Facultad de Ciencias Económicas y Empresariales
(ICADE)

Big Data en el sector inmobiliario: aplicación en los préstamos hipotecarios

Autor: Luis Martín Palla

Director: Raúl González Fabre

Madrid

Abril de 2023

RESUMEN

El Big Data y los modelos de *machine learning* han ganado un importante protagonismo en los últimos años. La totalidad de sectores económicos a nivel internacional se han visto influenciados por estas nuevas tecnologías, y todo parece indicar que en las próximas décadas este impacto será todavía de mayor dimensión. El sector inmobiliario y el bancario, generadores de más del 10% y del 4% del PIB en nuestro país respectivamente, también se han visto influenciados por estas herramientas, debido principalmente a la revolución que los sistemas de *credit scoring* han provocado en la concesión y gestión de préstamos hipotecarios. Ante este fenómeno, el presente trabajo de fin de grado estudia la posibilidad de construir un sistema de *credit scoring* basado en modelos de *machine learning* que permita alcanzar una precisión elevada en la predicción del acaecimiento del suceso de impago. Mediante la implementación de diferentes algoritmos en un ejercicio de clasificación, se obtienen unos resultados que no permiten confirmar la posibilidad de construir un sistema como el descrito. Sin embargo, sí se han obtenido *insights* significativos, como la calibración del algoritmo que mejor parece adaptarse al caso de estudio o la identificación de variables relevantes y de oportunidades de mejora de cara a futuros ejercicios de una índole similar.

Palabras clave: sector inmobiliario, sector bancario, préstamo hipotecario, riesgo de impago, *credit scoring*, Big Data, *machine learning.*

ABSTRACT

Big Data and machine learning models have gained significant prominence in recent years. All international economic sectors have been influenced by these new technologies, and everything seems to indicate that in the coming decades this impact will be even greater than that experienced to date. The real estate and banking sectors, which account for 10% and 4% of the GDP in our country respectively, have also been influenced by these tools, mainly due to the revolution that credit scoring systems have caused in the granting and management of mortgage loans. In view of this phenomenon, this thesis studies the possibility of building a credit scoring system based on machine learning models to achieve a high accuracy in the prediction of the occurrence of the default event. By implementing different algorithms in a classification exercise, results are obtained that do not allow confirming the possibility of building a system like the one described. However, significant insights have been obtained, such as the calibration of the algorithm that seems to be best suited to the case study or the identification of relevant variables and opportunities for improvement for future exercises of a similar nature.

Keywords: real estate sector, banking sector, mortgage lending, default risk, credit scoring, Big Data, machine learning.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Propósito y justificación del tema

A lo largo de la historia, la vivienda ha sido considerada como una de las necesidades básicas del ser humano, clave para nuestra supervivencia y determinante para nuestro bienestar. Sin embargo, la versatilidad y la capacidad de generación de ingresos recurrentes de los bienes raíces ha llevado a su percepción como un activo de inversión y un representativo de riqueza, más allá de su concepción como una necesidad fundamental para nuestra especie. Ante situaciones de incertidumbre o inestabilidad financiera, la vivienda resalta como una alternativa de moderado riesgo en la que buscar rentabilidad, acunando una importante porción del capital invertido tanto por empresas como por particulares cada año (Borja, s.f.).

Asimismo, la vivienda desempeña otra gran variedad de funciones relacionadas con la actividad económica que la convierten en un factor clave para las economías a nivel internacional, en las que puede llegar a representar más de un 10% del Producto Interior Bruto, como en el caso de nuestro país (Fernández, 2022). La importancia del sector inmobiliario en la transmisión de la política monetaria y sus fuertes relaciones intersectoriales lo establecen como una pieza clave para la garantía de la salud y el adecuado desarrollo económico, de tal forma que los ciclos experimentados por este sector pueden perjudicar fuertemente tanto a la economía real como a la estabilidad financiera (Zhu, 2014).

Sin embargo, la efectividad en el funcionamiento del sector inmobiliario es inconcebible sin el apoyo del sector bancario. La naturaleza de las operaciones inmobiliarias, por lo general de elevada cuantía económica, requiere en la mayoría de los casos del uso de la financiación, involucrando a los inversores y particulares como partes en contratos de préstamos o créditos inmobiliarios (Asociación Hipotecaria Española, 2014). Préstamos promotores, créditos constructores… estas son algunas de las formas en las que el sector bancario impacta directamente el devenir del sector inmobiliario. Sin embargo, la más conocida por su relevancia para las entidades bancarias, la economía y la sociedad son los préstamos hipotecarios (Santaella, 2022).

A través de estos préstamos los ciudadanos acceden a la financiación para afrontar el desembolso que la compra de una vivienda conlleva (Asociación Hipotecaria Española, 2014). Sin embargo, su relevancia adquiere una dimensión mucho más profunda, siendo

el principal instrumento por el cual el Banco Central Europeo, mediante la implementación de políticas de expansión o contracción económica, influencia los mercados inmobiliarios (Carbó Valverde y Rodríguez Fernández, 2021).

Más allá de la evidente relevancia de estos préstamos para la economía, antecedentes como la crisis financiera del 2008 iniciada en Estados Unidos, provocada por la concesión de hipotecas con alto riesgo de impago, han provocado un cambio de mentalidad de las entidades bancarias. Esta recesión, que tanto afectó al sector bancario y financiero de las grandes potencias a nivel internacional ha llevado a un considerable aumento de las precauciones adoptadas por las entidades bancarias para la minimización del riesgo de crédito en sus operaciones que ya se está viendo reflejada en el mercado hipotecario (Porras Castaño, 2018).

A este creciente blindaje del sistema bancario ante los riesgos de impago de los préstamos hipotecarios se suman un factor determinante. Hace ya algunos años desde que la digitalización mundial comenzó a afectar al sector financiero, que ha implementado exponencialmente en sus operaciones nuevas disciplinas como la Inteligencia Artificial o el *Data Analytics* con el objetivo de tomar decisiones cada vez más precisas (BBVA, 2022). Esto ha incrementado aún más las dificultades para el acceso a la financiación, que se suma a los obstáculos que las entidades bancarias llevan introduciendo desde hace unos años mediante el endurecimiento de sus políticas de concesión de préstamos hipotecarios. (Uría Menéndez, 2022).

En este nuevo escaparate, el empleo del Big Data y los modelos predictivos para la prevención del riesgo de crédito juegan un papel fundamental (Zárate, 2018). Si bien es cierto que este riesgo ha sido históricamente analizado en detalle por las entidades bancarias, las herramientas de *credit scoring* cobran especial relevancia en la actualidad (Puertas Medina y Martí Selva, 2013).

Por este motivo, el presente trabajo se centra en la situación y la aplicación del Big Data y los modelos predictivos en los préstamos hipotecarios, concibiendo estos como el vínculo sustancial entre dos de los principales sectores de la economía española, como son los sectores inmobiliario y bancario.

1.2. Objetivos del trabajo

El objetivo principal de este trabajo es el estudio de la influencia del Big Data y los modelos predictivos en la predicción del suceso de impago y la concesión de préstamos hipotecarios. Para ello, se busca verificar la hipótesis de que a través de un modelo de *machine learning*, que incorpore variables socioeconómicas (como la edad, el nivel de ingresos o el estado civil), se puede predecir el acaecimiento de un suceso de impago por parte del prestatario con una precisión superior al 80%.

Como se ha mencionado previamente, la concesión de hipotecas depende directamente del sector bancario e impacta de la misma manera en el sector inmobiliario, dos de los grandes sectores económicos de nuestro país. Por ello, para responder a este objetivo primordial, se contemplan una serie de objetivos secundarios en la investigación:

- Desarrollar el marco teórico de los sectores inmobiliario y bancario, así como sus puntos de conexión poniendo el foco en los préstamos hipotecarios como principal anexo entre ambos
- Estudiar las aplicaciones del Big Data y el *machine learning* en los préstamos hipotecarios, especialmente en la predicción del riesgo de crédito, así como la evolución de las técnicas empleadas a lo largo del tiempo.

1.3. Metodología de la investigación

El bloque teórico del trabajo se basará en la exploración de fuentes de información primarias y secundarias que permitan justificar los mensajes que se pretenden transmitir a lo largo de la investigación. Con este fin, las principales herramientas empleadas serán:

- La biblioteca *online* de la Universidad Pontificia Comillas
- Google Académico (*Google Scholar* en inglés)
- Artículos de páginas webs, informes corporativos y blogs, dada la escasez de información académica a la hora de establecer relaciones entre el sector bancario e inmobiliario con las disruptivas técnicas de Big Data y análisis predictivo de datos.

Una vez las fuentes de información hayan sido identificadas, se procederá a la selección de aquellas que mejor representen las ideas contempladas y aporten en un mayor grado a la calidad del contenido del trabajo.

Por su parte, el bloque práctico del trabajo se basará en la construcción de un modelo predictivo del suceso de impago en préstamos hipotecarios. Para ello, se ha seleccionado una base de datos de riesgo de crédito de un banco con sede en Países Bajos, a la cual se ha accedido por medio de kaggle, la conocida plataforma académica de bases de datos.

Como proceso posterior a la selección y obtención de la base de datos empleada para el análisis predictivo, se procederá al entendimiento y la adecuada interpretación de las variables de partida, tanto de índole cualitativa como cuantitativa. Consecutivamente se realizará una limpieza de la base de datos en cuestión, preparándola así para la fase de ajuste de modelos.

A continuación, mediante el entrenamiento de modelos predictivos de *machine learning* en RStudio, una de las principales herramientas de programación, se evaluará la hipótesis mencionada en el apartado anterior. En concreto, se aplicarán cuatro algoritmos a los datos de partida: LOGIT con método *Stepwise,* KNN, árbol de decisión y SVM.

Por último, con los datos obtenidos como fruto del análisis predictivo, se evaluará la adecuación de los diferentes modelos a la base de datos empleada, verificando o rechazando la hipótesis de la investigación.

1.4. Estructura del trabajo

El presente trabajo está estructurado en base a dos grandes bloques.

En primer lugar, un bloque teórico en el que se analizan los principales sectores y disciplinas influyentes en los préstamos hipotecarios. En este sentido, se comienza con un ejercicio de contextualización teórica de la relación entre este los sectores inmobiliario y bancario, obteniendo *insights* preliminares por medio del estudio de la evolución en la concesión de hipotecas. Además, se lleva a cabo un análisis descriptivo de la influencia del Big Data en la determinación del riesgo de crédito y por consecuencia, en la concesión de préstamos hipotecarios, haciendo una breve narración del origen y la evolución de estas técnicas y modelos de *credit scoring* a lo largo de la historia.

En segundo lugar, un bloque práctico en el cual se lleva a cabo un análisis predictivo del suceso de impago aplicado a los préstamos hipotecarios. En este bloque se incluye la descripción tanto del proceso de obtención, limpieza y preparación de los datos como del análisis descriptivo en función de los diferentes modelos empleados. Además, se incluye

una interpretación de los resultados obtenidos mediante el análisis, así como la evaluación de la adecuación de los modelos al caso de estudio.

Por último, se incluye un apartado final de anexos, en el cual se recogen las diferentes herramientas empleadas para la realización del análisis predictivo, como son el script de código de RStudio o las variables de la base de datos utilizada.

2. LA INTERDEPENDENCIA ENTRE EL SECTOR INMOBILIARIO Y EL SECTOR FINANCIERO

De acuerdo con la Real Academia Española (2022), el término inmobiliario/a en su acepción como sustantivo se refiere a: "Empresa o sociedad que se dedica a construir, arrendar, vender y administrar viviendas". Como podemos observar en la propia definición, la cadena de valor en este sector abarca numerosas fases y procesos que se distribuyen en el tiempo desde que se decide construir una vivienda hasta que esta es vendida al cliente final.

El amplio abanico de actividades, desde productivas hasta comerciales, invita a diferentes agentes a participar y beneficiarse de este sector. Con numerosos *stakeholders* intentando conseguir su porción de la tarta, la integración vertical de procesos es una alternativa muy recurrente en la industria, por lo que no es de extrañar que una empresa promotora o incluso constructora asuma actividades de gestión y comercialización. Sin embargo, la fragmentación de la industria y la versatilidad de sus participantes no son el principal motivo de complejidad del mundo inmobiliario. El factor regulatorio, tanto en lo que respecta a procedimientos y burocracia como a requerimientos de índole medioambiental, convierte al sector inmobiliario en un entorno con grandes barreras de entrada (Otero y Blanco, 2014).

Sin embargo, la trascendencia del sector inmobiliario va mucho más allá de las actividades realizadas por proveedores, clientes o trabajadores de empresas inmobiliarias. Este mundo depende de muchos otros sectores, como el de los materiales o el financiero, e influye de forma recíproca en ellos mediante la creación de empleos y la demanda de productos y servicios. Esta contribución al desarrollo de otros sectores de la economía convierte al sector inmobiliario en una importante fuente de crecimiento tanto para países desarrollados como para aquellos que aún están emergiendo (Karamelikli, 2016).

A continuación, se analizará la realidad del sector inmobiliario en España, describiendo su interdependencia con otros sectores de la economía española, especialmente el financiero.

2.1. Relación del sector inmobiliario con el sector financiero

Cuando la economía de un país prospera, el sector inmobiliario lo suele hacer de la misma manera. Por el contrario, en épocas de recesión económica, el sector inmobiliario acostumbra a encontrarse también en un tiempo de ralentización o paralización.

Remitiéndonos a la jerga financiera, cuando se construye un portafolio de inversión equilibrado existen dos conceptos importantes a considerar: valores cíclicos y valores defensivos. Mientras que los valores cíclicos son aquellos correlacionados positivamente con la economía, los valores defensivos o anticíclicos son menos sensibles a ella. El sector inmobiliario es un valor cíclico, pues su movimiento con respecto a la economía es armónico (Rey, 2020).

La compenetración y el movimiento acompasado entre el sector inmobiliario y la economía se debe a diferentes factores. En primer lugar, la relevancia del sector en nuestro sistema económico es notablemente elevada. En el año 2020 en España, las actividades inmobiliarias representaron un 11,7% del PIB (Fernández, 2022). Asimismo, la aportación del sector construcción a la creación de empleo es considerablemente elevada en comparación con la de otros sectores, dando trabajo a cerca de 1,5 millones de españoles (González Cuervo, 2021). En segundo lugar, la ya mencionada influencia macroeconómica del sector y las relaciones con otros sectores productivos y de servicios explican una parte importante de esta correlación. Sin embargo, a pesar de las numerosas conexiones del sector inmobiliario con otras áreas de la economía, indudablemente la relación más trascendente es la existente con el sector financiero (Daher, 2013).

A continuación, se analizará el sector financiero, su rol en la economía y su interacción con el sector inmobiliario considerando la situación macroeconómica y microeconómica española en la actualidad.

2.1.1. El papel del sector financiero

De acuerdo con fuentes como Guide To Business (2020): “Desde el punto de vista institucional, se puede definir el sistema financiero como el conjunto de entidades que

generan, recogen, administran y dirigen tanto el ahorro como la inversión, en un sistema político-económico" (p. 4).

