Anales del Instituto de Actuarios Españoles, 4ª época, 24, 2018/123-147

MACHINE LEARNING AND PREDICTIVE MODELING FOR AUTOMOBILE INSURANCE PRICING

MACHINE LEARNING Y MODELIZACIÓN PREDICTIVA PARA LA TARIFICACIÓN EN EL SEGURO DE AUTOMÓVILES

Montserrat Guillen[1] y Jessica Pesantez-Narvaez[1*]

Abstract

Historical records of insured policy holders constitute an ideal environment for the development of *machine learning* algorithms. These procedures are implemented on databases in order to extract knowledge. Here we explore different approaches to the prediction of claims and premiums in the automotive industry, comparing their implementation in a real sample, randomly divided into training and testing. We propose measures to help in the evaluation of the methods and their practical implication for the prediction of rare events and premium calculation. The main conclusion is that the dispersion of prices, and specifically the difference between the maximum pure and minimum premium, can become very different according to the predictive method used.

Keywords: data science, artificial intelligence, nonlife insurance, premiums, claims

Resumen

La información histórica de los asegurados constituye un entorno idóneo para el desarrollo de los algoritmos *machine learning*, cuya finalidad es extraer conocimiento a partir de bases de datos. En este artículo exploramos diversas aproximaciones a la predicción de la siniestralidad y las primas del ramo del automóvil, comparando su implementación en una muestra real, dividida aleatoriamente en muestras de entrenamiento y test. Proponemos medidas para ayudar en la valoración de los métodos y su implicación práctica para la predicción de eventos pocos frecuentes y el cálculo de primas. La principal conclusión es que la dispersión de precios, y

Se agradece la financiación recibida del proyecto del Ministerio de Economía y Competitividad y FEDER, ECO2016-76203-C2-2-P.

[1] Dept. Econometría, Riskcenter-IREA, Universidad de Barcelona, Av. Diagonal, 690, 08034 Barcelona.
*Autor para correspondencia (jessica.pesantez@ub.edu)

DOI 10.26360/2018_6

concretamente la diferencia entre la prima pura máxima y mínima, puede llegar a ser muy diferente según el método predictivo utilizado.

Palabras clave: ciencia de datos, inteligencia artificial, seguros generales, primas, siniestros

1. Introducción y motivación

La creciente disponibilidad de grandes bases de datos ha fomentado el desarrollo de algoritmos y procedimientos dedicados al análisis de volúmenes de información que hace unas décadas eran inimaginables. Padilla-Bareto *et al.* (2017) y Guillen (2016) resumen cómo el entorno asegurador se ha visto involucrado en este proceso, dado que no solo ha aumentado la capacidad de almacenaje y procesamiento de los registros sobre los asegurados, sino que además se han creado nuevas metodologías a fin de extraer conocimiento de dichos datos de una forma sistemática. A la generación de resultados, basados principalmente en hallar estructuras y patrones en un compendio de registros digitales, se le denomina *ciencia de los datos* y, más concretamente, cuando de ello se derivan resultados aplicables a las decisiones empresariales, se habla de *aprendizaje automatizado* o "machine learning" (Michalski *et al.,* 2013).

En este artículo abordamos qué potencial de aplicación tienen los métodos estándar del machine learning en la tarificación de seguros. Eligiendo una muestra real de pólizas del automóvil, analizamos la capacidad predictiva de distintas aproximaciones, y estudiamos cómo se pueden conseguir mejoras tanto en la precisión como en la robustez de la tarificación. Precisión, en el sentido de poder calcular una prima más acertada para cada asegurado, con un intervalo de confianza más reducido y menores tasas de error. Robustez, para que dichas predicciones sean lo más estables posibles cuando se produzcan modificaciones en la cartera, tales como algún siniestro de coste muy elevado, por ejemplo. De los resultados obtenidos, corroboramos que los modelos predictivos clásicos tienen la ventaja de proporcionar una interpretabilidad directa e intuitiva, es decir, se puede cuantificar fácilmente el impacto de cada factor de riesgo en el cálculo de la prima y justificar si cada uno de los factores juega a favor o en contra de un precio final (ver algunos casos aplicados en Guelman *et al.*, 2014 y Frees, 2009). Sin embargo, pese a que tradicionalmente se acepta que los métodos de la inteligencia artificial son cajas negras, algunas de las nuevas metodologías no resultan ser tan crípticas como se creía en sus inicios y pueden suponer

una nueva forma, mucho más ajustada, de obtener tarificadores en el entorno del machine learning (Witten *et al.,* 2016).

El elemento más innovador de esta presentación es el proporcionar medidas de comparación de la capacidad predictiva de los métodos, viendo además cómo abordar el cálculo de la prima bajo diferentes perspectivas y señalando las diferencias según el método de cálculo de la prima pura que se haya empleado. Además, mediante un análisis comparativo, se obtienen resultados sobre la estabilidad de la capacidad predictiva y el impacto en el precio final.

En primer lugar se realiza una revisión de los principales métodos existentes, exponiendo cómo ha ido incorporándose a la escena de la ciencia de los datos y se deja para el anexo una exposición más técnica de las metodologías que se utilizan en la parte empírica. En el apartado de datos y resultados se presenta una muestra de datos sobre siniestros del seguro del automóvil así como la exploración que se ha realizado de la misma. La comparativa e implementación de nuevas metodologías revela la importancia de los métodos de clasificación y su utilidad en el ámbito de la tarificación de los seguros. En nuestro caso se pretende buscar modelos que permitan discriminar entre las pólizas que han declarado algún siniestro frente al resto, de forma que puedan efectuarse predicciones posteriormente, y en consecuencia que se pueda tarificar. Se concluye que la elección del método predictivo puede tener una enorme relevancia para la distribución de las primas finales, siendo destacable que la diferencia entre la prima máxima y mínima puede ser muy distinta dependiendo del modelo implementado. Finalmente, se exponen algunas conclusiones y recomendaciones de utilidad práctica.

