CAMPUS
DE EXCELENCIA
INTERNACIONAL

Graduado en Matemáticas e Informática

Universidad Politécnica de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

TRABAJO FIN DE GRADO

Aplicación de Metodologías Machine Learning en la Gestión de Riesgo de Crédito

Autor:

David Trujillo Fernández

Director:

F. Águeda Mata Hernández

Daniel Ramos García

MADRID, JUNIO 2017

Agradecimientos

En primer lugar, me gustaría agradecer a mi familia por todo apoyo recibido durante estos años de carrera, por estar ahí en todo momento, apoyarme y motivarme, especialmente en los momentos más duros.

En segundo lugar, agradecer a mis tutores, Daniel Ramos, por confiar en mí, aportarme todos los conocimientos y ayuda necesaria y por su gran implicación a lo largo de todo el trabajo, así como a Águeda Mata, por ser mi tutora y hacerme ver el atractivo de las matemáticas.

Finalmente, querría agradecer a Management Solutions por brindarme la oportunidad de realizar este trabajo con ellos, en particular al equipo de I+D, con los que he podido aprender mucho, tanto en lo profesional como en lo personal.

¡Gracias!

Resumen

La complejidad en la gestión de los riesgos financieros ha aumentado durante los últimos años, lo cual ha generado, en algunos casos, grandes pérdidas económicas por parte de las instituciones financieras derivadas de fallos en los modelos de medición de riesgos y en los esquemas de toma de decisiones soportados por estos. En especial, el riesgo de crédito supone una de las principales preocupaciones para las entidades. Para reducir este tipo de riesgo, muchas entidades utilizan sistemas automáticos de clasificación de clientes denominados *credit scoring.* Como resultado de la evolución de la tecnología, los modelos de *machine learning* están cada vez más presentes en el sector financiero. En este contexto, este trabajo tiene como objetivo analizar la posibilidad de aplicación de estas metodologías en la gestión del riesgo crediticio. Además se estiman diversos modelos de *credit scoring* haciendo uso de algoritmos de *machine learning* y se realiza una comparación entre ellos. Por otra parte también se comparan los resultados de estos modelos con la regresión logística, técnica tradicional para el modelado del *credit scoring,* para poder analizar las ventajas y los avances que supone el uso de modelos *machine learning* frente a las técnicas de modelado tradicional.

Palabras clave: *riesgo de crédito, credit scoring, modelo, machine learning, clasificación, sector financiero.*

Abstract

The complexity of the financial risks management has increased for the past years, which has sometimes meant a huge amount of economic loss for financial institutions that comes from issues in financial risk forecasting models and decision-making schemes supported by these. Particularly, credit risk means the entity's main concern. In order to mitigate this kind of risk, many entities use automatic rating systems called credit scoring. As a result of the technology evolution, machine learning models are becoming even more present in the financial sector. In this context, the aim of this paper is to analyse the possibility of application of these methods in credit risk management. Besides, different credit scoring models are considered using machine learning algorithms and a comparison among them is done. Also the results of these models are compared to the logistic regression, a traditional technique for credit scoring modelling, to analyse the advantages and the progress that machine learning means over the traditional modelling techniques.

Key words: *credit risk, credit scoring, model, machine learning, classification, financial sector.*

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

Capítulo 1

1. Introducción

1.1. Introducción

En los últimos años, y gracias a los avances tecnológicos existentes en la ciencia computacional y de la información, se ha potenciado la búsqueda de herramientas o métodos de trabajo que faciliten y mejoren la realización de actividades relacionadas con la explotación, el tratamiento y el análisis de la información en una multitud de campos del conocimiento. De este modo, se están realizando avances dirigidos a mejorar el entendimiento de patrones, tanto en la naturaleza como de carácter social y humano, que conllevan avances en la toma de decisiones en campos dispares como la medicina, el marketing, la gestión pública y empresarial o la educación.

Hace simplemente poco más de veinte años, era difícil ver a las personas con teléfono móvil y con acceso a internet. Sin embargo, hoy en día, no es extraño ver a todo el mundo con un aparato electrónico con el que poder comunicarse y tener acceso a internet en cualquier lugar y en cualquier momento, lo que ha cambiado la forma de generar información, y que ha afectado también a los costes, el almacenamiento y el *time-to-market* de la misma.

Estos avances en la tecnología están beneficiando, entre otros, al sector financiero, donde la gestión de riesgos se está convirtiendo, hoy en día y debido al contexto que atraviesa el sector, en uno de los aspectos a tener más en cuenta por parte de las entidades financieras, con el objetivo de disminuir las pérdidas y disminuir la posibilidad de recaer en una nueva situación de crisis. Con este propósito, se está investigando nuevas líneas de actuación para identificar, cuantificar y mitigar, de la manera más eficaz posible, estos riesgos.

Actualmente, una de las líneas que está teniendo más acogida, no solo en la gestión de riesgo sino en otras áreas del mundo financiero, es la aplicación de las técnicas de *machine learning* (ML) o aprendizaje automático.

El uso de estas técnicas surge como consecuencia de diversos factores: la información de la que se dispone es mayor que antes, los cambios socioeconómicos ocurren con más frecuencia y la demanda de crédito es escasa. De esta forma, se intentan crear modelos con mayor vigencia (para adaptarse a los cambios en la

demanda), con mayor poder predictivo (para poder competir mejor en escasez) y con bajos costes por parte de la entidad (para poder gestionar la renovación de los modelos en un entorno de baja rentabilidad y ajustes de presupuestos).

No obstante, el *machine learning* no es un concepto nuevo, sino que surgió en el siglo XX, a finales de los años cincuenta y principio de los sesenta, coincidiendo con el inicio del desarrollo de la inteligencia artificial. Es, sin duda, uno de los conceptos de moda dentro del mundo del *big data* y de la inteligencia artificial.

Así, podríamos definir el ML o aprendizaje automático como una disciplina científica del entorno de la inteligencia artificial y la ciencia de la computación, cuyo objetivo reside en que los sistemas aprendan automáticamente (sin intervención humana).

Por el lado de la gestión de riesgos, uno de los riesgos que deben afrontar las entidades bancarias es el riesgo de crédito, definido por el BdE[1] como el riesgo de que una de las partes del contrato del instrumento financiero deje de cumplir con sus obligaciones y produzca en la otra una pérdida financiera.

En su origen, el encargado de gestionar este riesgo era un analista experto que decidía, basándose en ciertas reglas y en la experiencia, qué créditos se concedían y cuáles no. Si bien estos sistemas se han ido sofisticando, sistematizando y automatizando parcialmente, no sería hasta finales del siglo XX, coincidiendo con la globalización de los mercados financieros y la introducción de límites regulatorios, cuando se implementarán nuevos sistemas automatizados que realizarán la función del analista de una forma más eficaz y eficiente. Aun así, estos sistemas mostraban todavía recorrido de mejora para alcanzar mayor precisión, para mitigar los riesgos y reducir los costes. De esta forma y aprovechando el auge y evolución de las nuevas tecnologías y la mayor capacidad de almacenamiento y disponibilidad de datos, se ha generalizado en el sector las iniciativas dirigidas a introducir nuevas técnicas basadas en la inteligencia artificial y el *machine learning*.

No obstante, la aplicación de dichas técnicas debe afrontar diversos desafíos para ir adaptándose a la regulación del sistema financiero, ya que esta va evolucionando al ritmo que progresa el mercado y la economía, puesto que, tras la reciente crisis financiera, se hace aún más evidente la necesidad de la revisión de ciertos aspectos regulatorios.

La regulación financiera aplicable a este proceso se remonta a la década de los 70, cuando se comienzan a ver cambios en la dirección de las entidades bancarias a causa de la alta volatilidad de diversas variables que afectan a dichas entidades, causada principalmente por factores como la globalización de los mercados

[1] Circular 4/2004, de 22 de diciembre, a entidades de crédito, sobre normas de infromación financiera pública y reservada y modelos de estados financieros.

financieros, expectativas inflacionarias crecientes, y la desregularización financiera [1].

En 1974, se produce el cierre del Bankhaus Hersttat en Alemania y, como consecuencia, el Banco Internacional de Pagos (BIS) [2] , con sede en Basilea (Suiza), crea el Comité de Supervisión Bancaria de Basilea (BCBS), organismo conformado por los presidentes de los Bancos Centrales del grupo de los 10[3]. Es en esta situación del cierre imprevisto y con el propósito de restaurar la confianza y la estabilidad del sistema financiero internacional, cuando el BCBS desarrolla una serie de principios y reglas sobre prácticas de regulación y supervisión de los mercados, en vistas a evitar posteriores crisis.

El BCBS, consciente de la necesidad de crear un marco adecuado que sirva como referencia a las entidades en la gestión del riesgo de crédito, publica en 1988 el Acuerdo de Capital de Basilea (Basilea I), para regular el nivel de solvencia y disminuir el riesgo de crédito. Con el tiempo el marco regulatorio ha ido cambiando y evolucionando, haciéndose cada vez más exigente. Es por esto que los Acuerdos han estado sujetos a diversas reformas, publicándose más adelante Basilea II (2004) y Basilea III (2010), las cuales introducen nuevas restricciones de cara a los modelos, ya que incluyen la directriz de llevar a cabo un mayor seguimiento y control de los modelos, así como fomentar sistemas de estimación interna (IRB), etc.

No obstante, aunque el Comité no posee ninguna autoridad de supervisión formal y sus decisiones no son legalmente vinculantes, formula recomendaciones para la regulación de instituciones financieras y para poder afrontar las inestabilidades del mercado financiero mundial. A pesar de carecer de autoridad, sus directrices y recomendaciones de carácter vinculante fueron adoptadas por los reguladores de prácticamente todos los países con bancos internacionales activos que quisieron introducir estos acuerdos en su legislación.

A nivel europeo y actualmente, existen diversos documentos que regulan el uso de técnicas de *machine* learning en la estimación de modelos para la gestión de riesgos, relativos a su uso referido tanto a su relación con la gestión de riesgos y su impacto en la solvencia como referido a la protección de datos de los usuarios.

Referido al primer uso, el Parlamente Europeo y El Consejo publicaron la Directiva de Requisitos de Capital (CRD IV) y el Reglamento de Requisitos de Capital

[2] Ha servido como centro de discusión de temas financieros de forma internacional desde los años treinta.

[3] El G-10, cuyo origen se encuentra en los Acuerdos Generales para la Obtención de Préstamos (GAB) establecidos en 1962, lo forman los países miembros del G-7 que son Gran Bretaña, Italia, Japón y Estados Unidos, Canadá, Francia, Alemania a los que se añaden, Bélgica, Holanda, Suecia y Suiza. Suiza se incorporó en 1964, pero el nombre de G-10 no se cambió.

(CRR). El acuerdo (CRV IV/ CRR) regula los modelos de admisión de crédito e incorpora Basilea III a nivel Europeo.

Por otra parte, la revisión de la normativa y regulación sobre el acceso a datos personales se está llevando a cabo tras la publicación del el Reglamento General de Protección de Datos (GDPR)[4], que contiene la última y más profunda reforma de la regulación europea en materia de protección de datos, garantizando la protección de datos de personas físicas identificadas o identificables, datos generados por las máquinas y datos seudonimizados[5]. A pesar de que la información seudonimizada no permite la identificación directa del interesado, se trata de datos de carácter personal, y como tal, son objeto de protección del GDPR.

El tratamiento de datos por parte de las entidades debe obrar según lo establecido en el GDPR.

En lo que al uso de aprendizaje automático en la creación de modelos se refiere, el GDPR establece una serie de límites que impiden, ya sea directa o indirectamente, su aplicación.

Básicamente, el GDPR otorgará a los interesados el derecho a la explicación de cualquier decisión tomada mediante cualquier sistema automatizado o algoritmo. Como se sostiene en un análisis del GDRP [2]: "*es lógico que un algoritmo solo pueda ser explicado si el modelo entrenado puede ser expresado y entendido por una persona*", lo cual entra en conflicto con la naturaleza de los modelos de ML, de los cuales se conoce la entrada y la salida pero se desconoce todo el proceso asociado a la decisión obtenida (lo que se denomina habitualmente Black Box o Caja Negra). En el artículo "The Dark Secret at the Heart of AI" de Will Knight (MIT) [3] se explica la teoría de la Caja Negra de los algoritmos de ML, en que se sostiene que a medida que estos algoritmos se hacen más sofisticados, la posibilidad de que una persona entienda los procedimientos que utiliza para llegar a un resultado determinado disminuyen cada vez más, lo que nos deja en una situación en la que podemos evaluar los modelos en función de su calidad sin llegar a entender el procedimiento interno que realizan.

Además conforme con el artículo 22.1, todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte de modo similar, lo que limita el uso exclusivo de sistemas automatizados para la toma de decisiones.

[4] Entró en vigor el 25 de mayo de 2016 y comenzará a aplicarse el 25 de mayo de 2018.

[5] Según el GDPR; tratamiento de datos de manera tal que ya no puedan atribuirse a un interesado sin utilizar información adicional, siempre que esa información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable.

En España, la Agencia Española de Protección de Datos (AEPD) publicó a finales de enero de este año materiales y recursos con los que facilitar a las empresas su adaptación al GDPR [4]. Estos recursos contienen una *Guía del Reglamento General de Protección de Datos para responsables de tratamiento* [5], *Directrices para elaborar contratos entre responsables y encargados de tratamiento* [6] y una *Guía para el cumplimiento del deber de informar* [7].

Por su parte, el CRR establece que las entidades deberán contar con procedimientos para recopilar y almacenar los datos del modelo, calibrar los modelos, verificar sus resultados, y documentar todo lo relativo al diseño y construcción del mismo. Esto conlleva que las entidades deben recopilar y almacenar sus propios datos a fin de sustentar la veracidad de los del mismo en la medición y gestión del riesgo de crédito, al igual que debe garantizarse la completitud, precisión, consistencia y trazabilidad de los datos. En este sentido, la trazabilidad de los sistemas supone un inconveniente para estas técnicas, ya que, como consecuencia de la Caja Negra de la que se hablaba anteriormente, existe una gran falta de trazabilidad y seguimiento de estos modelos.

Por todos ello la aplicación de las técnicas de ML como método exclusivo en el tema que se trata en el trabajo no sería posible, puesto que debería haber un analista o gestor haciendo uso de estos sistemas como apoyo, sin llegar a utilizarlo como medio final de la decisión tomada.

En resumen, si bien la mayor disponibilidad de datos, su menor coste, así como un entorno propicio generado por la situación del sistema financiero suponen un marco adecuado para la proliferación de modelos de *machine* learning, la regulación actual limita el uso exclusivo y autónomo de estas técnicas, lo que lo sitúa como una herramienta soporte en procesos regulados o dedicado a aspectos concretos no regulados (i.e detección de fraude).

1.2. Objetivos del trabajo

El trabajo consistirá en la presentación del concepto del *machine learning* o aprendizaje automático y algunas de sus técnicas, así como sus aplicaciones en diversos sectores de actividad y, con base en ellas, crear modelos para la gestión del riesgo de crédito de una cartera con el fin de tener una idea más clara de sus posibilidades y avances.

Igualmente se dará a conocer en qué consisten estas técnicas y su grado de aplicación en diversas industrias, haciendo especial hincapié en el sector financiero, dónde su uso, aunque es relativamente limitado, está experimentando un crecimiento elevado.

Asimismo, la finalidad del trabajo reside en poder tener una visión general del sistema de la gestión del riesgo de crédito, así como plantear diversos modelos basados en el uso de aprendizaje automático sobre una cartera de riesgo de crédito, con el objeto de ver la actuación de estos modelos sobre un caso real, además del análisis y posterior comparación de los resultados obtenidos mediante cada una de las técnicas utilizadas.

Dichos modelos utilizarán las instancias y atributos particulares de una base de datos con antecedentes crediticios de una cartera de hipotecas de una entidad financiera. El objetivo de la aplicación de estas técnicas reside en poder predecir si un cliente va a poder hacer frente o no a la deuda. Para ello se utilizarán técnicas de *machine learning* que permitan clasificar las operaciones en buenas o malas a partir de la detección de patrones dentro de la información que contiene la base de datos.

Finalmente se verá la trascendencia y posterior ampliación a otras áreas del sector.

1.3. Estructura del trabajo

En el primer capítulo, al que pertenece este apartado, se hace una introducción a la evolución y desarrollo de la tecnología, así como una aproximación a la situación del mundo financiero. El objetivo de éste, es tener una primera visión global del estado de las nuevas tecnologías en sector financiero, así como la tesitura en la que se encuentra desde el punto de vista regulatorio. También se puede encontrar el estado del arte de la aplicación de las diferentes metodologías en la creación de modelos de *credit scoring*.

En el segundo capítulo se puede encontrar una visión más profunda y teórica sobre qué es el *machine learning*, qué aporta sobre los métodos de modelado tradicional, así como el estudio de algunas aplicaciones de estos métodos en las distintas industrias, con especial relevancia en el sector financiero, sobre el cual se desarrollará el trabajo.

En el tercer capítulo podemos encontrar los aspectos más técnicos del trabajo, donde se presentan las bases teóricas de diferentes técnicas de machine learning, así como su aplicación sobre una cartera de riesgo de crédito y la presentación de los resultados.

En el cuarto capítulo se presenta una interpretación más global sobre los resultados de los modelos y las conclusiones del trabajo, así como el análisis de las mejores técnicas que se pueden aplicar. Finalmente se verá el alcance del trabajo y su posible extensión a otras áreas del sector.

1.4. Estado del arte

En esta sección se realiza una revisión de algunos de los trabajos relacionados con la aplicación de nuevas técnicas en la gestión de riesgo de crédito[6], a través del *credit scoring*[7].

Las formas de hacer frente al problema de clasificación del *credit scoring* son varias, ya que existen una gran cantidad de técnicas que se soportan sobre el análisis estadístico, minería de datos, inteligencia artificial o *machine learning*.

La forma tradicional más empleada para estimar modelos de *credit scoring* y que generalmente ofrece buenos resultados estadísticos, ha sido la regresión logística. Asimismo, cuando no existe muestra histórica o esta presenta sesgos o problemas, es habitual la definición de reglas y árboles de decisión como método de *scoring* alternativo. No obstante, existen otras aplicaciones más novedosas utilizadas para crear estos modelos, como aquellas técnicas basadas en redes neuronales, *splines* de regresión adaptativa, máquinas de vector soporte o lógica difusa.

Como precursor del uso de métodos estadísticos en los modelos de *credit scoring* podemos mencionar a Durand (1941) [8]. Desde ese entonces se han desarrollado una multitud de trabajos. Entre los trabajos pioneros que ayudaron a asentar los modelos de *credit scoring* se encuentran Myers y Forgy (1963) [9], Orgler (1970) [10] y Bierman y Hauserman (1970) [11] , quienes formularon modelos de programación dinámica en los que las variables de decisión eran si otorgar o no créditos y qué cantidad.

Actualmente existen varias técnicas analíticas que han sido implementadas en la industria para el desarrollo de modelos de *scoring*, que pueden estar basadas en regresión multivariante [12], en el que construye un modelo de créditos al consumo utilizando cuatro variables continuas, o regresión de variable dependiente limitada como modelos *logit* [13], o *probit* [14]. En este último caso, Henley [14] encontró que una regresión logística no es mucho mejor que la lineal, no obstante, es común seguir realizando la modelización por medio de regresiones logísticas. Ello se atribuye a que una gran cantidad de créditos se encuentran entre los cuartiles 0.2 y 0.8, donde las distribuciones son casi idénticas. Por otra parte, aunque el objetivo sea el mismo, la principal diferencia entre ambas regresiones es el uso de una variable dependiente binaria. Por otra parte, si se utiliza una regresión lineal, los valores predichos pueden ser mayor que uno o menor que cero, aun siendo estos valores teóricamente no admisibles [15]. También, la myor popularidad de la regresión

[6] Se nombran algunas técnicas que serán explicadas y desarrolladas en el apartado 0

[7] Sistema de clasificación automática de solicitudes de operaciones crediticias, según su probabilidad de incumplimiento.

logística en estos temas se debe sus resultaos están acotados entre 0 y 1 y, por tanto, puede ser interpretada en términos de probabilidad. Además, la función que utiliza es la sigmoide, curva con forma de 'S', que se ajusta mejor a los datos, ya que en este caso solo pueden tomar valor 0 ó 1.

También hay técnicas de análisis estadístico multivariado como el análisis discriminante. En [16] y [17] Eisenbeis presenta una crítica al uso de análisis discriminante en estudios de negocio, economía y finanzas. También hay otros trabajos en los que se presentan estos modelos como los menos útiles por los resultados [18].

Además de estas, que son técnicas paramétricas, existen otras no paramétricas como la programación lineal, redes neuronales, árboles de decisión, máquinas de vector soporte, algoritmos evolutivos, redes bayesianas, modelos híbridos, etc.

Una explicación y aplicación práctica de varios métodos los podemos ver en [19] de Bonilla et al. (2003), donde haciendo uso de una base de datos con 690 registros con 14 atributos de los solicitantes de un crédito, estima siete modelos de *credit scoring*, dos paramétricos, análisis discriminante y regresión logística (LOGIT) y otros cinco no paramétricos: árboles de decisión (CART), árboles de regresión (C4.5), regresión lineal ponderada, *splines* de regresión adaptativa multivalente y redes neuronales.

En [20] se presenta un estudio acerca de la eficiencia referente a estudiar los modelos de *credit scoring* bajo enfoques paramétricos y no paramétricos, en el que se evidencia la capacidad de clasificación de los métodos de árboles de decisión. Por su parte Thomas en [21] también realiza un análisis de las técnicas estadísticas y de optimización empleadas en la construcción de modelos de concesión de créditos. En [22] Baesens realiza un estudio sobre ocho bases de datos utilizando varios métodos, concluyendo que el modelo menos eficiente estadísticamente hablando se consiguió con el método de Naïve Bayes, mientras que los mejores resultados se lograron con los modelos de regresión logística, redes neuronales y árboles de clasificación.

Sin embargo, en [23] hacen una recapitulación acerca de las diferentes técnicas utilizadas a lo largo de los años. En dicho trabajo ponen de evidencia la tendencia a modelar los scoring a través de técnicas de segmentación, como los árboles de decisión, lo cual ha perdurado a pesar del avance de nuevos métodos de aprendizaje automático como las redes neuronales, máquinas de vector soporte, etc.

Por otro lado, tanto Hung y Chen en [24] como Yu, Wang y Lai en [25], han demostrado que en la mayor parte de los casos estudiados, los modelos que consiguen una mayor precisión son los métodos *ensemble* o métodos híbridos. También en un estudio reciente realizado por Wang et al. [26] en el que se comparan regresión logística, arboles de decisión, redes neuronales y máquinas de vector

soporte con tres métodos *ensemble*: *Boosting*, *Bagging* y *Stacking*, se obtienen resultados más precisos cuando se utilizan los métodos *ensemble*, siendo *bagging* mejor que *boosting*.