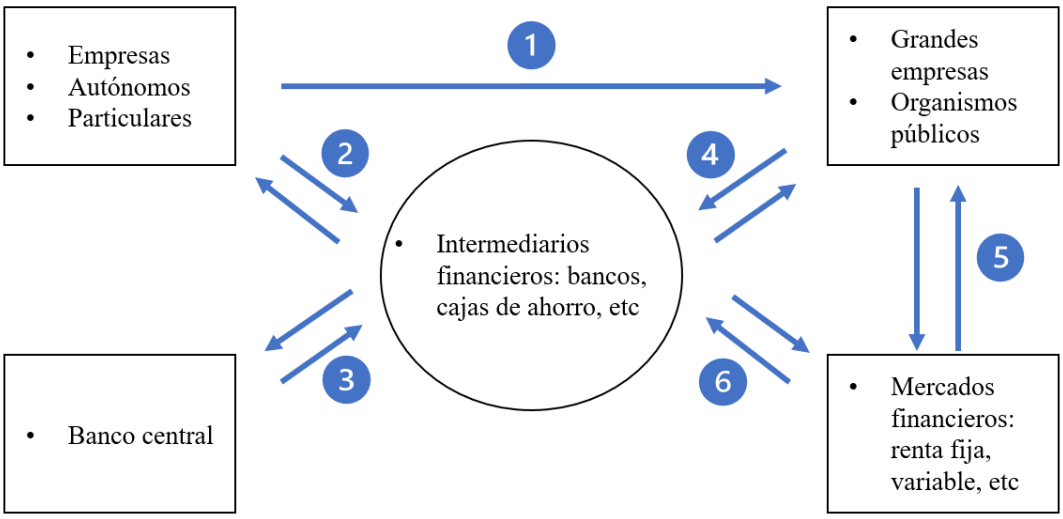
Los sistemas financieros involucran una amplia variedad de mercados y agentes: bancos centrales, fondos de inversión y de pensiones, autoridades regulatorias, aseguradoras y entidades del mercado de valores desempeñan un papel fundamental en la economía que va mucho más allá de la gestión de la relación ahorro-inversión. Las instituciones financieras, mediante su compenetración, contribuyen a la fluidez de los mercados y son determinantes en la implementación de políticas monetarias, favoreciendo así no solo a la creación de un sistema sólido y transparente, sino a la generación de empleo y al crecimiento económico global (Fondo Monetario Internacional, 2016).

Para entender el rol del sistema bancario en el sector inmobiliario, es importante desglosar el concepto de economía en dos vertientes: economía real y financiera. La economía real se basa en la producción, distribución y comercialización de bienes y servicios, generando empleos tanto de forma directa como indirecta. Por el contrario, la economía financiera se basa en activos y mercados financieros cubriendo una necesidad: el desplazamiento del dinero (López Millán, 2022).

A pesar de esta diferencia, ambas vertientes están muy relacionadas y son complementarias. Para que exista una economía real sólida, esto es que se produzcan y comercialicen bienes y servicios, la economía financiera juega un papel imprescindible ya que facilita y financia procesos que en muchas ocasiones requieren grandes cuantías de capital. La forma en la que los participantes de la economía real acceden a fondos con los que financiarse es mediante la figura de intermediarios financieros, es decir, del sistema bancario (Santaella, 2022).

En la economía existen una gran variedad de agentes, cuyos procesos y áreas de actuación son, en la mayoría de las ocasiones, completamente diferentes. Como podemos observar en la Figura Nº1, estos agentes económicos intercambian fondos entre sí, pero es el sistema bancario el que pone en contacto a todos ellos, dando armonía a sus actividades y facilitando la subsistencia de todos ellos (Carbó Valverde, 2020).

Figura Nº1: Papel del sistema bancario en la economía



Fuente: San Millán y Viejo (2008)

En primer lugar, empresas y particulares invierten sus ahorros en renta fija o variable, ya sea pública o privada. Esto lo pueden hacer tanto directamente [1] como indirectamente, con bancos o cajas de ahorro como intermediarios [2]+[4] o [2]+[6].

Por su parte, el estado y las grandes empresas buscan formas de financiar sus operaciones tanto de forma directa [5] como de forma indirecta [4]+[6], de nuevo gracias a la intermediación bancaria.

En lo que respecta a los intermediadores financieros, reciben fondos de todos los agentes mencionados [2], [4] y [6].

Por último, los bancos centrales implementan políticas monetarias para garantizar y controlar la liquidez del mercado [3] (San Millán y Viejo, 2008).

Uno de los ejemplos más ilustrativos de la relación que acabamos de explicar es mediante las hipotecas, con el sector inmobiliario como integrante de la economía real. Con el objetivo de financiar operaciones inmobiliarias, empresas y particulares buscan acceder a capitales de la economía financiera. El sistema bancario, intermediario financiero, actúa como puente ofreciendo a estas empresas e individuos fondos en forma de deuda, un capital que habrá conseguido de diferentes formas operando en los mercados financieros (Santaella, 2022).

Si analizamos las fuerzas que determinan el curso de los mercados inmobiliarios, encontramos tres pilares sobre los que se sustentan tanto las actividades productivas como comerciales de este sector, ya que afectan tanto al volumen como a los precios de las transacciones, y determinan volatilidad de la inversión existente y potencial (Nguyen, 2021). Estos pilares son:

- Demografía: el tamaño y el crecimiento de la población, la renta per cápita o el envejecimiento de la población son indicadores que tienen una importante influencia en los mercados inmobiliarios. Prueba de ello es el crecimiento en el volumen de transacciones y en el precio de la vivienda que ha traído consigo el crecimiento de la población en los últimos años (Marrero, 2022).
- Políticas fiscales: mediante créditos e incentivos fiscales, los gobiernos pueden manipular los mercados inmobiliarios. Gracias a estas medidas, se pueden impulsar y contraer tanto la oferta como la demanda, afectando así a los precios de la vivienda (Nguyen, 2021).
- Políticas monetarias: con el objetivo de lograr y mantener la estabilidad de precios, los bancos centrales manipulan los tipos de interés y la liquidez de los mercados. Estas variaciones en las tasas de interés afectan también al sector inmobiliario, expandiendo o contrayendo demanda y oferta (López, 2005).

En algunos de estos factores, sobre todo en la manipulación del coste del dinero, el papel que desempeña el sistema bancario es de carácter determinante, una muestra más de la interrelación entre en sector financiero y el inmobiliario.

2.1.2. Los préstamos hipotecarios

Debido a la importancia del capital y la normalmente elevada cuantía de las operaciones inmobiliarias, la evolución de la inversión en los bienes inmuebles ya sea por parte de empresas o particulares se fundamenta en la deuda. El papel crítico del apalancamiento al afrontar una transacción de este tipo posiciona a los contratos hipotecarios como la principal fuente de conexión entre el sector financiero y el inmobiliario (Asociación Hipotecaria Española, 2014).

Aunque en muchas ocasiones los conceptos préstamo y crédito se emplean indistintamente, existen ciertas características que los particularizan. Mientras que, en los préstamos, generalmente enfocados a particulares, el total del dinero prestado se entrega al inicio del contrato, los créditos suelen ser solicitados por PYMES (Pequeñas y Grandes

Empresas) o autónomos que pueden ir disponiendo de la cuantía acordada en base a sus necesidades en el tiempo. Sin embargo, tal y como se comentará a continuación, ambos tipos de financiación comparten la principal característica de las operaciones inmobiliarias de apalancamiento: la hipoteca (REALIA, 2018).

De acuerdo con el Banco de España (2013), los préstamos y créditos hipotecarios son contratos mediante los cuales un acreedor (entidades bancarias, cajas de ahorros…) presta una cuantía de dinero a un deudor (individuo, empresa o institución pública) durante un periodo de tiempo determinado, al final del cual la totalidad del capital prestado ha debido ser devuelto. En contraprestación, el deudor debe pagar unos intereses periódicos al acreedor, cuyas características dependen de las condiciones recogidas en el contrato. Sin embargo, la principal singularidad de estos contratos, diferente al resto de préstamos o créditos, es la concesión de una garantía adicional para el acreedor: el propio inmueble. De esta manera, la vivienda se hipoteca en favor del banco, asegurándole la recuperación del capital prestado en caso de que el deudor incumpla con sus obligaciones de pago.

Dada la importante inversión que generalmente supone la compra de un bien inmueble y al perfil del deudor en un contrato hipotecario (en la mayoría de las ocasiones individuos, especialmente si nos enfocamos en el sector inmobiliario residencial), los préstamos y créditos hipotecarios cuentan con plazos de mayor duración al de otras operaciones de apalancamiento. Asimismo, la existencia de la hipoteca como garantía complementaria reduce los tipos de interés de estos préstamos, convirtiendo la compra de inmuebles en operaciones de menor coste de financiación al de otras (Asociación Hipotecaria Española, 2014).

Los préstamos hipotecarios pueden clasificarse en base a diferentes variables. La cantidad de condiciones legales y financieras recogidas en un contrato de este tipo ha llevado a categorizar los préstamos hipotecarios en base a dos *drivers* principales: el tipo de interés y el tipo de cuota (Boedo Vilabella, 2008). La Asociación Hipotecaria Española (2014) determina que existen cuatro tipos principales de préstamos hipotecarios:

- Interés fijo: Como indica su nombre, el tipo de interés se mantiene constante durante la duración del préstamo. La estabilidad de la cuota a pagar por el prestatario aumenta su seguridad ante variaciones en los tipos en la economía. A cambio de ello, en los inicios del contrato el coste de financiación de este tipo de préstamos suele ser superior al de aquellos de interés variable.

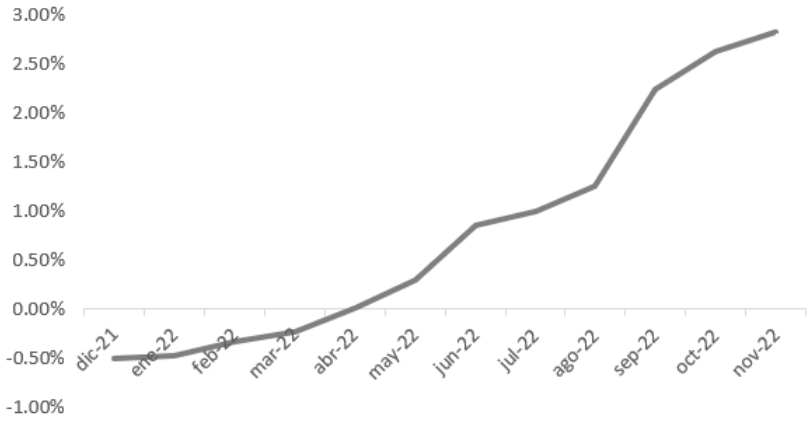
- Interés variable: Estos préstamos están referenciados a un índice, lo cual significa que varían en función de las fluctuaciones de dicho índice. El tipo de interés es revisado anual o semestralmente, y la forma de expresarlo es mediante la suma del índice más un porcentaje (X) fijo (p. ej. EURIBOR + X%). En situaciones de subida de los tipos, esta modalidad perjudica al prestatario, mientras que cuando los tipos bajan el coste de financiación del prestatario lo hace también, aunque esta tasa nunca puede ser negativa.
- Interés mixto: Es una combinación de los grupos mencionados anteriormente, de forma que se aplica un interés fijo durante un tiempo determinado para posteriormente cambiar a un tipo variable (Banco de España, 2019).
- Cuota fija: Se trata de préstamos con tipo de interés variable en los cuales la cuota pagada por el prestatario es constante en el tiempo, siendo el plazo de amortización el que se ve afectado si los tipos de interés fluctúan. La incertidumbre de esta clase de préstamos reside en que el prestatario no conoce la fecha de fin del contrato hipotecario, pues esta depende de factores externos.

2.1.2.1. Influencia de las políticas monetarias en los préstamos hipotecarios

Los mercados inmobiliarios se han visto fuertemente afectados por factores macroeconómicos en los últimos meses. Las políticas monetarias implementadas por el BCE han influido de forma directa en los contratos hipotecarios, un impacto que a su vez se ha visto reflejado en variaciones en la oferta y la demanda de viviendas (Carbó Valverde y Rodríguez Fernández, 2021).

Ante un inicio de 2022 con tipos de interés históricamente bajos, el aumento de la inflación provocado por las excesivas políticas monetarias expansivas implementadas para paliar la crisis de 2008 (Guidi, 2021) y amplificado debido a la guerra de Ucrania y la crisis en la cadena de suministro, ha provocado una importante subida de los tipos de interés por parte del Banco Central Europeo. Como podemos observar en la Figura Nº2, esto ha provocado que índices de referencia como el Euribor se disparen, afectando al precio de las hipotecas (Pulido et al., 2023).

Figura Nº2: Evolución del Euribor mensual

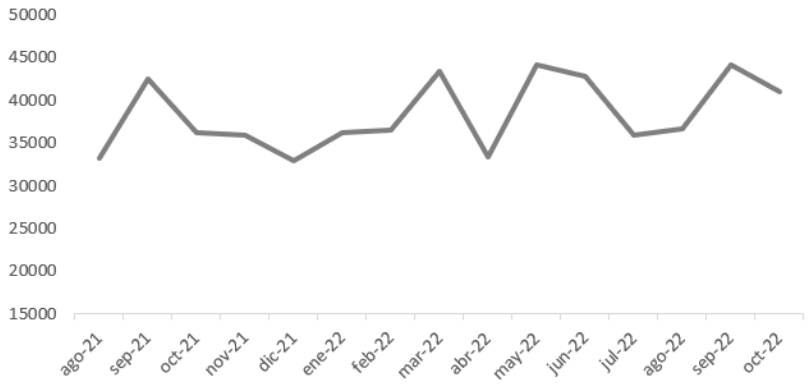


Fuente: Datosmacro (2022), Elaboración propia

A pesar de las históricas políticas monetarias de restricción económica implementadas, las intenciones del BCE de continuar endureciendo la política monetaria con el objetivo de frenar la inflación y reconducirla hacia su valor objetivo del 2% parecen indicar que la subida de tipos de interés se prolongará en el futuro (Schneeweiss, 2022).

Esta amenaza ha comenzado a intimidar a aquellos particulares y empresas que necesitan de un préstamo hipotecario para afrontar la compra de una vivienda. A pesar del incremento en los contratos hipotecarios durante finales de 2021 y principios de 2022, como podemos observar en la Figura Nº3 estos valores no han mostrado crecimiento en los últimos meses. Como se mencionará en los apartados siguientes, en parte se debe a un aumento de las dificultades para la contratación de hipotecas. Sin embargo, otra gran parte de esta expectación se debe al impacto que las medidas implementadas por los organismos gubernamentales y financieros están teniendo en la solicitud de este tipo de contratos. En cualquier caso, ambos factores indudablemente afectarán a la demanda del sector inmobiliario en los próximos meses (Fernández et al., 2021).

Figura Nº3: Hipotecas constituidas sobre viviendas



Fuente: Instituto Nacional de Estadística, INE (2022), Elaboración propia

Ante las negativas perspectivas económicas en el corto-medio plazo tanto particulares como empresas han comenzado a cuestionarse la forma más rentable de financiar sus operaciones. El aumento del esfuerzo hipotecario, entendido este como el importe de la cuota hipotecaria sobre la renta del hogar, es ya un hecho, lo que ha encendido el debate qué alternativa de financiación es la más acertada (Montoriol Garriga, 2022).

Prueba de ello es la transformación que está sufriendo el sector inmobiliario residencial. El crecimiento económico de España postpandemia, la rápida recuperación de los precios de la vivienda después de la crisis sanitaria o la atracción de capital extranjero con el objetivo de liquidar suelos en manos de entidades financieras como Sareb, han convertido la promoción inmobiliaria en España en un sector fuerte en capital. Esto está provocando que la actividad promotora de nuestro país empiece a estar en manos de fondos inmobiliarios cotizados, en los que la proporción de fondos apalancados es muy reducida (Valls, 2022).

Por parte de la demanda las opciones de financiación siguen en la mayoría de los casos centradas en torno a los bancos tradicionales. Si bien están surgiendo nuevas formas de financiación a través de entidades financieras que no dependen de los bancos y de mercados alternativos, donde tanto particulares como empresas pueden llegar a encontrar condiciones competitivas con el apalancamiento hipotecario tradicional, estas instituciones financieras parecen estar lejos de suplir a la deuda bancaria en el corto-medio plazo (KPMG, 2022).