2. Antecedentes

Es obligado referirse al Test de Turing, creado por Alan Turing en 1950, para situar cronológicamente el inicio de la era de la inteligencia artificial, ya que es en ese momento cuando se crea un procedimiento mediante el cual se contrasta si un humano puede distinguir entre otro humano o un ordenador, con el que interactúa. Desde entonces, la gran velocidad a la que se ha desarrollado la informática impide un relato detallado de cuáles han sido los diferentes hitos hasta alcanzar la actualidad (Turing, 2009). Algunos de los principales avances incluyen, en 1957 la creación de la primera red neuronal artificial por parte de Frank Rosenblatt, y en 1967, la propuesta del algoritmo del vecino más cercano, como forma para interpolar o para completar información faltante, usando el correspondiente dato que, en un

espacio multidimensional y con una determinada métrica, está más próximo al destinatario (Weinberger *et al*., 2006). En las siguientes décadas se dedican esfuerzos a la robótica y al desarrollo del lenguaje natural, pero es a partir de los 90, cuando se comienzan a crear programas para que los propios ordenadores analicen grandes cantidades de datos y extraigan conclusiones automáticamente, o "aprendan" de los datos para desvelar resultados. Se crea en definitiva un lenguaje paralelo al de la estadística y la econometría tradicional que no sólo persiste sino que está tomando fuerza hoy en día, y que incluso establece que los modelos estadísticos clásicos constituyen los primeros vestigios del machine learning, dado que se dice, por ejemplo, que la estimación de parámetros de un modelo lineal es en realidad un cómputo automatizado que permite encontrar patrones en los datos. En el año 2006, Geoffrey Hinton acuña el término "deep learning" para explicar nuevos algoritmos de reconocimiento de objetos e imágenes. En la década siguiente, aunque la mayoría de los avances se producen en este mismo ámbito, aparecen plataformas que permiten la implementación de métodos de inteligencia artificial en máquinas distribuidas, es decir, que facilitan el tratamiento de grandes bases de datos segmentando el problema de tener un volumen excesivo de datos para un solo ordenador a la resolución en subsecuencias en varios ordenadores diferentes.

3. Metodología: machine learning

Los métodos del machine learning se clasifican en tres grupos: aprendizaje supervisado, aprendizaje no supervisado y los más avanzados y cercanos a la realidad, que se han llamado de aprendizaje reforzado. Alpaydin (2009) escribe uno de los primeros libros sobre el tema, unificando en un único marco común los problemas y soluciones del machine learning.

En los métodos del aprendizaje supervisado, que incluyen los modelos predictivos clásicos, los árboles de decisión, las redes neuronales artificiales o el "support vector machine", el aprendizaje se realiza a través del análisis del pasado, tratando de reproducir la respuesta o cómo anticipar lo sucedido. Kotsiantis (2007) realiza una de las primeras revisiones de los métodos de clasificación supervisados.

En el caso de los seguros, en los métodos supervisados se trata de elegir una muestra de entrenamiento sobre la que establecer las bases para hallar una predicción. Dicha muestra puede consistir, por ejemplo, en la observación de la siniestralidad que los asegurados han sufrido en los años anteriores. Comparando la respuesta predicha por algún método supervisado con la observada, se puede decidir cuál de los procedimientos produce menos error.

Seguidamente se usa una muestra de contraste para evaluar si el procedimiento también es el mejor con unos datos no empleados en la primera fase. Como en el planteamiento de modelos predictivos se parte de un conjunto de pólizas para las que se conocen algunas de sus características y, por ejemplo, se sabe si han tenido o no un siniestro. Con el sistema de aprendizaje supervisado, se crea un modo de predecir si han tenido o no un siniestro, como se haría por ejemplo con un modelo clásico de regresión logística, y luego se compara lo observado con la predicción del modelo. Se concluye que es mejor aquel algoritmo que logra más aciertos. En este sentido, cabe señalar que pueden ponderarse los errores de manera distinta, si van en uno u otro sentido. Es decir, cuando el modelo predice que hay siniestro y no se ha declarado ninguno, o bien, al contrario, cuando el modelo predice que no hay siniestro y sí lo ha habido.

En los métodos de aprendizaje no supervisado, se persigue hallar estructuras o agrupaciones, en las que existan similitudes entre las unidades analizadas, en nuestro caso, las pólizas. Una vez se han encontrado dichos subconjuntos, se pueden tratar de forma homogénea, por ejemplo, otorgándoles la misma prima, o el mismo descuento para todas las pólizas del mismo subgrupo.

Respecto a los métodos de aprendizaje reforzado, Sutton y Barto (1998) remarcan la posible utilización de dichos modelos para realizar acciones específicas, optimizando un beneficio o recompensa final. Para ello se analiza la experiencia pasada y se intenta capturar el mejor conocimiento posible para tomar decisiones comerciales, por ejemplo, viendo cómo los clientes reaccionan ante un descuento, lo que puede sugerir aplicarlo en operaciones de renovación de pólizas parecidas.

Los métodos que van a emplearse en este trabajo son un total de nueve: Regresión Lineal, Regresión Logística, Árbol de Decisión, SVM (Support Vector Machine), Naive Bayes, kNN (K-Nearest Neighbors), Random Forest y dos algoritmos de Gradient Boosting (el GBM o Generalized Boosted Regression Modelling y el Discrete Adaboost).

A fin de no romper la exposición, se incluye una breve explicación de cada uno de los métodos en el Anexo, así como alguna referencia bibliográfica sobre ellos.

Para comparar el resultado predictivo que proporciona cada método se establece aquí una medida ampliamente utilizada como es el RMSE (*root mean squared error*), conocida como la raíz del error cuadrático medio, que se implementará tanto en la muestra de aprendizaje como en la de validación. Como se detallará en el apartado de los datos, se determina

aleatoriamente un 75% del total de las observaciones iniciales para formar una muestra de entrenamiento o aprendizaje sobre la cual se construye el modelo, y el 25% restante se utiliza para conformar la muestra de test o validación sobre la cual se mide la calidad del modelo. Para cada uno de los métodos se emplean las mismas muestras de aprendizaje y validación.

En la comparativa de los métodos, hay que valorar si cada uno de los procedimientos logra mejor o peor el objetivo de predecir si se produce o no un siniestro. Así, la respuesta toma dos posibles valores: cuando se produce al menos un siniestro el suceso respuesta se identifica con el valor 1, mientras que la ausencia del mismo se codifica con el valor 0. Las covariables, también llamadas factores de riesgo y empleadas para efectuar la predicción, son las mismas para todos los modelos. En nuestro caso, no hemos abordado un análisis del número de siniestros, ya que dicho análisis nos hubiera conducido a la implementación de modelos de tipo Poisson. La razón de distinguir solo entre dos tipos de eventos es que preferimos restringirnos a la presencia o ausencia de la declaración de siniestro para poder analizar así un mayor abanico de modelos del machine learning que tienen como objetivo clasificar en dos grupos.

Se define el error *E* como la diferencia entre el suceso observado y el predicho siendo tres los resultados posibles. Si la observación y la predicción coinciden, el error es cero. Si se produce un siniestro, pero el modelo predice que no, entonces el error vale 1, mientras que si no se produce el siniestro, pero la predicción establece que sí, entonces el error es -1. La Tabla 1 muestra la definición del error de predicción.

Tabla1. Valores que puede tomar el error *E*. En la observación y en la predicción, 1 significa que hay siniestro, y 0 que no lo hay

Observación	Predicción	Error (E)
1	0	1
0	1	-1
1	1	0
0	0	0

En la síntesis de los resultados que se mostrarán en las siguientes secciones se presentan las medidas estadísticas para valorar la capacidad predictiva y resultados de los modelos tanto para la muestra de entrenamiento como para la de test.

La raíz del error cuadrático medio (RMSE) se define como:

$$\sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}},$$

donde y_i es la variable dependiente observada para la observación *i* de la muestra y *N* el total de observaciones de dicha muestra, e $\hat{y}_i$ es el valor predicho para la observación *i* de la misma muestra, que puede ser un valor entre 0 y 1 si la predicción del modelo es una probabilidad de ocurrencia.