Por otra parte, desde su desarrollo en 1995 por Vapnik en [27] , las máquinas de vector soporte o *Support Vector Machines* (SVM), no han dejado de aplicarse en diversos trabajos relacionados con clasificación de clientes debido a los buenos resultados obtenidos. Algunos autores que han aplicado esta metodología con este fin son Yu et al. [28], Xu, Zhou y Wang [29] o Belloti y Cook [30], que además han comparado estas técnicas con otras. En este mismo contexto Moreno y Melo en [31] realizan un trabajo con SVM con dos bases de datos. En el documento los autores realizan un benchmark entre las técnicas de máquinas de vector soporte, regresión logística y análisis discriminante. Los resultados obtenidos arrojan un mejor desempeño en la predicción por parte de las SVM respecto a las otras dos metodologías.

Por tanto se observa que, si bien el uso de regresiones logísticas está más extendido que el resto de metodologías, los diversos trabajos citados encuentran otras metodologías útiles que podrían arrojar mejores resultados frente a las clásicas utilizadas. Esta idea, que se tratará de forma más exhaustiva en el capítulo 3 y que se conoce como *no free lunch theorem*, justifica la idea de analizar el uso de nuevas técnicas para la estimación de *credit scoring*.

Teniendo en cuenta todos estos estudios y las conclusiones alcanzadas en ellos, en este trabajo se crearán diferentes modelos de *credit scoring* basándose en diferentes técnicas que han resultado ser las mejores entre los demás trabajo, con el objetivo de sacar una conclusión clara. La idea principal reside en modelizar un clasificador mediante un método tradicional como una regresión logística y compararlo con otras técnicas más novedosas como las máquinas de vector soporte, el *random forest*, etc.

Capítulo 2

2. Marco Teórico

2.1. Machine Learning

2.1.1. ¿Qué es el machine learning?

El *machine learning* o aprendizaje automático es una disciplina científica del entorno de la inteligencia artificial y la ciencia de la computación, cuyo objetivo reside en que los sistemas aprendan automáticamente (sin intervención humana). Aprender en este sentido, no se refiere a lo que comúnmente se conoce como aprendizaje humano, basado en la experiencia y en la razón, sino a la identificación de patrones complejos dentro de una gran cantidad de datos obtenidos mediante ejemplos, la experiencia o las instrucciones predefinidas. Tras el concepto de aprendizaje se encuentra un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros, adaptándose a la incorporación de información adicional y recalibrando los resultados. Esto no quiere decir que las personas no tengan que desempeñar ningún papel. Los analistas son necesarios para participar en tareas como la revisión y confirmación de decisiones, y en casos especiales, en la toma de decisiones efectivas. Además, incluso los sistemas más automatizados, confían en expertos para crear y mantener reglas y monitorizar los resultados.

Como punto de partida, una de las definiciones formales de *machine learning* más citada y ampliamente aceptada en el campo es la de Tom M. Mitchell [32]:

> *"Se dice de un programa informático que aprende de la experiencia E con respecto algún conjunto de tareas T y la medida de rendimiento P si su rendimiento en las tareas en T, medido por P, mejora con la experiencia E"*
>
> *Tom M. Mitchell*

Aunque, como se ha dicho, el m*achine learning* pertenezca principalmente al campo de las ciencias de la computación, también posee un soporte estadístico detrás [33]. Por una parte, la informática se centra en resolver cómo podemos construir máquinas que resuelvan problemas y qué problemas son manejables y tratables y, por otra, la estadística se refiere a qué se puede deducir de los datos, junto a una serie de

suposiciones de modelado y con qué fiabilidad. Sin embargo, el aprendizaje automático se encuentra influenciado por ambos.

Por un lado, mientras la informática se ha centrado principalmente en cómo programar manualmente sistemas, el aprendizaje automático se centra en cómo conseguir que los programas se generen a sí mismos (por medio de la experiencia y alguna estructura o reglas iniciales). Por otro lado, mientras la estadística se ha centrado en el estudio de las conclusiones que se pueden deducir de los datos, el ML, además de este propósito, ha extraído la idea de incorporar preguntas adicionales acerca de las arquitecturas y algoritmos que se pueden emplear para capturar, almacenar, indexar, recuperar y fusionar esos datos.

De esta manera y haciendo uso del conocimiento de estas dos áreas, las técnicas de *machine learning* consisten en la detección de patrones presentes en los datos. Sin embargo, en lugar de extraer la información de forma comprensible para las personas, los sistemas de ML se centran exclusivamente en el uso de la información para detección de esos patrones y ajustar de manera adecuada el programa. Aun así, para algunos propósitos sigue resultando necesario que la predicción no solo sea adecuada, sino interpretable, de forma que pueda extraerse de la relación una explicación que sustente el patrón identificado, por lo que no todos los algoritmos serán útiles. En otras palabras, hay algunos entornos (como la medicina) en los que se requiere que el sistema encuentre unas reglas de predicción lo más sencillas posibles de interpretar [34], lo cual dependiendo del modelo es más o menos factible.

Además de la detección de patrones, su objetivo reside en el desarrollo de algoritmos con valor práctico. Estos algoritmos deben ser eficientes, para ello, las cuestiones a tener en cuenta, computacionalmente hablando, son el tiempo de ejecución y el espacio utilizado, ya que, al tratar con una gran cantidad de datos, estos puntos pueden suponer la diferencia entre un buen y un mal modelo.

En el ámbito del aprendizaje, un punto notable es la cantidad de datos necesarios para entrenar el modelo, ya que cuanto más datos se tengan, más robusto será éste, pudiéndose identificar patrones con mayor soporte.

Dentro del *machine learning*, se pueden diferenciar varios métodos, los dos más adoptados son el aprendizaje supervisado y el aprendizaje no supervisado.

2.1.1.1. Aprendizaje supervisado

El aprendizaje supervisado consiste en que, partiendo de una muestra

$$\mathcal{L} = \{(X_i, Y_i) | i = 1,2, \dots n\}$$

construida por n realizaciones de un par de variables (X, Y), se construye una función $f: \mathcal{X} \rightarrow \mathcal{Y}$ con la cual, dado un vector de entrada X, se pueda predecir con cierto grado de confianza la variable $Y = f(X)$. Para cada observación (X_i, Y_i) de $\mathcal{L}$, a la variable $X_i \in \mathcal{X}$ se le llama variable de entrada, explicativa o input y a $Y_i \in \mathcal{Y}$ variable dependiente u output [35].

A su vez, los problemas de aprendizaje supervisado se pueden dividir en problemas de clasificación y regresión:

- Clasificación: cuando la variable dependiente es discreta o categórica, se trata de un problema de clasificación. En estos problemas se intenta predecir a qué clase pertenece un conjunto de datos. Cuando solo hay dos clases se denomina clasificación binaria. Cuando hay más de dos categorías, se trata de un problema de clasificación multiclase.
- Regresión: cuando la variable dependiente es continua, estamos ante un problema de regresión. Tiene como objetivo predecir valores continuos.

A su vez, existen distintos algoritmos para poder resolver cada tipo de problemas. La Figura 2.1 representa un esquema sencillo sobre los problemas y los algoritmos más representativos y utilizados en el aprendizaje supervisado.

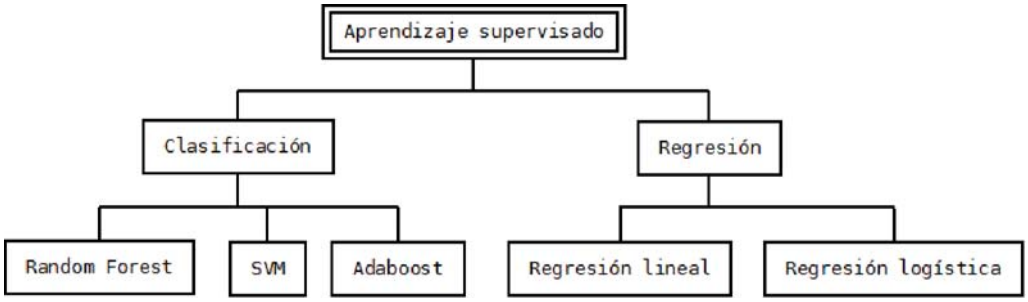

Figura 2.1: Diagrama aprendizaje supervisado

Como se verá más adelante, el problema al que se hace frente en este trabajo pertenece al aprendizaje supervisado, ya que la base de datos de la que se dispone está previamente etiquetada. Dentro del aprendizaje supervisado se trata de un problema de clasificación binario, puesto que el objetivo es clasificar los clientes en una de las dos clases, o buenos o malos en función de los resultados de la predicción de los modelos.

2.1.1.2. Aprendizaje no supervisado

En el aprendizaje no supervisado estamos ante el problema de que solo se tiene las variables de entrada (X) pero no la de salida. El objetivo de estos métodos consiste en modelar una estructura o una distribución subyacente a los datos con el fin de obtener más información sobre los mismos.

A su vez, estos problemas se pueden agrupar en Clustering o Asociación:

- Clustering: su objetivo es la división de los datos en grupos según algunas de sus características.
- Asociación: consiste en la descripción de reglas que describan grandes cantidades de los datos, como por ejemplo decir que las personas que suelen comprar a, también compran b.

Al igual que en el aprendizaje supervisado, en el aprendizaje no supervisado hay varios algoritmos con los que resolver los distintos tipos de problemas que existen. Los algoritmos más representativos se muestran en la Figura 2.2.

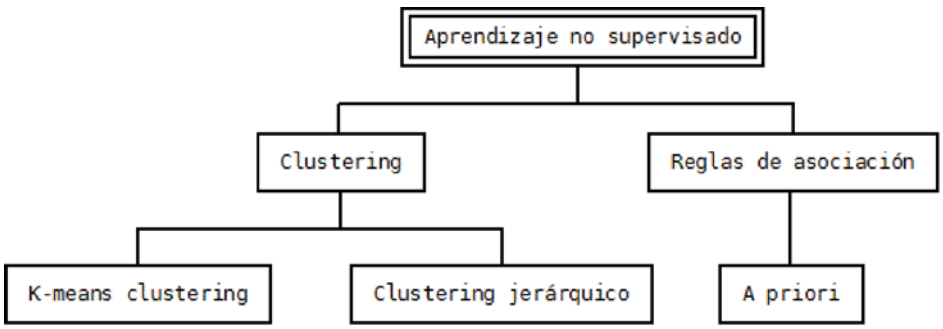

Figura 2.2: Diagrama aprendizaje no supervisado

2.1.1.3. Aprendizaje semi-supervisado

Además de estos dos métodos, se encuentra el aprendizaje semi-supervisado, aunque está menos extendido que los anteriores. Este método se utiliza cuando se dispone los inputs pero solo algunos están calificados con sus correspondientes outputs. En estos casos se pueden usar técnicas de aprendizaje no supervisado para aprender y obtener patrones de los datos de entrada. También se utilizan técnicas de aprendizaje supervisado para realizar mejores predicciones sobre los datos sin etiquetar, devolver esos datos al algoritmo de aprendizaje supervisado como datos de entrenamiento y utilizar el modelo para predecir los nuevos datos [36].

2.1.2. Diferencias entre modelos de Machine Learning y los modelos Estadísticos Clásicos

Ambas técnicas, *machine learning* y la estadística tradicional, están convergiendo cada vez más a formar parte de una misma metodología y, en algunos casos, se sostiene que no existen grandes diferencias entre ellas, como mantiene Larry Wasserman [37]:

> *"¿Cuál es la diferencia entre estos dos campos??La respuesta corta es: Ninguna. Ambos se centran en la misma pregunta: ¿cómo aprendemos de los datos?"*

No obstante, aunque ambos tienen la misma finalidad, aprender de los datos, difieren en las metodologías aplicadas para el tratamiento de los mismos, así como para la generación de los modelos. Las diferencias más claras son:

- Desde el punto de vista tradicional del análisis de datos, el *machine learning* consiste en el uso de algoritmos que aprenden de los datos sin necesidad de depender de la programación basada en reglas. Por otra parte, los modelos estadísticos consisten en la formalización de las relaciones entre las variables en forma de ecuaciones y fórmulas matemáticas [38].

- Otra diferencia reside en el dominio del que provienen, mientras el Machine Learning procede, en un principio, de la ciencia de la computación y la inteligencia artificial (como se ha expuesto en el punto anterior), el origen de los modelos estadísticos se encuentra en las matemáticas, ya que, por definición, la estadística es una rama de las matemáticas que utiliza datos, ya sea de toda la población o de una muestra de la población para llevar a cabo análisis y presentar inferencias, haciendo uso de algunas técnicas como la regresión, la varianza, la desviación, la probabilidad condicional, etc.

- Mientras los modelos estadísticos clásicos hacen uso de datos estructurados, el *machine learning* se abastece de datos tanto estructurados como no estructurados, además, la cantidad de datos del que hace uso el ML es mucho mayor, ya que pueden proceder desde BBDD públicas, redes sociales hasta Internet of Things (IoT). Ello conlleva unos requerimientos computacionales mayores, desde más espacio de disco hasta mayor velocidad de procesamiento y de memoria interna (RAM) [39].

- En los modelos estadísticos existe una limitación en los patrones encontrados debido a la suposición de hipótesis previas para la creación de modelos. El ML evita la suposición de hipótesis previas, por lo que será capaz de descubrir patrones ocultos en los datos, lo cual deriva en modelos con un mayor poder predictivo.

- La interpretación de resultados de los modelos estadísticos resulta más sencilla. Sin embargo, debido a la cantidad de datos del ML y la no asunción de hipótesis para la creación de modelos, los resultados de éstos suelen ser una especie de caja negra, sobre la cual sabemos los resultados pero no su interpretación, ya que la metodología interna aplicada a la creación del modelo es desconocida.

- La mejora de los modelos estadísticos se basa en hipótesis predefinidas por el conocimiento o suposición previa de las relaciones entre variables. Los modelos de ML a su vez, al nutrirse de tanta cantidad de datos, buscan patrones y relaciones entre las variables sin restricciones previamente definidas.

Además de la diferencia entre ML y los modelos estadísticos clásicos a fin de ver la aportación del ML sobre éstos, es preciso tener claras las diferencias existentes entre inteligencia artificial (IA), *machine learning* y *deep learning* o aprendizaje profundo (DP), ya que al desarrollarse y evolucionar de manera simultánea y coexistente, pueden existir ciertas dificultades a la hora de diferenciarlas al ser la línea que los separa muy delgada.

La manera más sencilla de pensar en su relación y a su vez distinguirlos, es visualizarlos como círculos concéntricos con la IA como el mayor de todos (concepto que surgió en primer lugar), luego el ML, siendo una de las ramas más destacadas de la inteligencia artificial y, el menor, el DL o aprendizaje profundo.

La inteligencia artificial es un concepto más amplio, cuyo objetivo es que las máquinas sean capaces de realizar tareas de una manera que consideramos inteligente aunque, en muchos casos, consista en la ejecución de reglas previamente programadas. Por su parte, el *machine learning*, es una rama de la IA basada en la idea de ser capaz de dar a las máquinas acceso a los datos y dejar que aprendan por sí mismas. Finalmente, el *deep learning* o aprendizaje profundo, está asociado con un algoritmo de ML, las redes neuronales artificiales, que hace uso del concepto del cerebro humano para facilitar el modelado de funciones, como definió el científico Andrew Ng[8] en una entrevista en Recode [40]:

> *"Es una tecnología de aprendizaje que funciona simulando libremente el cerebro […]. Deep learning funciona mediante la simulación de cientos de miles de millones de neuronas, simuladas por la computadora, hablando entre sí"*

Por último, cabe destacar que también existen diferencias entre el *machine learning* y el *data mining* (DM) o minería de datos, conceptos que se suelen confundir. El ML se dedica al estudio, diseño y desarrollo de distintos algoritmos que aportan a los sistemas la capacidad de aprender de forma automática sin ser explícitamente programados. Por su parte, el DM se refiere al proceso completo que, a partir de los datos intenta extraer algún conocimiento desconocido. Durante este proceso se emplean algoritmos de ML. Es decir, el *data mining* trata con la búsqueda de información específica a través del uso de técnicas de ML y el *machine learning* se centra únicamente en realizar una tarea determinada.

Todo ello puede visualizarse de una manera más clara en el diagrama de Venn de la Figura 2.3.

[8] Científico jefe de Baidu (el Google chino), profesor de Stanford y fundador de Google Brain

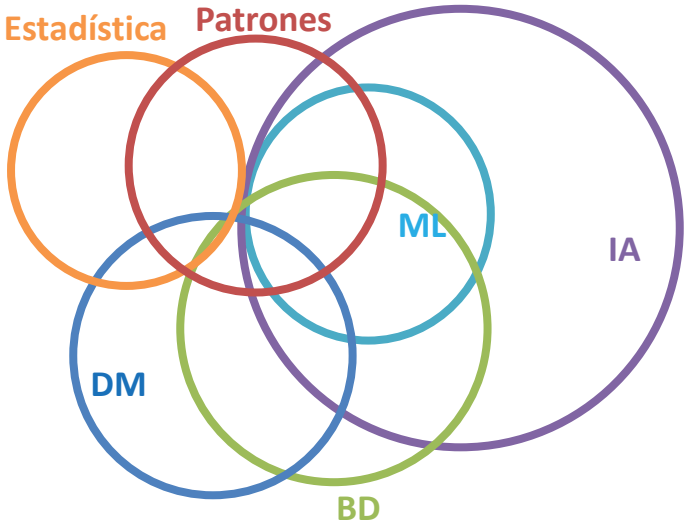

Figura 2.3: Diagrama de Venn. Fuente: Elaboración propia

Como se puede observar y ya se ha comentado, el *machine learning* y la estadística comparten campos y cada vez convergen más a un punto común.

Por otra parte, dentro del círculo del ML se encontraría el *deep learning*, ya que surgió más tarde y se centra en una de sus técnicas.

Por último, cabe destacar que el *data mining* o la minería de datos, hace uso del conocimiento de todos los ámbitos anteriores para poder dar solución a una amplia serie de problemas.

2.2. Aplicaciones del Machine Learning

En este apartado se estudiará el grado de aplicación de las técnicas de aprendizaje automático en diversas industrias, así como algunas de sus aplicaciones.

2.2.1. Sectores de actividad

Actualmente, debido a la gran cantidad de datos que se disponen, la velocidad a la que se generan y los avances tecnológicos, las técnicas de *machine learning* cada vez tienen una mayor presencia en ciertos sectores en los que, por la cantidad de datos, la relevancia de éstos y la necesidad de modelos cada vez más fiables, están experimentando un gran desarrollo y avance. Esto se debe a que hoy, una máquina

puede realizar algunos trabajo con mayor eficacia y eficiencia que una persona ya que sus capacidades no solo se limitan a hacer lo que le hemos especificado, sino que va mucho más allá y progresa más rápido, precisamente porque una de las cosas que estos sistemas pueden hacer mejor que las personas es aprender de los datos, puesto que tiene mayores posibilidades de combinarlos e iterarlos hasta encontrar una respuesta que satisfaga el problema inicial con la mayor eficiencia posible. Uno de los factores que influyen, es la capacidad de entender las restricciones asignadas e ir más allá a la hora de buscar soluciones, de manera que para un humano le sería casi imposible, debido a sus propias limitaciones.

Un ejemplo de estos avances se puede observar el juego del Go. En marzo de 2016 se realizó una competición entre Lee Sedol, uno de los jugadores más competitivos en este juego, y AlphaGo, un sistema desarrollado por Google. La victoria fue para el sistema. Teniendo en cuenta que el Go se considera el juego con más combinaciones posibles, esto nos da una idea de la capacidad de la máquina, que es capaz de concebir jugadas nunca antes llevadas a cabo por un ser humano, y es capaz de llegar a ellas sin seguir ningún ejemplo, sino jugando partidas contra sí mismo e iterando para encontrar el mejor movimiento posible, lo cual pone de manifiesto también la gran capacidad de los procesos de toma de decisiones.

Algunas de las industrias con un mayor grado de implantación de estas técnicas son, entre otras, la medicina, la tecnología o el marketing.

2.2.1.1. Medicina

En los últimos años ha habido un aumento drástico del uso de métodos computacionales para el análisis de señales biomédicas [41]. El enfoque general de estos métodos reside en la evolución de la inteligencia artificial y el machine learning. Una de las muchas aplicaciones consiste en crear clasificadores que pueden separar sujetos en dos clases (normal, inusual) o más, basadas en los atributos de cada sujeto.

La implantación de estos nuevos sistemas tiene su origen en el desarrollo tecnológico del que ya se ha hablado anteriormente, ya que, con el rápido crecimiento de las tecnologías, es posible disponer de una gran cantidad de datos en un tiempo reducido como la determinación de genotipos, proteínas, el desarrollo de obtención de imágenes médicas, etc.

Debido a este gran aumento del número de parámetros y de la importancia de los mismos para el diagnóstico de alguna enfermedad, el desafío consiste en hacer uso de toda esta nueva información para poder aumentar la precisión de estos diagnósticos gracias al empleo de nuevas técnicas como el ML, la IA o la visión

artificial puesto que los diagnósticos aportados mediante los métodos tradicionales contienen gran cantidad de fallos.

Inicialmente, la dependencia de información a escala macro mantenía el número de variables lo suficientemente pequeño como para que los métodos estadísticos tradicionales o incluso, a veces, la intuición de médico, pudieran utilizarse para obtener un resultado. Sin embargo, ahora nos encontramos con decenas de nuevos parámetros moleculares, celulares y clínicos, quedando los antiguos métodos, para el manejo de esta cantidad de atributos, obsoletos. Por consiguiente, una solución plausible reside en la utilización de nuevas técnicas basadas en el uso de *machine learning*. No obstante, estos sistemas tienen una serie de requerimientos [42] entre los que destacan la transparencia del diagnóstico y la capacidad de explicación, ya que las decisiones tomadas por el sistema deben ser transparentes para el profesional al cargo y debe ser capaz de explicar el porqué de las decisiones tomadas a la hora de diagnosticar a un nuevo paciente; de lo contrario no se consideraría un resultado serio ni consistente. La única situación en la que se aceptaría una caja negra sería aquella en la que el clasificador supere, con un gran margen, el rendimiento del resto de clasificadores médicos, incluyendo la propia interpretación del médico. Sin embargo, esta situación es altamente improbable.

Hoy en día, el *machine learning* se está aplicando en la detección y clasificación de tumores por medio de Rayos-X e imágenes CRT, detección de cáncer, cirugía robótica (robot da Vinci), detección de drogas, etc.

2.2.1.2. Marketing

El desarrollo de las operaciones llevadas a cabo en Internet influye en la creciente oferta y demanda de nuevas tecnologías. Uno de los sectores que está sufriendo una mayor transformación, es el *marketing*. Esto es debido a que el uso de internet y la aplicación del *machine learning* están permitiendo que surja cada vez un mercado más personalizado, ya que las empresas se enfrentan al problema de ajustar su oferta a las perspectivas individuales. Esta personalización del mercado es también posible debido, en parte, a la gran disponibilidad de información sobre cada individuo.

