Europea." *Diario Oficial de Las Comunidades Europeas* 325(33):33–184. doi: http://www.europa.eu.int/eur-lex/lex/es/treaties/index.htm.

Flores, Alvaro, Sebastián Maldonado, and Richard Weber. 2015. "Selección de Atributos y Support Vector Machines Adaptado Al Problema de Fuga de Clientes´Alvaro." *Revista Ingeniería de Sistemas* 87–109.

ICEA. 2021. *El Seguro de Salud a Marzo. Año 2021*.

Medina-Merino, Rosa Fátima, and Carmen Ismelda Ñique-Chacón. 2017. "Bosques Aleatorios Como Extensión de Los Árboles de Clasificación Con Los Programas R y Python." *Interfases* 0(010):165. doi: 10.26439/interfases2017.n10.1775.

Meyers, Gert, and Ine Van Hoyweghen. 2020. "'Happy Failures': Experimentation with Behaviour-Based Personalisation in Car Insurance:" *Https://Doi.Org/10.1177/2053951720914650* 7(1). doi: 10.1177/2053951720914650.

Montero, Roberto. 2016. "Modelos de Regresión Lineal Múltiple." *Documentos de*

Trabajo En Economía Aplicada 60.

Montillo, Albert, and Haibin Ling. 2009. "Age Regression from Faces Using Random Forests." *Proceedings - International Conference on Image Processing, ICIP* (May):2465–68. doi: 10.1109/ICIP.2009.5414103.

Nicolas, Patrick R. 2014. *Scala for Machine Learning*.

OECD/European Observatory on Health Systems and Policies. 2019. "State of Health in the EU. España: Perfil Sanitario Nacional 2019." *OECD/European Observatory on Health Systems and Policies* 1–24.

Pimpler, Eric. 2017. *Data Visualization and Exploration with R*.

Porrini, Donatella. 2017. "Regulating Big Data Effects in the European Insurance Market." *Insurance Markets and Companies* 8(1):6–15. doi: 10.21511/INS.08(1).2017.01.

Vanegas, Jairo, and Fabián Vásquez. 2017. "Multivariate Adaptative Regression Splines (MARS), Una Alternativa Para El Análisis de Series de Tiempo." *Gaceta Sanitaria* 31(3):235–37. doi: 10.1016/j.gaceta.2016.10.003.

Wickhman, Hadley ; Grolemund, Garrett. 2016. *R for Data Science*. Vol. 47.