En definitiva, nos encontramos en un momento de enorme incertidumbre para todos los agentes del sector inmobiliario, pero en especial para aquellos envueltos o con intención

de comprometerse en un contrato hipotecario. Sin embargo, mientras que los prestatarios tienen como principal preocupación la fluctuación de los tipos de interés y cómo esta puede afectar a la cuantía de su préstamo, los prestamistas (entidades bancarias) tienen una preocupación que los primeros, excepto en caso de quiebra del banco, no tienen: el riesgo de crédito o impago. Ante las recientes subidas de los tipos de interés, este riesgo parece haber aumentado, con particulares y empresas involucrados en contratos de préstamo de condiciones cada vez más complicadas de afrontar (Riera, 2022).

Ante esta situación, y con el precedente de la crisis financiera del año 2008 desencadenada por el colapso del sistema bancario debido a la titulación de las conocidas como "hipotecas basura" y su posterior impago por parte de los prestatarios, la digitalización, la recopilación y el análisis de datos del cliente son ahora imprescindibles para reducir el riesgo crediticio y favorecer así a la dinamización y la seguridad del sistema bancario.

En el siguiente apartado se estudiarán los préstamos hipotecarios desde la perspectiva del prestamista, es decir, de la entidad bancaria, describiendo las tecnologías y modelos de datos empleados en el análisis de su concesión.

3. BIG DATA EN LOS PRÉSTAMOS HIPOTECARIOS

Desde hace muchos años nos encontramos en un ecosistema en el que el *Data Analytics*, es decir, el dato y la capacidad de tomar decisiones mediante su análisis son factores imprescindibles y presentes en cada una de las acciones del día a día. Aunque no en todos los países y culturas la relevancia de los datos es de la misma dimensión, la globalización y la digitalización han convertido al ser humano en una especie que no solo genera datos de forma masiva, sino que es cada vez más dependiente de ellos.

La ebullición de las redes sociales como fuente de comunicación, información y entretenimiento es uno de los grandes justificantes de la importancia de los datos en el mundo actual. Mediante actividades tan simples y fundamentales para el ser humano como las mencionadas, nuestra especie registra infinidad de información que posteriormente es empleada por empresas e instituciones con finalidades muy variadas. Ahorro, profundidad de conocimiento, mayor seguridad, éxito en la toma de decisiones… estas son algunas de las grandes ventajas que los datos aportan a numerosas compañías y organizaciones a nivel internacional, que ven como sus modelos de negocio y propuestas de valor requieren gradualmente de los datos para satisfacer a sus clientes y diferentes *stakeholders* (Prieto, 2022).

Los beneficios intrínsecos al análisis de datos y el potencial de crecimiento de esta disciplina ante el aumento de la población y la evolución tecnológica, potenciada como consecuencia del COVID-19, han atraído a la gran mayoría de participantes de los diferentes sectores económicos. La salud, el deporte, los seguros o la publicidad son algunos de los sectores que más revolucionados se han visto debido a la aplicación del análisis de datos en sus operaciones (Delgado, 2021).

Sin embargo, el sector bancario no ha quedado rezagado en este proceso de revolución provocado por el *Data Analytics*, ya que es uno de los sectores económicos que más ha implementado este enfoque en su forma de operar. Los bancos se han reinventado, cambiando por completo su funcionamiento y modelo de negocio en comparación con el modelo bancario tradicional. Gracias a los datos generados por sus clientes, la personalización del servicio y la sencillez en el *customer journey* son ahora los grandes pilares de la actividad bancaria a nivel internacional. Tal es la adaptación de este sector a la emergente relevancia de los datos que actualmente casi la totalidad de funciones bancarias se basan, bien de forma completa o bien de forma parcial, en el análisis predictivo (BBVA, 2022).

En la actividad de concesión de préstamos hipotecarios, una de las principales funciones de las entidades bancarias, esta capacidad de recolección y análisis de datos para la predicción de sucesos futuros juega un papel fundamental.

3.1. Revolución tecnológica del proceso hipotecario: el *Hipotech*

La implementación por parte del sistema bancario de disciplinas de análisis de datos se ha visto potenciada por una importante ola de digitalización experimentada por este sector en los últimos años. Como se ha mencionado en apartados anteriores, los sectores económicos han sufrido una transformación digital que ha marcado un antes y un después para un gran número de compañías, que o bien se han visto reforzadas por este fenómeno o han fracasado en su adaptación. La evolución de la tecnología y el cambio de mentalidad del consumidor son los principales factores que han empujado a empresas de todos los sectores a adoptar herramientas digitales, adaptándose a una nueva realidad (Nanney, 2017).

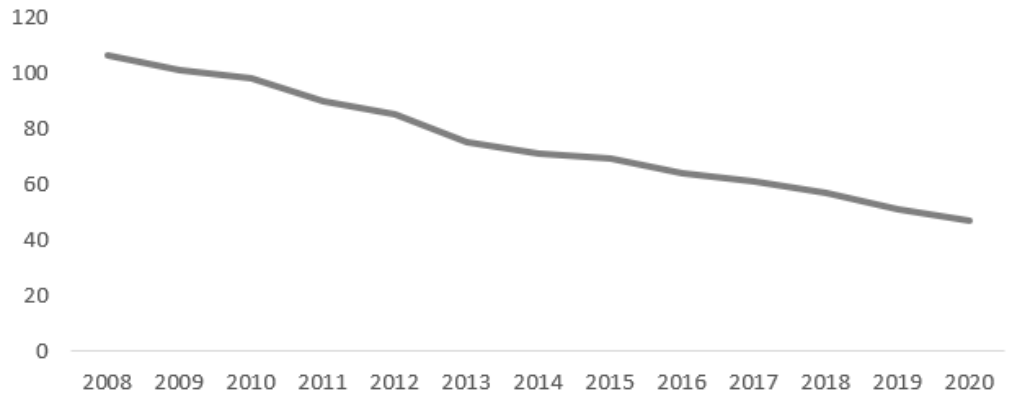
Si bien es cierto que tarde o temprano todos los sectores han ido dejando atrás mentalidades y metodologías tradicionales, no todos se han sumado a la ola digital a la misma velocidad. Probablemente provocado por la tardía adopción y demanda de la

tecnología por parte del consumidor de servicios de banca, el sector bancario ha sido uno de los que más ha tardado en unirse al movimiento en masa de implementación tecnológica en las operaciones. Hasta hace poco tiempo, la tasa de digitalización del consumidor bancario en España era de entre el 35% y el 40%, una de las más bajas en comparación con los principales sectores de nuestro país. Sin embargo, en los últimos años este número ha crecido exponencialmente, alcanzando en el año 2021 valores cercanos al 60% (KPMG, 2021).

La importante polarización cultural de nuestra sociedad, debida a la convivencia entre una fuerte porción de población envejecida y las nuevas generaciones, educadas en un entorno digital, ha supuesto otro gran obstáculo para la transformación tecnológica de este sector. La notable diversidad de nuestro país ha puesto en un conflicto a las principales entidades bancarias, que han tenido que tomar decisiones estratégicas sobre hacia qué modelo de negocio decantarse, sabiendo que tanto un completo enfoque digital como una focalización en la banca tradicional tendrían efectos muy opuestos para los clientes en función de su edad y mentalidad, lo que a su vez afectaría de forma directa en sus resultados económicos. A pesar de este dilema, el sistema bancario de nuestro país ha seguido la estela de aquellas entidades financieras líderes en las grandes potencias económicas a nivel internacional, la transformación hacia el fenómeno conocido como *online banking* (Toloba y del Río, 2021).

Esta revolución en las operaciones del sistema bancario se ha visto reflejada en un cambio en el *front office* del sector. Mientras que tradicionalmente el servicio proporcionado por las entidades financieras se basaba en medios presenciales, su modelo de negocio se ha visto tremendamente modificado. Prueba de ello es la drástica caída en el número de sucursales bancarias en nuestro país, mostrada en la Figura Nº4. Como se puede observar, el número de sucursales bancarias por cada 100.000 adultos se vio reducido en más de la mitad desde el año 2008 hasta el año 2022, una tendencia que se ha visto fomentada debido a los efectos del COVID-19.

Figura Nº4: Nº de sucursales bancarias por cada 100.000 adultos (España)



Fuente: Instituto Nacional de Estadística, INE (2022), Elaboración propia

Ante esta transición tecnológica han surgido numerosos modelos de negocio enfocados en servicios online, con una propuesta de valor clara para clientes jóvenes que desean hacer sus gestiones bancarias en tan solo un *click*. Este aumento en la competencia ha provocado una mayor fragmentación del sector, llevando a gigantes financieros a competir y aliarse con pequeñas y medianas empresas en un entorno regulatorio que parece no acabar de adaptarse a las demandas de los diferentes *players* (KPMG, 2021).

Uno de los modelos de negocio más sonados debido a su rápido crecimiento y aceptación por parte del usuario es el de aquellas empresas conocidas como *Fintech*. Este término, procedente de la abreviación de las palabras *Finance* y *Technology,* engloba a todas las prácticas relacionadas con la implementación de la tecnología y la innovación en la oferta de productos y la prestación de servicios del ámbito financiero, y desde hace tiempo supone una gran amenaza para las entidades bancarias más afianzadas en casi la totalidad de economías a nivel internacional (Comisión Nacional del Mercado de Valores, s.f.).

Bajo el paraguas de las *Fintech* se recoge un gran número de empresas que llevan la innovación tecnológica al sector financiero, agilizando las operaciones para los clientes y ofreciendo un servicio más cómodo. Sin embargo, no todas cubren las mismas necesidades, de forma que en función de la rama financiera que cubren pueden clasificarse en siete categorías (Comisión Nacional del Mercado de Valores, s.f.):

- Gestión patrimonial y asesoramiento: actividades de asesoramiento automatizado, aplicación de la inteligencia artificial y modelos algorítmicos en la toma de decisiones financieras, *social trading*, etc.

- Finanzas personales: digitalización de procesos y de la administración de las finanzas personales del cliente.
- Financiación alternativa: microcréditos, préstamos de corta duración, financiación a través de *crowdfunding*, etc.
- Servicios de pago
- Big Data: aplicación de técnicas de Big Data para la extracción de *insights* y su aplicación a la toma de decisiones para el cliente.
- Identificación *online* de clientes: transformación de parámetros de identificación, véase la inclusión del reconocimiento facial para el acceso a una plataforma.
- Criptoactivos: inversión en activos con tecnología *blockchain* como criptomonedas.

El sector inmobiliario se ha visto afectado por el surgimiento de las *Fintech* y la trasformación que estas han traído consigo. Los servicios hipotecarios, de los más importantes para las entidades bancarias y que tanta burocracia han tenido siempre, han comenzado a proveerse también de forma digital, dando lugar a un nuevo tipo de *Fintech* conocido como *Hipotech* (Utrera, 2019).

Este modelo surgió en el año 2012 en Estados Unidos, y fue adoptado por Reino Unido años después. En estos años, el crecimiento de este tipo de empresas ha sido exponencial, representando el 10% y el 5% respectivamente de las transacciones de los mercados hipotecarios en ambos países. Ante su indudable éxito, en nuestro país han surgido numerosas empresas que replican el modelo de origen americano permitiendo al cliente la contratación de una hipoteca desde su móvil (Utrera, 2019).

El *Hipotech* ofrece a los clientes potenciales de un contrato hipotecario la posibilidad de comparar digitalmente las condiciones hipotecarias ofrecidas por los diferentes bancos en base a sus características, apoyándose en técnicas de Big Data e Inteligencia Artificial. Aunque este servicio tiene un coste para el cliente, este es inferior al ahorro proveniente de la comparación y selección de las mejores condiciones hipotecarias. De esta manera los clientes ahorran, siendo los bancos los que pagan la mayor parte del coste a las *Hipotech* (entre el 0,5% y el 1% de comisión) por recibir clientes de estas (Gastesi, 2022).

A pesar de su beneficiosa propuesta de valor desde el punto de vista del cliente, la contratación de una hipoteca mediante un canal 100% online tiene ciertas implicaciones que han abierto el debate sobre la fiabilidad de este nuevo tipo de *Fintech*. Sin embargo,

la regulación establecida por la Ley 05/2019 de 15 de marzo, que reformó la Ley de Crédito Inmobiliario, obliga a estas empresas a cumplir con una serie de condiciones de cara a contar con la autorización para poder operar de forma legal y segura para el cliente. Gracias a esta reforma legal, la seguridad de este tipo de empresas queda acreditada; prueba de ello es la alianza con los principales bancos de nuestro país, que han visto una oportunidad de atraer clientes confiando en este nuevo modelo de negocio (Helloteca, 2022).

Las cada vez más asentadas *Fintech* y las recientemente surgidas *Hipotech* son un ejemplo del impacto de la digitalización y el Big Data en las operaciones bancarias, y más concretamente en la contratación de hipotecas. Sin embargo, si bien estos modelos de negocio han supuesto una revolución en el *front office* del sistema bancario español, la verdadera transformación del sector bancario y los procesos hipotecarios se produce en el *back office*, algo que los bancos llevan practicando desde mucho antes mediante la implementación de modelos predictivos de inteligencia artificial (García Montalvo, 2014).

3.2. El riesgo de crédito y su papel en los préstamos hipotecarios

Son muchos los casos en los que el sistema bancario ha implementado el *Data Analytics* y los modelos predictivos como motor en la toma de decisiones. La prevención de fraude, la optimización de procesos, la adquisición de clientes potenciales y la búsqueda de oportunidades mediante venta cruzada se han sofisticado notablemente en los últimos años. Sin embargo, una de las grandes áreas que más robustecida se ha visto en los últimos tiempos es la gestión de riesgos (Porras Castaño, 2018).

El papel protagonista del sistema bancario en la financiación en nuestro país es tal que alrededor del 70% de los ingresos de las entidades bancarias españolas corresponde a intereses de préstamos o créditos a empresas y particulares. Esta dependencia de los préstamos y los ingresos asociados a ellos convierte al riesgo de crédito en una de las grandes preocupaciones de las entidades participantes de este sector (Maudos, 2022).

Cuando nos referimos al riesgo de crédito, son varios los tipos que nos podemos encontrar:

- Riesgo de incumplimiento del crédito o riesgo de impago: como su propio nombre indica, es el riesgo asociado al incumplimiento del contrato por motivos comerciales por parte del prestatario. Este se puede ver afectado por diferentes

motivos, como variaciones económicas o alteraciones en la situación financiera del mismo (Kagan, 2020).

- Riesgo de concentración: este riesgo surge ante la falta de diversificación de destinatarios a los cuales los fondos son destinados. En estos casos, el prestamista tiene una alta dependencia de unos pocos o incluso de un único tipo de deudor, aumentando así sus probabilidades de incurrir en pérdidas si el deudor fracasa en el cumplimiento del contrato firmado (Corporate Finance Institute, 2022).
- Riesgo-País: surge ante la insolvencia de los prestatarios de un determinado país provocada por circunstancias ajenas a motivos comerciales, véase factores sociales, políticos, macroeconómicos o naturales. Se produce en contratos de deuda extranjera (Iranzo, 2008).

Como se ha mencionado anteriormente, el riesgo de impago por parte del prestatario cobra especial relevancia en los préstamos hipotecarios. La variedad de perfiles del solicitante, la elevada cuantía financiada y las características contractuales de este tipo de préstamos hace que la adecuada clasificación de los clientes en base a la evaluación de su capacidad de pago sea determinante para garantizar la seguridad de una operación de tal dimensión (Galeano y Peña, 2019).

Por este motivo, son muchas las variables que las entidades bancarias consideran para la concesión de fondos mediante un contrato de préstamo. En concreto, los principales criterios de juicio en el caso de préstamos hipotecarios son (Banco de España, 2013):

- Datos personales del prestamista: edad, estado civil, situación personal y laboral...
- Situación económica: los ingresos percibidos, tanto principales como adicionales, información de otros préstamos activos del prestamista o el grado de endeudamiento del prestamista son datos determinantes para las condiciones de la concesión del préstamo hipotecario.
- Registro de morosidad: para ello, las entidades bancarias consultan el historial del prestamista en la Central de Información de Riesgos del Banco de España (CIRBE).
- Registro de bienes en propiedad: a través del Registro de la Propiedad, las entidades bancarias consultan los bienes muebles e inmuebles registrados bajo la identidad del solicitante del préstamo con el objetivo de completar la construcción de su perfil hipotecario.