Una de las mayores aplicaciones del aprendizaje automático en el sector se da en los sistemas de recomendación, que tratan de encontrar el contenido adecuado para cada usuario como: la respuesta a alguna información requerida, qué te gustaría leer, escuchar, ver y, también, comprar. Los motores de búsqueda están dedicados a ello, puesto que hay veces que un usuario ni siquiera sabe dónde o qué buscar cuando quiere leer las noticias, ver una película… De esta manera, el objetivo consiste en

recomendar algo que puede resultar útil al usuario, basándose en sus intereses demostrados con su actividad anterior relacionada con el mismo ámbito.

Los datos que se necesitan para crear uno de estos sistemas, capaces de basarse en la actividad de los usuarios, provienen de los clics que hace en un *link* o producto, de las *cookies*, de las páginas que se visita, etc. Toda esta información es almacenada para poder ser utilizada en la próxima búsqueda y utilizada de diferentes formas [43], por ejemplo en crear sistemas de *rating* sobre productos, basándose, como se ha mencionado anteriormente, en los productos seleccionados, los marcados como "me gusta", los añadidos al carro… De esta manera, los sistemas de clasificación pueden asignar calificaciones implícitas basadas en las acciones de los usuarios.

Tras esto, el siguiente paso es el de filtrado, que consiste en la criba de los productos basándose en los *ratings* y los datos sobre otros usuarios. Hay varios métodos de filtrado, uno de los más conocidos y utilizados se basa en la comparación de las elecciones de los usuarios y las recomendaciones. Por ejemplo si un usuario X compra los productos A, B y C y otro usuario Y compra A, B, C y D; es probable que el producto D le sea recomendado al usuario X. Este es el sistema empleado por empresas como Facebook, Twitter, Amazon, Spotify, etc. Estos son conocidos como modelos de asociación que se nombraron en el apartado 2.1.1 dentro de los modelos de aprendizaje no supervisado.

Asimismo, como observamos, basándose en el histórico de movimiento de los usuarios, se puede predecir qué es probable que les interese ver a continuación y recomendarlo. Por ello, los resultados de la aplicación de estas nuevas técnicas en este sector, se verán reflejados en una mayor lealtad, compromiso y gasto del cliente.

2.2.1.3. Tecnología

El auge y el avance de las nuevas tecnologías hacen que éstas se autoabastezcan y, así, poder evolucionar con mayor rapidez. Como ya se ha comentado, el ML no es nuevo, sin embargo, es ahora, cuando la aplicación de nuevos algoritmos sobre volúmenes de datos elevados y que van en aumento, cada día es más factible. Es por ello que, el ML o aprendizaje automático cada vez está más presente en la tecnología que usamos a diario.

Algunas de sus aplicaciones en la tecnología del día a día son:

- Procesamiento del lenguaje natural: Hasta ahora, las diferentes herramientas y sistemas tenían grandes dificultades para poder interpretar la información transmitida por medio del lenguaje. La información difundida a través de las redes sociales y los medios, requiere de herramientas automáticas capaces de

entender ese lenguaje para, posteriormente, poder clasificar la información, recomendar o incluso responder, por ejemplo, en un sistema de asistencia al cliente.

- Reconocimiento de imágenes: Esta tarea requiere la clasificación de una serie de objetos dentro de una fotografía, como un conjunto de objetos previamente conocidos. Es decir, para entrenar el modelo, previamente se tiene que alimentar con una gran cantidad de objetos, y así poder establecer una serie de características presentes en los objetos, para que, al introducir posteriormente una imagen nueva, el modelo sea capaz de reconocer los diferentes objetos integrantes de la imagen, basándose en esas propiedades previamente observadas.

- Clasificador de correos spam: Esta técnica se basa en la búsqueda de características y patrones sobre los correos etiquetados anteriormente como spam y no spam por parte de usuario. Una vez creado el modelo de aprendizaje automático, se pueden determinar patrones para poder decidir si un correo se trata o no de spam. Por ejemplo, si un correo contiene una cierta relación entre imágenes y texto, si proceden de determinadas IPs, los correos que contienen ciertas palabras, etc.

2.2.2. Sector financiero

Como se ha podido ver en el apartado anterior, las técnicas de aprendizaje automático están muy presentes en ciertas industrias, entre las que destacan la medicina, el marketing a través de los sistemas de recomendación y las nuevas tecnologías. De estas últimas hay varias a las cuales damos un uso diario, como sistemas de reconocimiento de imágenes, de reconocimiento de voz, procesamiento, reconocimiento y producción de lenguaje natural, entre otros.

Asimismo, se va a estudiar el grado de aplicación de estos métodos en el ámbito financiero, donde su uso, actualmente, es relativamente limitado, aunque está experimentando importantes procesos de crecimiento e integración de nuevas técnicas al negocio.

Hoy en día, una gran cantidad de datos de clientes está disponible y accesible para los Bancos. Una mayor capacidad computacional y más barata, permite a las entidades bancarias aprovechar la nueva información, como el comportamiento de los clientes frente al gasto, la presencia en las redes sociales, la navegación en línea, el uso de las tarjetas, el acceso a datos externos no estructurados; para la mejora de toma de decisiones en la gestión de riesgo crediticio [44].

Aunque algunas instituciones financieras están experimentando y aplicando estas técnicas en múltiples áreas, que se comentarán más adelante, la adopción de modelos de autoaprendizaje en la gestión de riesgo de crédito no se encuentra tan desarrollada, puesto que se enfrenta a algunos desafíos regulatorios y de gestión. Por una parte, existen ciertos aspectos regulatorios y normativos que impiden, de forma directa, el uso de las metodologías de ML en este ámbito. Por otra parte, desde el punto de vista de la gestión, la aplicación generalizada de estos métodos resulta complicada por la dificultad que conlleva la integración de nuevos modelos y la formación de los analistas y gestores de riesgo encargados de su uso. La dificultad del seguimiento del seguimiento y la nula trazabilidad de los modelos constituye otro inconveniente a la hora de la implantación de dichos modelos en las entidades.

En las áreas de la industria financiera en las que se están aplicando, las empresas suelen hacer uso de *data warehouses, data lakes* y otras herramientas para informar y analizar el comportamiento de un cliente [45], con el fin de anticipar mejor sus movimientos optimizando las operaciones. Mediante el uso del *big data* y el *machine learning* se pueden lograr mayores beneficios a medida que el negocio gana capacidad predictiva y se vuelve más ágil, produciendo un indicador de crédito más robusto y amplio. La aplicación de estas técnicas permite a los bancos obtener niveles de conocimiento más exhaustivo de los datos a la vez de una reducción de costes, mayor precisión en la toma de decisiones y una respuesta más rápida a los clientes.

Aun cuando la automatización completa de uno de estos procesos es posible, algunas cuestiones fiduciarias, legales o éticas puede que requieran la presencia de una persona responsable que desempeñe un papel activo supervisando el resultado final de la técnica aplicada. A continuación se exponen los principales retos asociados a la introducción de las técnicas de *machine learning* en la estimación de estos modelos.

En el sector bancario, por ejemplo, se ofrece a los clientes un crédito basándose en sus historiales de crédito y pago y, dependiendo, en alguna medida, de un analista humano apoyado en la política y los procesos de la entidad que decide qué atributos se tienen en cuenta y cuáles no a la hora de generar el modelo, que después se utilizará en la valoración de la operación. En este caso, si se aplicaran las técnicas de aprendizaje automático para crear el modelo y se rechazara una operación, se encontraría ante el dilema de conocer el porqué, ya que como ya se ha mencionado anteriormente, la salida de estos modelos es una caja negra difícil de interpretar y con falta de trazabilidad, de manea que dicha resolución no se podría justificar con argumentos de negocio de manera precisa.

No obstante, las tomas de decisión mediante sistemas automatizados pueden ayudar a las entidades a realizar tareas rutinarias de una forma más eficiente y rápida, aunque también conlleva una gestión activa de las mismas. No importa cuántas tareas

puedan ser automatizadas, los gestores todavía tendrían la responsabilidad de definir tanto el contexto como los límites de estos sistemas para la toma de decisiones finales. Esto requiere monitorizar las aplicaciones para asegurarse de que no se sale de los límites de actuación definidos, así como de los niveles de riesgo del modelo[9], porque si no se limita su radio de actuación, se podrían estar cometiendo una gran cantidad de errores de manera generalizada y totalmente automatizada, lo que supondría fatales consecuencias para las empresas.

Los gestores, además de la supervisión de los niveles de riesgo de modelo, deben desarrollar procesos estables para la gestión de las excepciones. Éstas se dan cuando el sistema tiene muy pocos datos sobre los que entrenar el modelo o cuando hay muchos *missings* en los atributos, de manera que el modelo no puede ser robusto. Una de las soluciones que se podría llevar a cabo en el problema de los *missings*, sería eliminando esos atributos con falta de datos. Si se actuase de esta forma de manera generalizada, el modelo no se ajustaría del todo a la realidad resultando así un modelo poco robusto y con un bajo poder predictivo. Por ello, resulta de gran importancia establecer un sistema de actuación frente a estas situaciones.

Otra preocupación de las instituciones financieras es cómo responderán los reguladores ante el uso de estas técnicas. Los reguladores financieros han emitido una amplia guía de supervisión (Basilea) sobre el uso de las tecnologías de la información y la seguridad, la privacidad y la gestión que requieren las instituciones financieras para evaluar el riesgo y desarrollar controles adecuados. Además de ello, a medida que aumenta la aplicación y las solicitudes de la inclusión de estas nuevas técnicas, es probable que los reguladores se centren en el uso de las mismas e identifiquen deficiencias en los controles, desarrollando así nuevas regulaciones que contemplen su uso.

Por todo ello, la implantación de estas técnicas en los modelos de *scoring*[10] y en general en los modelos de detección de riesgos se está viendo demorada frente al resto de ámbitos del sector, puesto que uno de los mayores retos a los que se enfrenta es a cuestiones regulatorias, debido a que los reguladores alientan a las entidades a estandarizar sus procesos de toma de decisiones como una forma de asegurar el cumplimiento de los aspectos regulatorios [46].

Aparte de la aplicación de técnicas de aprendizaje automático en la creación de un modelo de scoring automático en la industria financiera, hay instituciones que están aplicando estas técnicas en diversas áreas. En los siguientes apartados se describen

[9] De acuerdo con la Fed y la OCC, se define como *el conjunto de posibles consecuencias adversas derivadas de decisiones basadas en resultados e informes incorrectos de modelos, o de su uso inapropiado*.

[10] Un Scoring es un sistema de clasificación automática de solicitudes de operaciones crediticias que permiten ordenarlas en función de su probabilidad de incumplimiento (préstamos al consumo, hipotecas, concesiones de tarjetas de crédito, etc.).

algunas de las prácticas más conocidas dentro de los límites de actuación de la banca tradicional y las nuevas empresas cuyo desarrollo ha derivado del avance tecnológico (Fintech).

2.2.2.1. Banca tradicional

2.2.2.1.1. Detección de fraude

A medida que aumenta el comercio y las transacciones por Internet, aumenta también la amenaza de fraude. Esto no solo lleva consigo las pérdidas económicas pertinentes, sino también la reputación de la empresa y la confianza de los consumidores, por lo que resulta un problema de gran importancia a depurar. El fraude se ha convertido en un gran negocio que afecta a un gran número de industrias. Según una encuesta realizada por PWC en 2016 sugiere que más de uno de cada tres (36%) organizaciones son víctimas de estos delitos [47].

Actualmente, la lucha contra el fraude y demás los delitos financieros son uno de los retos más duros a los que se enfrenta el sector financiero.

La banca maneja grandes volúmenes de datos, no obstante, los sistemas previos de detección de fraude financiero dependían en gran medida, de conjuntos de reglas rígidos y complejos, suponiendo que todos los fraudes seguían un patrón similar y ocurrían de la misma forma, buscando patrones, tales como las direcciones IP sospechosas o inicios de sesión inusuales. Con el avance de la tecnología y el conocimiento de la misma, los defraudadores tienen nuevas posibilidades de cometer fraude, por ello el BD y el ML está cambiando este enfoque, yendo más allá de seguir una lista de factores de riesgo. Haciendo uso del aprendizaje automático los sistemas pueden detectar actividades o comportamientos pocos frecuentes pero sistemáticos, encontrando casos de fraude que no son obvios y de baja incidencia.

No obstante, la detección de fraude no se limita solo a las transacciones financieras, sino también a la detección de fraude interno dentro de las empresas por parte de los empleados. Es decir, si algún trabajador está utilizando sus privilegios para sustraer dinero, mercancías, utilizar información privilegiada, etc.

Para poder hacer uso de estas técnicas de detección de fraude es necesario conseguir tantos datos como sea posible, y tras la recopilación de diversas fuentes, se entrenan los modelos para poder predecir las acciones fraudulentas. Para ello también se manejan datos no estructurados, que a veces contienen la información más útil [48]. El desafío de estos sistemas de detección de fraude radica en evitar en la medida de lo posible los falsos positivos (situaciones en las que se señalan que hay algún riesgo sin serlo) y el *time-to-market* de los resultados.

Diversas empresas que utilizan técnicas de ML para la detección de fraude son:

- PayPal*:* emplea diversas técnicas de machine learning de forma combinada para obtener el mejor resultado posible en diversos entornos como realización de pagos, transacciones, etc. Su interés por el fraude se debe a las innumerables estafas que han afectado a los usuarios.

- Guardian Analytics: empresa estadounidense pionera y proveedora líder de soluciones de análisis de comportamiento y aprendizaje automático para prevenir fraude bancario, implementa un modelo no supervisado que puede identificar anomalías entre la sesión del banco online, la actividad durante el inicio de sesión y las transacciones que se realizan [49].

Algunas otras son: FICO, Sift Science, Feedzai, Brighterion, CoreLogic, etc.

2.2.2.1.2. Trading automático

Los servicios de trading asistidos por ordenador se remontan a la década de los 70 y permiten a los inversores realizar una determinada operación cuando un activo financiero alcanza un precio determinado. La automatización de estas funciones facilita el comercio para inversores. Incluso, se pueden hacer recomendaciones basadas en el análisis automatizado de las tendencias en el mercado (conocido como *roboadvisor*), si bien existen ciertas limitaciones, puesto que estos sistemas necesitan ser monitorizados, ya que puede ser que interese más realizar una operación o esperar a algún acontecimiento en vez de dejarle total autonomía a estos sistemas.

Asimismo, gran parte de los sistemas de trading automático necesitan una gestión activa, al ser preciso modificar alguno de los parámetros, según vayan variando las condiciones de los mercados.

En los últimos años, los fondos de inversión se han alejado cada vez más de los métodos tradicionales, adoptando algoritmos de aprendizaje automático para predecir las tendencias de los mercados. Los gestores de estos fondos, haciendo uso del aprendizaje automático, pueden identificar los cambios en el mercado antes que con los modelos tradicionales.

Es por esto que el potencial de las técnicas de ML está siendo tomado en cuenta por alguna de las principales empresas para irrumpir en la banca de inversión, desarrollando asesores de inversión automatizados basados en el aprendizaje automático. Algunas de estas empresas son:

- Cerebellum Capital. Es una empresa de gestión de fondos de cobertura, cuyos programas de inversión son continuamente diseñados, ejecutados y mejorados por un software basado en técnicas de machine learning.

- Sentient Investment Management. Se trata de una empresa de inversión que utiliza la IA para desarrollar estrategias de inversión y trading. Al igual que Cerebellum Capital, emplea el *machine learning* para evolucionar y optimizar sus algoritmos.

Estas aplicaciones son algunas de las más populares y relevantes. No obstante, existe un sinfín de prácticas en el sector. Así, para poder hacernos una idea de las posibilidades y la difusión de las técnicas, se enumeran otras de sus muchas aplicaciones:

- Segmentación de clientes según su comportamiento
- Definición de perfiles de gasto/ahorro
- Definición del portfolio de productos óptimos para cada segmento
- Venta cruzada de productos y servicios personalizados
- Recomendación personalizada
- Prevención de fuga de clientes
- Prevención de consultas e incidentes, y aceleración en su resolución
- Modelos semánticos de gestión recuperatoria
- Optimización de geolocalización de recursos
- Modelos de estimación de renta
- Evolución de los sistemas de alerta
- Valoración de activos
- […]

De manera paralela a la proliferación y extensión de estas técnicas basadas en *machine learning*, han surgido también escépticos en el sector que critican los riesgos y límites del uso de estos nuevos métodos. Algunos expertos no creen que las nuevas tecnologías vayan a ser el futuro del sector financiero, e incluso ponen de manifiesto que, tal vez, el aprendizaje automático y la inteligencia artificial podrían no ser del todo adecuados para este sector.

Esta idea se sustenta sobre el hecho de que en el mundo financiero el cambio y los avances tecnológicos tienden a realizarse lentamente. Bharath Kadaba, Director de Innovación de Intuit, sostiene que esto se debe a que las instituciones quieren protegerse de las posibles consecuencias imprevistas por el rápido avance tecnológico, que pueden llegar a ser catastróficas, afectando tanto a los clientes individualmente como a la economía global [50]. Según Ramneek Gupta, Director de

Citi Ventures, "*en esta industria altamente regulada, que contiene algunos de los datos más sensibles de los consumidores, es primordial salvaguardarse contra las consecuencias imprevistas de la tecnología, que avanza rápidamente*".

Por su parte, Douglas Greenig, Ph.D en matemáticas, que creó el fondo de cobertura Florin Court Capital, dice que "*Cuando la gente habla de aprendizaje automático, un 65 por ciento es una estrategia de marketing y un 35 es sustanciosa. Pero eso 35 por ciento puede ser muy bueno*"

Anthony Ledford, principal científico y coordinador de Man AHL y trabajador de Man Research Laboratory (Oxford), sostiene que "*Los fondos de cobertura, afligidos por ocho años de bajo rendimiento, están aferrándose al aprendizaje automático como una respuesta de alta tecnología para sus problemas. Pero la embriagadora búsqueda de Wall Street por la máquina de dinero perfecta se ha topado con la realidad. La tecnología, que aprende por su cuenta a encontrar nuevas ideas de inversión mediante la búsqueda de datos de gran valor, requiere un gran compromiso de tiempo y dinero, y una alta tolerancia al fracaso, ya que la mayoría de los algoritmos resultan ser defectuosos*".

También David Harding, fundador y CEO de Winton Capital Management, se muestra escéptico ante las expectativas puestas en el aprendizaje automático y la inteligencia artificial aplicados al sector financiero y el *trading*. Recuerda que, en los años 90, hubo un auge similar por el interés en las redes neuronales, dando como resultado el principio de muchas *startups*. "La *gente empezó diciendo: ≪Hay una nueva técnica de computación increíble que va a hacer desaparecer todo lo que había pasado antes≫. Puedo decir que ninguna de esas empresas existe hoy".*

Por tanto, y como en todos los ámbitos que incluyen las tecnologías en su funcionamiento, no solo hay partidarios de la aplicación de estas técnicas en el sector que sostengan que van a suponer un avance importante, sino también detractores que piensan todo lo contrario. Llegado a este punto, lo único que cabe es esperar a ver en qué desemboca esta situación en el sector, qué ritmo y qué tendencia de crecimiento se desarrolla, puesto que una gran cantidad de empresas en todo el mundo están destinando grandes cantidades de recursos y tiempo en la investigación y desarrollo de estos sistemas.

2.2.2.2. FinTech (Financial Technology)

Hoy en día existen diversas vías por las cuales una empresa o individuo puede conseguir financiación, sin ser necesario acudir a una sucursal bancaria tradicional. Esto se debe a que, en ciertas áreas del sector bancario, existes productos y servicios financieros basados en el uso de la web y los datos a los que los clientes no tienen acceso. Esto da lugar a un nuevo entorno competitivo. Los proveedores no bancarios, principalmente impulsados por la tecnología, están entrando en los mercados para ofrecer servicios financieros sencillos. Aquí es donde entran en juego las *FinTech*. Empresas que hacen uso de las tecnologías de la información y comunicación para desarrollar servicios financieros de la forma más eficaz y menos costosa, sustituyendo los canales de distribución tradicionales, la red de oficinas, el contacto personal, los canales telefónicos y electrónicos. Esto permite, por un lado, minimizar los costes operativos una vez que el número de clientes alcanza un tamaño crítico y, por otro lado, satisfacer la demanda de un determinado grupo de clientes; frenando así el avance de la competencia.

Esta situación ha provocado que muchas empresas online creen un ambiente competitivo con el sector financiero, puesto que, aunque gran parte de los clientes se acogen a los patrones de consumo financiero tradicional, en unos años una parte creciente de la sociedad estará en "línea", como resultado del cambio demográfico, la comodidad de realizar gestiones sin moverte de casa, etc. De la misma manera, le exigirán a su entidad que se adapte a estos cambios, ya que las sucursales con localización y horario fijo son, para el usuario, cada vez un inconveniente mayor.

Basándose en el hecho de que cada vez más utilizamos los medios digitales para comparar productos, reservar entradas, pagar productos y realizar otras transacciones bancarias diarias; no es extraño que surjan servicios y aplicaciones financieras inteligentes, centradas en el uso de la web, que nos proporcionan información y servicio las 24 horas. Aquí es donde reside el auge de estas empresas, puesto que es, en este punto, donde el sector financiero tradicional tiene mayores carencias.

Sabiendo esto, las ofertas de las *FinTech* se centran en el trato personalizado de los clientes, y sus actuaciones van desde el pago digital y servicios de información, ahorros y depósitos hasta la banca en línea, servicios de asesoría multicanal, así como, soluciones de financiación y uso de software financiero compatible.

Uno de los rasgos fundamentales del desarrollo de las *FinTech* ha sido el uso de algoritmos de aprendizaje automático, obteniendo así una ventaja sobre la forma tradicional.

Cuantas más técnicas de inteligencia artificial y *machine learning* sean aplicadas al mundo financiero, más evolucionará éste, dejando a un lado la banca tradicional tal

como la conocemos, para dejar paso a la aparición de un sector financiero más tecnológico, barato, rápido, disponible y con una mayor aceptación social.

Las aplicaciones del *machine learning* en el entorno de las FinTech son numerosas. Algunas de las más comunes son:

- Análisis predictivo de la calificación crediticia y préstamos dudosos. Las empresas del área de préstamos están utilizando el *machine learning* para predecir los préstamos incobrables y para construir modelos de riesgo de crédito. Algunas de las empresas que hacen uso del aprendizaje automático con estos fines son:
  - Lending Club. Empresa pionera y con el mayor mercado en préstamos *peer-to-peer*[11], que hace uso de *machine learning* para poder analizar la calidad de un préstamo y poder predecir aquellos que incurrirán en impago.
  - Kabbage. Proporciona fondos directamente a pequeñas empresas y consumidores través una plataforma de préstamos automatizada que emplea *machine learning*. Algunos de sus partners en España son Santander, ING y MasterCard entre otros.