El promedio de error es una medida que se establece como la media del error *E* tal como está definido en la Tabla 1. El umbral seleccionado para la probabilidad de ocurrencia en el caso de que esta sea la que permita que la predicción $\hat{y}_i$ tome el valor de 1 se elige aquí como la media aritmética de la variable dependiente del conjunto de datos global.

Alternativamente, se define la raíz del error cuadrático medio RMSE* basado en la definición de la Tabla 1. Por lo tanto se trata de la raíz cuadrada del sumatorio de los errores *E* de cada uno de los valores predichos en la muestra elevados al cuadrado dividido por *N.* Por lo tanto aquí no se usan predicciones probabilísticas sino simplemente la respuesta binaria predicha.

Para finalizar, decir que la precisión de cada método se mide como la proporción de observaciones correctamente clasificadas por dicho modelo.

4. Datos y resultados

Los datos que hemos empleado para ilustrar los algoritmos de machine learning proceden de una muestra de asegurados del año 2011 del seguro del automóvil. Se trata de una muestra de asegurados jóvenes que disponen de un seguro para el que se conoce una mínima información telemática, que se han utilizado en otros trabajos como por ejemplo en Boucher *et al.* (2017) para estudiar la transformación de los factores de riesgo, o bien en Ayuso *et al.* (2016a) y (2016b) para analizar los efectos de la distancia recorrida hasta que se produce un primer accidente o la diferencia entre hombres y mujeres. Se dispone de la variable respuesta (Y) que denota la existencia de una declaración de accidente con culpa durante el año de observación, que toma el valor 1 si ha habido al menos una declaración de accidente con culpa, y 0 en caso contrario. Se tienen además las características de los asegurados: edad, sexo, antigüedad del permiso de conducir (antigc), antigüedad del vehículo (antigv), total de kilómetros recorridos durante el año (Km_anuales), porcentaje de kilómetros recorridos en zona urbana (porc_urba), porcentaje de kilómetros recorridos por encima de los niveles máximos permitidos de velocidad (porc_velo) y porcentaje de kilómetros

recorridos en horario nocturno (porc_noct). La media de coste en las pólizas que han tenido algún siniestro con culpa es igual a € 1937,13.

Tabla 2. Descripción de los datos sobre accidentes con culpa (2011). Media de las variables según ausencia o presencia de siniestro con culpa. Para la variable sexo, frecuencia absoluta y proporción por fila

Variables		**No hay declaración de accidente (Y=0)**	**Hay declaración de accidente con culpa (Y=1)**	**Total**
Edad (años)		25,10	24,55	25,06
Sexo	Mujer	1263 (93,21%)	92 (6,79%)	1355
	Hombre	1309 (92,71%)	103 (7,29%)	1412
Antigüedad del permiso de conducir (años)		4,98	4,46	4,94
Antigüedad del vehículo (años)		6,37	6,17	6,35
Total de kilómetros recorridos en el año		7094,63	7634,97	7132,71
Porcentaje de kilómetros recorridos en zona urbana		24,60	26,34	24,72
Porcentaje de kilómetros recorridos por encima de los niveles permitidos de velocidad		6,72	7,24	6,75
Porcentaje de kilómetros recorridos en horario nocturno		6,88	6, 66	6,86
Coste medio por póliza		0,00	1937,13	136,52
Total de asegurados		2572	195	2767

En la descripción de los datos de la Tabla 2, se puede apreciar que el 7,05% del total de asegurados declaran al menos un accidente con culpa. No se presenta una brecha significativa por sexo, es decir que en promedio hombres y mujeres tienen una tasa muy similar de accidentes con culpa. La edad promedio de los asegurados oscila en torno a los 25 años aproximadamente. El promedio de los años con permiso de conducir oscila alrededor de 5 años aproximadamente para los asegurados que no han tenido accidentes, y 4,5 años aproximadamente para los que sí tuvieron al menos

una declaración de accidente con culpa. Los asegurados que han declarado un accidente con culpa han recorrido 7634,96 km anuales en promedio, mientras que los que no, tan solo han recorrido 7094,63 km anuales en promedio. El porcentaje de kilómetros recorridos en zona urbana y por encima de los niveles permitidos de velocidad es mayor en los asegurados que han sufrido un accidente con culpa.

A fin de analizar la capacidad predictiva de los modelos implementados, los resultados de la modelización se muestran en las Tablas 3 y 4. La obtención de los resultados se realizó a través de R y los algoritmos, así como ejemplos para su utilización están disponibles a partir de las autoras.

En términos generales, lo que se esperaría de un método de machine learning con alta efectividad en la predicción, es que presente un valor de *RMSE* lo suficiente pequeño, una precisión muy alta y el promedio del error *E* sea lo más cercano a cero. Por lo tanto, se entiende que un modelo ideal (de máxima capacidad predictiva) tendrá una precisión del 100%, un *RMSE* igual a 0, y por ende el promedio del error *E* también igual a 0.

Las medidas estadísticas de las Tablas 3 y 4 han sido calculadas tanto para la muestra de entrenamiento como la muestra de test. Los resultados indican que el comportamiento entre las dos muestras es muy similar, por lo que se podría descartar el problema de sobreaprendizaje u overfitting en inglés, el cual haría que en muestras de datos muy complejas, el modelo tendiera a ajustarse excesivamente a la muestra de entrenamiento, y fuera menos capaz de clasificar correctamente nuevos datos como la muestra de test que con la muestra de entrenamiento.

Sin embargo, extraer las conclusiones tomando en cuenta únicamente las medidas estadísticas proporcionadas por la Tabla 3 y 4 resulta limitante, en el sentido de que ciertos métodos a excepción del Discrete Adaboost poseen parámetros o son técnicas paramétricas, y dependiendo de las combinaciones de los valores de dichos parámetros llevarán a resultados distintos, pudiendo así también mejorar la capacidad predictiva del modelo.

Tabla 3. Valores de la raíz del error cuadrático medio (RMSE), media del error, raíz del error cuadrático medio considerando predicciones dicotómicas (RMSE*) y precisión para cada método utilizado en la muestra de entrenamiento (2075 casos) y de test (692 casos)

Muestra de entrenamiento				
Método	**RMSE**	**Media error**	**RMSE***	**Precisión**
Regresión Lineal	0,26	-0,43	0,69	0,52
Regresión Logística	0,26	-0,39	0,66	0,56
Árboles de decisión	0,25	-0,01	0,35	0,88
SVM	0,26	-0,79	0,94	0,12
Naive Bayes +	0,27	0,07	0,27	0,93
KNN +	0,25	0,05	0,25	0,94
Random Forest ** +	0,00	0,00	0,00	1,00
GBM	0,25	-0,34	0,61	0,62
Discrete Adaboost +	0,27	0,07	0,27	0,93
Muestra de Test				
Método	**RMSE**	**Media error**	**RMSE***	**Precisión**
Regresión Lineal	0,25	-0,44	0,71	0,50
Regresión Logística	0,25	-0,40	0,68	0,53
Árboles de decisión	0,25	-0,01	0,37	0,86
SVM	0,25	-0,80	0,91	0,18
Naive Bayes+	0,26	0,06	0,26	0,93
KNN+	0,29	0,06	0,29	0,91
Random Forest+	0,26	0,07	0,26	0,93
GBM	0,25	-0,35	0,66	0,57
Discrete Adaboost+	0,26	0,07	0,26	0,93

** El método Random Forest no proporciona ningún error en la muestra de entrenamiento.

+ La predicción de declaración de siniestros de estos métodos es de tipo {0,1} directamente.