Si bien el riesgo de crédito siempre ha constituido uno de los principales motivos de preocupación del sector bancario, la pandemia del COVID-19 y las medidas implementadas por los bancos centrales para paliar sus efectos han empujado a las entidades bancarias a adoptar medidas de prevención aún más estrictas (PwC, 2021).

Esta forma de actuación adoptada por el sistema bancario no es de extrañar. Durante épocas de expansión, las dificultades y la dureza de las condiciones crediticias de los prestamistas son limitadas. Cuando la economía se encuentra en un momento de bonanza, la confianza en el positivo devenir de las inversiones, así como en la capacidad de los prestatarios de devolver la cuantía prestada, es máxima. Esto provoca que proyectos menos atractivos desde el punto de vista del prestamista sean financiados de forma más liberal que en situaciones de desaceleración económica. Precisamente, ante desaceleraciones como la que afecta al mundo occidental en la actualidad, las entidades de crédito tienden al conservadorismo, endureciendo las condiciones de los préstamos, así como los criterios para su concesión. Durante estos tiempos de retroceso en la economía, únicamente aquellos perfiles considerados como crediticiamente seguros acceden a la reducida oferta de fondos (Jiménez y Saurina, 2006).

3.3. El *credit scoring* como herramienta de determinación de la calidad crediticia del cliente de un préstamo hipotecario

Son muchas las técnicas y herramientas que el sistema bancario ha empleado a lo largo de su historia en la toma de decisiones de financiación. Sin embargo, la herramienta empleada en la actualidad es conocida como *credit scoring*. De acuerdo con su definición, el *credit scoring* es "todo sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito" (Puertas Medina y Martí Selva, 2013). Tal y como su nombre indica, durante este proceso de análisis de riesgo se simplifican los datos del cliente en un único valor, el *credit score*, que representa la solvencia del cliente en base a la probabilidad de que este no cumpla con el contrato crediticio, incurriendo en impago (García Montalvo, 2014).

De esta forma, la medición de la capacidad de pago del prestamista se puede expresar empleando la expresión matemática que se muestra a continuación (Puertas Medina y Martí Selva, 2013):

P = f(x1, x2,..., xk) + e , donde:

- xi: son las características del sujeto, recogidas por la entidad bancaria en base a las dimensiones mencionadas en el apartado *3.2. El riesgo de crédito y su papel en los préstamos hipotecarios.*
- f(x): es la función mediante la cual se determina la relación entre las variables xi utilizadas.
- e: referido a la perturbación aleatoria.
- P: la variable dependiente, que representa la probabilidad de que el crédito resulte fallido, también conocido como *credit score.*

Los modelos de *credit scoring* son ampliamente utilizados por el sistema bancario en la toma de decisiones financieras. Mediante esta técnica, se puede llevar a cabo la evaluación crediticia de cualquier tipología de cliente. Si bien es cierto que particulares, Pequeñas y Medianas Empresas o incluso clientes corporativos pueden ser financieramente mapeados con estos modelos, en el caso de clientes *corporate* no es la técnica más utilizada para su calificación crediticia. Las diferencias existentes entre las variables a considerar, más cualitativas y de difícil estandarización para clientes corporativos que para comerciales, ha llevado a las entidades bancarias a emplear diferentes herramientas para el primer grupo de clientes, que en lugar de tratarse de una clasificación otorgan una calificación al prestatario mediante sistemas de *rating* (Gutiérrez Girault, 2007).

3.3.1. Origen y evolución del *credit scoring*: el impacto del Big Data

La evaluación del riesgo crediticio del prestatario no siempre se ha basado en las mismas características usadas por el sistema bancario en la actualidad. Antes de la invención de los sistemas de *credit scoring*, los prestamistas evaluaban a los consumidores de créditos en base a aspectos de carácter cualitativo y subjetivo (Paul, 2023).

La inexistencia de medidas legales de protección del consumidor permitía a las entidades de crédito la recopilación y la comercialización de toda información a la que pudieran acceder, algo que hoy en día está mucho más controlado. De esta forma, el prestamista contaba con una posición de clara superioridad, pudiendo basar sus decisiones financieras en criterios como la raza, la religión, la nacionalidad, el sexo o incluso rumores sobre la vida personal del consumidor (Gallo, 2022).

La discriminación según el perfil del prestatario era mucho más notable que en la actualidad. Sin embargo, estas características no eran suficientes para la correcta

identificación del nivel de riesgo crediticio y la adecuada clasificación de los clientes en base a su capacidad de cumplimiento del contrato de préstamo. La tardía digitalización de los procesos y herramientas de este sector convertía la recolección de una masa de datos suficiente como para asegurar el éxito en la toma de decisiones en una tarea de gran complicación, lo cual provocaba una notable probabilidad de fracaso desde el punto de vista del prestamista (Porras Castaño, 2018).

No fue hasta mediados del siglo XX cuando la industria crediticia empezó a implementar el enfoque estadístico en la toma de decisiones. Las herramientas de computación comenzaron a evolucionar, y la cantidad de información disponible empezaba a ser lo suficientemente abundante y variada como para llevar a cabo un mejor diagnóstico del perfil crediticio del cliente (Paul, 2023).

Ante el crecimiento exponencial de las solicitudes de crédito, en el año 1956 William Robert Fair y Earl Judson Isaac, un ingeniero y un matemático estadounidenses, decidieron asociarse para la creación de Fair, Isaac and Company, actualmente conocida como FICO. Su objetivo era crear una empresa que estandarizase la determinación del riesgo crediticio de los consumidores de cuentas de crédito, considerando variables objetivas y aplicando modelos estadísticos en el proceso (Altman, 2023).

De esta forma, FICO proponía la creación del hoy en día conocido como FICO Score, que es calculado con la información disponible en los informes crediticios del cliente. Sin embargo, esto no significa que los prestamistas solo consideren la información recogida por este parámetro, ya que el FICO Score puede ser complementado con información adicional que el prestamista considere. De acuerdo con FICO (2018), los criterios considerados y su importancia en la determinación de la fiabilidad crediticia del consumidor son:

- Historial de pagos (35%): la información histórica sobre el comportamiento del cliente en sus contratos de préstamo es determinante para la predicción de su comportamiento futuro. Por este motivo, este es el principal factor considerado para el cálculo del FICO Score.
- Nivel de endeudamiento (30%): aunque tener una mayor cantidad de deuda no siempre indica un mayor nivel de riesgo de impago que el que pueda tener un cliente con menor cantidad de deuda, sí debe ser un motivo a considerar dado que puede indicar sobreendeudamiento.

- Duración del historial de crédito (15%): aunque no es un requisito necesario para obtener un buen FICO Score, los clientes con historiales más alargados cuentan con una ventaja sobre aquellos con historial más corto. La antigüedad de las cuentas de crédito del cliente (desde la más antigua hasta la más nueva) así como la duración de aquellas cuentas abiertas de forma puntual son consideradas para el diagnóstico del perfil crediticio del cliente.
- *Mix* crediticio (10%): todas aquellas cuentas, operaciones o contratos financieros vigentes y pasados son consideradas para el cálculo del FICO Score. Cuentas de crédito, cuentas en entidades financieras, préstamos de consumo e hipotecarios son incluidos en dicho *mix*.
- Nuevo crédito (10%): el número y velocidad con la que se abren nuevas cuentas determinan el riesgo crediticio de un cliente. El FICO Score considera este factor como el de menor importancia, sin embargo, su efecto puede ser notable.

A pesar del inicial rechazo de la industria crediticia ante el revolucionario sistema planteado por Fair e Isaac, durante la década de los 90 la mayoría de las entidades crediticias de Estados Unidos ya empleaban los estándares de FICO en la determinación del perfil de riesgo de sus clientes (Altman, 2023).

Sin embargo, no fue hasta años más tarde cuando estas técnicas de *credit scoring* fueron implementadas en la gestión de préstamos hipotecarios, cuando en 1990 dos de las principales empresas de préstamos en EEUU, Freddie Mac y Fannie Mae, decidieron aplicar estas metodologías a sus sistemas de concesión de hipotecas (Paul, 2023).

Desde entonces el *scoring* de crédito no ha hecho más que evolucionar. El crecimiento exponencial de la digitalización y, consecuentemente, la generación masiva de datos por el ser humano ha permitido a las entidades de crédito sofisticar sus modelos con el tiempo. Como se ha mencionado anteriormente, esto ha tenido un fuerte impacto en el mercado hipotecario, revolucionando los procesos de concesión de hipotecas y aumentando las complicaciones para acceder a ellas desde el punto de vista del consumidor (Ganuza Fernández, 2020).

3.3.1.1. Implicaciones reglamentarias del Big Data y el *social scoring*

A pesar del papel indispensable que el Big Data ha jugado para el desarrollo de los sistemas de *credit scoring* hasta lo que conocemos en la actualidad, la existente cantidad de datos es de tal dimensión que está comenzando a complicar este proceso para las

entidades bancarias. Geolocalización, redes sociales, transacciones bancarias… Son muchas las fuentes que hoy en día nutren a estas entidades de información de sus clientes, una variedad que, sumada a la alta velocidad con la que nuevos datos son generados, está provocando un exceso de información que es necesario clasificar y seleccionar (Zárate, 2018).

Las redes sociales e internet son dos de los principales justificantes del exceso de información existente en la actualidad. Mediante nuestra aceptación de las políticas de todas las plataformas y portales que utilizamos, damos acceso a numerosas empresas a datos tanto personales como de comportamiento. Las consecuencias de esto son cada vez más tangibles y el control sobre nuestras preferencias es cada mayor. Prueba de ello es la clasificación de los consumidores de cara a practicar la venta online, por medio de la cual las empresas nos segmentan en grupos con consumidores de características, preferencias y comportamientos similares (Botella, 2021).

Aunque el sector financiero ya emplea este tipo de datos para prácticas similares, como la oferta y publicidad de productos o servicios, han empezado a surgir movimientos y disciplinas que pretenden implementar la información que estas plataformas recogen en la determinación del riesgo crediticio de los ciudadanos (Ganuza Fernández, 2020).

Tal y como hemos mencionado anteriormente, los datos que generamos en las redes sociales quedan fuera de las variables consideradas para la determinación de nuestro *credit scoring*. Sin embargo, ante la propuesta del *social scoring* las interacciones y comportamientos online se tendrían en cuenta como variables de predicción de la fiabilidad del cliente y su riesgo de impago. Ante el peligro que propuestas como esta presentan para la sociedad las entidades reguladoras han comenzado a movilizarse con el objetivo de aumentar la protección del ciudadano. En este sentido, la Comisión Europea presentó en abril de 2021 la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de Inteligencia Artificial (ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión (Galeano, 2021).

Mediante esta propuesta, la Comisión Europea pretende establecer un marco legal inexistente actualmente, que regule las aplicaciones de la inteligencia artificial. En ella, se establecen una serie de supuestos en los que la implementación de la inteligencia

artificial no es legal o es considerada de alto riesgo, en los cuales se recoge la práctica del *credit scoring*:

> En concreto, deben considerarse de alto riesgo los sistemas de IA usados para evaluar la calificación crediticia o solvencia de personas físicas, ya que deciden si dichas personas pueden acceder a recursos financieros o servicios esenciales como la vivienda, la electricidad y los servicios de telecomunicaciones. (Comisión Europea, 2021)

La reacción de la Unión Europea en la limitación del *credit scoring* y la regulación de las aplicaciones de la inteligencia artificial es solo un ejemplo de los retos y dificultades que el Big Data y la implementación tecnológica van a plantear en el futuro. En este sentido, la agilidad de las entidades reguladoras ante los rápidos cambios vividos en el sector financiero, entre otros, será determinante para la seguridad y la adecuada convivencia entre el Big Data y las actividades desempeñadas por el ser humano en el día a día (Galeano, 2021).

Tal y como se ha podido comprobar, el protagonismo de los modelos de *credit scoring* en la determinación del riesgo de impago del prestatario de un crédito lleva creciendo de forma notable en los últimos años. Además, la repercusión que este proceso tiene en los mercados inmobiliarios y la economía de nuestro país es muy notable. Por ello, resulta de gran relevancia el estudio y la aplicación práctica de algunos de los conceptos explicados con el objetivo de materializar mediante datos el impacto y la efectividad que el Big Data ha tenido en los préstamos hipotecarios.

4. APLICACIÓN PRÁCTICA DEL MACHINE LEARNING EN LA PREDICCIÓN DEL SUCESO DE IMPAGO EN PRÉSTAMOS HIPOTECARIOS

Como se ha mencionado en apartados anteriores, la revolución de los procesos de gestión y concesión de préstamos se fundamenta en dos factores principales: la evolución de las técnicas de computación y la creciente disponibilidad de los datos. Esto convierte a las técnicas de *Data Analytics* en las protagonistas en un entorno cada vez más digitalizado y conservador ante las incertidumbres.

Más allá de la explicación teórica del impacto de estas herramientas en los préstamos hipotecarios y la transmisión de su impacto en los mercados inmobiliarios, en el presente

apartado se procederá a la aplicación de la práctica para el análisis de su efectividad. Las herramientas de *credit scoring* llevan liderando la toma de decisiones en cuanto a gestión del riesgo de crédito desde hace ya décadas, pero aún existe gran opacidad en las características de los modelos empleados, que varían en función de la entidad bancaria estudiada. Por ello, se pretende construir un modelo de *credit scoring* que ofrezca una alta precisión en la predicción del suceso de impago mediante el apoyo en modelos de *machine learning.*

4.1. Objetivos

En este apartado se pretende cubrir el objetivo principal de este trabajo, el cual consiste en llevar a cabo un estudio empírico del impacto del Big Data en la concesión de préstamos hipotecarios, midiendo este como la efectividad y precisión de los modelos que se apoyan en los datos masivos en la identificación del suceso de impago por parte del solicitante de una hipoteca.

Para ello, son dos los principales objetivos técnicos que se pretenden lograr en este ejercicio:

- El estudio de la interrelación entre las principales variables socioeconómicas del solicitante de un préstamo hipotecario, así como su influencia en la gran preocupación por parte de los prestamistas: la determinación del acaecimiento del suceso de impago. Como se ha descrito en apartados anteriores, si bien es cierto que las variables predictivas empeladas por las entidades bancarias pueden diferir en cierta manera de una a otra entidad, en la inmensa mayoría de las ocasiones estas suelen ser las mismas. En este sentido, el análisis descriptivo de estas variables, su correlación y su interacción en la determinación de la variable de salida del modelo, el acaecimiento o no del suceso de impago, cobra un gran interés.
- El desarrollo de un modelo predictivo de *machine learning* que emule la acción desempeñada por los modelos de *credit scoring* empleados por las entidades bancarias en la determinación del acaecimiento del suceso de impago de un préstamo hipotecario. En el proceso de concesión de un préstamo hipotecario, las entidades crediticias llevan a cabo un proceso similar al que se llevará a cabo en el presente apartado. En este sentido, se busca lograr la construcción de un algoritmo predictivo que ofrezca una precisión superior a un 80% en la

determinación del acaecimiento del suceso mencionado, tal y como se contempla en la hipótesis del presente trabajo.

4.2. Metodología

Son varias las herramientas y técnicas de procesamiento de datos que serán empleadas en la consecución de los objetivos mencionados. Sin embargo, todas ellas se basarán en el empleo de R como lenguaje de programación fundamental para el análisis práctico. En concreto, se utilizará RStudio, el entorno de desarrollo integrado para el lenguaje de programación mencionado, que ofrece una alternativa gratuita de software *open source* de una elevada capacidad de procesamiento. El motivo de la elección de RStudio como herramienta computacional para la construcción del modelo predictivo es, más allá de mi conocimiento sobre el funcionamiento del programa, la amplia oferta de recursos estadísticos e instrumentos de representación gráfica que este sistema acoge, que serán determinantes para el adecuado reflejo de las ideas que se pretenden transmitir.