- Toma de decisiones. Un tema importante para la eficacia de las empresas financieras es la toma de decisiones, lo cual puede contribuir en el ahorro de costes y la mayor rapidez. La toma de decisiones se puede mejorar mediante tecnologías que utilicen el *machine learning* como soporte para permitir a los sistemas procesar datos y tomar decisiones más rápidas y eficientes. Algunas de empresas que han incorporado el *machine learning* en este sentido son:
  - BillGuard. Es una empresa de seguridad financiera personal que alerta a los usuarios de posibles estafas, errores de facturación o cargos fraudulentos tras el escaneo de las transacciones con tarjetas de crédito y débito. Emplea algoritmos de *machine learning* para mejorar estos servicios.
  - ZestFinance. Es una de las empresas tecnológicas financieras de más rápido crecimiento. Se dedica al *lending*. Utilizan el *machine learning* y gran cantidad de datos para identificar, con mayor precisión, a los buenos prestatarios, lo que permite tasas de reembolso más altas para los prestamistas y créditos de bajo coste para los consumidores.

[11] Los préstamos *peer-to-peer* o préstamos entre particulares son una forma novedosa de financiarse a invertir sin la intervención de una institución financiera tradicional. Se refiere a préstamos ofertados de particulares a particulares.

- Extracción de información. Otro práctica del *machine learning* en estas empresas, consiste en la extracción de información y contenido de páginas web como artículos, publicaciones, documentos, etc. Las diferentes empresas que lo practican son:
    - Dataminr. Empresa líder en la extracción de información en tiempo real. Transforma, en tiempo real, el flujo de datos de Twitter y otra serie de datos públicos, en señales procesables, proporcionando información útil a clientes financieros, sector público, noticias, etc. Para que esto sea posible, hace uso de algoritmos de aprendizaje automático.
    - AlphaSense. Es un motor de búsqueda especializado en servicios financieros. Para desarrollar sus tareas, aprovecha el procesamiento de lenguaje natural, así como los algoritmos de *machine learning* para proporcionar un servicio más potente y fidedigno.

Además de las aplicaciones de machine learning mencionadas en el apartado anterior, en las *FinTech* también se llevan a cabo funciones de detección de fraude, gestión de información y trading automático de manera muy similar a las entidades financieras tradicionales.

2.3. Conclusión

El interés de la aplicación de estas tecnologías en diversas industrias se debe a la actual capacidad de los sistemas de procesamiento y almacenamiento de datos y el incremento de los datos tanto estructurados como no estructurados, la disponibilidad de nuevas tecnologías y cantidad de datos disponibles (*big data*), así como las necesidades del negocio.

Uno de los detonantes para la popularidad alcanzada por el *machine learning* es el hecho de su capacidad de mejora con el tiempo y a medida que se procesan mayor cantidad de datos. A pesar del éxito de su aplicación en diversas industrias, la banca ha adoptado un enfoque algo más cauteloso. Pese a ello, las crecientes presiones por parte de la nueva competencia (*FinTech*), el aumento de la regulación y las expectativas de los clientes han creado un ambiente perfecto para que el uso del aprendizaje automático se haya propagado a varias áreas del sector financiero, aun así, su uso en los modelos de *scoring* se está un poco atrasado, en parte como consecuencia de imitaciones regulatorias, de gestión y el escepticismo de la plantilla.

Por otra parte, a pesar de su amplio abanico de posibilidades y su gran potencial, hay que tener en cuenta que, como cualquier otro sistema, tiene algunas desventajas y limitaciones. A se vez, hay que tener en cuenta que estos algoritmos, al igual que cualquier otro, dependen en gran parte de los datos, por lo que sin una buena base de datos con información de gran valor los sistemas no serán capaces de obtener un modelo fiable, de manera que hay veces sus limitaciones están condicionadas por otros factores que no tiene que ver con la capacidad de procesamiento, el coste computacional o limitaciones en el propio algoritmo.

Capítulo 3

3. Aplicación sobre un caso real

Para todo el análisis y tratamiento de los datos, así como la creación de los modelos se hará uso de la herramienta R, entorno y lenguaje de programación enfocado al análisis estadístico y de SAS, software de análisis estadístico.

3.1. Credit Scoring

El "*Institute of Internal Auditors (The IIA)*" define el riesgo como la posibilidad de que ocurra un acontecimiento que genere un impacto (severidad) sobre el alcance de los objetivos de una empresa. El riesgo se mide en términos de impacto y probabilidad.

El riesgo puede clasificarse como sistemático o no sistemático [51]. El riesgo sistemático, a veces también denominado riesgo de mercado, se refiere al riesgo inherente a todo el mercado o la economía y no puede evitarse mediante la diversificación. El riesgo no sistemático, también conocido como riesgo específico, riesgo residual o riesgo diversificable, es el riesgo asociado a los activos individuales y, por tanto, se puede evitar o disminuir mediante la diversificación. Estos son los dos componentes en los que se separa el riesgo total de un activo financiero.

Los riesgos en el ámbito del mundo financiero siguen apareciendo en los últimos años, lo cual influye en las pérdidas de las entidades. Entre los riesgos a los que están expuestas las entidades, se encuentra el riesgo crediticio, considerado como el riesgo por excelencia de las entidades financieras, hace referencia a la probabilidad de impago por parte del prestatario y al incumplimiento de las condiciones pactadas en el contrato [52].

Para disminuir el riesgo de crédito, se usa un sistema de calificación de créditos de forma que automatice la toma de decisiones en lo que a conceder o no un determinado crédito se refiere. A este sistema de calificación se le llama *credit scoring*, que consiste en el uso de diferentes técnicas que permitan estimar la

probabilidad de *default*[12] de los solicitantes de un crédito, así como colaborar a diseñar políticas crediticias conforme con el nivel de riesgo que puede asumir la entidad. Para ello el *scoring* se basa en el historial de pago y toma de decisiones de los clientes y así poder predecir el comportamiento futuros de los créditos.

Como define Mark Schreiner[13] en [53]:

> *"Credit scoring se refiere al uso de conocimiento sobre el desempeño y características de préstamos en el pasado para pronosticar el desempeño de préstamos en el futuro"*

Como podemos ver, el scoring de carteras de crédito se trata de un problema de clasificación en el que se diferencian buenos y malos prestatarios.

En un principio, la clasificación de los clientes y la decisión sobre la concesión de un préstamo recaía en un analista de crédito que, basado en la experiencia, decidía si otorgar un crédito o no, estimando el riesgo de caer en default. No solo se tenía en cuenta la información histórica, sino que se intentaban hacía proyecciones de la situación probable en el futuro del prestatario y su capacidad de hacer frente a la amortización del crédito. Estas técnicas tradicionales consistían en realizar una evaluación de los candidatos a recibir el crédito basándose en lo que se conoce en el mundo financiero como las 5 C's:

1) ***Carácter del solicitante de crédito***. Se califica al solicitante dependiendo, en gran medida, de su historial de crédito. Es decir, con base en los antecedentes que generó al administrar créditos anteriores y efectuar pagos a largo plazo. Esta información es suministrada por los prestamistas que le otorgaron un crédito.

2) ***Capacidad de pago***. Los analistas de crédito deben determinar si el solicitante puede administrar sus pagos con comodidad, de manera que no suponga un problema el pago del nuevo préstamo. Tanto su pasado como su historial de empleo son indicadores de la capacidad para hacer frente a la deuda.

3) ***Capital disponible como respaldo***. Aunque se espera que los ingresos sean la principal fuente de pago, también se tiene en cuenta el resto de capital del que dispone el prestatario, ya sean ahorros, inversiones y otros activos que puedan servirle para hacer frente al pago del préstamo.

4) ***Colateral como garantía del préstamo***. Los préstamos que se soliciten pueden estar garantizados o no. El valor del colateral se evaluará y, en el caso de

[12] Término utilizado para hacer referencia a una situación en la que el prestatario no hace frente a las obligaciones legales y contractuales que tiene con sus acreedores.
[13] Consultor en *Microfinance Risk* Management e investigador en el Center for Social Development en Washington University de Saint Louis.

que exista cualquier deuda garantizada por ese colateral, se restará el valor sobre el total del préstamo. El valor neto restante será un factor a tener en cuenta para la decisión final de la concesión o denegación del préstamo.

5) ***Condiciones de la economía en general***. Teniendo en cuenta el propósito del crédito, es posible que los prestamistas quieran saber a qué se va a destinar el dinero. Asimismo, se pueden tener en cuenta las condiciones económicas relativas al ámbito de inversión.

Este procedimiento no era del todo uniforme, dependiendo en gran medida de la subjetividad del analista y con falta de transparencia, debido a que además de la valoración personal y empírica del experto, dependía, en parte, de las reglas impuestas por cada entidad. Esto unido al aumento de demandantes de crédito, hacía que el sistema no fuera del todo sostenible. De tal manera que se pusieron en marcha nuevos modelos automatizados llamados *credit scoring*.

Las técnicas de *credit scoring* son algoritmos que, de manera automática, evalúan el riesgo de crédito de un solicitante, dejando atrás el sistema en que era el analista el que decidía la concesión final o no del crédito. Según Matías Alfredo Gutiérrez [54], el uso de estos modelos de scoring en el mundo financiero comenzó en los 70's pero no fue hasta los 90's cuando, finalmente, se generalizaron, siendo los que están en vigor en la actualidad. Esto fue debido tanto por la evolución de los recursos computacionales y estadísticos, como por la creciente necesidad por parte de sector financiero de hacer más eficaz y eficiente la concesión de financiaciones. Desde la década de los 90 hasta el año 2007 (crisis financiera), el crédito experimentó un incremento significativo, debido en parte al reducido coste de la financiación ya los nuevos sistemas de *credit scoring* que permitían gestionar la demanda y monitorización de la solicitud de créditos con costes reducidos.

A pesar del desarrollo de estos modelos, y aun siendo automáticos, la opinión del analista continúa teniendo una especial relevancia en la concesión de créditos. Puede ser de manera directa, decidiendo si una préstamo se concede o no y en forma de reglas, aplicadas para filtrar las solicitudes y que se introducen en los modelos. Como ya se ha comentado con anterioridad, no es extraño que ambos ejercicios se complementen, de manera que con el modelo se obtenga una primera puntuación y dependiendo de ésta, sea el experto el que realmente haga efectivo la concesión o no, especialmente cuando el *scoring* arroja un resultado no concluyente.

Hoy en día nos encontramos en una situación similar a la que favoreció que los nuevos sistemas de *credit scoring* emergieran; un gran avance tecnológico en lo que a capacidad de procesamiento, almacenamiento y velocidad se refiere.

Junto con estos nuevos avances, también siguen apareciendo riesgos financieros, lo cual influye en las pérdidas de las instituciones bancarias. Por esta razón la previsión del riesgo de crédito es una tarea más importante hoy en día, en especial tras la crisis. Con el fin de disminuir lo máximo posible las tasas de incumplimiento por parte de sus cliente, los gestores de riesgo crediticio deben considerar medios innovadores, para lo que deben sopesar el uso de nuevas tecnologías y algoritmos para mejorar la precisión con la que se emite el crédito, puesto que habrá un momento en el que las medidas tradicionales se quedarán obsoletas.

Uno de estos medios es el uso del *big data* y de herramientas que permitan la construcción de modelos basados en el uso de toda la información de la que se dispone hoy en día, como el *machine learning* o el *deep learning.*

De esta manera, el trabajo se enfocará en el desarrollo de varios modelos de *scoring* automático para poder gestionar el riesgo de crédito sobre una cartera real de una entidad financiera, haciendo uso de diversas técnicas de *machine learning*, pudiendo predecir si un cliente incurrirá en impago en caso de ser concedido el crédito, de manera que se pueda ver cuán eficiente pueden llegar a ser estos nuevos modelos y el uso de las nuevas técnicas, así como las diferencias metodológicas y de resultados entre los distintos algoritmos empleados.

3.2. Metodología de los modelos

3.2.1. Modelado

En el proceso de construcción de un modelo se debe utilizar una estrategia para que el sistema de clasificación sea lo más óptimo y robusto posible. Aunque los argumentos del modelado de este trabajo se vean en detalle más adelante, funcionalmente se comportan de una forma similar a la definida a continuación.

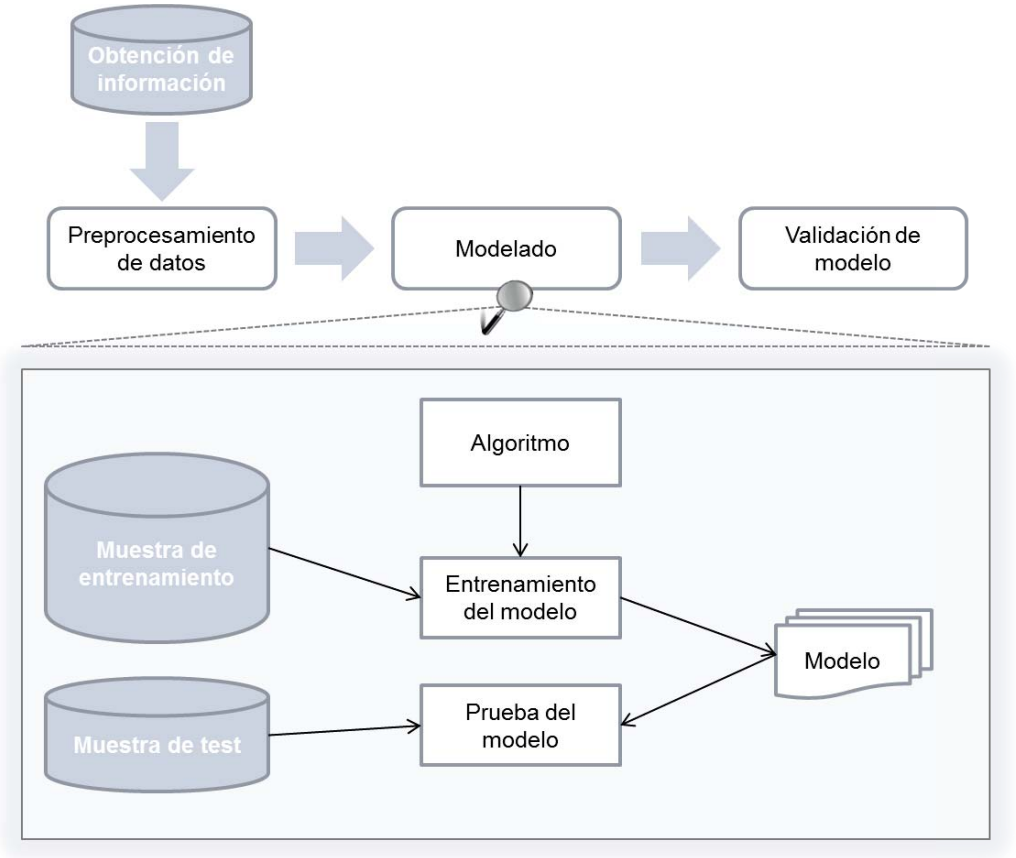

Figura 3.1: Flujo de trabajo modelo machine learning. Fuente: Elaboración propia

Como se puede ver en la Figura 3.1, la primera fase consiste en la búsqueda y análisis de las distintas fuentes de donde se va a obtener la información para poder construir el modelo. Estas fuentes pueden ser tanto internas de las entidades, entre las que están las BBDD internas y cualquier otra información en posesión de la propia entidad, como externas, entre las que se encuentran cualquier base de datos de acceso público o privado (listas de morosos). Además de estas, y cada vez con más presencia en los modelos, está la información extraída de las redes sociales aunque, acorde con el GDPR, los datos deben ser seudonimizados.

Además, la importancia del *data quality* resulta esencial para poder obtener el mayor potencial de la información disponible, puesto que un modelo basado en el análisis de datos se verá afectado si estos datos no son de calidad. Esto tiene efecto directo sobre el entorno actual, en el que la complejidad de las nuevas tecnologías y las grandes cantidades de datos requieren un mayor esfuerzo para garantizar la calidad de los datos. Esto quiere decir que los datos sean capaces de aportar su máximo potencial, asegurando la precisión, integridad, consistencia no duplicidad de los mismos.

La siguiente fase consta del preprocesamiento de datos y la selección de atributos. La mayor parte del trabajo de la preparación los datos para su uso en la construcción de modelos de ML consiste en la obtención de un conjunto de datos consistente y en decidir cuál es la mejor manera de sacar provecho esos datos para resolver el problema. Posteriormente hay que preparar los datos y que estén listos es el preprocesamiento. En este momento se transforman los datos limpios y válidos de la manera que mejor se adapte a las necesidades del modelo [55], asegurando el *data quality*.

Adicionalmente al tratamiento de datos, es necesario hacer una selección de los atributos más relevantes para entrenar el modelo, se realizan a priori del entrenamiento del modelo. A veces ésta selección se realiza manualmente, pero a menudo requieren experiencia. Más frecuentemente, los atributos son seleccionados automáticamente por un conjunto de algoritmos.

Las razones principales por las que se recurre a la selección de atributos son que los datos contienen un gran número de atributos redundantes o irrelevantes [56], por lo tanto pueden ser eliminados sin acarrear una gran pérdida de información pero mejorando el coste computacional, hecho que nos interesa teniendo en cuenta que en la construcción de modelos de ML muchas veces se tratan grandes volúmenes de datos, lo cual lleva asociado un gran coste computacional.

Puede que sea necesario realiza esta fase varias veces y sin ningún orden preestablecido, como consecuencia del modelo que estamos entrenando, puesto que no todos los modelos tienen el mismo nivel de aceptación de *missings*, *outliers* o valores atípicos, etc.

Entre las diferentes estrategias que se pueden seguir para crear un modelo en la fase del desarrollo del modelo, la más generalizada es aquella en la que, partiendo del conjunto de datos inicial, se obtienen dos tablas, una de entrenamiento (*training*) y otra de testeo (*test*). Se consiguen haciendo una partición del *dataset* inicial con una proporción generalmente de $\frac{2}{3}$ y $\frac{1}{3}$ respectivamente, aunque se puede emplear la segmentación que se crea más conveniente, puesto que si se dispone de una base de datos con pocas entradas es recomendable tener una *training set* lo más grande posible para poder entrenar mejor el modelo. Para ello se puede seguir un muestreo aleatorio o seguir una estrategia determinada, aunque con la restricción de que cada registro puede aparecer solo en uno de los *datasets*.

Cuanta más información contenga la base de datos mejor modelo se puede obtener.

El *dataset* de *training* es utilizado para entrenar los modelos de *machine learning*, es decir, para estimar los parámetros del modelo. Una vez tengamos nuestro modelo, el conjunto de datos de testeo se utiliza para comprobar el comportamiento del modelo, ya que si esta operación la lleváramos a cabo con el mismo conjunto con el que

hemos entrenamos el modelo, los resultados no serían correctos sino espurios, ya que es normal que sobre este conjunto tuviera un mejor comportamiento que el que debiera, en parte porque puede que el modelo esté sobreentrenado.

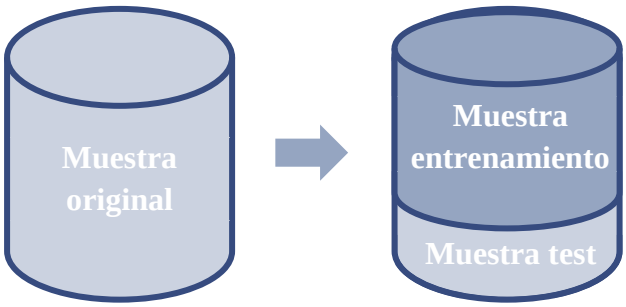

Figura 3.2: División conjunto de datos. Fuente: Elaboración propia

Estas técnicas de clasificación son enfoques sistemáticos para la construcción de modelos de clasificación a partir de un conjunto de datos de entrada (*training*). Cada técnica utiliza un algoritmo de aprendizaje para poder identificar el modelo que mejor ajusta la relación entre el conjunto de atributos y la clase objetivo. El modelo creado a partir de este algoritmo debe ajustarse bien a los datos de entrada y predecir correctamente las clases de los nuevos registros [57].

Por último, hay que llevar a cabo la validación del modelo. Para ello se hará uso de distintas métricas que permitan cuantificar la calidad del modelo y así poder decidir entre los distintos modelos teniendo en cuenta cual tiene mayor poder predictivo. Estas métricas de evaluación del rendimiento del modelo se verán en el apartado 3.2.5.

3.2.2. Sobreentrenamiento

El sobreentrenamiento u *overfitting* se refiere a cuando el modelo tiene un buen rendimiento en los datos de entrenamiento, pero mala generalización a otros datos.

Esto ocurre cuando, en la fase de entrenamiento, el algoritmo intenta aumentar el poder predictivo y como consecuencia puede que el modelo se ajuste excesivamente a la muestra de entrenamiento de forma que al comprobar el poder predictivo sobre los datos de prueba, este caiga.

En la Figura 3.3 podemos ver un ejemplo gráfico, en el que la línea negra sería el modelo que buscamos, mientras que la roja sería el resultado de un modelo sobreentrenado.

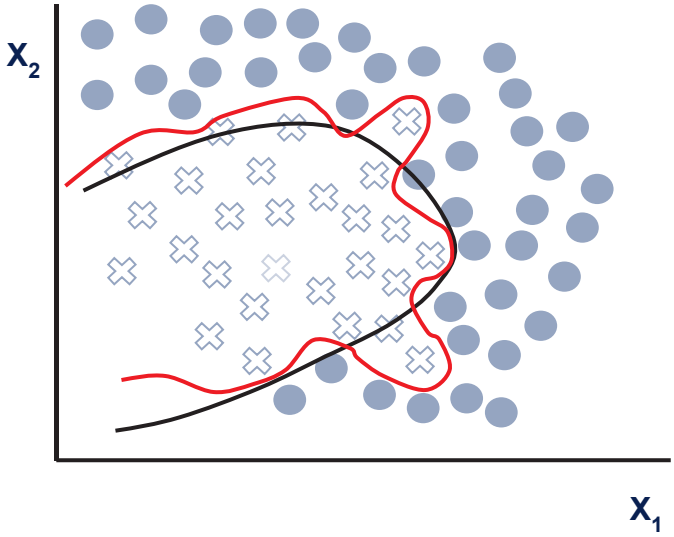

Figura 3.3: Sobreentrenamiento. Fuente: Elaboración propia

3.2.3. Missings values

La presencia de valores de los que no se dispone información (*missing values*) es un problema común cuando tratamos con bases de datos, cuestión que se acrecienta cuanto disponemos de una gran cantidad de datos, como en nuestro caso.

A la hora de poder entrenar un modelo, es importante tener en cuenta estos *missing values*, puesto que ignorarlos o no tratarlos adecuadamente puede tener graves repercusiones sobre los modelos que van desde pérdida de precisión hasta la aparición de sesgos importantes.

La ausencia de ciertos datos se puede deber a diversas razones como los fallos en los instrumentos de medida, en el caso particular del *scoring:* el individuo no aporta una información requerida, se deja en blanco parte de un cuestionario, hay un error a la hora de agregar varias bases de datos, hay cambios en las bases de datos y en los cuestionarios a lo largo del tiempo, etc.