Tabla 4. Matriz de confusión para los métodos analizados en la muestra de entrenamiento (2075 casos) y de test (692 casos)

Muestra de Entrenamiento						
Método	$Y_i = 0, \hat{Y}_i = 0$	$Y_i = 1, \hat{Y}_i = 0$	$Y_i = 0, \hat{Y}_i = 1$	$Y_i = 1, \hat{Y}_i = 1$	Sensibilidad	Especificidad
Regresión Lineal	983	47	943	102	0,68	0,51
Regresión Logística	1073	54	853	95	0,64	0,56
Árboles de decisión	1792	121	134	28	0,19	0,93
SVM	200	93	1726	56	0,36	0,10
Naive Bayes	1922	149	4	0	0,00	1,00
kNN	1964	123	9	24	0,16	1,00
Random Forest	1926	0	0	149	1,00	1,00
GBM	1182	40	744	109	0,73	0,61
Discrete Adaboost	1926	149	0	0	0,00	1,00
Muestra de Test						
Método	$Y_i = 0, \hat{Y}_i = 0$	$Y_i = 1, \hat{Y}_i = 0$	$Y_i = 0, \hat{Y}_i = 1$	$Y_i = 1, \hat{Y}_i = 1$	Sensibilidad	Especificidad
Regresión Lineal	321	21	325	25	0,54	0,50
Regresión Logística	347	24	299	22	0,48	0,54
Árboles de decisión	596	44	50	2	0,04	0,92
SVM	85	7	561	39	0,85	0,13
Naive Bayes	644	46	2	0	0,00	1,00
kNN	590	47	9	1	0,02	0,98
Random Forest	646	46	0	0	0,00	1,00
GBM	375	28	271	18	0,39	0,58
Discrete Adaboost	646	46	0	0	0,00	1,00

Sensibilidad es el porcentaje de casos de respuesta 1 correctamente clasificados y Especificidad es el porcentaje casos de respuesta 0 bien clasificados.

Hemos propuesto considerar las medidas estadísticas de la Tabla 3 y 4 para establecer las conclusiones tomando en cuenta los objetivos del análisis, en este caso, la detección de declaración de accidente con culpa, pues a partir de este evento las compañías aseguradoras podrán prever con más exactitud el coste de la siniestralidad que les corresponde asumir por este concepto.

Entonces, partiendo de los métodos que poseen los valores más altos de precisión y los *RMSE* más bajos en la muestra de test, detectamos que Naive Bayes, Random Forest y Discrete Adaboost podrían tener el mejor comportamiento predictivo. Y los que en este sentido tendrían bajo rendimiento serían en primer lugar el SVM, seguido por la regresión lineal y logística. No obstante, para nuestro objetivo es importante que el modelo presente una alta sensibilidad, y los primeros métodos tienen una sensibilidad del 0 % en la muestra de test, lo que les coloca en una mala posición.

Por lo tanto, se consideran más potentes los métodos cuyos valores de sensibilidad en la muestra de test sean mayores aun cuando se sacrifique una alta especificidad. En este caso, tanto la regresión logística como la lineal demuestran ser mucho más efectivas en la muestra de test para detectar aproximadamente la mitad de las pólizas que declararon un siniestro con culpa, cosa que no son capaces de hacer los tres métodos que inicialmente eran mejores en la muestra de entrenamiento según los valores de la Tabla 3.

Una posible justificación sobre las diferencias en capacidad predictiva de los diferentes métodos es que tanto el modelo de regresión lineal como en el logístico buscan modelizar linealmente la respuesta o mediante un predictor lineal directamente, mientras que los algoritmos de clasificación de los otros métodos efectúan combinaciones más complejas entre los factores y hacen que el modelo prediga de manera más exacta la mayor cantidad de observaciones, pero al parecer omiten con mayor facilidad la ocurrencia de eventos raros cuando se implementan en una nueva muestra.

Por lo tanto, si la prioridad fuese mejorar la predictibilidad de la no ocurrencia de eventos, por ejemplo, la declaración de no accidente por parte del asegurado, entonces se diría que Naive Bayes, Discrete Adaboost o Random Forest son modelos apropiados.

Aunque es muy complejo generalizar para qué tipo de muestra los métodos pueden ser más efectivos, la característica de esta base de datos es que cuenta con una variable dependiente binaria (declaración afirmativa o negativa de al menos un accidente con culpa), donde sólo un porcentaje muy bajo de los asegurados sufren siniestralidad, por lo que la ocurrencia de este evento es poco frecuente.

En la Tabla 5 se resumen los resultados para cada tipo de modelización respecto a qué variables son las más relevantes en la muestra de entrenamiento. Para llevar a cabo este objetivo, se utilizó la función **VarImp** del paquete **caret** del Software estadístico R, que calcula la importancia de

las variables en modelos de regresión y clasificación (Kuhn y Johnson, 2013).

A pesar de cada método tiene una capacidad predictiva diferente en cada modelo en concreto, los resultados de la Tabla 5 son importantes para detectar que los kilómetros anuales recorridos por el asegurado parecen ser el factor que tiene un mayor impacto en la declaración de accidentes con culpa, por lo que ello convierte a esa variable en un factor de mucho interés al momento de explicar la siniestralidad de los asegurados, y por ende en la tarificación del seguro. A esta variable le sigue el porcentaje de kilómetros recorridos por encima de los niveles permitidos de velocidad que podría manifestarse como en un segundo lugar de importancia. En tercer lugar se situaría la variable que recoge el porcentaje de kilómetros recorridos en vía urbana, aunque en algunos modelos, no aparece este factor de forma tan clara como los dos anteriores, y se da más relevancia la edad o la antigüedad del vehículo.

Tabla 5. Orden de importancia de las variables en los modelos de regresión o clasificación al predecir la variable dependiente sobre declaración de accidente con culpa

Método	**Primera**	**Segunda**	**Tercera**
Regresión Lineal	Km_anuales	porc_velo	porc_urba
Regresión Logística	Km_anuales	porc_velo	porc_urba
Árboles de decisión	porc_velo	antigv	Km_anuales
SVM	Km_anuales	edad	porc_velo
Naive Bayes	porc_velo	Km_anuales	edad
KNN	Km_anuales	edad	porc_velo
Random Forest	porc_urba	porc_velo	Km_anuales
GBM	edad	Km_anuales	porc_velo
Discrete Adaboost	Km_anuales	edad	porc_velo

4.1 Cálculo de primas puras según el coste esperado

La gran oferta de seguros para vehículos motiva a las aseguradoras a fijar tarifas cada vez más competitivas, por lo que es necesario mejorar la precisión en el cálculo de la prima. En este apartado se busca identificar el impacto que tendría cada método en el cómputo de la prima pura del seguro al aplicar el mismo.