De esta manera, se seguirá el siguiente proceso:

1. Selección de una base de datos: por medio de kaggle, la plataforma colaborativa de bases de datos más grande del mundo, se accederá a una muestra de datos de solicitantes de préstamos hipotecarios. Las dimensiones y las características de esta muestra deberán ser tales que permitan la aplicación de técnicas y modelos predictivos de *machine learning*.
2. Preprocesamiento: tras la identificación de la muestra, se procederá al preprocesamiento de los datos, un paso determinante para la facilitación de los procesos que se llevarán a cabo en adelante. En este sentido, las variables de la base de datos original que no aporten al objetivo del análisis serán eliminadas o transformadas.
3. Análisis exploratorio: en esta etapa del estudio práctico se procederá al análisis de las variables que se emplearán en los modelos predictivos, identificando conexiones y extrayendo *insights* de su naturaleza. Asimismo, por medio de este análisis se obtendrá un mayor conocimiento sobre el perfil del solicitante de un préstamo hipotecario en la actualidad, los segmentos bajo los cuales se agrupan estos perfiles y el riesgo de crédito que estos suelen llevar asociado.
4. Transformación de variables: tras le ejecución del análisis exploratorio, se procederá a la transformación de variables. Los modelos predictivos de *machine learning* requieren de cierta estructura y estandarización en las variables de

entrada, por lo que procesos como dicotomizar ciertas variables serán realizados en esta etapa.

5. Partición de los datos en *training* y *test* sets y selección de predictores: en esta etapa se dividirá la base de datos de entrada de los modelos en dos sets, uno con el cual se entrenarán y otro sobre el que se implementarán los modelos entrenados y se medirá la precisión de los mismos. Además, se seleccionarán aquellas variables consideradas como más representativas y de mayor relevancia para el entrenamiento de los modelos predictivos.
6. Entrenamiento de modelos predictivos, *testing* y evaluación de resultados: una vez conocidas las variables de mayor relevancia, se procederá al entrenamiento de los modelos predictivos de *machine learning*, cuyas funcionalidades se explicarán a medida que se vayan implementando. Con los modelos ajustados a los datos de entrenamiento, se procederá a su aplicación al *test* set. De esta manera, se evaluará la precisión de cada uno de ellos y se confrontarán los errores obtenidos.
7. Comparación de modelos: En esta última etapa los modelos y sus resultados serán comparados, determinando cuál de ellos se adapta mejor al caso de estudio.

4.3. Aplicación práctica del *machine learning* en los préstamos hipotecarios

4.3.1. Selección de la base de datos

Tras analizar diferentes fuentes y bases de datos de posible aplicación, se ha seleccionado un *dataset* de Home Credit, una entidad bancaria europea con sede en Países Bajos y que, entre otros servicios financieros, ofrece financiación hipotecaria a sus clientes.

La base de datos seleccionada cuenta con 122 variables y 307.5011 observaciones de solicitudes de hipotecas. Entre las variables recogidas, podemos encontrar la edad, la cuantía del préstamo, la ocupación del solicitante o la puntuación crediticia recibida por entidades externas (lo cual podría ser un equivalente del FICO Score). Sin embargo, la variable de mayor importancia para los análisis que se llevarán a cabo a continuación es la variable denominada como TARGET, que es el acaecimiento o no del suceso de impago por parte del solicitante del préstamo hipotecario. Esta variable, de carácter dicotómico, cobra el valor 0 cuando el solicitante es capaz de afrontar sin dificultades los pagos de la hipoteca contratada, mientras que para los casos encontrados en la situación contraria el valor de esta variable será igual a 1.

4.3.2. Preprocesamiento de los datos

Como paso previo al tratamiento de los datos, se procede a la realización de un preprocesamiento o limpieza de los datos, eliminando las variables repetidas y aquellas consideradas de poco valor o de difícil interpretación. Más allá de la complejidad de entrenar un modelo predictivo empleando variables de dudoso significado y valor, el procesamiento de 120 variables y más de 300.000 observaciones resulta de un tremendo peso incluso para programas computacionales como RStudio.

Por ello, es necesaria una reducción considerable tanto de las variables como de las observaciones contempladas. En este sentido, un gran número de variables son consideradas de reducida utilidad. Es el caso de las más de 30 variables indicadoras del tipo de documentación entregada o las más de 50 variables descriptivas del área de domicilio del solicitante en el momento previo a la obtención de la hipoteca. En cuanto a la reducción de observaciones, se ha decidido llevar a cabo una reducción de registros de la base de datos de entrada por medio de una selección aleatoria de alrededor de un décimo de las observaciones totales (se empleará una muestra de 35.962 observaciones), disminuyendo así la densidad de la base de datos de manera considerable.

Además de la reducción de variables y observaciones, existe la necesidad de llevar a cabo un proceso de transformación de variables. En concreto, son dos aquellas que requieren de este proceso: *AGE* y *YEARS_ID_PUBLISH*, variables de carácter temporal que deben ser convertidas de frecuencia diaria a frecuencia anual.

Una vez llevada a cabo la eliminación y la transformación de las variables mencionadas contamos con una base de datos de un tamaño más reducido. El *dataset* está ahora formado por 19 variables tanto socioeconómicas como crediticias, una masa de una practicidad mucho mayor a la de la base de datos inicial, lo cual facilitará la aplicación de modelos de *machine learning*.

4.3.3. Análisis exploratorio

Como paso posterior al preprocesamiento de datos descrito en el apartado anterior, se procede a la realización de un análisis exploratorio de la base de datos final con el objetivo de comprender las variables recogidas y las interconexiones existentes entre ellas.

Como se ha mencionado anteriormente, el *dataset* de entrada que será empleado para el entrenamiento y test de modelos predictivos está constituido por 19 variables cuyo significado y tipología son explicados en la Tabla Nº1:

Tabla Nº1: Variables del dataset de entrada

Variable	Tipo	Descripción
TARGET	Cuantitativa Binaria	Variable objetivo (1 - cliente con dificultades de pago: tuvo retrasos en el pago de más de X días en al menos uno de los primeros Y plazos del préstamo en nuestra muestra, 0 - todos los demás casos)
NAME_CONTRACT_TYPE	Cualitativa Binaria	Identificación si el dinero concedido es un préstamo o un crédito
CODE_GENDER	Cualitativa Binaria	Género del solicitante del préstamo
FLAG_OWN_CAR	Cualitativa Binaria	Y- si el cliente tiene un coche. N- si el cliente no tiene un coche
FLAG_OWN_REALTY	Cualitativa Binaria	Y- si el cliente ya tiene una casa a su nombre. N- si el cliente no tiene aún una casa a su nombre
AMT_INCOME_TOTAL	Cuantitativa Discreta	Ingresos del cliente
AMT_CREDIT	Cuantitativa Discreta	Importe del crédito
AMT_ANNUITY	Cuantitativa Discreta	Anualidad del préstamo
AMT_GOODS_PRICE	Cuantitativa Discreta	Precio del bien inmueble para el cual es concedido el préstamo
NAME_INCOME_TYPE	Cualitativa No Ordenable	Tipo de ingresos del cliente (empresario, trabajador, baja por maternidad…)
NAME_EDUCATION_TYPE	Cualitativa No Ordenable	Nivel de estudios más alto alcanzado por el cliente
NAME_FAMILY_STATUS	Cualitativa No Ordenable	Situación familiar del cliente
AGE	Cuantitativa Discreta	Edad del cliente en años en el momento de la solicitud
YEARS_ID_PUBLISH	Cuantitativa Discreta	Cuántos años antes de la solicitud cambió el cliente el documento de identidad con el que solicitó el préstamo
CNT_FAM_MEMBERS	Cuantitativa Discreta	Cuántos miembros de la familia tiene el cliente
REGION_RATING_CLIENT	Cuantitativa Discreta	Calificación de Home Credit de la región donde vive el cliente (1,2,3)

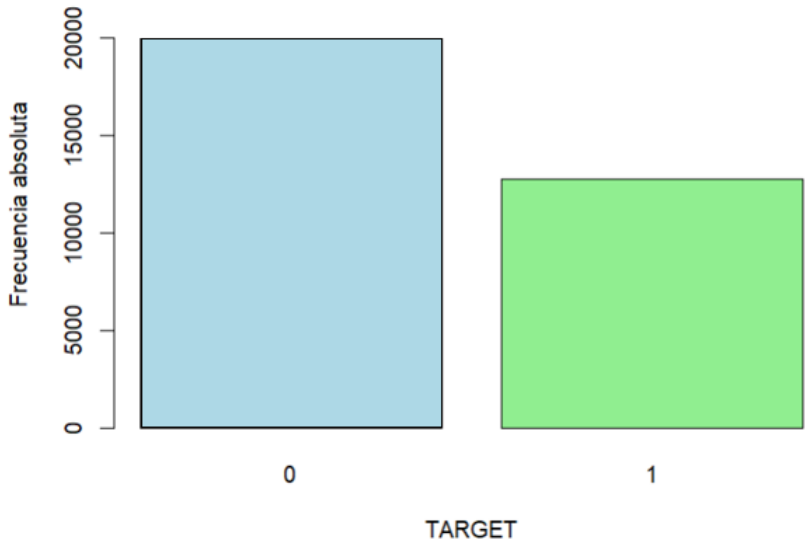
WEEKDAY_APPR_PROCESS_START	Cualitativa No Ordenable	En qué día de la semana solicitó el cliente el préstamo
HOUR_APPR_PROCESS_START	Cuantitativa Discreta	A qué hora aproximadamente solicitó el préstamo el cliente
EXT_SOURCE_2	Cuantitativa Continua	Puntuación normalizada de una fuente de datos externa sobre el riesgo de crédito del solicitante del préstamo (mayor número=mayor riesgo)

Fuente: Elaboración propia

Tal y como se ha adelantado previamente, de todas las variables explicadas en la tabla, 18 de ellas serán independientes mientras que la variable dependiente será aquella denominada como TARGET.

Como siguiente fase del análisis exploratorio, se efectúa un estudio de esta variable dependiente. Tras un primer análisis, tal y como se puede observar en la Figura Nº5, la distribución de las 35.962 observaciones entre los valores que esta variable puede tomar (0 y 1) cuenta con una proporción desbalanceada. De todas las observaciones del *dataset* alrededor del 64% toman el valor 0 (el solicitante del préstamo no tiene dificultades de pago) y el 36% el valor 1 (acaecimiento del suceso de impago). Esta desproporción en el reparto de las observaciones será considerada en adelante, sobre todo a la hora de llevar a cabo la partición de la base de datos entre el *training* y el *test* set.

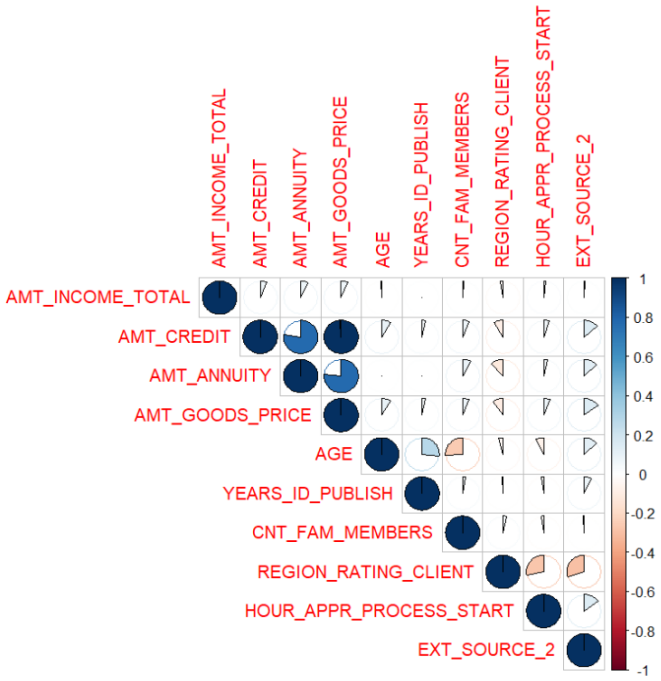
Figura Nº5: Proporción de sucesos de impago en el *dataset* de entrada



Fuente: Elaboración propia

Una vez realizado el análisis exploratorio de la variable TARGET, que complementaremos más adelante, resulta de gran relevancia el análisis de correlaciones entre las diferentes variables contenidas en la base de datos. Únicamente aquellas de carácter numérico / cuantitativas podrán ser empleadas para este análisis.

Figura Nº6: Mapa de correlaciones



Fuente: Elaboración propia

Como podemos observar en la Figura Nº6, en la mayoría de las ocasiones la correlación existente entre las diferentes variables es, aunque leve, positiva (color azul). Así, variables como la cuantía del préstamo (AMT_CREDIT), el bien inmueble financiado (AMT_GOODS_PRICE) y la anualidad del préstamo (AMT_ANNUITY) están correlacionadas fuerte y positivamente entre sí, de forma que cuanto más alto es el valor de una de las variables también lo es el de las restantes.

Sin embargo, en esta figura podemos observar cómo hay ciertas variables negativamente correlacionadas entre sí. Es el caso de la valoración del domicilio del solicitante (REGION_RATING_CLIENT) con la valoración crediticia de fuentes externas (EXT_SOURCE2) o la hora en la que se efectuó la aplicación al préstamo (HOUR_APPR_PROCESS_START), que crecen a medida que la primera disminuye.

Precisamente este análisis nos permite tomar consciencia de la coherencia y las relaciones entre las variables del *dataset*, llevando a cabo un primer estudio que será de gran utilidad para la posterior aplicación de los modelos.

4.3.4. Adecuación de la base de datos

Tras la realización del análisis exploratorio de la base de datos, se continúa con la limpieza del *dataset* para la facilitación del proceso de selección de las variables predictoras de los modelos.

Este proceso está constituido de tres etapas:

- **Tratamiento de valores ausentes (NA):** Tal y como podemos observar en la Tabla Nº2, ciertas variables tienen observaciones con valores ausentes. En concreto, existen 115 valores ausentes distribuidos en 2 variables de la base de datos. Cuando nos encontramos con un problema de este tipo existen dos principales alternativas: inferir los valores perdidos en función del resto de observaciones o eliminarlos. Dada la densidad de observaciones de nuestro *dataset* de entrada se decide eliminar los NA.

Tabla Nº2: Variables con valores ausentes

Variable	Número de NA
AMT_GOODS_PRICE	28
EXT_SOURCE_2	87

Fuente: Elaboración propia

Tras la eliminación de los NA nuestra base de datos se ha visto reducida en 115 observaciones, estando ahora constituida por 35847 filas y 19 columnas.

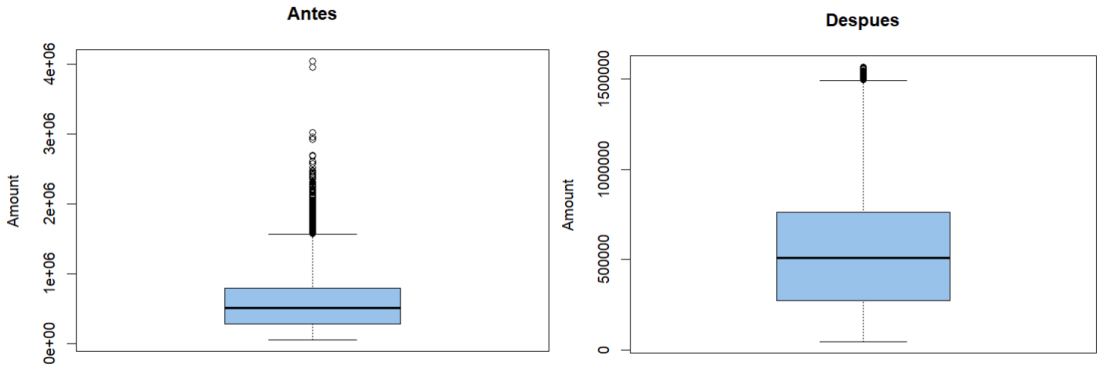
- **Tratamiento de *outliers:*** Los *outliers*, comúnmente conocidos como valores atípicos, distorsionan la imagen y las métricas de un dataset en concreto. Sin embargo, si bien es cierto tienen una gran fuerza para influir en valores como la media o la mediana, estos valores no deben ser eliminados. Por el contrario, se deben reemplazar estos valores en función del resto de registros, con el objetivo de quitarles peso.