Existen varias soluciones para tratar las bases de datos con *missings*. La opción por defecto es el análisis exclusivo de los casos con información completa en el conjunto de variables. Aunque es cierto que esta solución destaca por su simplicidad, no siempre es adecuada, puesto que se pueden excluir más casos de los necesarios y

perder mucho poder en el análisis estadístico. Otras alternativas para imputar son los datos son las sustitución de los valores *missing* por el valor de la media de la variable, o la moda si es categórica, o existe también la imputación mediante regresión múltiple, mediante la cual se asignan a los *missing* los valores predichos por una ecuación de regresión estimada a partir de los casos con información completa.

En nuestro caso, se utilizarán distintos métodos de imputación de valores a los *missing* dependiendo de las variables y las instancias que sean tratadas.

3.2.4. Outliers

Los *outliers*, valores atípicos o valores extremos son aquellos cuya disposición es diferente a la de los otros valores. Su tratamiento a la hora de construir los modelos resulta esencial, dado que suelen afectar al rendimiento de los modelos.

La mayoría de las estadísticas paramétricas, como la media, la desviación estándar y las correlaciones, así como todas las medidas estadísticas de tendencia central y todas las estadísticas derivadas de estas son altamente sensibles a los *outliers* [58]. Por ello, algunos modelos son más sensibles que otros a los mismos. Por ejemplo, AdaBoost trata los *outliers* como casos especiales y les asigna mucho peso, mientras que los árboles de decisión simplemente toman cada *outlier* como un caso falso [59]. De manera que si el *dataset* contiene una cantidad significativa de *outliers*, es importante utilizar algoritmos de modelado lo menos sensible posible a los *outliers* o filtrar los posibles valores outliers antes de entrenar el modelo.

Sin embargo, las causas de la aparición de valores atípicos en los datos pueden provenir de sucesos fuera de lo común, lo cual debe ser estudiado y analizado por si puede aportarnos alguna información de gran valor y otros se deben a errores en los datos, ya sea por errores a la hora de introducir los datos o por errores técnicos de aparatos mal calibrados, errores de transmisión o fallos a la hora de traspasar la información.

En la Figura 3.4 se muestra el efecto de los outliers sobre un modelo de regresión.

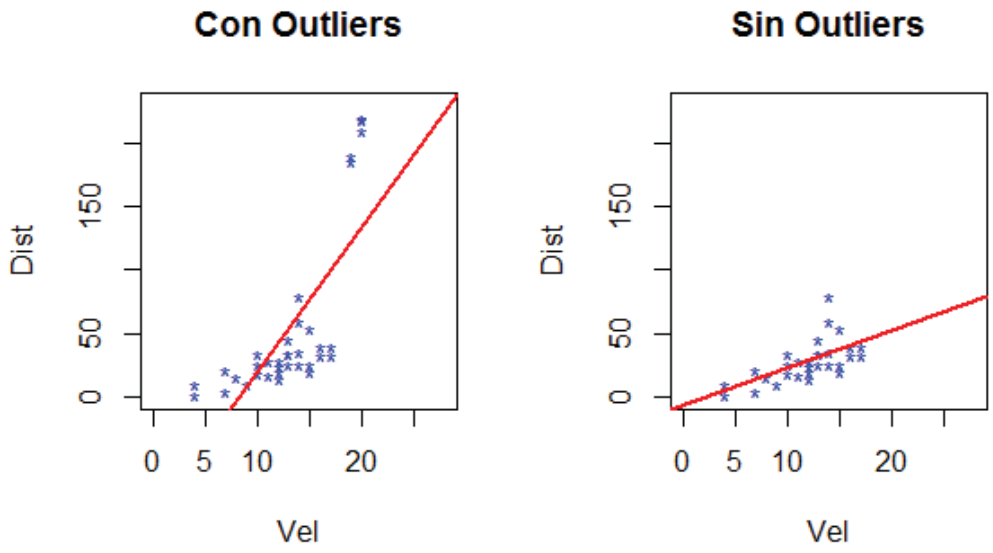

Figura 3.4: Influencia de outliers en un modelo. Fuente: Elaboración propia

En este sentido, los modelos de ML no supondrían diferencia alguna en el coste computacional comparado con las técnicas utilizadas actualmente, puesto que dependiendo del modelo, el tratamiento de éstos debe ser distinto.

3.2.5. Evaluación de modelos

Tras el proceso de modelado resulta primordial el poder estimar la calidad del modelo que hayamos construido.

En nuestro caso (*credit scoring*), el objetivo consiste en clasificar a los clientes como 'buenos' (pagan) o 'malos' (morosos). No obstante, algunos clientes buenos serán clasificados como malos (error de tipo II) y habrá casos en los que clientes morosos serán clasificados como buenos (error de tipo I). De manera que el mejor modelo será aquel que consiga minimizar estos errores a la hora de clasificar los clientes. El análisis tanto del error del tipo I como del tipo II es importante en el caso del *credit scoring*, puesto que pueden suponer en la entidad morosidad en la cartera (al otorgar un crédito a un cliente potencialmente moroso) y pérdida de oportunidad de negocio (al denegar un crédito a un cliente con baja probabilidad de mora).

Para poder estimar cuál de los modelos es mejor, existen varias métricas de evaluación. A continuación se desarrollan los métodos de evaluación que serán utilizados a lo largo del trabajo para calificar los modelos.

Para poder estudiar las diferentes métricas de evaluación es necesario definir primero la matriz de confusión, una forma más gráfica de resumir la información sobre los éxitos y fracasos en la predicción de los datos. Categoriza las predicciones de acuerdo a si coinciden o no con el valor real.

Como se puede ver en la siguiente tabla, cada fila de la matriz representa las instancias de la clase real, mientras que cada columna refleja el número de predicciones de cada clase. De esta forma, la matriz nos permite tener una visión más clara sobre la distribución del error a lo largo de las clases.

		Clase predicha	
		Clase = Sí	Clase = No
Clase real	Clase = Sí	TP	FN
	Clase = No	FP	TN

Tabla 3.1: Matriz de confusión

- *True Positive* (TP): Correctamente clasificado como positivo
- *True Negative* (TN): Correctamente clasificado como negativo
- *False Positive* (FP): Incorrectamente clasificado como positivo
- *False Negative* (FN): Incorrectamente clasificado como negativo

Aunque la tabla representa una matriz de 2x2, esto se debe a que nuestro problema es un problema de clasificación binaria de dos clases, no obstante también hay matrices para modelos que predicen cualquier número de categorías.

En las secciones posteriores se utilizará esta matriz para poder entender de manera más fácilmente cada una de las medidas del rendimiento del modelo, ya que la matriz es la base de la mayor parte de ellas.

3.2.5.1. Accuracy

La precisión o *accuracy* (AC) se define como el cociente entre el número de casos clasificados correctamente entre el total de instancias. Representa el porcentaje de acierto del modelo.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

Esta es la métrica más sencilla y la más utilizada, pero a su vez resulta algo inexacta, ya que no tiene en cuenta la distribución del error entre las clases. Ello conlleva que en una muestra donde el 90% de los casos son de una 'buena' y el resto de la clase 'mala', cualquier modelo que obtenga una *accuracy* alrededor de 0,9 tenderá a ser considerado como un modelo aceptable, cuando en realidad puede que este clasificando todos los ejemplos como buenos. Por lo que en estos casos en los que el valor de la *accuracy* sea menor que el mayor porcentaje de la clase más repetida, se dirá que el modelos no aporta ningún tipo de conocimiento.

Es por esto por lo que la evaluación de un modelo no se puede hacer con respecto a una sola métrica, por lo que el mejor procedimiento radica en hacer un estudio de diferentes métricas.

3.2.5.2. Sensibilidad y especificidad

Encontrar un clasificador útil a menudo implica un equilibrio entre las predicciones que son demasiado conservadoras y aquellas que son excesivamente agresivas. Por ejemplo, un filtro de los mensajes de correo electrónico, podría garantizar el eliminar todos los mensajes de spam eliminando agresivamente casi todos los mensajes 'malos'. No obstante, garantizar que ningún mensaje 'bueno' se filtra por error requiere que el filtro permita que una reducida cantidad de correos spam se cuelen. Un par de medidas que capturan este balance son la sensibilidad y la especificidad.

La sensibilidad de un modelo (S) mide la proporción de los casos positivos que han sido clasificados correctamente.

$$S = \frac{TP}{TP + FN} \qquad S \in [0,1] \tag{3.2}$$

La especificidad del modelo (E) mide la proporción de casos negativos que han sido clasificado correctamente.

$$E = \frac{TN}{TN + FP} \qquad E \in [0,1] \tag{3.3}$$

El valor deseado para ambos es lo más cerca posible del 1, pero es importante el balance entre ambos valores. En nuestro ejemplo, en el que todos los casos se clasifican como buenos, el valor de la sensibilidad sería 1, por su parte el de la especificidad sería 0, por lo que el equilibrio entre ambos es inexistente. De esta manera se podría detectar el problema que no se pudo en un principio solo con el valor de la *accuracy*.

3.2.5.3. Precisión y recall

Otras dos medidas estrechamente relacionadas con la sensibilidad y la especificidad son la precisión y el *recall*. Empleados principalmente en la recuperación de información, estos dos estadísticos indican el interés y la relevancia de los resultados de un modelo, o si las predicciones se ven afectadas por el ruido.

La precisión (P) se define como la proporción de casos positivos que son realmente positivos, dicho de otra manera, cuando un modelo predice un caso positivo, con qué frecuencia lo hace correctamente. De esta manera un modelo preciso será aquel que solo predice casos positivos cuando son muy probables de ser positivos.

$$P = \frac{TP}{TP + FP} \qquad P \in [0,1] \tag{3.4}$$

Por otra parte, el *recall* (R) indica la completitud de los resultados.

$$R = \frac{TP}{TP + FN} \qquad R \in [0,1] \tag{3.5}$$

Como se puede observar, la fórmula el *recall* es la misma que la de la sensibilidad. Sin embargo la interpretación varía ligeramente. Un modelo con un alto *recall* identifica gran parte de los casos positivos, lo que significa que tiene amplitud. En nuestro caso, un modelo con un recall alto identificaría de manera correcta la mayoría de los no defaults.

Existe una medida del rendimiento del modelo que combina la precisión y el recall en un solo valor, se conoce como el *F-measure* o *F-score*.

La *F-measure* combina las anteriores medidas utilizando la media armónica, un tipo de media que se utiliza para las tasas de cambio. Esta media se utiliza en lugar de la media aritmética común puesto que tato a precisión como el recall se expresan como proporciones entre 0 y 1, por lo que pueden interpretarse como tasas.

$$F - measure = \frac{2 \cdot Precisión \cdot Recall}{Precisión + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (3.6)$$

3.2.5.4. Estadístico Kappa

El estadístico Kappa (κ) es un índice muy utilizado que ajusta la *accuracy* teniendo en cuenta la posibilidad de que una predicción sea correcta por medio del azar. Esto resulta especialmente importante en la situación que se ha comentado anteriormente en el apartado de *accuracy*.

El valor puede variar entre 0 y 1, y dependiendo del uso o la finalidad del modelo, la interpretación del valor puede diferir, siendo subjetiva en cada caso, aunque una interpretación más generalizada es la siguiente:

- Mala concordancia ≡ 0.00 – 0.20
- Concordancia razonable ≡ 0.20 – 0.40
- Concordancia moderada ≡ 0.40 – 0.60
- Buena concordancia ≡ 0.60 – 0.80
- Muy buena concordancia ≡ 0.80 – 1.00

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \quad (3.7)$$

P(a) se refiere a la proporción real que concuerda con la realidad y P (b) se refiere a la concordancia esperada entre los valores predichos y la realidad.

Siendo:

$$P(a) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.8)$$

$$P(e) = \frac{(TP + FN) \cdot (TP + FP)}{2 \cdot (TP + TN + FP + FN)} + \frac{(TN + FP) \cdot (TN + FN)}{2 \cdot (TP + TN + FP + FN)} \quad (3.9)$$

3.2.5.5. Curva ROC

El análisis de curvas ROC constituye una de las formas más eficaces y utilizadas en la evaluación del rendimiento de un modelo orientado a la clasificación binaria.

La curva ROC (*Receiver Operation Characteristic*) representa la tasa de verdaderos positivos (Sensibilidad) en función de la tasa de falsos positivos (1 - Especificidad) para diferentes puntos de corte.

Las curvas se definen en una gráfica con la proporción de verdaderos positivos en el eje de ordenadas y la proporción de falsos positivos en el eje de abscisas, ya que, como se ha comentado, estos valores son equivalentes a la Sensibilidad y (1 – Especificidad) respectivamente.

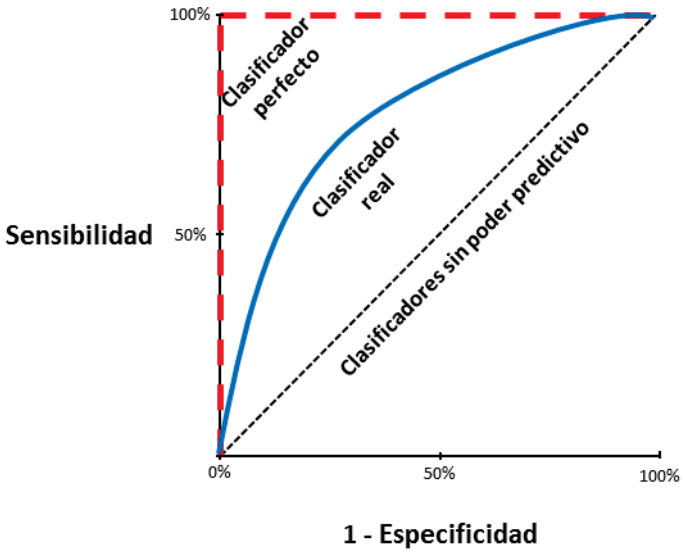

Figura 3.5: Curva ROC. Fuente: Elaboración propia

Para ilustrar este concepto, se contrastan tres clasificadores hipotéticos en el anterior gráfico. La línea diagonal de puntos negros representa un clasificador sin poder predictivo, esto significa que el clasificador detecta verdaderos positivos y falsos positivos con igual tasa, lo que implica que el modelo no puede discriminar entre ambos. Esta es la línea por la que los modelos serán juzgados, ya que sus respectivas curvas están por debajo de ésta, indica que los modelos no son de utilidad. El clasificador perfecto tiene una curva que pasa a través del punto con una tasa de verdaderos positivos del 100% y un 0% de falsos positivos, es decir, es capaz de clasificar de manera correcta todos los casos positivos. La mayor parte de los

modelos aplicados a casos reales tendrán curvas similares a la azul, por encima del clasificador sin poder predictivo y por debajo del clasificador perfecto.

Cuanto más cerca esté la curva del clasificador perfecto, mejor identificará los valores el modelo.

Como se puede observar en la figura XX, una curva ROC es una representación bidimensional del rendimiento del modelo de clasificación. Para poder comparar diferentes modelos, debemos reducir el rendimiento mostrado por la ROC a un único valor escalar que represente el rendimiento esperado. Un método bastante común consiste en calcular el área bajo la curva ROC (AUC) [60], [61].

Debido a que el AUC es una porción del área de la unidad del cuadrado, su valor estará entre [0,1]. Sin embargo, como se ha dicho anteriormente, ningún modelo realista debe estar por debajo de la diagonal del cuadrado unidad y por tanto tener un AUC menor de 0.5.

El AUC posee una propiedad estadística importante [62], el AUC de un modelo es equivalente a la probabilidad de que el modelo clasifique una instancia positiva más alta que una instancia negativa. Por otra parte, el AUC también está estrechamente relacionado con el índice de Gini, que es el doble del área entre la diagonal y la curva ROC. Hand y Till (2001) señalan en [63] que $Gini + 1 = 2 * AUC$.

Una convención para la interpretación de la puntuación del AUC utiliza una escala similar a la siguiente:

- Excepcional ≡ 0.9 – 1.0
- Excelente/ bueno ≡ 0.8 – 0.9
- Aceptable/ razonable ≡ 0.7 – 0.8
- Pobre ≡ 0.6 – 0.7
- Fallido ≡ 0.5 – 0.6

Aunque la mayoría de las escalas son similares, los niveles pueden funcionar mejor en algunos casos que en otros; aun así, la categorización es subjetiva.

3.2.5.6. Cross-Validation

La forma idónea de actuar al crear un modelo reside en seguir una estrategia que, como se ha comentado anteriormente, consiste en entrenar el modelo con un conjunto de datos de entrenamiento y con otro conjunto comprobar la evaluación del mismo. Sin embargo esta situación no siempre es viable, ya sea por falta de datos, imposibilidad de obtener un muestreo adecuado o cualquier otro impedimento. Para ello se suele aplicar una estrategia posterior a la de la creación del modelo

denominado validación cruzada o *cross-validation* (también llamado *k-fold* CV), técnica empleada para evaluar el rendimiento del modelo y garantizar que los resultados son independientes de la partición entre los datos de entrenamiento y prueba.

Este procedimiento consiste en la división del conjunto de datos inicial en k particiones aleatorias llamadas *folds*. Tras esto se utilizan k-1 *folds* para entrenar el modelo y la partición restante se emplea en la evaluación del modelo. Este procedimiento se repite k veces, cambiando cada vez la partición de prueba de la evaluación. Como medida final de la performance del modelo se toma la media de las medidas de las k evaluaciones.

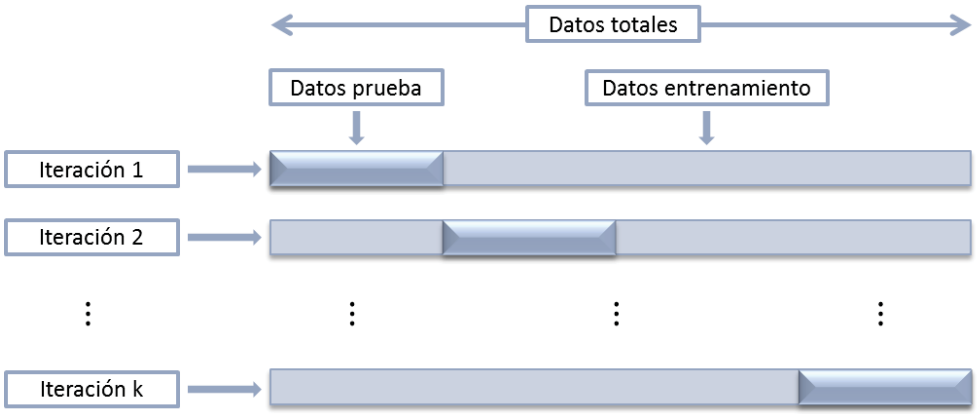

Figura 3.6: Estructura cross-validation. Fuente: Elaboración propia

En la FiguraXX se puede ver gráficamente la división del conjunto total de datos y cómo, en cada iteración, se selecciona una partición nueva como conjunto de prueba, utilizando el resto de la muestra para entrena el modelo.

3.3. El desbalanceo de clases

El desbalanceo de clases se produce cuando una de las clases se encuentra presente en menor mayor medida que otras. Este problema puede surgir en cualquier conjunto de datos y no existe ningún límite que nos indique a partir de qué valor se puede afirmar que una conjunto de datos está desbalanceado.

Esto resulta un gran inconveniente de cara a la resolución de problemas de clasificación puesto que puede suponer una disminución de la eficiencia del

modelo. En particular, ejercerá un buen trabajo sobre la clase mayoritaria en perjuicio de la minoritaria debido a que al detectar una mayor proporción de una clase en particular, el modelo se ajustará a esta.

El desbalanceo de clases es un problema bastante común en los casos en los que se utilizan el histórico de datos crediticios de una entidad bancaria, puesto que por lo general habrá muchos más casos en los que el cliente ha pagado que aquellos en los que no. Como se podrá ver a continuación, los datos utilizados en este trabajo presentan un desbalanceo de clases bastante alto.

Para poder hacer frente a este problema se pueden utilizar técnicas de remuestreo o *resampling* que lo que hacen es, en cierta manera, balancear lo más posible las clases. Algunas de estas técnicas de *resampling* son:

- *Oversampling*: Consiste en el incremento del número de observaciones de la clase minoritaria.
- *Undersampling*: Consiste en la disminución del número de observaciones de la clase mayoritaria.
- Técnicas híbridas: Consisten en la combinación de las técnicas de *oversampling* y *undersampling* con el objetivo de balancear lo máximo posible las clases.

A lo largo de la construcción de los modelos se realizarán distintas técnicas de *resampling* atendiendo a los resultaos iniciales obtenidos, con el objetivo de obtener el mejor clasificador posible.

3.4. Base de datos

3.4.1. Descripción de la base de datos

Para poder realizar el trabajo y aplicar diversas metodologías, se va a hacer uso de una base de datos de una entidad financiera que contiene datos de sus clientes del período comprendido entre enero de 2001 y diciembre de 2007 con un total de 104.961 registros. De ellos, 100.216 devolvieron el crédito y 4.745 operaciones han registrado algún evento de default material a lo largo de la vida de las mismas. Cada una de las instancias se compone de 108 atributos, tanto numéricos como categóricos. Estos atributos proporcionan información relacionada tanto con el crédito pedido (destino del mismo, importe concedido, nº de titulares, nº de avalistas, etc.) como con los titulares de este (antigüedad como cliente, edad, edad hijo menor, estado civil, deudas de los titulares, etc.).

Puesto que la base de datos utilizada en el trabajo contiene información real perteneciente a una entidad financiera, y con el objetivo de mantener la privacidad, las variables y algunos datos expuestos en el trabajo se han anonimizado.

Como se puede observar, la base de datos tiene un problema de desbalanceo de clases que, probablemente, afecte al rendimiento de los modelos.

Algunas de las variables más representativas de la base de datos son:

Variables numéricas:

Antigüedad_cliente: La mayor de las antigüedades como cliente de todos los intervinientes.

Antigüedad_empleo: La mayor de las antigüedades en el empleo de todos los intervinientes.

Edad_hijo_menor: La menor de las edades de los hijos de todos los titulares.

Edad: La mayor de las edades de todos los titulares.

Endeudamiento: Ratio de las deudas totales de todos los intervinientes entre el patrimonio de todos los intervinientes.

Estado_civil: Estado Civil del primer titular.

Fecha: Fecha de apertura del contrato.

Importe_inicial: Importe por el que se concede la operación.

Importe_patrimonio: Suma del importe total del patrimonio, agrupado para todos los titulares y todos los avalistas.

Ingresos: Ingresos anuales de todos los intervinientes de la operación, entendidos como la suma de los Ingresos del trabajo y de los Ingresos de otras rentas de la declaración de bienes.

LTV: El ratio del límite de la operación sobre la tasación de las garantías.

Numero_deudas: Número total de deudas informadas en la declaración de bienes del titular.

Numero_aval: Número de avalistas.

Numero_tit: Número de titulares.

Numero_hijos: Número máximo de hijos de todos los titulares.

Deudas: Suma del importe total de las deudas declaradas en la declaración de bienes del titular.

Tasa_esfuerzo: Ratio de la cuota anual de la operación sobre los ingresos brutos anuales de todos los titulares.