La prima pura se fija en esta ilustración a partir de la probabilidad de que un asegurado tenga al menos un accidente con culpa multiplicado por el coste medio obtenido en esta misma base de datos para las pólizas que han sufrido

al menos un accidente. No se consideran accidentes cuya culpa no sea del asegurado analizado. Por esta razón, el ejercicio que se lleva a cabo en ese apartado no está exento de limitaciones. De hecho, para poder calcular la prima real también se deberían tener en cuenta los costes de los siniestros cuya culpa no es del asegurado pero que acarrean costes a la entidad. Dichos recargos se sumarían al resto de gastos generales que completan el coste final. Además, desde un punto de vista de la implementación práctica, para poder implementar los modelos especificados en este trabajo se necesitaría conocer el total de kilómetros que va a recorrer el asegurado y el resto de las variables que se han utilizado en la modelización.

El Gráfico 1 muestra la distribución de las primas puras calculadas a partir de cada método en la muestra de test.

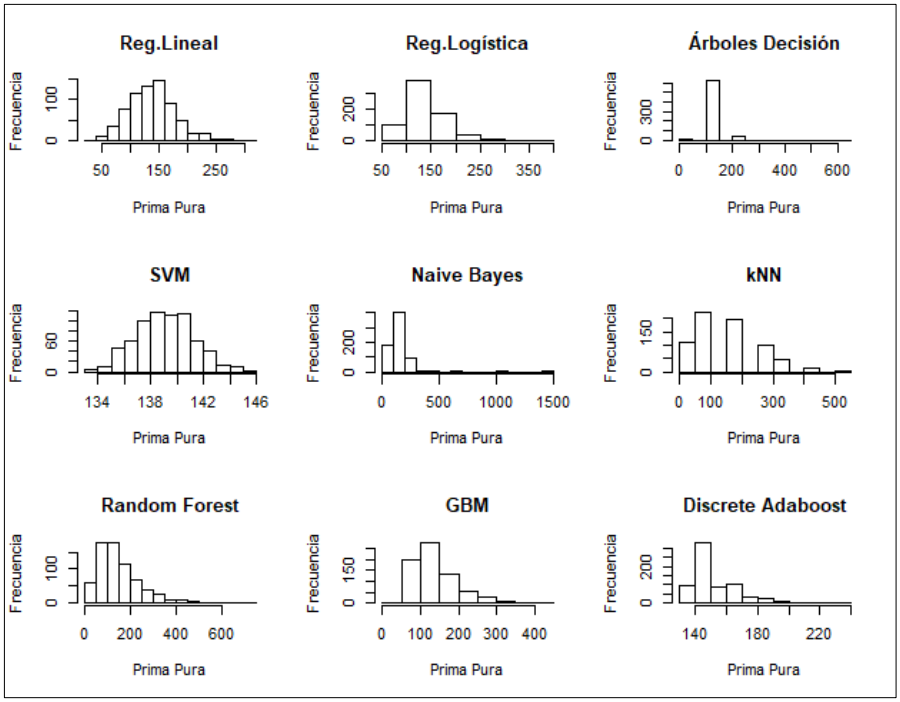
De acuerdo al Gráfico 1, métodos como Random Forest y los árboles de decisión posibilitan la obtención de primas puras superiores a € 500, aunque dichos valores sean muy poco frecuentes. Una posible justificación es debida a los bajos niveles de sensibilidad que poseen ambos modelos, así que no les es posible predecir eficientemente los casos más inusuales, y por ende el valor de la prima puede sobreestimarse para los valores extremos.

Adicionalmente, según el método SVM las primas se acumulan excesivamente entre los valores de € 100 y € 150, lo que quiere decir que la frontera de decisión está tomando un criterio totalmente distinto al de los otros métodos. Aunque al igual que la regresión logística se toma un umbral para determinar la clasificación, el método SVM parece ser ineficiente para predecir observaciones lejanas a la media, por lo que para este conjunto de datos no podría ser un buen método.

De acuerdo al método de Naive Bayes, la distribución de las primas se encuentra acumulada principalmente bajo el valor de € 300, y en casos muy concretos se acerca a € 400. En este caso, el cálculo de la prima puede estar influenciado por el hecho de que las variables que están altamente correlacionadas cuentan doble (votan dos veces) y en consecuencia no estarían bien representadas en este modelo. Además, si alguna variable en concreto no fue analizada en la muestra de entrenamiento, entonces el modelo le asignará una probabilidad de cero y será incapaz de hacer la predicción. Para solucionar este problema, se puede utilizar técnica de suavizado como la estimación de Laplace, aunque aquí no se ha implementado.

Tanto la regresión logística como la lineal acumulan el valor de las primas por debajo de los € 300.

Gráfico 1. Histogramas de la prima pura para cubrir siniestros con culpa según los diferentes métodos de machine learning



Los métodos SVM, Naive Bayes, Random Forest y Discrete Adaboost poseen una predicción de declaración de siniestros de tipo [0,1].

Un resultado interesante es que el método KNN no tiene prácticamente primas con un valor entre €100 a €150 ni de €200 a €250, lo que querría decir que en realidad observaciones de muy poca frecuencia aparecen en estos intervalos, por lo que dichas observaciones debieron ser clasificadas en grupos de una probabilidad mayor o bien menor que la real. Esta técnica tampoco parece ser efectiva puesto que la idea principal es identificar eventos poco frecuentes, pero es más eficiente agrupando observaciones en conjuntos homogéneos, por lo que podría ser recomendable para predecir situaciones con muchas más características o variables que las que se han contemplado aquí.

4.2 Cálculo de primas usando un precio por kilómetro

Tal como se ha visto en la Tabla 5, el kilometraje recorrido por un asegurado juega un papel importante en la propensión a declarar un siniestro con culpa y por lo tanto, debe serlo en la fijación de una tarifa, puesto que además se entiende que a mayor distancia recorrida, mayor será la probabilidad de sufrir un accidente. Este fenómeno ha sido reconocido por diferentes autores (ver por ejemplo, Guillen et al. 2018).