  Aunque existen diferentes formas de llevar a cabo este reemplazo, en este caso se ha empleado el rango intercuartílico para la ejecución de esta tarea, otorgando el

valor de la mediana de cada variable a los *outliers* mayores al límite superior de este rango y el valor de la media de cada variable a los valores menores al límite inferior.

En la Figura Nº7 se puede observar el impacto de este proceso con un ejemplo de la variable representativa de la cuantía del préstamo.

Figura Nº7: Análisis *outliers* AMT_CREDIT



Fuente: Elaboración propia

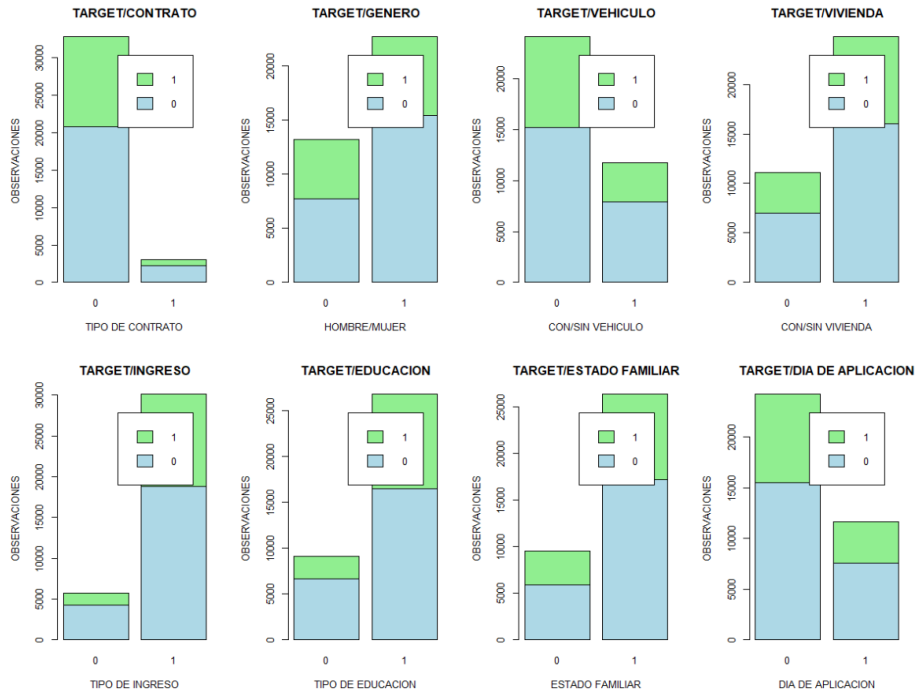
- **Transformación de variables:** Posterior a la eliminación de valores ausentes y atípicos se procede a la transformación de variables del *dataset*. Para ello se han llevado a cabo dos procesos: dicotomizar variables cualitativas y normalizar y centrar las variables numéricas.

  En concreto han sido 8 variables las que han sido transformadas en binarias, al ser todas ellas de carácter cualitativo con una gran variedad de opciones para las observaciones de la base de datos. Por ello, se ha optado por dicotomizar todas ellas, agrupando el abanico de alternativas de cada variable en dos valores: 0 y 1. Las variables transformadas han sido:

  1. NAME_CONTRACT_TYPE: Consituida por *Cash* (0) y *Revolving* (1) loans
  2. CODE_GENDER: Constituida por Hombres (0) y Mujeres (1)
  3. FLAG_OWN_CAR: Constituida por Sin coche (0) y Con coche (1)
  4. FLAG_OWN_REALTY: Constituida por Sin casa (0) y Con casa (1)
  5. NAME_INCOME_TYPE: Constituida por Sin empleo (0) y Con empleo (1)
  6. NAME_EDUCATION_TYPE: Constituida por Nivel Secundaria (1) y Nivel mayor a secundaria (0)
  7. WEEKDAY_APPR_PROCESS_START: Constituida por Entre semana (0) y Fin de semana (1)
  8. NAME_FAMILY_STATUS: Constituida por Soltero (0) y Con pareja (1)

Dicotomizar estas variables permite, por un lado, llevar a cabo un análisis de mayor profundidad de la variable TARGET, tal y como podemos observar en la Figura Nº8. De acuerdo con este nuevo análisis, podemos observar cómo la mayoría de los sucesos de impago se concentran en *cash loans*, mujeres, solicitantes sin vehículo, solicitantes con casa, solicitantes sin empleo (entre los que se incluyen pensionistas o personas en baja de maternidad), clientes que solo han estudiado hasta secundaria, clientes con pareja y solicitudes realizadas entre semana. Por otro lado, esta transformación facilitará la aplicación de modelos predictivos al reducir la complejidad de las variables contempladas.

Figura Nº8: Proporción de sucesos de impago por variable



Fuente: Elaboración propia

Además, se han estandarizado y centrado las variables numéricas del *dataset*, algo imprescindible para la aplicación de modelos predictivos. Por medio de este proceso conseguimos que las variables cuantitativas tengan media 0 y se encuentren todas en la misma escala.

4.3.5. Partición de los datos en *training* y *test* sets y selección de predictores

A continuación, se procede a la partición del dataset de entrada en el set de *test* y el set de entrenamiento. Aunque no existe una forma exacta de llevar a cabo esta división, se suele emplear un método porcentual por el cual una porción aleatoria de los datos se toma

como set de entrenamiento y otra porción como set de *testing*, sobre el cual se aplicarán los modelos entrenados y se medirá la precisión de los resultados obtenidos.

Si bien no existe una regla sobre qué proporción de datos asignar a uno u otro set, se recomienda llevar a cabo una división que otorgue un mayor número de observaciones al set de entrenamiento. En este caso se ha optado por hacer una división 70%-30% para el set de entrenamiento y de *testing* respectivamente.

Una vez llevada a cabo la partición de datos, se continúa con la selección de las variables predictoras de los modelos de *machine learning* que se implementarán más adelante.

Para ello, tal y como podemos observar en la Figura Nº9 se ha implementado un modelo LOGIT que nos permite identificar qué variables son significativas en la determinación de nuestra variable TARGET y qué otras lo son menos. Esta significación puede ser medida mediante:

- Error estándar: Refleja cuan de diferente habría sido la estimación si se hubieran empleado otros datos de entrada. Cuanto menor es, más significativa es la variable.
- Z valor: Determina la precisión relativa de la estimación. Cuanto mayor es, más significativa es la variable.
- P valor: Refleja la probabilidad de imitar los resultados obtenidos en caso de que beta fuese igual a 0. Cuanto menor, mayor grado de significación (Manterola y Pineda, 2008).

Así, las variables con valores cercanos a 0 en el error estándar o el p valor y valores cercanos a 1 en el z valor son significativas (resaltadas con estrellas) y viceversa. Por ejemplo, la variable NAME_CONTRACT_TYPE parece tener una gran relevancia en la determinación de nuestra variable TARGET, mientras que la variable HOUR_APPR_PROCESS_START tiene poco peso en ella.

Figura Nº9: Variables significatividad de las variables

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.558243   0.059190  -9.431  < 2e-16 ***
NAME_CONTRACT_TYPE1        -0.446931   0.047401  -9.429  < 2e-16 ***
CODE_GENDER1               -0.413179   0.026205 -15.767  < 2e-16 ***
FLAG_OWN_CAR1              -0.290147   0.027317 -10.622  < 2e-16 ***
FLAG_OWN_REALTY1            0.036374   0.025781   1.411   0.1583
AMT_INCOME_TOTAL           -0.009456   0.013520  -0.699   0.4843
AMT_CREDIT                 -0.048720   0.029119  -1.673   0.0943 .
AMT_ANNUITY                 0.149903   0.018138   8.264  < 2e-16 ***
AMT_GOODS_PRICE            -0.063737   0.027895  -2.285   0.0223 *
NAME_INCOME_TYPE1           0.101627   0.042986   2.364   0.0181 *
NAME_EDUCATION_TYPE1        0.441857   0.028978  15.248  < 2e-16 ***
NAME_FAMILY_STATUS1        -0.175018   0.035725  -4.899 9.63e-07 ***
AGE                        -0.214216   0.015932 -13.445  < 2e-16 ***
YEARS_ID_PUBLISH           -0.120496   0.012327  -9.775  < 2e-16 ***
CNT_FAM_MEMBERS             0.018967   0.016151   1.174   0.2403
REGION_RATING_CLIENT        0.090494   0.012835   7.051 1.78e-12 ***
WEEKDAY_APPR_PROCESS_START1 -0.053195   0.025031  -2.125   0.0336 *
HOUR_APPR_PROCESS_START     0.005197   0.012403   0.419   0.6752
EXT_SOURCE_2               -0.490691   0.012446 -39.427  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fuente: Elaboración propia

Sorprendentemente, al correr el modelo LOGIT, observamos cómo además de la variable mencionada en el párrafo anterior, la variable AMT_INCOME_TOTAL carece de una significación alta. Mientras que el sentido común nos indicaría que esta es una de las variables que más relevancia tienen en la predicción del acaecimiento del suceso de impago, nuestro modelo dice lo contrario.

Con el objetivo de determinar si esto se debe al ruido que puede estar causando la alta correlación entre las variables AMT_CREDIT, AMT_ANNUITY y AMT_GOODS_PRICE, procedemos a correr el modelo eliminando dos de ellas, concretamente la primera y la tercera. Sin embargo, tal y como se puede observar en la Figura Nº10, la variable representativa del nivel de ingresos del solicitante es aún menos significativa en la predicción de la variable TARGET que antes.

Figura Nº10: Variables significatividad de las variables (omitiendo las variables AMT_CREDIT y AMT_GOODS_PRICE)

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.542024   0.059102  -9.171  < 2e-16 ***
NAME_CONTRACT_TYPE1        -0.445109   0.047382  -9.394  < 2e-16 ***
CODE_GENDER1               -0.426473   0.026108 -16.335  < 2e-16 ***
FLAG_OWN_CAR1              -0.293443   0.027291 -10.752  < 2e-16 ***
FLAG_OWN_REALTY1            0.045270   0.025727   1.760   0.0785 .
AMT_INCOME_TOTAL           -0.015555   0.013477  -1.154   0.2484
AMT_ANNUITY                 0.074345   0.013531   5.494 3.92e-08 ***
NAME_INCOME_TYPE1           0.094939   0.042952   2.210   0.0271 *
NAME_EDUCATION_TYPE1        0.442261   0.028958  15.272  < 2e-16 ***
NAME_FAMILY_STATUS1        -0.184655   0.035668  -5.177 2.25e-07 ***
AGE                        -0.226254   0.015816 -14.305  < 2e-16 ***
YEARS_ID_PUBLISH           -0.120474   0.012318  -9.780  < 2e-16 ***
CNT_FAM_MEMBERS             0.018848   0.016137   1.168   0.2428
REGION_RATING_CLIENT        0.089112   0.012826   6.948 3.71e-12 ***
WEEKDAY_APPR_PROCESS_START1 -0.053554   0.025016  -2.141   0.0323 *
HOUR_APPR_PROCESS_START     0.003087   0.012389   0.249   0.8032
EXT_SOURCE_2               -0.492186   0.012437 -39.575  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fuente: Elaboración propia

Tras la implementación de este modelo hemos identificado qué variables de nuestra base de datos son más importantes para la adecuada predicción de la variable TARGET. Sin embargo, ante la incertidumbre recién descrita con la variable AMT_INCOME_TOTAL y su relevante significado, no eliminaremos ninguna variable. A la hora de implementar los modelos predictivos introduciremos todas las variables, ya que aquellas que en el modelo LOGIT pueden parecer de poca relevancia pueden tener un rol contrario en un algoritmo diferente.

4.3.6. Entrenamiento de modelos predictivos, *testing*, evaluación de resultados

El entrenamiento de modelos es probablemente la fase más importante de todo el proceso predictivo. Aunque la realización de las fases anteriores a esta es de vital importancia, su labor principal es perfeccionar y adecuar el input de los modelos predictivos. En este sentido, la adecuada elección de algoritmos es determinante.

En este caso, se entrenarán cinco algoritmos: modelo de regresión lógica con método *Stepwise*, KNN, árbol de decisión y *Support Vector Machine* (SVM). A su vez, estos modelos serán empleados para la predicción sobre el conjunto de datos de *testing*, lo cual permitirá comparar los modelos en base a los resultados obtenidos.

4.3.6.1. Modelo LOGIT- *Stepwise*

Los modelos de regresión, basados en la identificación de relaciones entre variables para la toma de decisiones, es uno de los más utilizados en ejercicios predictivos. Una de las técnicas más conocidas para construir una regresión es la conocida como S*tepwise,* la cual selecciona de forma automática las variables independientes que mayor significación tienen en la variable objetivo. Este proceso iterativo puede llevarse a cabo en diferentes direcciones (Mendez, 2019):

- *Forward*: el modelo se construye incluyendo variables de forma iterativa, manteniendo únicamente aquellas que son significativas.
- *Backward*; en este caso, de todas las variables existentes se van eliminando las menos significativas.
- Bidireccional: se combinan los métodos anteriores. En nuestro caso, el modelo de regresión se ha construido por este método.

Como podemos observar en la Figura Nº11, el método *stepswise* empleado ha construido un modelo de regresión en el que no están contempladas como predictoras todas las variables de la base de datos. Únicamente han sido consideradas como significativas 14 de las 18 variables, olvidando así las variables FLAG_OWN_REALTY1, AMT_INCOME_TOTAL, CNT_FAM_MEMBERS y HOUR_APPR_PROCESS_START.

Figura Nº11: Modelo *Stepwise*

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.56141    0.05275 -10.642  < 2e-16 ***
NAME_CONTRACT_TYPE1        -0.44217    0.04723  -9.362  < 2e-16 ***
CODE_GENDER1               -0.40783    0.02586 -15.770  < 2e-16 ***
FLAG_OWN_CAR1              -0.28939    0.02719 -10.644  < 2e-16 ***
AMT_CREDIT                 -0.04945    0.02908  -1.700   0.0890 .
AMT_ANNUITY                 0.14881    0.01770   8.409  < 2e-16 ***
AMT_GOODS_PRICE            -0.06505    0.02785  -2.335   0.0195 *
NAME_INCOME_TYPE1           0.10248    0.04274   2.398   0.0165 *
NAME_EDUCATION_TYPE1        0.44502    0.02873  15.491  < 2e-16 ***
NAME_FAMILY_STATUS1        -0.14552    0.02685  -5.419 6.00e-08 ***
AGE                        -0.21674    0.01534 -14.132  < 2e-16 ***
YEARS_ID_PUBLISH           -0.11945    0.01221  -9.779  < 2e-16 ***
REGION_RATING_CLIENT        0.09028    0.01233   7.321 2.46e-13 ***
WEEKDAY_APPR_PROCESS_START1 -0.05411    0.02501  -2.164   0.0305 *
EXT_SOURCE_2               -0.49093    0.01239 -39.619  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fuente: Elaboración propia

Tras el entrenamiento del modelo, se ha procedido a su aplicación al *test* set. Para ello, se debe establecer un punto de corte o *cutoff* que determine a partir de qué probabilidad una predicción es considerada como negativa (acaecimiento del suceso de impago) o positiva (solidez crediticia del solicitante). Aunque es común asignar un valor de 50% a este punto de corte, en ciertas ocasiones, como en el caso del *credit scoring*, es conveniente emplear el valor que mayor sensibilidad y especificidad aporte (Costa Cor et al., 2017). Por este motivo, en nuestro caso el valor asignado al *cutoff* ha sido de 35%.