Total_tas: Suma del valor de las últimas tasaciones de los bienes asociados al contrato.

Variables categóricas:

IND_MORA: Indica si la operación ha registrado algún evento de default material a lo largo de la vida (variable que queremos predecir al final del ejercicio).

Identificador: Número que identifica al contrato.

Destino: Código de destino de la financiación del contrato categorizado. El código puede hacer alusión a: Nuevos inmuebles, Refinanciaciones u Otros

Incidencias_en_sistema: Indica si han existido incidencias en el sistema en los últimos 12 meses anteriores a la formalización para alguno de los intervinientes.

Contratato_indefinido: Indica si alguno de los dos primeros titulares tiene contrato indefinido..

Incumplimiento: Indica si han existido incumplimientos medios superiores a 20 días en los 12 meses previos a la formalización para alguno de los titulares.

Nomina_domic: Indica si alguno de los intervinientes tiene un abono por concepto de nómina domiciliada en alguna cuenta a la vista.

Profesión: Profesión del primer titular categorizada. Las varibales están clasificadas en:

- Titulados superiores.
- Funcionarios civiles.
- Jefe administrativo o taller.
- Rentista. Clero o religioso. Otras profesiones. Trabajador no clasificado por cuenta ajena
- Ama de casa. No activo.
- Oficios varios.
- Empresario, director o gerente de empresas
- Resto

Como ya se ha comentado, estas son algunas de las variables más representativas e intuitivas de la base de datos con la que se va a trabajar.

Algunas de las variables contienen información que no es de utilidad o repetida, por lo que antes de utilizar las variables para la construcción del modelo se ha realizado un estudio detallado de cada una de ellas. En el siguiente apartado se muestran algunos de estos estudios y las decisiones tomadas respecto a *missings*, *outliers*, variables repetidas y correlacionadas, etc.

A lo largo del trabajo los términos '*features'*, 'atributos' y 'variables' se han utilizado de forma indistinta.

3.4.2. Preprocesamiento de datos

Cuando hablamos de modelizar haciendo uso de *big data* se asume que si se utilizan todos los atributos o incluso si, además, se incluyen otros nuevos, se obtiene el mejor modelo posible puesto que no se está perdiendo información y cuanta más información disponible mejor será nuestro modelo. No obstante, esta asunción está equivocada, ya que, como se ha explicado en el apartado XX, la calidad de los datos (*data quality*) es de gran trascendencia. En este caso, puede ocurrir que haya variables correlacionadas que terminen introduciendo ruido en el modelo en vez de aportar información de valor.

Para ello se ha realizado un análisis de correlación la correlación entre las variables a partir del coeficiente de Pearson que determina el grado de relación lineal entre las variables. De esta manera, las variables con una alta correlación han sido eliminadas. En el apéndice X se muestra la matriz de correlaciones entre las variables.

Al realizar el estudio de la correlación, se ha descubierto que hay un par de variables cuyo valor es estático y no varía («*SelectionProb»* y «*SamplingWeight»*), de manera que no aporta ningún valor añadido al modelo, por lo que ha sido eliminadas.

También hay variables como «*Identificador»* que solo introducen ruido y mayor carga computacional al modelo, ya que el número que identifica a una operación no aporta ninguna información adicional relevante sobre el problema de clasificación ante el que se está, por lo que la variable se ha eliminado de a base de datos.

Por otro lado, como criterio de tratamiento de *missings* se ha realizado, por una parte, de manera similar a [64], eliminando las variables con más de un 20% de *missings* (17 en total) y, por otra, en aquellas variables con un porcentaje de *missings* inferior al 20%, se han sustituido, dependiendo de la variable, por la media, mediana, moda o incluso se ha llegado a eliminar la instancia si ésta tenía gran número de atributos sin información, de manera que tras su tratamiento resultaría ser una variable distinta,

por lo que se ha decidido eliminar para poder mantener la mayor veracidad posible de los datos.

Por ejemplo, los *missings* de la variable *«Endeudamiento»* se han sustituido por la media de los valores de todos los endeudamientos. De la misma manera se ha hecho con las variables *«Antigüedad_empleo »*, *«Tasa_esfuerzo »* y *«LTV».*

También se ha realizado un tratamiento de *outliers* haciendo un estudio de los valores atípicos de cada variable mediante los diagramas de caja. La siguiente figura ilustra el diagrama de caja perteneciente la variable *«Edad».* Como se puede observar, no tiene sentido que una persona con menos de diez años pida un crédito, por lo que se puede asumir que esos datos outliers y eliminarlos. No obstante, como se ha comentado en el apartado específico sobre los outliers, no se puede suponer que todos los valores por encima del límite superior sean incorrectos, por lo que se ha tomado como medida eliminar las instancias con un valor de la edad mayor de 85.

Otras variables han sido discretizadas atendiendo a diversos criterios personales con el fin de poder englobar la información de algunas variables en otras con menor desviación típica.

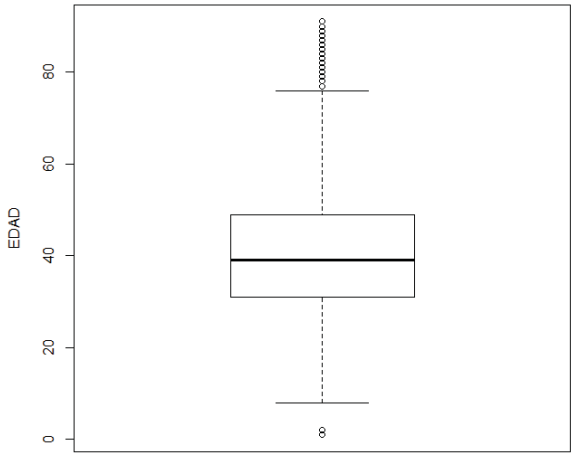

Figura 3.7: Diagrama de caja y valores atípicos. Fuente: Elaboración propia

Para la construcción del modelo, se ha realizado una partición de la base de datos tratada mediante un muestreo aleatorio simple con 70% de los datos para entrenamiento y un 30% para testeo.

Finalmente, tras todo el procesamiento y filtrado anteriormente descrito, la base de datos resultante tiene 56 variables, habiéndose reducido así prácticamente hasta la mitad.

Adicionalmente al tratamiento de la base de datos llevado a cabo, en algunos casos, previamente a la construcción del modelo, se ha vuelto a efectuar esta fase para poder realizar un tratamiento más exhaustivo de ciertos temas (*outliers*, *missings*,…) dependiendo de la permisividad de los modelos frente éstos. Este tratamiento específico será indicado con más detalle en los apartados pertenecientes a cada modelo.

3.5. Elección de modelos

Una parte fundamental del *machine learning* trata de diseñar diferentes modelos y algoritmos para poder adaptarlos de manera que se obtenga el que mejor se ajusta al problema. Para poder elegir de manera empírica el mejor modelo, podemos utilizar varios métodos. Sin embargo, no hay un modelo en particular que de forma general sea el que mejor se ajusta a todo tipo de problemas (también conocido como *no free lunch theorem* [65]). Esto se debe a que un conjunto de suposiciones que funcionan bien en un dominio puede no ser efectivo en otro [66].

Consecuentemente, para poder obtener el modelo más robusto, es necesario desarrollar diferentes modelos para poder tratar la gran variedad de datos que se producen en el mundo real. Para cada modelo puede haber varios algoritmos que podemos utilizar para entrenar el mismo y ajustar la velocidad-exactitud-complejidad del mismo.

No obstante, para poder escoger el modelo más preciso entre un grupo, primero hay que seleccionar los modelos que se van a entrenar, puesto que existen cientos de algoritmos y no es factible utilizar todos, por lo que tenemos que seleccionar los que mejor puedan ajustarse a nuestro problema. Para ello lo primero es categorizar el problema. Este es un proceso que consta de dos partes con el que identificamos el tipo de problema al que estamos haciendo frente. Por un lado está la categorización por entrada, por la que se decide si estamos ante un problema de aprendizaje supervisado o no supervisado basándonos en si los datos están etiquetados o no. Por otra parte está la categorización por salida, dependiendo de si la salida del modelo es un número o una clase estaremos ante un problema de regresión o clasificación.

Una vez tenemos claro al problema al que nos enfrentamos, el segundo paso es seleccionar los algoritmos más adecuados para resolverlo. Por ejemplo, si estamos ante un problema de clasificación, hay que tener en cuenta si se trata de un problema

de clasificación binaria, que tendrá algoritmos que funcionen mejor (Regresión logística, SVM) que otros.

La siguiente fase consiste en el entrenamiento de los modelos más adecuados y evaluarlos de cara a seleccionar el mejor. A la hora de escoger un modelo, no solo debemos tener en cuenta las métricas de evaluación anteriormente explicadas, sino también otros aspectos como el coste computacional, la interpretabilidad, el comportamiento frente a los *missings* y *outliers*, etc.

Todo este proceso supone una gran diferencia frente al modelado tradicional del *credit scoring*, en el cuál no hay que decidir entre diversos modelos, ahorrándose todo el proceso de selección del modelo más adecuado, lo cual conlleva un gran coste computacional por trabajar con grandes volúmenes de datos, debido a que, de forma general, se utiliza la regresión logística para entrenar el modelo, pudiendo invertir más tiempo y recursos al ajuste de éste. No obstante los modelos de credit scoring tienen un mayor poder predictivo, por lo que puede que sacrificar algo más de tiempo y recursos a la larga resulte mejor.

Siguiendo con esta estrategia, en el trabajo se van a desarrollar distintos modelos para poder decidir cuál de ellos tiene mejores resultados sobre una cartera de riesgo de crédito.

3.6. Aplicación de algoritmos

En esta sección se presentan los distintos algoritmos que se van a utilizar para la aplicación de las técnicas de ML en la gestión de riesgo de crédito.

3.6.1. Regresión Logística

3.6.1.1. Introducción regresión logística

Como se ha visto en el estado del arte y se ha comentado a lo largo el trabajo, históricamente, el modelo de regresión logística ha sido el método estadístico más empleado a lao hora de determinar la probabilidad de que un cliente entre en default en los modelos de *credit scoring*. Este método trata de explicar la probabilidad de que se devuelva el crédito o no en función de una serie de variables explicativas.

El valor principal de éste método reside en la buena interpretabilidad y la fácil explicación [67]. Otra ventaja reside en la modelización de probabilidades y en el hecho de que los modelos de regresión logística son menos sensibles a los *outliers*.

La variable dependiente presenta dos categorías que representan la ocurrencia y la no ocurrencia del acontecimiento definido por la variable, en este caso el pago o no de la deuda, codificándose con los valores uno y cero respectivamente. En lo que se refiere a las variables explicativas, pueden ser tanto cuantitativas como categóricas sin límite de categorías.

Como se ha comentado en el primer párrafo, el modelo de regresión logística expresa la variable dependiente como la ocurrencia o no de un acontecimiento en términos de probabilidad, haciendo uso de la función logística para estimar la probabilidad de que ocurra el acontecimiento mediante la formulación:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \quad \textbf{(3.10)}$$

siendo π_i la probabilidad de pertenecer a la clase buena y x_i las variables explicativas.

Puesto que el modelo anterior no es lineal respecto a las variables independientes, se considera la inversa de la función logística, a lo que se llama logit, definiéndose como en cociente entre la probabilidad de que ocurra un acontecimiento y la probabilidad de que no ocurra, que es su complementaria, como puede observarse a continuación:

$$g(x) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad \textbf{(3.11)}$$

Esta formulación facilita la interpretación del modelo y de sus coeficientes, que reflejan el cambio en el logit correspondiente a un cambio unitario en la variable independiente.

La probabilidad π_i obtenida por la ecuación (3.11) es el límite de la clasificación. El cliente es considerado como propenso a hacer default si es mayor es 0.5 o no propenso al contrario. No obstante, según Lin [68], no es del todo apropiado tomar 0.5 como punto de corte cuando existe un desbalanceo de clases.

En un *credit scoring*, se puede asociar $\beta_i x_i$ a la calidad crediticia del individuo. La calidad crediticia del individuo se supone como el resultado de una función lineal en sus parámetros y X contiene la información específica de los deudores. Es importante mencionar que las estimaciones β_i no tiene una interpretación directa, ya que solo representan el efecto que un cambio en x_i tiene sobre el resultado final, a la vez que su signo muestra si la relación con la probabilidad de incumplimiento es directa o inversa.

En la siguiente tabla se describen algunas ventajas y desventajas a la hora de construir una regresión logística.

REGRESIÓN LOGÍSTICA
Ventajas
• Es un modelo de fácil interpretación. • Tiene un coste computacional bajo. • Permite el uso de múltiples variables con relativamente pocos casos. • No se ve afectado por outliers o datos que no aportan valor frente a un conjunto.
Desventajas
• Mayoritariamente utilizada para modelos binomiales, por lo que si queremos clasificar en más de dos clases, habría que hacer lo dos en dos. • Bajo soporte a variables muy correlacionadas.

Tabla 3.2: Ventajas y desventajas de los modelos de regresión logística

3.6.1.2. Modelo regresión logística

Como se destacó anteriormente, la regresión logística es el algoritmo más empleado en la construcción de modelos de credit scoring, de manera que se ha implementado en este trabajo para de poder comparar sus resultados con los del resto de algoritmos.

La tabla XX presenta los resultados de las predicciones del modelo.

Muestra	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	0,656	0,941	0,371	0,24	0,809
Test	0,651	0,941	0,36	0,232	0,799

Tabla 3.3: Resultados regresión logística

Los resultados sobre la muestra de entrenamiento son claramente mejores, esto se debe a que el modelo se ha entrenado con estos datos y, a la hora de predecir, es más probable que el modelo se ajuste mejor a los datos con los que se han entrenado. Por esta razón la muestra inicial se divide en dos, una de entrenamiento para construir el modelo y otra de test sobre la cual evaluar el mismo.

A primera vista, parece que los resultados del modelo (muestra de test) son satisfactorios, puesto que se obtiene una AUC bastante alta (cerca de 0,8) y una

precisión (*accuracy*) de un 0,91. No obstante la especificidad es algo baja, lo que quiere decir que los casos en los que el cliente no paga no son predichos correctamente.

Por ello, se ha decidido introducir en la tabla la precisión balanceada (*Balanced Accuracy*) en lugar de la *Accuracy*, puesto que al predecir un gran porcentaje de las entradas como buenos clientes se obtendrá una *accuracy* muy alta, pudiendo llevar a equivocación. La *Balanced Accuracy* se obtiene sumando la sensibilidad y la especificidad y dividiéndolo por 2, de forma que represente de una manera más real la precisión del modelo.

$$BA = \frac{Sensibilidad + Especificidad}{2} \qquad \textbf{(3.12)}$$

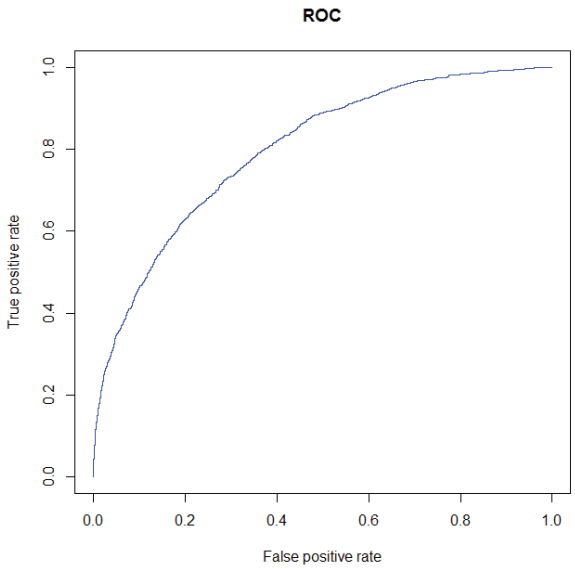

Figura 3.8: Curva ROC regresión logística

La Figura XX representa la curva ROC asociada al AUC de la regresión logística. Como se puede ver, y comparando con la curva de la figura XX, se encuentra entre un clasificador perfecto y uno sin poder predictivo, lo cual indica que el modelo está dentro de los límites permitidos en la evaluación del rendimiento.

La regesión logística toma como infromación más relevante para poder predecir el comportamietno de los clientes el fin al que se destinará el préstamo, la fecha de concesión del préstamo, si alguno de los titulares ha impagado durante más de diez días en el último año y la profesión, dándole mayor importacia a la profesión de

‘oficios varios’ (se puede presuponer que la importancia es negativa) y los empresarios, directores o gerentes de empresas.

Por otra parte, el modelo te permite obtener el peso relativo de todas la variables y, aun habiendo hecho correctamente el filtrado de datos, el análisis de la corelación y todos los requisitos para asegurar el data quality, sigue habiendo variables con un peso mínimo y, como es de suponer, si eliminamos estas variables, los resultados del modelo son similares. No obstante, este útimo filtro no merece la pena puesto que la regesión logística es un modelo que se entrena bastante rápido, por lo que realmente el tiempo que se emplea en realizar este segundo análisis y filtro es mayor que el de entrenar de nuevo el modelo.

3.6.2. Métodos ensemble

Dentro de las técnicas de *machine learning* se pueden distinguir entre los métodos individuales, donde interviene un único modelo y métodos *ensemble*, que entrenan varios modelos para resolver el mismo problema. En contraste con los enfoques de aprendizaje ordinario que tratan de construir un solo modelo, los métodos *ensemble* intentan construir un conjunto de modelos y combinarlos de manera que tengan mayor precisión. Esos modelos utilizados para su posterior combinación se denominan modelos base o débiles.

Los métodos *ensemble* son atractivos porque son capaces de entrenar modelos débiles que, unidos actúan, por lo general, mejor que los modelos individuales con un gran poder predictivo.

La figura XX muestra la arquitectura de estos modelos.

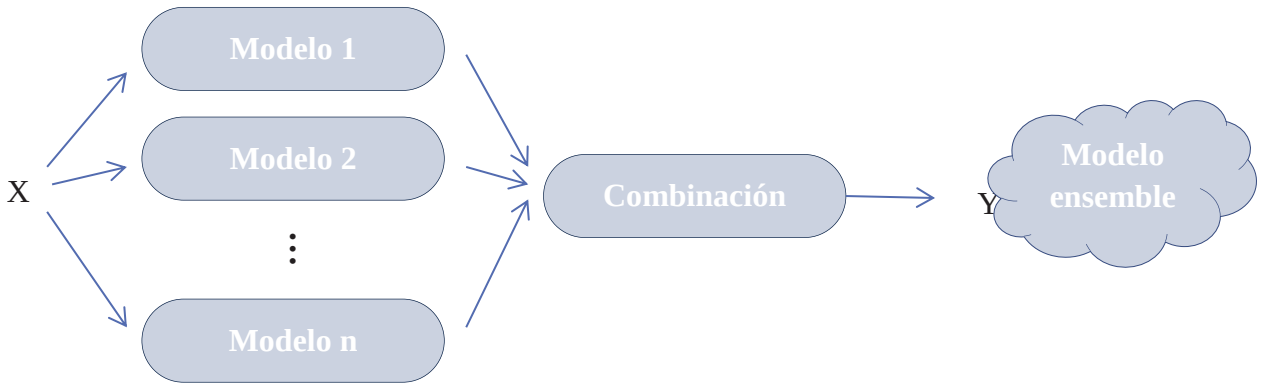

Figura 3.9: Arquitectura métodos ensemble

Dentro de los métodos *ensemble* existen dos técnicas, el *boosting*, cuyo máximo exponente es el *AdaBoost*, y el *bagging*, con el *random forest* como máximo representante.

Ambos métodos tiene similitudes, como que ambos combinan clasificadores básicos para obtener uno más robusto, pero difieren en la forma de obtener los clasificadores, ya que en el *bagging* se realiza un entrenamiento en paralelo y en *boosting* lleva a cabo un entrenamiento secuencial de los clasificadores.

3.6.2.1. Bagging (Random Forest)

3.6.2.1.1. Introducción *random forest*

El *bagging* es uno de los primeros métodos *ensemble*. Utiliza una técnica denominada *bootstrap aggregating*, de ahí su nombre de *bagging*. Según lo describió Leo Breiman en 1994, el *bagging* genera una serie de conjuntos de datos que son utilizado para generar un conjunto de modelos por medio de clasificadores individuales. Las predicciones de los modelos se combinan por medio de una votación o por la media aritmética.

Aunque es una técnica relativamente simple, puede funcionar bastante bien si se utiliza con modelos inestables. Esto son aquellos que tienden a cambiar cuando los datos de entada cambian ligeramente. El uso de modelos inestables resulta esencial para asegurar la diversidad del conjunto a pesar de haber pequeñas variaciones entre los conjuntos de entrenamiento [69]. Por ello, se utilizan a menudo los árboles de decisión, que tienden a variar dado un cambio menor en los datos de entrenamiento.

El *random forest* es uno de los modelos más representativos del *bagging*. Es un algoritmo desarrollado por el propio Leo Breiman [70] combinando su propia idea de *bagging* e inspirándose en el anterior trabajo de Amit y German [71] sobre la selección aleatoria de atributos.

El método de *random forest* se basa en un conjunto de árboles de decisión como clasificadores.

El algoritmo del *radom forest consta de los* siguientes pasos:

- Se crean aleatoriamente n conjuntos de la muestra de entrenamiento con remplazamiento del mismo tamaño que la original. Estos subconjuntos serán los conjuntos de entrenamiento para cada árbol. Al seleccionarse de esta forma, no todos los datos de la muestra original estarán necesariamente en los subconjuntos de entrenamiento, estos datos se denominan *out of bag*.

- Para crear cada nodo del árbol se seleccionan aleatoriamente una cantidad de variables de entrada. Normalmente se un número de variables iguala la raíz cuadrada de los atributos del conjunto original. Este número es contante a lo largo del entrenamiento de todo el bosque, no obstante, en cada nodo se seleccionan aleatoriamente otras variables nuevas de entre todas las variables explicativas.
- Se construye cada árbol con la máxima extensión posible.
- Si estamos ante un problema de clasificación, el bosque obtendrá como resultado la votación mayoritaria de los árboles, esto es que se queda con el resultado que se haya elegido más veces. Si se trata de un problema de regresión, el resultado final será la media aritmética de los resultados de los árboles.