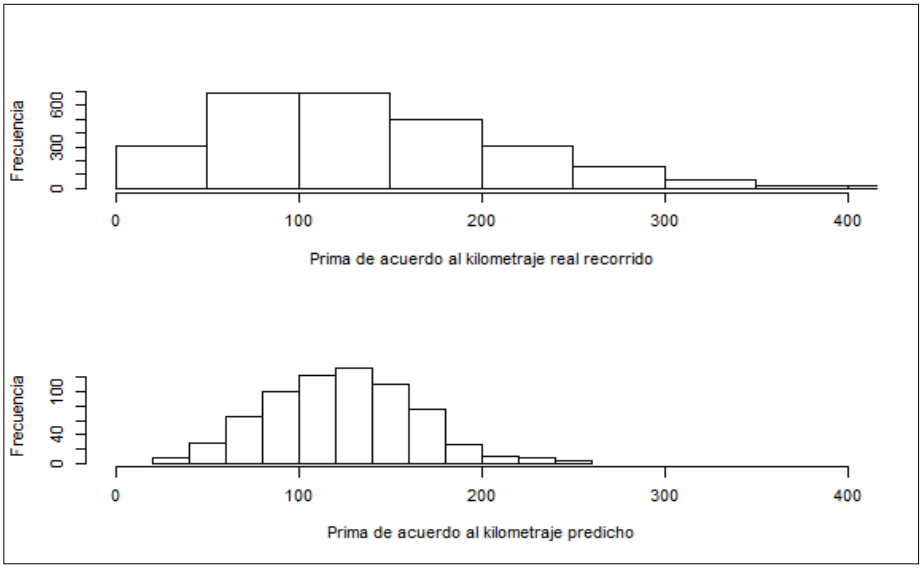
En este apartado, hemos considerado interesante calcular el coste total de los accidentes considerados en la muestra, que es igual a € 377 740,40 (1937,13 * 195), así como el kilometraje total recorrido por todos los asegurados, que es igual a 19 736 209 km (7132,71 * 2767). Y con ello podemos obtener una aproximación al coste medio de la siniestralidad observada por kilómetro recorrido, que es equivalente a € 0,02/km y que resulta de dividir el anterior total de costes entre el anterior total de kilómetros. La pregunta ahora es cómo fijar un precio de la prima, que pueda distinguir entre aquellos asegurados que tienen mayor o menos propensión a declarar un siniestro.
Por lo tanto, hemos propuesto cuatro métodos para diferenciar el cálculo de la prima considerando los kilómetros recorridos. Los dos primeros no utilizan los métodos de clasificación del machine learning. La prima pura se calcularía según cada uno de los cuatro procedimientos como se describe a continuación:

1) Coste medio por kilómetro recorrido multiplicado por los kilómetros recorridos anuales observados. El objetivo es analizar cuánto debieron pagar en relación al kilometraje recorrido, con un precio estimado de 2 céntimos por kilómetro.
2) Coste medio por kilómetro recorrido multiplicado por la predicción de kilometraje anual. Dicha predicción se obtiene tras emplear un modelo log-normal cuya variable dependiente es el total de kilómetros anuales recorridos explicada por las covariables detalladas en la Tabla 1 sin contar con sexo. El objetivo es hacer un pronóstico de cuánto deberían pagar los asegurados considerando los kilómetros que se estiman que recorrerán, utilizando el precio medio de 2 céntimos por kilómetro.
3) Coste medio por kilómetro recorrido multiplicado por el kilometraje real recorrido y por un índice de riesgo que se calcula como la predicción del modelo de machine learning con las características del asegurado y un total de kilómetros anuales promedio dividida por el nivel medio de siniestralidad observada (7,05%). El índice de riesgo es superior a uno si el asegurado tiene una mayor propensión

a declarar un accidente con culpa suponiendo que recorriera un total de kilómetros medio, y es inferior a uno si su riesgo es inferior.

4) Coste medio por kilómetro recorrido multiplicado por la predicción del kilometraje y el índice de riesgo que se calcula como la predicción del modelo de machine learning con las características del asegurado y un total de kilómetros anuales promedio dividida por el nivel medio de siniestralidad observada (7,05%).

Gráfico 2. Comparativa de primas de acuerdo al kilometraje real y predicho, sin tener en cuenta la predicción de la siniestralidad mediante un algoritmo machine learning



El Gráfico 2 muestra las distribuciones de las primas considerando el kilometraje predicho y el kilometraje real. En adición, los resultados indican que la media de la distribución de la prima de acuerdo al kilometraje real recorrido es de 142,65 y su cuantil 0,90 es de 251,16, mientras que la media de la distribución de la prima de acuerdo al kilometraje predicho es de 122,50 y su cuantil 90 es de 173,54. Esto sugiere que el cálculo de la prima prediciendo el kilometraje podría estar muy por debajo de los valores que en realidad le corresponderían asumir a los asegurados más riesgosos, puesto que la discrepancia más fuerte se halla en el cuantil 0,90 de la distribución de primas. Aunque este método es importante para denotar el impacto de los kilómetros recorridos en la siniestralidad, podría ser mejor valorado para

asegurados con bajos niveles históricos de siniestralidad, pues así el cobro de la prima tiende a ser más justo, lo que consecuentemente motiva a los asegurados a renovar o mantener sus pólizas.

Tabla 6. Media y cuantil 90% de la prima en la muestra de test según el método

Método	**Predicción siniestralidad*coste medio**		**€ 0,02*Km observados* índice riesgo**		**€ 0,02*Km predichos*índice riesgo**	
	Media	**Cuantil 0,90**	**Media**	**Cuantil 0,90**	**Media**	**Cuantil 0,90**
Regresión Lineal	138,03	187,65	139,56	242,73	121,16	178,78
Regresión Logística	137,93	184,81	138,47	241,88	120,03	173,27
Árboles de decisión	133,67	129,22	137,11	244,52	118,79	169,99
SVM	139,09	141,93	145,23	253,03	124,93	176,54
Naive Bayes	148,35	217,18	157,93	292,11	138,87	217,97
KNN	149,49	252,67	137,70	296,46	118,66	236,67
Random Forest	152,23	278,55	140,84	298,88	121,05	224,42
GBM	136,80	212,04	129,20	232,42	112,40	181,32
Discrete Adaboost	149,02	169,73	154,77	274,67	133,39	192,24

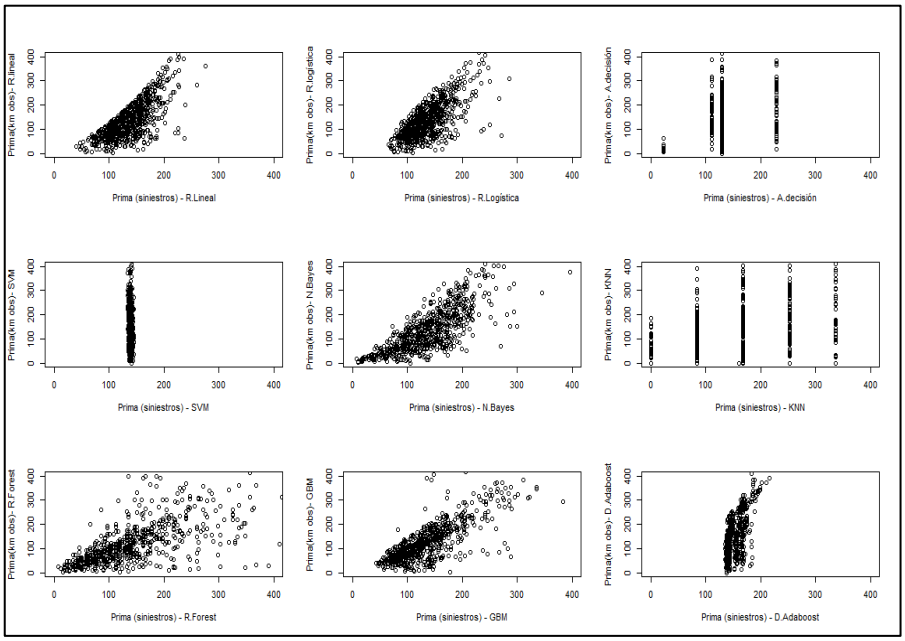
El índice de riesgo que se calcula como la predicción del modelo de machine learning con las características del asegurado y un total de kilómetros anuales promedio dividida por el nivel medio de siniestralidad observada (7,05%).