En la Tabla Nº3, podemos observar algunas métricas de los resultados obtenidos en la predicción del acaecimiento del suceso de impago:

- *Accuracy*: refleja la precisión de la predicción del modelo entrenado.
- Sensibilidad: refleja la fracción de verdaderos positivos, es decir, la precisión en la predicción en los casos en los que se da el suceso de impago.
- Especificidad: refleja la fracción de verdaderos negativos, es decir, la precisión en la predicción en los casos en los que no se da el suceso de impago.
- AUC (*Area Under the Curve*): un AUC cercano a 1 indica que el modelo es perfecto, mientras que un AUC cercano a 0,5 indica que el modelo predice de forma aleatoria.

Tabla Nº3: Métricas modelo *Stepwise* bidireccional

Accuracy	Sensibilidad	Especificidad	AUC
64,07%	63,10%	64,61%	0,694

Fuente: Elaboración propia

4.3.6.2. Algoritmo KNN

El algoritmo KNN, también conocido como algoritmo de K vecinos más cercanos, es un método de aprendizaje supervisado no paramétrico muy utilizado para la construcción de modelos predictivos. Aunque este algoritmo es una alternativa de gran utilidad tanto para ejercicios de regresión como de clasificación, en el segundo de los casos cobra especial protagonismo (IBM, s.f.).

El funcionamiento del algoritmo KNN (*K Nearest Neighbors*) se basa en la clasificación de observaciones del test set en función de los valores de las K observaciones más cercanas del set de entrenamiento. De esta forma, los datos se agrupan en base a la

distancia existente entre ellos, estableciéndose el tamaño de estos grupos mediante el hiperparámetro K.

De esta manera K desempeña un papel determinante puesto que, si es muy pequeño, existe el riesgo de cometer sobreajuste, mientras que si es muy grande aumenta considerablemente el sesgo de la predicción. En nuestro caso, hemos llevado a cabo una validación cruzada con el objetivo de eliminar este riesgo. Para ello se han establecido 2 repeticiones, 2 folds y 30 valores a probar para el hiperparámetro K, lo cual indica que el algoritmo se ajusta 120 veces.

Como se puede observar en la Figura Nº12, el modelo que mejor se ajusta al set de entrenamiento es aquel en el que el hiperparámetro k toma el valor 63, es decir, aquel en el que se clasifican observaciones en base a los 63 vecinos más cercanos. Si bien podríamos seguir con este ejercicio hasta que el ROC comenzase a disminuir, su crecimiento es tan reducido en este valor que lo tomaremos como el óptimo.

Figura Nº12: Precisión en función de los valores del hiperparámetro K

k	ROC	Sens	Spec
5	0.6118141	0.7763989	0.3729445
7	0.6219623	0.8014979	0.3585094
9	0.6324014	0.8189527	0.3473758
11	0.6378783	0.8305893	0.3398232
13	0.6439938	0.8409879	0.3336691
15	0.6486593	0.8475489	0.3247176
17	0.6519472	0.8544504	0.3194583
19	0.6545226	0.8593402	0.3150386
21	0.6571325	0.8655298	0.3133605
23	0.6579647	0.8691817	0.3088847
25	0.6592952	0.8707601	0.3028981
27	0.6608625	0.8748143	0.2990378
29	0.6621515	0.8787138	0.2973597
31	0.6631849	0.8793018	0.2941706
33	0.6645648	0.8823657	0.2927160
35	0.6652806	0.8839750	0.2903102
37	0.6661139	0.8851201	0.2881842
39	0.6666085	0.8868841	0.2859463
41	0.6678460	0.8893600	0.2837642
43	0.6683427	0.8910002	0.2816943
45	0.6694472	0.8920525	0.2806315
47	0.6700363	0.8935380	0.2788970
49	0.6706544	0.8952711	0.2762672
51	0.6710102	0.8970661	0.2738056
53	0.6720665	0.8984278	0.2744210
55	0.6730188	0.8995420	0.2726307
57	0.6738415	0.8996658	0.2735818
59	0.6743472	0.9018940	0.2718475
61	0.6746443	0.9024202	0.2707846
63	0.6751941	0.9031010	0.2703928

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 63.

Fuente: Elaboración propia

Una vez entrenado el algoritmo con el set de entrenamiento, se aplica el modelo de mayor precisión al set de entrenamiento. De nuevo, se ha aplicado un *cutoff* de 35% para clasificar las predicciones de la variable TARGET en 1 o 0, obteniendo las métricas mostradas en la Tabla Nº4.

Tabla Nº4: Métricas modelo KNN

Accuracy	Sensibilidad	Especificidad	AUC
63,44%	62,30%	64,06%	0,678

Fuente: Elaboración propia

4.3.6.3. Árbol de decisión

Los árboles de clasificación son técnicas no paramétricas y no lineales de aprendizaje supervisado muy utilizadas tanto en ejercicios de regresión como de clasificación. Mediante la medida de impurezas (con el índice de Gini o la entropía) y la combinación de valores de la base de datos de entrada, este algoritmo permite realizar una clasificación categórica de observaciones en un proceso nodal. El funcionamiento del algoritmo se basa en la partición recursiva de un espacio de p dimensiones (p=número de variables del modelo) priorizando aquellas particiones que reducen en mayor medida la heterogeneidad, construyendo así un modelo con la mayor pureza posible (IBM, 2023).

En cuanto al hiperparámetro empleado en este algoritmo, llamado Alpha, tiene como objetivo la penalización de la complejidad del modelo predictivo, tomando valores entre 0 (nula penalización: *full grown*) y 1 (penalización máxima: árbol mínimo).

Para la definición de este hiperparámetro Alpha, se decide realizar un proceso de *crossvalidation* similar al realizado en los algoritmos anteriormente explicados. En este caso, se han establecido 10 particiones del set de entrenamiento, 4 repeticiones y 146 valores de prueba para el hiperparámetro Alpha (con un abanico de valores entre 0,008 y 0,3), lo cual significa que el algoritmo se ajusta en 5840 ocasiones. Tras el entrenamiento del modelo, se obtiene que aquel que mejor se adapta al set de entrenamiento (ofrece un mayor AUC) es en el que el hiperparámetro Alpha toma el valor 0,008.

A continuación, se procede a analizar la importancia de los predictores y la estructura del árbol de decisión del modelo entrenado. Como se puede observar en la Figura Nº13, el modelo obtenido considera como significativos únicamente 9 predictores, ordenados de mayor a menor relevancia en la predicción de la variable TARGET.

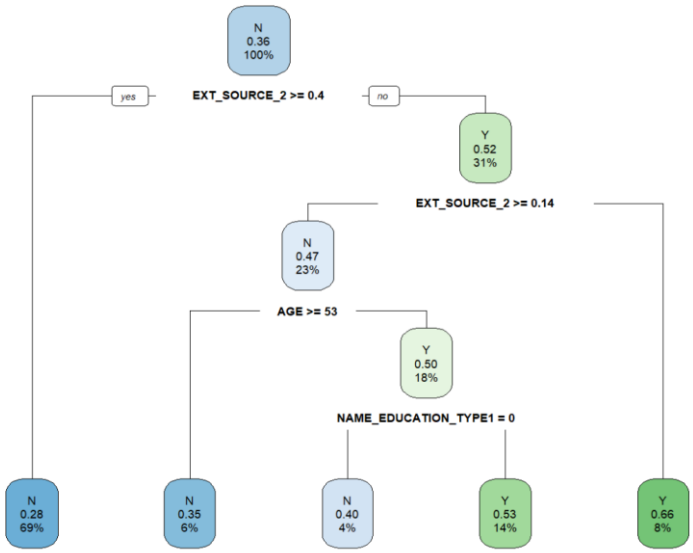
Figura Nº13: Significatividad de predictores

	Overall
EXT_SOURCE_2	100.000
AGE	46.398
CODE_GENDER1	27.748
NAME_EDUCATION_TYPE1	20.786
REGION_RATING_CLIENT	12.477
NAME_INCOME_TYPE1	12.332
YEARS_ID_PUBLISH	10.291
AMT_ANNUITY	5.652
NAME_CONTRACT_TYPE1	2.345
FLAG_OWN_REALTY1	0.000
HOUR_APPR_PROCESS_START	0.000
AMT_INCOME_TOTAL	0.000
FLAG_OWN_CAR1	0.000
AMT_GOODS_PRICE	0.000
AMT_CREDIT	0.000
NAME_FAMILY_STATUS1	0.000
CNT_FAM_MEMBERS	0.000
WEEKDAY_APPR_PROCESS_START1	0.000

Fuente: Elaboración propia

Estas variables significativas se ven ilustradas en la estructura del árbol de clasificación obtenido, tal y como se puede observar en la Figura Nº14.

Figura Nº14: Estructura árbol de decisión



Fuente: Elaboración propia

Si bien son muchas las interpretaciones extraíbles de esta gráfica, algunos ejemplos son:

El 69% de las solicitudes de préstamo del set de entrenamiento recibieron una valoración externa (variable EXT_SOURCE_2) superior a 0,4, y solo el 28% de ellas terminó en impago.

El 6% de solicitudes de préstamo del set de entrenamiento recibieron una valoración externa (variable EXT_SOURCE_2) comprendida entre 0,4 y 0,14 y el fueron realizadas por una persona mayor de 53 años. De este grupo de solicitudes, el 35% terminaron en impago.

Por último, se procede a la aplicación del modelo entrenado al set de *testing*. Manteniendo el *cutoff* aplicado en los algoritmos previamente explicados se han obtenido las métricas mostradas en la Tabla Nº5:

Tabla Nº5: Métricas árbol de decisión

Accuracy	**Sensibilidad**	**Especificidad**	**AUC**
65,14%	44,61%	76,45%	0,612

Fuente: Elaboración propia

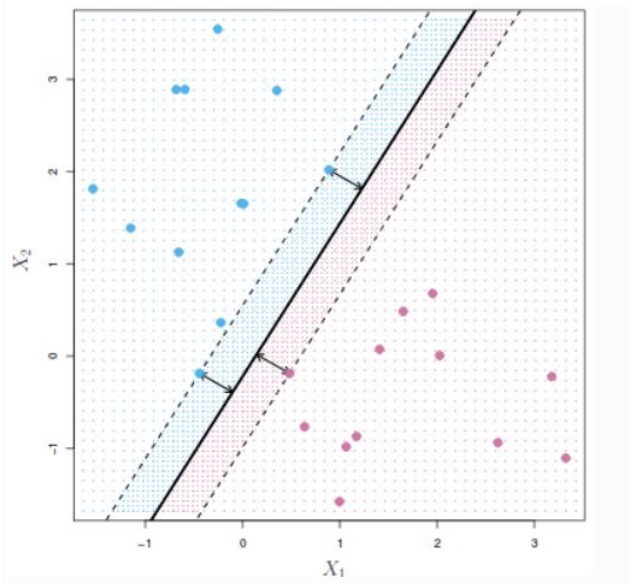
4.3.6.4. *Support Vector Machine* (SVM)

El *Support Vector Machine* es un algoritmo de aprendizaje supervisado de gran utilidad en problemas de regresión y clasificación.

Mediante la generación de hiperplanos, el algoritmo busca encontrar aquél que mejor separe las clases de datos, es decir, que ofrezca un mayor margen o distancia con respecto a todas las observaciones del set de entrenamiento (MathWorks, 2023).

Para poder obtener este hiperplano, conocido como hiperplano óptimo de separación, se debe calcular la distancia perpendicular de cada una de las observaciones con respecto a cada uno de los hiperplanos probados. La menor de estas distancias determina el margen del hiperplano probado, y aquel que ofrezca el mayor margen es el considerado como óptimo (Amat, 2017). En la Figura Nº15 podemos observar un ejemplo de este proceso.

Figura Nº15: Hiperplano óptimo de separación



Fuente: Amat, 2017

Existen diferentes vertientes del algoritmo SVM: kernel radial, lineal o polinómico. En nuestro caso, se ha decidido emplear el primero de ellos para el entrenamiento de nuestro modelo.

Este algoritmo cuenta con un hiperparámetro principal: C.

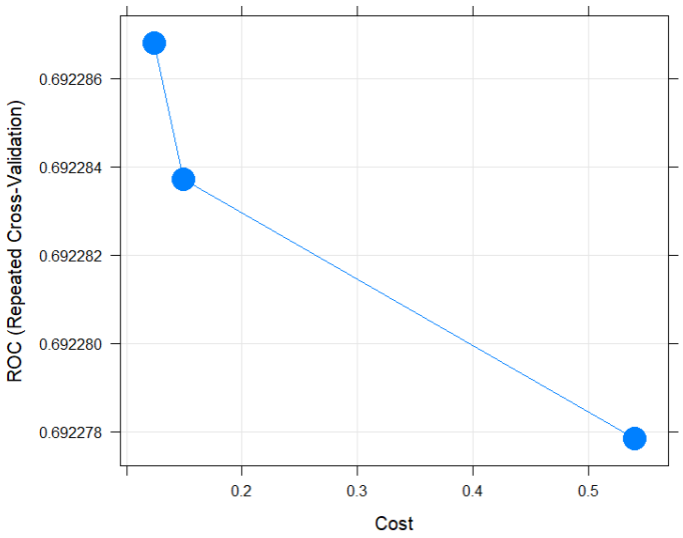
C es el hiperparámetro que permite el control entre la varianza y el sesgo, tomando valores entre 0 e infinito en función de la penalización aplicada a cada observación clasificada de forma errónea por el modelo. Cuando C=0, todas aquellas observaciones que violen el margen (mínima distancia entre cada una de las observaciones y el hiperplano) no serán penalizadas. A medida que el valor de C aumenta, las observaciones erróneamente clasificadas son penalizadas con mayor severidad. En este sentido, cuando C=∞ estamos ante el llamado *Maximal Margin Classifier*, únicamente alcanzable cuando las clases se pueden separar de manera perfecta (Amat, 2017).

Ante casos en los que la separación lineal no es viable es necesario emplear una transformación dimensional. Para ello, se incluyen nuevos hiperparámetros que facilitan el proceso. Por ejemplo, el empleo de un kernel radial implica la introducción del hiperparámetro sigma que determina la complejidad o flexibilidad del modelo (Amat, 2018).

De la misma manera que en los algoritmos anteriores, hemos aplicado la técnica de *crossvalidation* para obtener el valor del hiperparámetro que mejor se ajusta a nuestro modelo.

Como se puede observar en la Figura Nº16, el modelo entrenado que mayor precisión ofrece es el que otorga a C un valor de 0,1236218, logrando un ROC de 0,6922868.

Figura Nº16: ROC en función de hiperparámetro C



Fuente: Elaboración propia

A continuación, se procede a la implementación del modelo entrenado en el set de *testing*. Empleando el mismo *cutoff* que el aplicado anteriormente, se han obtenido las métricas mostradas en la Tabla Nº6:

Tabla Nº6: Métricas SVM

Accuracy	**Sensibilidad**	**Especificidad**	**AUC**
63,95%	61,33%	65,39%	0,691

Fuente: Elaboración propia

4.3.7. Comparación de modelos

Tras el entrenamiento y la aplicación de los algoritmos predictivos de *machine learning* a la base de datos de entrada, procedemos a la comparación de los resultados obtenidos en cada uno de ellos.