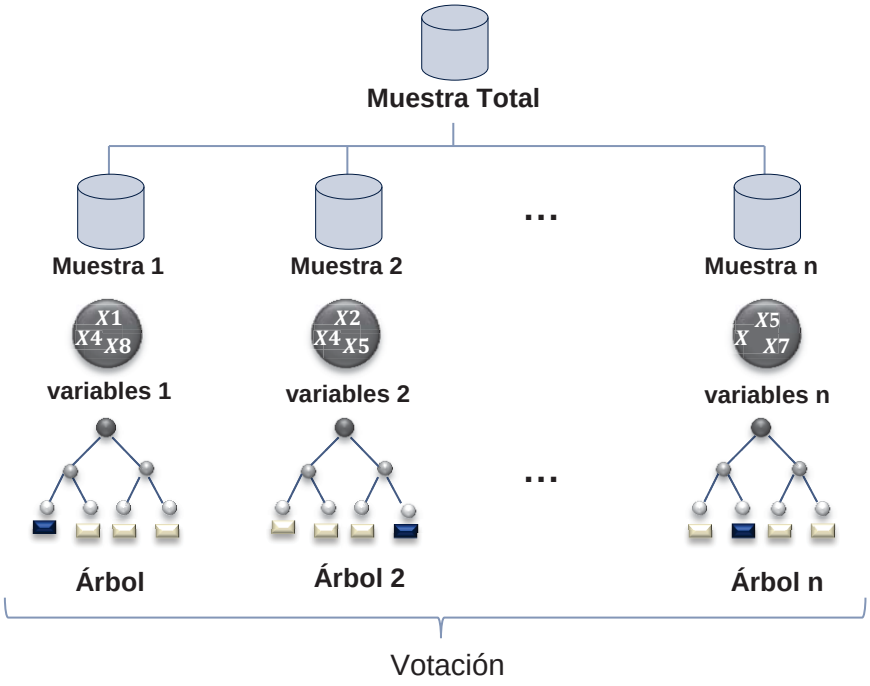

Figura 3.10: Flujo del modelo *random forest.* Fuente: Elaboración propia

A la hora de construir estos modelos hay que tener en cuenta las ventajas y desventajas para poder realizar un mejor tratamiento de los datos y ajustar los parámetros con el fin de obtener el clasificador más preciso.

RANDOM FOREST
Ventajas
• Sirve para realizar cualquier tipo de problema (clasificación y regresión). • Soportan un nivel elevado de *missings*, valores atípicos sin que la predicción se vea afectada. • Acepta variables tanto discretas como continuas. • Selecciona las variables más importantes, además de devolver una estimación de la importancia relativa de cada variable en la clasificación. • Pueden ser utilizados sobre un gran número de datos. Para una muestra lo suficientemente grande produce un clasificador muy certero. • Son capaces de soportar un gran número de variables sin excluir ninguna. • No se sobreentrenan con el incremento de árboles. • Las variables explicativas categóricas con un gran número de categorías pueden llegar a generar un sesgo sobre la importancia de las variables.
Desventajas
• Al contrario que los árboles de decisión, los RF son modelos de difícil interpretación. • Puede requerir un cierto procesado de los datos para ajustar el modelo lo máximo posible. • Los modelos de *random forest* tienden a sobreajustarse sobre ciertos problemas de clasificación con un elevado ruido. • Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento del modelo, los grupos más pequeños salen más favorecidos.

Tabla 3.4: Ventajas y desventajas de los modelos de *random forest*

3.6.2.1.2. Modelo random forest

Si bien los modelos de *random forest* poseen un alto soporte sobre los *missings* y *outliers*, se ha realizado una revisión de la muestra de entrenamiento con el fin de eliminar el ruido de la misma, puesto que el modelo se podría ver afectado sobreentrenándose. Además, para poder probar el modelo sobre la muestra de test, es necesario eliminar los *missings* de esta última, puesto que los modelos de *random forest* creados no son capaces de evaluar muestras con *missings*.

Adicionalmente, para poder crear este modelo, hay que tener en cuenta un par de detalles. Hay dos parámetros que intervienen en la construcción de un modelo de

random forest, el número de variables empleadas para cada árbol y el número de árboles totales. De esta forma, en la tabla XX se recogen los resultados asociados a distintas pruebas realizadas con el fin de obtener el mejor valor de cada parámetro.

Muestra	Nº variables	Nº árboles	AUC
Entrenamiento	2	20	0,997
		50	0,998
		100	0,999
		500	0,999
	5	20	0,999
		50	0,999
		100	0,999
		500	0,999
	7	20	0,999
		50	0,999
		100	0,999
		500	0,999
	10	20	0,999
		50	0,999
		100	0,999
		500	0,999
Test	2	20	0,77
		50	0,801
		100	0,807
		500	0,813
	5	20	0,77
		50	0,792
		100	0,8
		500	0,807
	7	20	0,767
		50	0,7884
		100	0,797
		500	0,805
	10	20	0,769
		50	0,794
		100	0,791
		500	0,802

Tabla 3.5: Análisis parámetros random forest

Según los resultados, se pueden observar que la precisión del clasificador aumenta a medida el número de árboles crece. Cabe destacar que la construcción de los árboles se realiza en paralelo, por lo que no se ve altamente afectado el coste computacional.

Como se puede ver, los mejores resultados se obtienen al entrenar el modelo con 2 variables y 500 árboles. Para poder hacer un mejor estudio el modelo, los resultados de las métricas de evaluación del *random forest* con estos parámetros están en la tabla XX.

Muestra	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	0,990	0,99	1	0,99	0,99
Test	0,528	0,997	0,056	0,1	0,813

Tabla 3.6: Resultados del random forest

Se puede apreciar que este modelo arroja buenos resultados sobre la muestra de entrenamiento, lo que puede ser debido a que el modelo esté sobreentrenado. Si bien el rendimiento se reduce significativamente en la muestra de test, sigue arrojando valores bastante aceptables, con una AUC superior a 0,8.

Sin embargo, la especificidad es demasiado baja, incluso comparada con la del modelo de la regresión logística. Realizando técnicas de *resampling*, el mejor resultado obtenido es el que se muestra en la tablaXX. Aunque aumente la especificidad, la AUC disminuye notablemente, al igual que la sensibilidad, por lo que aun haciendo uso de estas técnicas, el modelo no es del todo satisfactorio.

Muestra	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Test	0,561	0,766	0,356	0,039	0,711

Tabla 3.7: Resultados del random forest tras técnicas de resampling

La Figura 3.11 muestra la ROC del modelo que, como se aprecia a simple vista, es mayor que la de la regresión logística.

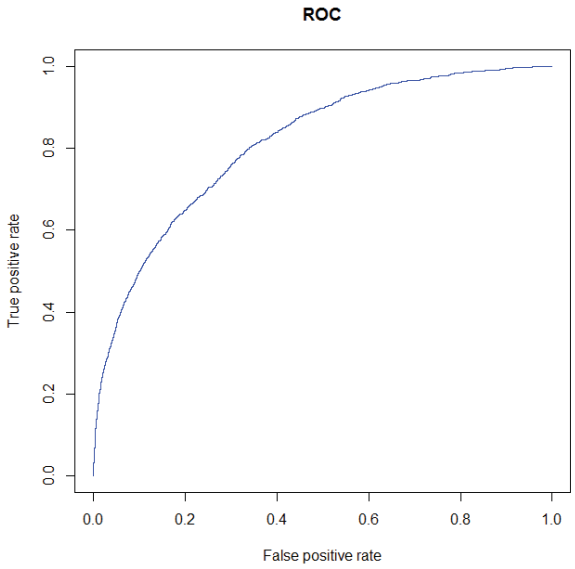

Figura 3.11: Curva ROC random forest

El modelo devuelve una estimación de la importancia relativa de cada variable en la clasificación. Las variables con mayor importancia son aquellas que aparecen por las ramas superiores de los árboles. Esto no sólo puede ser utilizado para reentrenar el modelo con las variables más importantes ganando en coste computacional, sino que también puede servir de apoyo a otros modelos a la hora de realizar una selección de variables.

3.6.2.2. *Boosting (AdaBoost)*

3.6.2.2.1. Introducción adaboost

El *boosting* es un multiclasificador similar al *bagging* que surgió en el año 1989 de la mano de Robert Schapire. Sin embargo, lo que lo hace diferente es que el entrenamiento de los clasificadores base se lleva a cabo de manera secuencial, donde cada clasificador sucesivo se construye teniendo en cuenta los residuos de la predicción del clasificador anterior, de manera que cada modelo simple generado se centra en predecir los fallos de los modelos anteriores.

El *AdaBoost* (*Adaptative Boosting*) es el algoritmo de *boosting* más utilizado, ya que se centra en aquellos casos que son más difíciles de clasificar. Es un modelo que entrena secuencialmente regresiones logísticas sobre el mismo conjunto de datos, dando como resultado la composición de todos los modelos secuenciales anteriores.

El procedimiento de este algoritmo para construir el modelo final es el siguiente:

- Inicialmente a todos los datos de la muestra de entrenamiento con la que se va a estimar el primer clasificador se les asigna un peso idéntico, $w_i = \frac{1}{n}$, con $i = 1,2 \ldots n$, donde n es el tamaño de la muestra..
- Se entrena el primer clasificador utilizando la muestra de entrenamiento y se calcula el error cometido por el modelo (entradas mal clasificadas) y la ponderación α_1 asociada al modelo $g_1(x)$. Se incrementan los pesos en los casos de entrenamiento en los que el modelo calcula erróneamente.
- Se entrena un nuevo modelo usando el conjunto de pesos modificados.
- Se repite iteradamente este proceso hasta que se alcance el número de iteraciones seleccionadas en un principio o la función del error sea baja (criterio de convergencia).
- El modelo final será un promedio ponderado de los distintos resultados obtenidos.

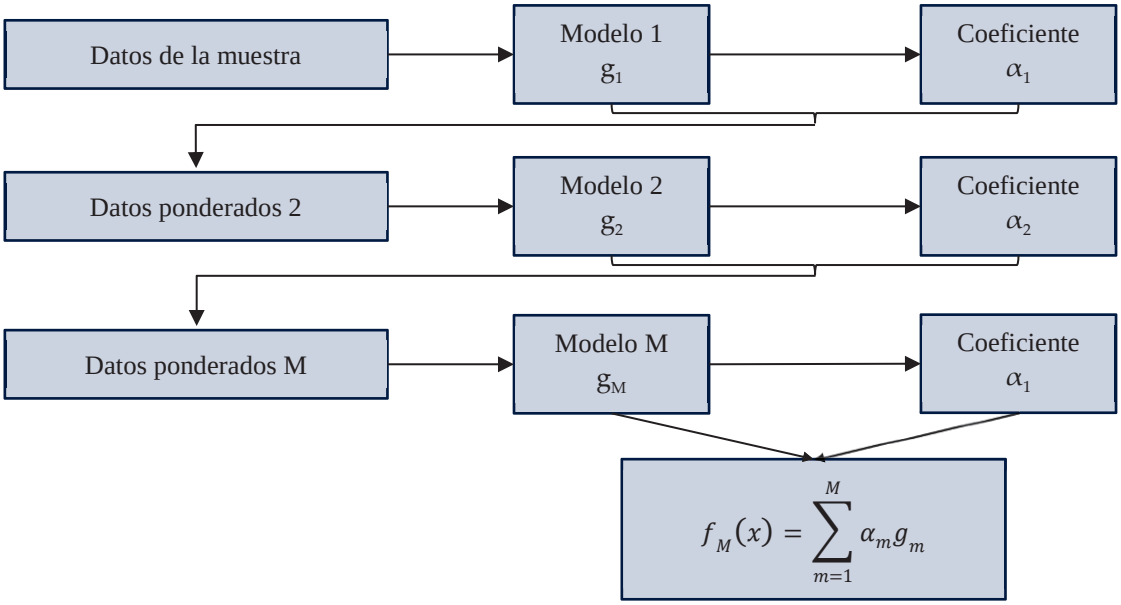

Figura 3.12: Flujo del modelo adaboost. Fuente: Elaboración propia

A continuación se muestra una tabla con las principales ventajas y desventajas a tener en cuenta a la hora de crear estos modelos.

ADABOOST
Ventajas
• Es un modelo relativamente fácil de interpretar comparado con otros. • Posee pocos parámetros para ajustar a la hora de la construcción del modelo, la cantidad de iteraciones. • El coste computacional asociado a su construcción es bajo.
Desventajas
• Es sensible al ruido y la presencia de *outliers* en la muestra de entrenamiento debido a la actuación de los pesos. • Pueden llegar a sobreentrenarse si el número de iteraciones es muy elevado.

Tabla 3.8: Ventajas y desventajas de los modelos *adaboost*

3.6.2.2.2. Modelo *adaboost*

Ya que uno de los pocos parámetros a ajustar en la fase de construcción de estos modelos es el número de iteraciones, y a priori no sabemos cuál es el valor óptimo, en la Tabla 3.9 se presentan diversos modelos con distinto número de iteraciones.

Muestra	Iteraciones	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	20	0,563	0,993	0,132	0,189	0,836
	50	0,573	0,992	0,154	0,214	0,855
	100	0,581	0,991	0,168	0,226	0,882
Test	20	0,551	0,993	0,11	0,158	0,802
	50	0,556	0,993	0,12	0,176	0,806
	100	0,567	0,989	0,146	0,191	0'808

Tabla 3.9: Análisis de los parámetros del *adaboost*

Como cabía esperar, a medida que aumentan las iteraciones aumenta la precisión del modelo, ya que en cada iteración el modelo se ajusta más a los datos que previamente se habían clasificado erróneamente. No obstante, también aumenta el tiempo de ejecución, puesto que en este caso, el entrenamiento se realiza de manera secuencial, pasando de unos escasos minutos con 20 iteraciones a algo más de media hora con 100. Además a mayor número de iteraciones disminuye mínimamente la sensibilidad mientras aumenta la especificidad, lo cual confirma lo dicho en la teoría sobre que en cada paso el modelo da mayor importancia frente aquellos casos

clasificados erróneamente con el clasificador anterior. Esta leve disminución de la sensibilidad y aumento de especificidad hacen que también aumente ligeramente la AUC.

Adicionalmente, se puede observar que, al igual que en el resto de algoritmos, la especificidad es bastante baja, por lo que se han realizado técnicas de *resampling* sobre los datos y entrenado el modelo con 20 iteraciones obteniendo los resultados de la tabla 3.10.

Muestra	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	0,689	0,916	0,463	0,257	0,836
Test	0,68	0,913	0,447	0,241	0,806

Tabla 3.10: Resultados del *adaboost* tras técnicas de *resampling*

A diferencia que el resto de algoritmos, además de obtener una AUC, aumenta bastante la especificidad sin disminuir demasiado la sensibilidad.

Por otra parte, cabe destacar que el modelo obtiene el mayor valor de la kappa.

La curva ROC perteneciente a este modelo en particular es la de la Figura 3.13.

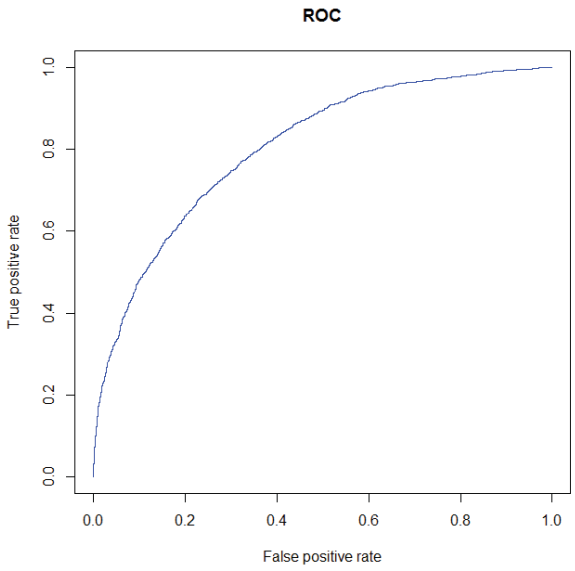

Figura 3.13: Curva ROC adaboost

3.6.3. Máquinas de Vector Soporte (SVM)

3.6.3.1. Introducción SVM

Dentro de las técnicas de *machine learning* el uso de las máquinas de vector soporte como clasificador se ha visto incrementado en los últimos años debido a que sirven para resolver problemas de clasificación y regresión y que su rendimiento en los diferentes campos en los que se utilizan suele ser bastante alto, resultando ser uno de las técnicas más precisas.

El algoritmo de las máquinas de vector soporte o SVM fueron inventadas por Vladimir N. Vapnik y más tarde expuesto por Cortes y Vapnik en 1993 pero publicado en 1995 [27]. Son modelos pertenecientes al aprendizaje supervisado que son igualmente utilizados para clasificación y regresión.

La finalidad de estos modelos es la construcción de un hiperplano o conjunto de hiperplanos que disten lo más posible de los puntos más cercanos de cada clase. A la distancia total se le llama margen y los puntos que definen la distancia del margen se denominan vectores soporte.

Mientras que el problema inicial puede presentarse en una dimensión finita, a menudo sucede que los conjuntos no son linealmente separables en ese espacio. Por esta razón, se propuso mapear el espacio de dimensión original en una dimensión mucho mayor, haciendo la separación más fácil en ese espacio y haciéndolo linealmente separable.

Para hacer posible la transformación del espacio original de las variables de entrada en otro de mayor dimensión donde la separación entre clases sí es lineal se hace uso de unas funciones *kernel*.

Funciones *kernel*.

A la hora de entrenar estos modelos hay que tener en cuenta que la dimensión del espacio necesario para separar las variables puede ser muy grande y aumentar considerablemente el coste computacional. Una forma bastante efectiva de transformar un espacio de una dimensión en uno de una dimensión superior son las funciones núcleo o *kernel*.

Estas funciones mapean es espacio de entradas X a un nuevo espacio de características de mayor dimensionalidad (Hilbert).

Entonces, una función *kernel* es una función $K : \mathbb{X}\, x\, \mathbb{X} \longrightarrow \mathbb{R}$ que asigna a cada par de elementos del espacio de entrada, $\mathbb{X}$, un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio F, es decir,

$$K(x, x') =< \Phi(x), \Phi(x') >= (\phi_1(x)\phi_1(x') + \cdots + \phi_m(x)\phi_m(x')),$$

donde ϕ es una transformación de $\mathbb{X}$ en un espacio de Hilbert, F.

Además, existen varias posibles funciones núcleo que pueden ser utilizadas para crear tal espacio de características de mayor dimensionalidad:

- Polinómica: $K(x_i, x_j) = (x_i \cdot x_j)^n$
- Perceptrón: $K(x_i, x_j) = \| x_i - x_j \|$
- Base radial Gaussiana: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2(\sigma)^2)$
- Sigmoidal: $K(x_i, x_j) = \tanh(x_i \cdot x_j - \theta)$

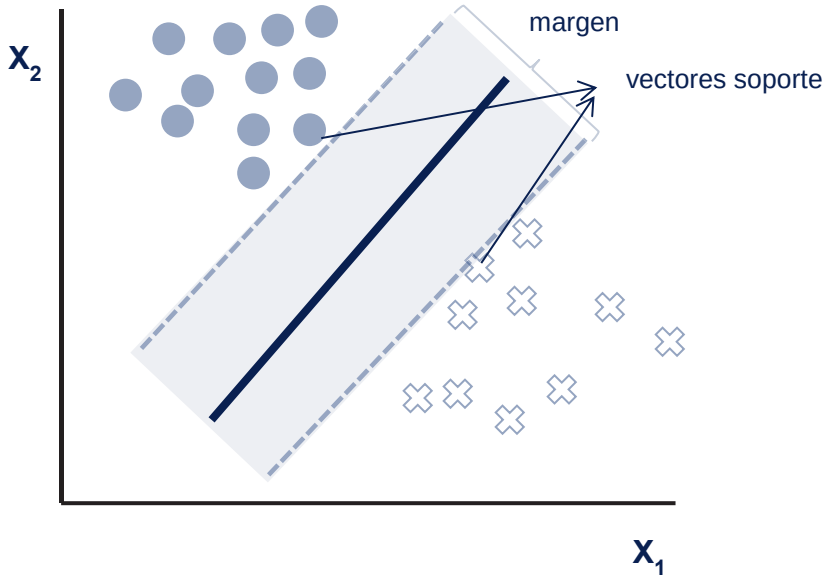

Figura 3.14: Gráfica SVM. Fuente: Elaboración propia

Algunas de las ventajas y desventajas a tener en cuenta en estos modelos son:

SVM
Ventajas
• Se pueden utilizar para problemas de clasificación y regresión. • Soporta bastante bien el ruido en los datos y no son muy propensos al sobreentrenamiento. • Posee una alta precisión. • El uso de *kernels* en la búsqueda del hiperplano de separación perite transformar el espacio de las variables en otro de mayor dimensión de manera que la separación de las clases sea lineal.
Desventajas
• Encontrar el mejor modelo a menudo requiere diversas pruebas de varias combinaciones de *kernels* y parámetros. • Pueden llegar a ser bastante lentos en la fase de entrenamiento, especialmente si el conjunto de datos posee un gran número de variables. • Los resultados del modelo resultan ser una caja negra difícil, sino imposible, de interpretar.

Tabla 3.11: Ventajas y desventajas de los modelos de SVM

3.6.3.2. Modelo SVM

Para poder encontrar el mejor modelo se han tenido que probar varios tipos de *kernel*. Tras esto, se ha estimado que el mejor clasificador era aquel que tenía como función *kernel* una función de base radial gaussiana. Adicionalmente a esto, para poder estimar el parámetro de regularización (C), que nos permite penalizar en mayor o menor medida la clasificación errónea, se han realizado varias pruebas con varios modelos con el *kernel* comentado. Los resultados de estos modelos están en la tabla XX.

Aunque parece ser que según los resultados en la muestra de entrenamiento cuanto mayor C mejor será el modelo, sobre la muestra de test no es igual, por lo que se utilizará C = 5 como parámetro de coste, puesto que es el que mejor rendimiento tiene.

Aunque en un principio parezca que el mejor coste a la hora de entrenar el algoritmo sea C=10 en la muestra de entrenamiento, se observa que realmente a la hora de predecir sobre los datos de testeo, el modelo que mejor rendimiento tiene coste C=5.

Al contrario de lo que cabía esperar, los resultados del modelo (tabla XX) no son del todo satisfactorios, resultando en un principio ser el peor de los modelos hasta ahora cuando, teóricamente, el método de SVM es uno de los que mejor se comporta como clasificador binario, como ya se ha comentado y citado en el apartado del estado del arte.

Aun así, estos resultados son consecuencia del desbalanceo de clases, ya que SVM es un modelo que necesita muchos datos de cada clase para poder entrenarse y la escasa proporción de clientes morosos hace que la especificidad sea casi nula.

Para poder obtener un mejor clasificador se han llevado a cabo técnicas de *resampling* para poder solucionar un poco el problema del desbalanceo de clases. Al volver a entrenar el modelo tras estas medidas con una función de base radial gaussiana y coste C=5, los resultados obtenidos se muestran en la Tabla 3.13.

Como se puede observar, los resultados han mejorado notoriamente, destacando el aumento de la especificidad y la AUC competente. Aun así, el aumento de coste computacional a la hora de entrenar el modelo también ha aumentado significativamente.

Muestra	Coste	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	1	0,507	1	0,015	0,027	0,923
	5	0,585	0,999	0,171	0,281	0,935
	10	0,632	0,998	0,265	0,405	0,943
Test	1	0,501	0,999	0,004	0,006	0,706
	5	0,521	0,998	0,045	0,08	0,724
	10	0,529	0,997	0,062	0,103	0,718

Tabla 3.12: Análisis parámetros SVM

Muestra	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	0,879	0,939	0,819	0,495	0,953
Test	0,642	0,921	0,363	0,189	0,771

Tabla 3.13: Resultados SVM tras resampling

La ROC resultante del modelos de las máquinas de vector soporte está en la Figura 3.15.