La Tabla 6 hace una comparativa de la media y el cuantil 90% de las primas de acuerdo a las propuestas mencionadas anteriormente. En primer lugar, la prima como resultado de la predicción del siniestro por el coste medio que fue previamente ilustrado y analizado en el gráfico 1 y en la sección anterior. En segundo lugar, la prima como resultado de multiplicar € 0,02 por el número de kilómetros observados y por un índice de riesgo, tal como se planteó en el tercer método de la sección 4.2. Y en tercer lugar, la prima como resultado de multiplicar € 0,02 por el número de kilómetros observados y por un índice de riesgo, tal como se planteó en el cuarto método de la sección 4.2.

Los resultados de la Tabla 6 indican que la media de la prima que considera la predicción de la ocurrencia de al menos un siniestro con culpa se asemeja mucho a la que considera el kilometraje real observado, lo que querría decir que para observaciones con una propensión de siniestralidad media, ambos

métodos serían equivalentes. Sin embargo, si se usara el kilometraje recorrido, se obtienen primas con un cuantil 90% más elevado, lo que indica que habría asegurados que pagarían mucho más. Por otro lado, aunque el cuantil 90% de la prima que considera los kilómetros observados no es cercano al de las otras dos alternativas de la Tabla 6, la prima que considera el kilometraje predicho es la que más se le aproxima, lo que quiere decir que bajo esta metodología se podrían obtener cuantías de la prima moderadas.

Gráfico 3. Gráfico de dispersión entre la prima calculada mediante la predicción de siniestros y la prima calculada mediante el coste medio por kilómetro recorrido multiplicado por el kilometraje real recorrido y por el índice de riesgo



La Gráfica 3 permite visualizar la correlación y relación de las primas basadas en la predicción de la siniestralidad de acuerdo a cada uno de los métodos de machine learning frente la prima basada en el kilometraje real recorrido donde el precio por kilómetro se modifica según un índice de riesgo del asegurado, de forma que pagan más por kilómetro los que más riesgo de accidente tienen. En efecto, se entiende que la prima cobrada más justa correspondería a los kilómetros recorridos observados, por lo que a más de ser una propuesta para tarifar las primas de vehículos, es también un referente de comparación con las otras propuestas.

En este caso, tanto la regresión logística como la lineal parecen correlacionarse bastante bien en la mayoría de las observaciones, no obstante, se observa una ligera dispersión en un número menor de observaciones, pudiendo ser estas las observaciones poco frecuentes.

Por otro lado, el GBM y Naive Bayes poseen comportamiento similar que el de los dos métodos previos, sin embargo, presentan mayor dispersión. Adicionalmente, Random Forest presenta también una tendencia parecida, pero su dispersión se eleva para valores de la prima superiores a €100, entonces hay discrepancias importantes principalmente para los cuantiles altos del valor de la prima.

Tanto el Discrete Adaboost como el SVM indican un comportamiento particular puesto que la prima no toma valores en un rango extendido o más amplio, por el contrario tienen a concentrarse en valores cercanos. El método de Discrete Adaboost presenta una dispersión ligeramente mayor que el SVM.

Si se usaran tanto los árboles de decisión como kNN para realizar los índices de riesgo y luego se calculara un precio por kilómetro, los valores de primas tomarían unos rangos en intervalos más amplios, a diferencia de los puntuales que se obtendrían usando el método clásico que se basa en multiplicar el coste medio por la estimación de la probabilidad de siniestro.

5. Conclusiones

Se concluye que existen grandes similitudes entre los diferentes métodos en cuanto a la capacidad de predecir la siniestralidad en una muestra de pólizas del seguro de automóviles y que la decisión sobre cuáles implementar depende más del deseo de interpretabilidad, reproducibilidad y elementos de la supervisión, que de una ganancia excesiva en la precisión y robustez de cada uno de ellos al implementarlos en la muestra de test.

La utilización de los algoritmos de machine learning revela que estos tienen capacidades predictivas muy similares, aunque a nivel de las primas finales que se obtendrían, se acaban viendo mayores o menores dispersiones de los precios según cual sea el algoritmo usado y la aproximación al cálculo de la prima. En el largo plazo, la decisión de qué método utilizar puede incidir en las políticas comerciales que quieran emprenderse.

En este trabajo se ha propuesto una nueva manera de abordar el cálculo de la prima teniendo en cuanta la posibilidad de un pago por kilómetro recorrido, ya sea anticipando el total de kilómetros o bien, contabilizándolos una vez ya observados. Las diferencias que se aprecian no son excesivas, aunque sí se aprecia una mayor dispersión en las primas más altas al emplear los kilómetros reales observados según cuál sea el método implementado. Por ello concluimos que la elección del modelo predictivo sigue siendo mucho más determinante incluso que la decisión sobre cómo tener en cuenta los factores telemáticos en la tarificación.

Referencias

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

Ayuso, M., Guillen, M. y A. M. Pérez-Marín (2016a). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies* 68, 160-167.

Ayuso, M., Guillen, M. y A. M. Pérez-Marín (2016b). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks* 4(2), 1-10.

Boucher, J-P., Côté, S. y M. Guillen (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models, *Risks* 5(4), 54; https://doi.org/10.3390/risks5040054.

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

Guelman, L., Guillen, M. y A. M. Pérez-Marín (2014). A survey of personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics* 58, 68-76.

Guillen, M. (2016). Big data en seguros. *Indice: revista de estadística y sociedad*, 28-30.

Guillen, M., Nielsen, J.P., Ayuso, M. y A. M. Pérez-Marin (2018). The use of telematics devices to improve automobile insurance rates, *Risk Analysis*, accepted, in press.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica* 31, 3-24.

Kuhn, M. Y K. Johnson (2013). *Applied predictive modelin*, vol. 26. Springer, New York.

Michalski, R. S., Carbonnel, J. G. y T. M. Mitchell (2013). Machine learning: an artificial intelligence approach. En R. Milchalski, J. Carbonnel, y T. Mitchell (eds.) *Machine learning: An artificial intelligence approach,* Springer Science & Business Media.

Padilla-Bareto, A., Guillen, M. y C. Bolancé (2017). Big-data Analytics en seguros. *Anales del Instituto de Actuarios Españoles*, 23, 1-19.

Sutton, R. S. y A. G. Barto (1998). *Introduction to reinforcement learning*. MIT press, Cambridge.

Turing A.M. (2009) Computing Machinery and Intelligence. En Epstein R., Roberts G. y G. Beber (eds) *Parsing the Turing Test*. Springer, 23-65.

Weinberger, K. Q., Blitzer, J. y L. K. Saul (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 1473-1480.

Witten, I. H., Frank, E., Hall, M. A. y C. J. Pal (2016). *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Anexo

En este apartado se resumen las principales características de los métodos empleados en el artículo. En cuanto a la modelación predictiva en general, cabe destacar que se ocupa principalmente de minimizar un error definido en base a la diferencia entre la observación y la predicción. En otras palabras, puede decirse que pretende hacer las predicciones lo más precisas posibles a expensas de la interpretación de dichos resultados y su significado, llegando a ser a veces excesivamente engorrosa para algunas aplicaciones prácticas y demasiado distante de las explicaciones cualitativas.

- *Regresión Lineal.* Es el método clásico para datos en los que se desea explicar una variable dependiente en función del resto. Su principal característica es la simplicidad e interpretabilidad. Su objetivo es minimizar la suma de errores cuadrados considerando las deviaciones de los valores medios y tomados estos como la diferencia entre la observación y una combinación lineal de factores. La inferencia de este modelo se realiza en base a suponer que la variable respuesta sigue una distribución normal.
- *Regresión Logística*. Es una técnica también utilizada en machine learning que se conoce desde hace varias décadas y es fundamental en muchas áreas aplicadas, tales como pruebas de medicamentos, calificación crediticia, análisis de fraude y un gran número de situaciones en los que interviene una clasificación. Hoy en día, es el método de referencia para problemas de clasificación binaria. La regresión logística opera de manera similar a la regresión lineal al encontrar los valores de los coeficientes que multiplican al valor de cada variable explicativa. A diferencia de la regresión lineal, la predicción de la respuesta se transforma utilizando una función no lineal, de modo que los resultados se pueden interpretar como una puntuación o una probabilidad estimada. Desafortunadamente, poco se ha dicho sobre los valores atípicos en este contexto y el papel de las observaciones extremas.
- *Árboles de decisión (clasificadores y regresión).* Como clasificadores, los árboles son algoritmos que actúan como modelos predictivos en machine learning y admiten una representación gráfica del modelo de árbol de decisión como un árbol binario, que muestra cómo los datos se dividen paso a paso, en base a qué factor explicativo y cuál es el orden secuencial de la ordenación organizada. Cada nodo representa una sola variable de entrada y un punto de división en esa variable cuando la variable es cuantitativa. Los nodos de la hoja del árbol contienen una variable de salida que

se usa para hacer la predicción. Los valores extremos causan una enorme inestabilidad en la construcción de árboles, algo que puede evaluarse fácilmente mediante bootstrapping y que ha dado lugar a algunos de los métodos más sofisticados que se describen a continuación.

- *Naive Bayes.* Este procedimiento funciona como un procedimiento directo de modelización predictiva. El modelo se compone de dos tipos de probabilidades que se pueden calcular directamente a partir de una muestra de entrenamiento: 1) La probabilidad de pertenencia a cada una de las clases y 2) la probabilidad condicional de ocurrencia del suceso de interés para cada clase, que da lugar a un valor predictivo. Una vez calculado, el modelo de probabilidad se puede usar para hacer predicciones para nuevos datos utilizando el Teorema de Bayes. Asumir que cada variable de entrada es independiente y normalmente distribuida es una suposición fuerte y es bastante inviable para datos reales, sin embargo, la técnica es muy efectiva en una amplia gama de problemas complejos.
- *K-Nearest Neighbors*. Las predicciones sobre el método kNN se hacen buscando a través de todo el conjunto de entrenamiento cuáles son los K casos más similares (los vecinos) y resumiendo la variable de salida para esos casos. Para los problemas de regresión, esta podría ser la media de la variable de salida, para los problemas de clasificación este podría ser el valor de la moda (o más común). En este método es fundamental determinar qué medida se empleará para medir la distancia entre las características de los datos. El enfoque más simple es encontrar la distancia euclídea, pero esto puede requerir una gran cantidad de memoria o espacio para almacenar todas las similitudes de datos y debe actualizarse cuando entre un nuevo caso. La idea de distancia o cercanía acaba provocando muchas dificultades en dimensiones muy altas (muchas variables de entrada) que pueden afectar negativamente el rendimiento del algoritmo. Esto se conoce como la maldición de la dimensionalidad, que sugiere usar aquellas variables de entrada que son más relevantes para predecir la variable de salida.
- *Support Vector Machine.* Las máquinas de vectores de soporte son algoritmos de machine learning que encuentran un hiperplano que divide el espacio de la variable de respuesta y lo separa de tal manera que toma los mejores puntos del espacio de la variable de respuesta, utilizando su clase, ya sea 0 o 1. El algoritmo de aprendizaje SVM encuentra los coeficientes que dan como resultado una mejor separación de las clases por el hiperplano. La distancia entre el hiperplano y los puntos de datos más cercanos se denomina

margen. El hiperplano mejor u óptimo que puede separar las dos clases es la línea que tiene el margen más grande. A los puntos más cercanos se les llama vectores de soporte, ya que apoyan o definen el hiperplano. En la práctica, se usa un algoritmo de optimización para encontrar los valores de los coeficientes que optimizan el margen.

- *Bagging y Random Forest.* Es un algoritmo de machine learning potente que se basa en Bootstrap Aggregation o bagging. Bootstrap toma muestras de los datos, calcula el valor de interés (o variable de respuesta) y luego promedia todos los valores para obtener una mejor estimación del valor real. El método bagging tiene el mismo enfoque, pero se usa para estimar modelos estadísticos completos en lugar de valores únicos, como lo hacen comúnmente los árboles de decisión. Los modelos creados para cada muestra de datos son, por lo tanto, más diferentes de lo que serían de otra manera, pero igual de precisos. La combinación de sus predicciones da como resultado una mejor estimación del verdadero valor de salida.
- *Boosting and AdaBoost.* Boosting es una técnica que intenta crear un clasificador fuerte a partir de una serie de clasificadores débiles mediante la construcción de un modelo a partir de los datos de entrenamiento, y luego la creación de un segundo modelo que intenta corregir los errores del primer modelo. Los modelos se agregan hasta que el conjunto de entrenamiento se predice perfectamente o se agrega un número máximo de modelos. AdaBoost y Discrete Adaboost fueron los primeros algoritmos realmente exitosos desarrollados para la clasificación binaria. Para el caso de GBM (generalized boosting regression method), la función link es logit y se introduce para predecir variables de tipo binario.