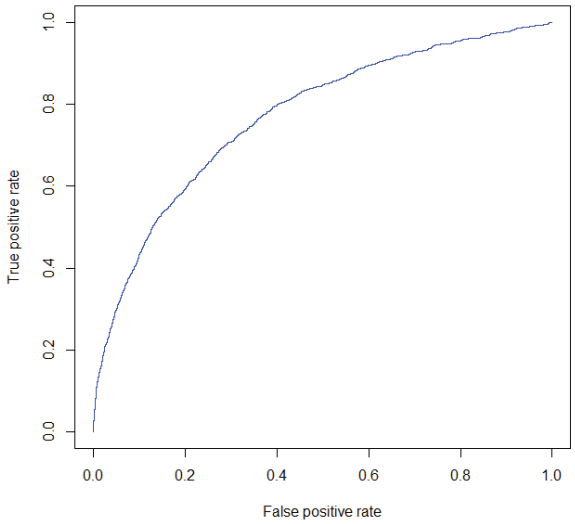

Figura 3.15: Curva ROC SVM

3.6.4. Naïve Bayes

3.6.4.1. Introducción naïve bayes

En el contexto de *machine learning*, los clasificadores *naïve bayes* son una familia de clasificadores probabilísticos basados en la aplicación del teorema de Bayes asumiendo la independencia entre los atributos. A esto se le llama una asunción ingenua o *naïve*, de ahí el nombre del algoritmo. Esto es, un clasificador *naïve bayes* asume que el valor de un atributo en particular no está relacionado con la presencia o ausencia de otro, un clasificador de este tipo considera que cada una de las variables contribuye de forma independiente al modelo.

Para algunos tipos de modelos de probabilidad, estos clasificadores pueden resultar muy eficientes en un entorno de aprendizaje supervisado, como en nuestro caso.

Una de las principales ventajas de *naïve bayes* es que sólo requiere de una pequeña cantidad de datos para entrenar un buen modelo.

El modelo de probabilidad para estos clasificadores es un modelo condicional

$$p(C|F_1, \dots, F_n)$$

Sobre una variable dependiente C con un pequeño número de salidas, dependiendo de una serie de atributos $\{F_1, \dots, F_n\}$. En el caso de tener un gran número de atributos o algún atributo que pueda tomar un gran número de valores, calcular la probabilidad no es viable. Por lo tanto se reformula el modelo para hacerlo más manejable.

Haciendo uso del teorema de Bayes, se puede escribir de la siguiente manera

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)} \quad \textbf{(3.13)}$$

Al igual que en resto de modelos, a continuación se presenta una tabla con algunos aspectos a tener en cuenta antes de construir un modelo.

NAIVE BAYES
Ventajas
• Es un modelo simple, rápido y bastante efectivo. • Soporta bastante bien los *missings* y el ruido. • Requiere de pocos ejemplos para poder entrenarse, aunque también funciona bien con un gran número de datos. • Fácil de obtener la probabilidad estimada de cada predicción.
Desventajas
• Se basa a menudo en suposiciones defectuosas como la igualdad de la importancia de las variables y su independencia, lo cual puede llevar a una falta de precisión. • No es ideal para *datasets* con demasiados valores numéricos. • Las probabilidades estimadas son menos fiables que las clases previstas.

Tabla 3.14: Ventajas y desventajas de los modelos *naïve bayes*

3.6.4.2. Modelo naïve bayes

Para poder construir el modelo, se ha hecho un ligero tratamiento previo, adicional al de preprocesamiento, de la base de datos. Como se ha visto antes, es un modelo que supone la igualdad de variables y su independencia, por lo que las variables correlacionadas tienden a empeorar un poco el modelo, por ello se ha realizado la revisión de la correlación de las variables, disminuyendo el límite de correlación establecido en un principio de 0,8 a 0,7. Además de este tratamiento específico, no ha hecho falta realizar ningún otro más exhaustivo más allá el preprocesamiento

inicial ya mencionado. Esto se debe a que el modelo soporta bastante bien los *missings* y el ruido, de forma que no se ha tenido tan en cuenta.

Los resultados de clasificador bayesiano se muestran en la tabla 3.15, donde se puede observar que el clasificador bayesiano resulta ser un modelo robusto frente al sobreentrenamiento una vez se ha realizado el tratamiento oportuno de las variables con alta correlación.

Adicionalmente, se puede apreciar que el modelo dispone de una precisión bastante aceptable, prácticamente al nivel del *AdaBoost*, el mejor clasificador hasta el momento. No obstante, es el modelo que mejor predice los clientes morosos, de forma que el modelo con mayor especificidad. Por ello, aunque no posee la mejor AUC, se podría considerar como uno de los mejores modelos, puesto que el resto de métricas son bastante satisfactorias.

Muestra	Accuracy	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Entrenamiento	0,817	0,679	0,831	0,528	0,144	0,771
Test	0,814	0,675	0,827	0,522	0,138	0,765

Tabla 3.15: Resultados naïve bayes tras *resampling*

Un punto que llama la atención y cabe destacar de este modelo es la velocidad de entrenamiento del mismo, puesto que mientras los demás modelos tardan en la fase de entrenamiento varios minutos e incluso horas (como en el caso de SVM), estos modelos no ha necesitado más de unos segundos para construirse. De este modo, aunque no necesita de grandes cantidades de datos para entrenarse, funciona bastante bien también con grandes cantidades de datos (big data) sin suponer un coste computacional demasiado elevado.

Por último, decir que a pesar de su diseño ingenuo y sus asunciones aparentemente simples, estos modelos han trabajado muy bien en situaciones reales de gran complejidad.

En este caso la ROC que sale es la cura de la figura de abajo.

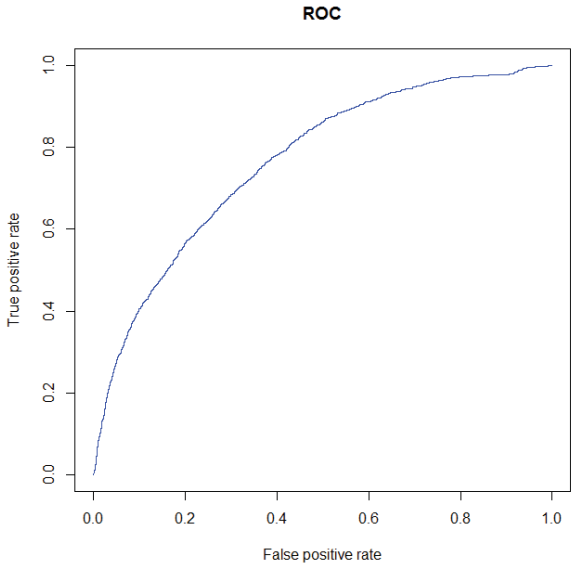

Figura 3.16: Curva ROC naïve bayes

3.6.5. Comparación final

A continuación se muestra una tabla resumen con os mejores resultados de todos los modelos. En rojo se destacan los mejores resultados de cada métrica.

Modelo	BalancedAccuracy	Sensibilidad	Especificidad	Kappa	AUC
Regresión logística	0,651	0,941	0,36	0,232	0,799
AdaBoost	**0,680**	0,913	0,447	0,241	**0,806**
Random Forest	0,528	**0,997**	0,056	0,1	**0,813**
SVM	0,642	0,921	0,363	0,189	0,771
Naive Bayes	0,675	0,827	**0,522**	0,138	0,765

Tabla 3.16: Comparación de resultados

Los resultados de la tabla ponen de manifiesto que modelos basados en algoritmos de *machine learning*, sobre la regresión logística, suponen alternativas valorables en la estimación de un *scoring* crediticio para esta cartera, ya que como se ve, las mejores métricas pertenecen a estos modelos, destacando unos más que otros en las distintas métricas.

En este caso en particular, el modelo más regular de todos es el *adaboost* puesto que tiene los mejores resultados en 3 de las métricas observadas y posee una sensibilidad bastante alta y una especificidad aceptable dado el gran desbalanceo existente en los datos.

Capítulo 4

4. Conclusiones del trabajo

4.1. Conclusión

Como resultado de haber realizado un estudio profundo sobre el *machine learning*, sus técnicas más destacables y haber construido los diferentes modelos utilizados en este trabajo, se pueden extraer una serie de conclusiones, que se detallan a continuación:

- En este caso y a pesar de que la técnica que finalmente se utilizó en este caso fue la regresión logística, se observa que el *adaboost* mejora el poder predictivo del modelo avalado por una alta precisión, una sensibilidad y especificidad regular y una alta precisión a través del área bajo la curva ROC.

- Basándose en este ejercicio y en la bibliografía analizada y citada en los capítulos relativos al estado del arte del *machine learning*, se puede afirmar que las técnicas de *machine* learning pueden aportar soluciones más eficientes a la estimación de *credit scoring.*

- Si bien estas técnicas pueden resultar útiles, necesitan incorporar algunas técnicas complementarias para obtener los mejores resultados, como el *resampling* y otras, como las relativas al *data quality*, se vuelven más importantes y decisivas.

- Con todo ello, y en un contexto donde la gestión de la morosidad y la competencia por la demanda solvente, y aunque el marco regulatorio no permite el uso de estas técnicas, puede ser útil evaluar el uso complementario de estas técnicas.

4.2. Futuras líneas de investigación

Existen varias líneas de investigación adicionales sobre el uso de las diversas técnicas de *machine learning* en el *credit scoring*:

- Análisis de las técnicas computacionales para automatizar el cálculo, la comparación, el entrenamiento y el seguimiento de los modelos.
- Análisis de las técnicas computacionales para reducir el coste de computación de estas técnicas.
- Análisis de la incorporación de técnicas de *machine learning* en los actuales *frameworks* de construcción, seguimiento y validación en las entidades financieras.
- Como complemente a este trabajo, puede realizarse un estudio del coste de oportunidad y la diferencia de morosidad en función del uso de distintos modelos para la cartera analizada.

Bibliografía

[1] F. O. Pérez y H. F. Castaño, «Las redes neuronales y la evaluación del riesgo de crédito,» Revista Ingenierías niversidad de Medellin, Medellin (Colombia), 2007.

[2] B. Goodman y S. Flaxman, «European Union regulations on algoritmic decision-making and a "right to explanation",» Oxford, 2016.

[3] W. Knight, «MIT Techonology Review,» 11 Abril 2017. [En línea]. Available: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. [Último acceso: 7 Mayo 2017].

[4] «agpd,» Agencia Epañola de Protección de Datos, 26 Enero 2017. [En línea]. Available: http://www.agpd.es/portalwebAGPD/revista_prensa/revista_prensa/2017/notas_prensa/news/2017_01_26_01-ides-idphp.php. [Último acceso: 5 Mayo 2017].

[5] «Guía del Reglamento General de Protecciónde Datos para Responsables de Tratamiento,» Agencia Española de Protección de Datos, 2017.

[6] «Directrices para la elaboración de contratos entre responsables y encargados del tratamiento,» Agencia Epañola de Protección de Datos, 2017.

[7] «Guía para el cumplimiento del deber de informar,» Agencia Española de Protección de Datos, 2017.

[8] D. Durand, «Risk Elements in Consumer Instalment Financing,» *National Buereau of Economic Research,* 1941.

[9] J. Myers y E. Forgy, «The Development of Numerical Credit Evaluation System,» *Journal of the American Statistical Association,* vol. 58, nº 303, pp. 799-806, 1963.

[10] Y. E. Orgler, «A Credit Scoring Model for Commercial Loans,» *Journal of Money, Credit and Banking,* vol. 2, nº 4, pp. 435-445, 1970.

[11] H. Bierman y W. Hausman, «The Credit Granting Decision,» *Management Science,* vol. 16, pp. 519-532, 1970.

[12] Y. E. Orgler, «Evaluation of bank consumer loans with credit scoring models,» *Journal of Bank Research,* vol. 2, nº 1, pp. 31-37, 1971.

[13] J. C. Wiginton, «A note on the comparison of logit and discriminant models,» *Journal of Financial an Quantitative Analysis,* vol. 15, nº 3, pp. 757-770, 1980.

[14] W. E. Henley, «Statistical aspects of credit scoring,» *Open University,* 1994.

[15] C. Bolton, «Logistic regression and its application in credit scoring,» University of Pretoria, 2009.

[16] R. A.Eisenbeis, «Pitfalls in the application of discriminant analysis in business, finance and economics,» *The Journal of Finance,* vol. 32, nº 3, pp. 875-900, 1977.

[17] R. A. Eisenbeis, «Problems in applying discriminant analysis in credit scoring models,» *The Journal of Banking and Finance,* vol. 2, nº 3, pp. 205-219, 1978.

[18] D. Hand y W. Henley, «Statistical classification methods in consumer credit scoring,» *Journal of the Royal Statistical Society,* vol. 160, nº 3, pp. 523-541, 1997.

[19] M.Bonilla, I.Olmeda y R.Puertas, «Modelos paramétricos y no paramétricos en problemas de credit scoring,» *Revista Española de Financiación y Contabilidad,* vol. XXXII, nº 118, pp. 833-869, 2003.

[20] V. Srinivasan y Y. Kim, «The Bierman-Hausman credit granting model: A note,» *Management Science,* vol. 33, nº 10, pp. 1361-1362, 1987.

[21] L. C. Thomas, «A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers,» *International Journal of Forecasting,* vol. 16, nº 2, pp. 149-172, 2000.

[22] Baesens y Bart, «Developing Intelligent Systems for Credit scoring using Machine Learning Techniques,» *Katholieke Universiteit Leuven (Tesis),* 2003.

[23] J. Crook, D. Edelman y L. Thomas, «Recent development in consumer credit risk assessment,» *European Journal of Operational Research,* vol. 183, nº 3, pp. 1447-1465, 2007.

[24] C. Hung y J. Chen, «A selective ensemble based on expected probabilities for bankruptcy prediction,» *Expert Systems with Applications,* vol. 36, nº 3, pp. 5297-5303, 2009.

[25] L. Yu, S. Wang y K. K. Lai, «Credit risk assessment with a multistage neural network ensemble learning approach,» *Expert System with Applications,* vol. 34, nº 2, pp. 1434-1444, 2008.

[26] G. Wang, J. Hao, J. Ma y H. Jiang, «A comparative assessment of ensemble learning for credit scoring,» *Expert Systems with Applications,* vol. 38, nº 1, pp. 223-230, 2011.

[27] V.Vapnik y C.Cortés, «Support-Vector Networks,» *Kluwer Academic Publishers,* nº 20, pp. 273-297, 1995.

[28] L. Yu, W. Yue, S. Wang y K. Lai, «Support vector machine based multiagent ensemble learning for credit risk evaluation,» *Expert Systems with Applications,* vol. 37, nº 2, pp. 1351-1360, 2010.

[29] X. Xu, C. Zhou y Z. Wang, «Credit scoring algorithm based on link analysis ranking with support vector machine,» *Expert Systems with Applications,* vol. 36, nº 2, pp. 2625-2632, 2009.

[30] T. Bellotti y J. Cook, «Support vector machines for credit scoring and discovery of significant features,» *Expert Systems with Applications,* vol. 36, nº 2, pp. 3302-3308, 2009.

[31] J. Moreno y L. Melo, «Pronóstico de incumplimiento de pago mediante máquinas de vectores soporte: una aproximación inicial a la gestión del riesgo de crédito,» *Borradores de Economía,* nº 667, 2011.

[32] T. M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.

[33] T. M.Mitchell, «The Discipline of Machine Learning,» School of Computer Science, Carnegie Mellon Univerity, Pittsburgh, 2006.

[34] R. Schapire, «Theoretical Machine Learning,» 2008.

[35] M. Bourel, «Métodos de agregación de modelos y aplicaciones,» 2012.

[36] J. Brownlee, «machinelearningmasery,» 16 Marzo 2016. [En línea]. Available: http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [Último acceso: Febrero 2017].

[37] L. Wasserman, «normaldeviate.wordpress.com,» [En línea]. Available: https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/.

[38] Shah y Aatash, «Edvanced Eduventures,» 8 Enero 2016. [En línea]. Available: http://www.edvancer.in/machine-learning-vs-statistics/.

[39] J. H. Friedman, «Data Mining and Statistics: What's the connection?,» Stanford University, Stanford.

[40] A. Ng, Interviewee, *Deep-Learning AI is taking over tech. What is it?*. [Entrevista]. 15 Julio 2015.

[41] J. A. Cruz y D. S. Wishart, «Applications of Machine Learning in Cancer Prediction and Prognosis,» NCBI, 2007.

[42] I. Kononenko, «Machine Learning for Medical Dignosis: History, Statae of Art and Perpective,» Ljubljana, 2001.

[43] K. Pal, «KDnuggets,» Octubre 2015. [En línea]. Available: http://www.kdnuggets.com/2015/10/big-data-recommendation-systems-change-lives.html. [Último acceso: Febrero 2017].

[44] P. Härle, A. Havas, A. Kremer, D. Rona y H. Samandari, «The future of bank risk management,» McKinsey&Company, 2015.

[45] Oracle, «Big Data en Financial Services and Banking,» Oracle Enterprise, 2015.

[46] T. H. Davenport y J. G. Harris, «MIT Sloan Management Review,» 15 Julio 2005. [En línea]. Available: http://sloanreview.mit.edu/article/automated-decision-making-comes-of-age/. [Último acceso: 14 Marzo 2017].

[47] pwc, «Adjustign the Lens on Economic Crime: Preparation brings opportunity back into focus,» pwc, 2016.

[48] Rajvanshi y Aastha, «LinkedIn,» 7 Febrero 2016. [En línea]. Available: https://www.linkedin.com/pulse/predictive-analyticsmachine-learning-fraud-detection-aastha-rajvanshi. [Último acceso: Marzo 2017].

[49] «guardiananalytics.com,» [En línea]. Available: http://guardiananalytics.com/. [Último acceso: Marzo 2017].

[50] R. Gupta y B. Kadaba, Interviewees, *Slow And Steady: The Financial Services Industry Can´t Rush AI Adoption.* [Entrevista]. 12 Enero 2017.

[51] A.-T. Hussein y A.-M. Faris, «Banks' risk management: a comparison study of UAE national and foreign banks,» *The Journal of Risk Finance,* vol. 8, nº 4,

pp. 394-409, 2007.

[52] C. Ruza y Paz-Curbera, El riesgo de crédito en perspectiva, Madrid: UNED, 2013.

[53] M. Schreiner, «Ventajas y Desventajas del Scoring Estadístico para las Micrfinanzas,» Microfinancerisk Management, Washington University (St.Louis), 2002.

[54] M. A. Gutiérrez, «Modelos de Credit Scoring - Qué, Cómo, Cuándo y Para Qué -,» Octubre 2007. [En línea]. Available: http://www2.bcra.gob.ar/Pdfs/Publicaciones/CreditScoring.pdf.

[55] «Google Cloud Platform,» 15 Marzo 2017. [En línea]. Available: https://cloud.google.com/ml-engine/docs/concepts/ml-solutions-overview. [Último acceso: 10 Mayo 2017].

[56] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro y C. S. Haley, «Scientific Reports,» 19 Mayo 2015. [En línea]. Available: https://www.nature.com/articles/srep10312. [Último acceso: Mayo 2017].

[57] P.-N. Tan, M. Steinbach y V. Kumar, «Classification: Basic Concepts, Decision Trees and Model Evauation,» de *Introduction to Data Mining*, Boston, Addison-Wesley Longman Publishing, 2006, pp. 145-205.

[58] «The analysis factor,» [En línea]. Available: http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/.

[59] C.-T. Chung, «KDnuggets,» Marzo 2015. [En línea]. Available: http://www.kdnuggets.com/2015/03/machine-learning-data-science-common-mistakes.html. [Último acceso: Mayo 2017].

[60] J. A. Hanley y J. B. McNeail, «The meaning and use of the area under a receiver operating characteristic (ROC) curve,» *Radiology*, vol. 143, nº 1, pp. 29-36, 1982.

[61] Bradley y P. Andrew, «The use of the area under the ROC curve in the evaluation of machine learning algorithms,» *Pattern Recognition,* vol. 30, nº 7, pp. 1145-1159, 1997.

[62] T. Fawcett, «ROC Graphs: Notes and Practical Considerations,» *HP Laboratoiries,* 2003.

[63] D. J. Hand y R. J. Till, «A simple generalization of the area under the ROC curve to multiple class classification problems,» *Machine Learning,* vol. 45, nº 2, pp. 171-186, 2001.

[64] O. Troyanskaya, M. Cantor, G. Sherlock, P.Brown, T. Hastie, R. Tibshirani, D. Botstein y R. B. Altman, «Missing value estimation methods for DNA microarrays,» *Bioinformatics,* vol. 17, nº 6, pp. 520-525, 2001.

[65] D. H. Wolpert y W. G. Macready, «No Free Lunch Theorems for Optimization,» *IEEE Trasaction on Evolutionary Computation,* vol. 1, nº 1, pp. 67-82, 1997.

[66] K. P. Murphy, Machine Learning: a probabilistic perspective, Cambridge, Massachusetts: The MIT Press, 2012.

[67] D. J. Hand y M. G. Kelly, «Superscoreards,» *IMA Journal of Management Mathematics,* vol. 13, nº 4, pp. 273-281, 2002.

[68] S. L. Lin, «A new Two-Stage Hybrid Approach of Credit Risk in Banking Industry,» *Expert Systems with Applications,* vol. 36, nº 4, pp. 8333-8341, 2009.

[69] B. Lanz, Machine Learning with R, Birmingham: Packt Publishing Ltd, 2015.

[70] L. Breiman, «Random Forests,» *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[71] Y. Amit y D. German, «Shape quantification and recognition with randomized trees,» *Neural Computation,* vol. 9, nº 7, pp. 1545-1588, 1997.

[72] C. Y.-W. (. Chiu), Machine Learning with R Cookbook, Birmingham: Packt Publishing Ltd, 2015.

[73] P. B. Bocigas, «FinTech spain,» 9 Febrero 2017. [En línea]. Available: http://fintechspain.com/2017/02/09/alternativas-financiacion-de-empresas/. [Último acceso: Marzo 2017].

[74] P. Chaman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, «Step-by-step data minig guide,» CRISP-DM, 2000.

[75] W. Yang y M.-h. Tsai, «Google Research Blog,» 8 Octubre 2015. [En línea]. Available: https://research.googleblog.com/2015/10/improving-youtube-video-thumbnails-with.html. [Último acceso: Febrero 2017].

[76] V. N. Vapnik, The Nature of Statistical Learning Theory, New York: Springer, 2000.

Este documento esta firmado por

Firmante	CN=tfgm.fi.upm.es, OU=CCFI, O=Facultad de Informatica - UPM, C=ES
Fecha/Hora	Sun Jun 11 18:29:31 CEST 2017
Emisor del Certificado	EMAILADDRESS=camanager@fi.upm.es, CN=CA Facultad de Informatica, O=Facultad de Informatica - UPM, C=ES
Numero de Serie	630
Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)