A continuación, se muestra un resumen de las métricas obtenidas con cada uno de los modelos:

Tabla Nº7: Métricas de los modelos

Modelo	***Accuracy***	**Sensibilidad**	**Especificidad**	**AUC**
Stepwise	64,07%	63,10%	64,61%	0,694
KNN	63,44%	62,30%	64,06%	0,678
Árbol de decisión	65,14%	44,61%	76,45%	0,612
SVM	63,95%	61,33%	65,39%	0,691

Fuente: Elaboración propia

Como se puede observar en la Tabla Nº7, el modelo con mayor precisión (65,14%) en la predicción del acaecimiento del suceso de impago es el obtenido mediante la aplicación de árboles de decisión. Por el contrario, el modelo que peor resultados ofrece es el entrenado con el algoritmo KNN, algo que no es de extrañar, puesto que este algoritmo tiene grandes dificultades en el proceso predictivo cuando la variable objetivo tiene mucha asimetría, como precisamente ocurre en nuestro caso.

En cuanto a la sensibilidad y la especificidad, los modelos que mejor resultado han obtenido son el *Stepwise* y el árbol de decisión respectivamente. Como se ha mencionado anteriormente, la sensibilidad es la probabilidad de clasificar de forma correcta los verdaderos positivos, es decir, la precisión en la predicción del acaecimiento del suceso de impago. Por el contrario, la especificidad es la probabilidad de clasificar correctamente los verdaderos negativos, es decir, la precisión en la predicción del no acaecimiento del suceso de impago. Ambas métricas se ven influenciadas por dos factores principales:

- El algoritmo empleado: tal y como podemos observar en la Tabla Nº7
- El punto de corte empleado en los modelos, también llamado *cutoff*. A medida que aumenta el punto de corte, la sensibilidad disminuye y la especificidad aumenta.

En función de las prioridades de la entidad bancaria emisora del préstamo hipotecario puede resultar de mayor utilidad tener una especificidad o una sensibilidad mayor, es decir, puede resultar más conveniente predecir con mayor precisión qué operaciones pueden terminar en impago por encima de predecir las que tendrán éxito, y viceversa. En nuestro caso, se ha buscado lograr un balance entre ambos, lo cual nos ha llevado a aplicar un punto de corte (*cutoff*) de 0,35 en los modelos. Si bien en ocasiones esto puede

perjudicar a la precisión del modelo, en este ejercicio en concreto se ha antepuesto lograr la mayor precisión posible para ambos casos por encima de la precisión global.

Por último, analizamos la métrica de precisión AUC (*Area Under the Curve*) obtenida en cada modelo. De acuerdo con esta métrica, es el modelo *Stepwise*, y no el árbol de decisión, el que mayor precisión logra. Ante un conjunto de datos con asimetría en la variable objetivo, es recomendable el empleo de la métrica AUC para determinar qué modelo tiene una mayor precisión en el ejercicio predictivo (Chugh, 2022). En este sentido, los modelos entrenados con los algoritmos *Stepwise* y SVM ofrecen la mayor precisión para el conjunto de datos estudiado.

5. CONCLUSIONES

La digitalización y la implementación del *Data Analytics* han revolucionado la gran mayoría de sectores económicos provocando una transformación de modelos de negocios y de los *drivers* de la toma de decisiones.

El sector inmobiliario no ha quedado exento. Si bien la implementación tecnológica ha llegado a este sector de muchas formas, uno de los principales cambios ha sido el experimentado por los préstamos hipotecarios, imprescindibles para el devenir tanto de este sector y como de la economía en nuestro país.

En este sentido, el Big Data es ya una realidad en el proceso de concesión de hipotecas. Su implementación favorece al estudio de la capacidad crediticia de los clientes, reduciendo así las probabilidades de impago en las operaciones financiero-inmobiliarias. Sin embargo, la precisión de los modelos predictivos y las variables a considerar varían de unas entidades bancarias a otras, dejando un gran espacio para la mejora y la optimización de estas técnicas revolucionarias.

Del presente trabajo se han obtenido las siguientes conclusiones:

El sector inmobiliario es uno de los principales sectores en nuestro país. Su aportación a la economía española, su capacidad de crear empleos cada año y su relevancia macroeconómica son algunos de los motivos que convierten a este sector en una de las grandes preocupaciones de las instituciones gubernamentales. Sin embargo, el factor más determinante en la explicación de esta importancia son las fuertes relaciones que este sector guarda con otros sectores, sobre todo con el financiero.

Los préstamos hipotecarios suponen el nexo entre el sector inmobiliario y el financiero. A través de ellos y mediante la manipulación de los tipos de interés, los gobiernos y entidades financieras son capaces de transmitir sus pretensiones e influir de manera directa en los mercados inmobiliarios, afectando no solo a la economía sino a las tendencias y el estilo de vida de los consumidores. En este sentido, este tipo de préstamos es un instrumento muy poderoso en el condicionamiento del curso de la economía en España, pero a su vez muy peligroso si no recibe una gestión adecuada.

Desde hace décadas, el *Data Analytics* y el *credit scoring* han ido integrándose en la toma de decisiones de concesiones hipotecarias por parte de las entidades financieras con el objetivo de aumentar el éxito de sus operaciones. Sin embargo, la crisis financiera de 2008 aumentó la preocupación ante el impago de préstamos, lo cual ha supuesto un avance exponencial de la sofisticación y la robustez de los modelos predictivos empleados. En este sentido, el volumen de datos y variables considerados en este proceso ha experimentado un gran crecimiento, abriendo un dilema moral sobre si debe existir un límite mediante el cual la privacidad y el estilo de vida del solicitante sean respetados. Con el objetivo de proteger al consumidor, deben establecerse legislaciones claras que establezcan qué datos deberían ser considerados en la determinación del riesgo de crédito y qué otros no.

En lo que respecta a los resultados obtenidos en el estudio empírico relativo al impacto del Big Data en la concesión de préstamos y la predicción del suceso de impago, cabe mencionar que:

- Sobre la principal hipótesis de nuestro trabajo: a través de un modelo de *machine learning*, que incorpore variables socioeconómicas se puede predecir el acaecimiento de un suceso de impago por parte del prestatario con una precisión superior al 80%; no se ha logrado encontrar ningún algoritmo que ofrezca esta precisión con la base de datos empleada. De hecho, los resultados obtenidos de los modelos empleados son notablemente lejanos a este valor, siendo algunos más cercanos a la precisión correspondiente al azar (50%). Concretamente, el modelo predictivo LOGIT con técnica *Stepwise* y los árboles de decisión ofrecen la mayor precisión de entre los modelos probados.
- El número de variables consideradas en la predicción del suceso de impago es muy alto, convirtiendo esta tarea en un proceso de cierta complicación por los

programas necesarios para el procesamiento de datos, así como por la profundidad de los análisis necesaria para la obtención de *insights*.

- La asimetría de la variable que se pretende predecir tiene efectos negativos en diferentes aspectos del proceso predictivo. En primer lugar, la distribución irregular de los valores de esta variable puede afectar a la sensibilidad y especificidad de los modelos, algo experimentado en nuestro caso ya que ha resultado difícil encontrar un balance entre estas métricas. En segundo lugar, esta asimetría puede limitar la precisión en función del algoritmo empleado, como en el caso del algoritmo KNN, el que peor resultado ha ofrecido durante este estudio empírico con un 60,45% de accuracy y 0,641 de AUC.
- El tratamiento de la base de datos es determinante para la implementación de algoritmos predictivos de *machine learning*. Si bien la etapa principal del proceso de predicción del suceso de impago es el entrenamiento y *testing* de modelos, los pasos previos de preparación de la base de datos como el tratamiento de NA´s y *outliers* son imprescindibles.
- La variable AMT_INCOME_TOTAL no es significativa en la predicción de la variable TARGET. Esto indica que, sorprendentemente, el nivel de ingresos del solicitante apenas tiene relevancia en el riesgo de impago del préstamo, al menos con la base de datos empleada.
- De entre todos los algoritmos empelados, los árboles de decisión podrían resultar la mejor alternativa para la realización de un ejercicio como este por dos principales motivos. En primer lugar, ofrecen una de las mejores precisiones de entre todos los algoritmos probados. En segundo lugar, ofrecen la capacidad de interpretar los resultados de manera muy intuitiva, permitiendo obtener *insights* con rapidez y profundidad.

En resumen, tras la realización del estudio empírico no se puede aceptar la hipótesis inicial, es decir, no se puede confirmar que mediante algoritmos de *machine learning* se puede entrenar un modelo que prediga el acaecimiento del suceso de impago con una precisión superior al 80%.

Sin embargo, son muchos los factores que han influido en este estudio y que si fueran modificados podrían llevar a un resultado diferente o incluso a confirmar la hipótesis inicial. Algunos de estos factores son:

- La base de datos empleada: son muchas las variables a considerar en la determinación del riesgo de crédito de un cliente. En nuestro caso, hemos empleado una pequeña parte del total de variables relevantes por motivos de simplicidad y accesibilidad a los datos, pero si se hubieran considerado otras variables quizás se habría obtenido una mayor precisión de los modelos.
- Los algoritmos empleados: durante el estudio empírico realizado en este trabajo se han empleado algunos de los principales algoritmos para tareas de predicción binaria. Sin embargo, existen otras muchas alternativas que se podrían haber empleado, como el algoritmo *random forest* o los ensembles de modelos.
- El punto de corte empleado (*cutoff*) en la implementación de los modelos: como se ha mencionado con anterioridad, suele emplearse un valor de 0,5 para este factor, a pesar de no existir ninguna regla sobre si este debe ser siempre el valor que emplear. En nuestro caso se ha empleado un *cutoff* de 0,35, debido a la asimetría de la variable TARGET y la intención de alcanzar la máxima sensibilidad y especificidad. Sin embargo, si se hubieran empleado otros valores para el punto de corte se podrían haber obtenido precisiones diferentes para los distintos modelos.

De cara a futuros estudios sería relevante tener en cuenta esos aspectos, ya que son oportunidades de mejorar los resultados obtenidos en este trabajo y, por lo tanto, de acercarse a los modelos de *credit scoring* empleados por las principales entidades bancarias a nivel internacional.

6. BIBLIOGRAFÍA

Altman, T. (2023). *When Were Credit Scores Invented? A Brief Look At History*. OppU. https://www.opploans.com/oppu/articles/a-brief-history-of-credit-scores/

Amat, J. (2017). *Máquinas de Vector Soporte (Support Vector Machines, SVMs).* Ciencia de datos. https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Amat, J. (2018). *Machine Learning con R y caret*. RPubs. https://rpubs.com/Joaquin_AR/383283

Asociación Hipotecaria Española. (2014). *Guía Hipotecaria*. https://www.consumo.gob.es/sites/consumo.gob.es/files/consumo_masinfo/guia_hipotecaria.pdf

Banco de España. (2013). *Guía de acceso al préstamo hipotecario*. https://www.bde.es/f/webbde/Secciones/Publicaciones/Folletos/Fic/Guia_hipotecaria_2013.pdf

Banco de España. (2019). *Hipoteca a tipo fijo o variable*. Cliente bancario BDE. https://clientebancario.bde.es/pcb/es/menu-horizontal/productosservici/financiacion/hipotecas/guia-textual/primerospasoscon/Hipoteca_a_tipo_2ada24e53ab1d51.html

BBVA. (2022). *El científico de datos, un profesional clave para el sector bancario… y para sus clientes*. BBVA. https://www.bbva.com/es/el-cientifico-de-datos-un-profesional-clave-para-el-sector-bancario-y-para-sus-clientes/

Boedo Vilabella, L. (2008). *Las fuentes de financiación y su coste*. Netbiblo. https://books.google.es/books?hl=es&lr=&id=cYEVXfYpVF8C&oi=fnd&pg=PA117&dq=las+fuentes+de+financiacion+y+su+coste+boedo&ots=f31rLKZGX6&sig=xuBUWSw0CEzUGEz6UfeIMQ8eZlc#v=onepage&q=las%20fuentes%20de%20financiacion%20y%20su%20coste%20boedo&f=false

Borja, R. (s. f.). *Los bienes raíces en la historia de las grandes culturas*. Real Estate Market & Lifestyle. https://www.realestatemarket.com.mx/articulos/mercado-inmobiliario/19562-los-bienes-raices-en-la-historia-de-las-grandes-culturas

Botella, M. I. (2021). *Marketing relacional, un camino de la mano de los datos*. Puromarketing. https://www.puromarketing.com/44/35758/marketing-relacional-camino-mano-datos.html

Carbó Valverde, S. (2020). *Gobiernos y bancos en la crisis del Covid-19*. Santander. https://www.santander.com/es/sala-de-comunicacion/insights/autores/gobiernos-y-bancos-en-la-crisis-del-covid-19

Carbó Valverde, S., & Rodríguez Fernández, F. (2021). *El mercado de vivienda tras la Covid*. https://www.funcas.es/wp-content/uploads/2021/07/XX-2107-Carbo.indd_.pdf

Chugh, V. (2022). *Which Metric Should I Use? Accuracy vs. AUC*. KDnuggets. https://www.kdnuggets.com/2022/10/metric-accuracy-auc.html#:~:text=AUC%20is%20the%20go%2Dto,model's%20performance%20across%20different%20thresholds.

Comisión Europea. (2021). *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de Inteligencia Artificial (ley de inteligencia artificial) y se modifican determinados actos legislativos de la unión*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

Comisión Nacional del Mercado de Valores. (s.f.). *¿Qué es fintech?* https://www.cnmv.es/DocPortal/Publicaciones/Fichas/GR03_Fintech.pdf

Corporate Finance Institute. (2022). *Credit Risk Analysis Models*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk-analysis-models/

Costa Cor, T., Boj del Val, E., & Fortiana Gregori, J. (2012). *Bondad de ajuste y elección del punto de corte en regresión logística basada en distancias. aplicación al problema de credit scoring*. https://actuarios.org/wp-content/uploads/2017/02/anales2012_2.pdf

Daher, A. (2013). *El sector inmobiliario y las crisis económicas*. https://www.scielo.cl/pdf/eure/v39n118/art03.pdf

Delgado, A. (2021). *Big data, una herramienta fundamental en la etapa post-pandemia*. El Economista. https://www.eleconomista.es/especial-tecnologia-startups/noticias/11414631/10/21/Big-data-una-herramienta-fundamental-en-la-etapa-postpandemia-.html

Fernández, E., Díez, J. R., Aspachs, O., Jódar, S., & Montoriol Garriga, J. (2021). *Inmobiliario: Informe Sectorial*. https://www.caixabankresearch.com/sites/default/files/content/file/2021/12/24/91184/is-inmobiliario-1-2022-cast_web.pdf

Fernández, R. (2022). *Evolución anual del peso de las actividades inmobiliarias sobre el PIB en España desde 2005 hasta 2020*. Statista.

https://es.statista.com/estadisticas/549634/aportacion-de-las-actividades-inmobiliarias-al-pib-en-espana/#:~:text=De%20esta%20forma%2C%20mientras%20que,%2C%20con%2011%2C7%25

FICO. (2018). *To Score or Not To Score?.* FICO. https://www.fico.com/sites/default/files/inline-files/FICO_70_Insights_To_Score_or_Not_To_Score_3009WP_US_EN.pdf

Fondo Monetario Internacional. (2016). *La solidez del sistema financiero*. Fondo Monetario Internacional. https://www.imf.org/es/About/Factsheets/Financial-System-Soundness

Galeano, P., & Peña, D. (2019). *Las nuevas oportunidades del Big Data para las instituciones financieras*. https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_PEE/162art07.pdf

Galeano, S. (2021). *La UE prohibirá el uso de la inteligencia artificial en sistemas de social scoring*. Marketing4Ecommerce. https://marketing4ecommerce.net/la-ue-prohibira-el-uso-de-la-inteligencia-artificial-en-sistemas-de-social-scoring/

Gallo, N. (2022). *When Were Credit Scores Invented? A Brief History.* Finmasters. https://finmasters.com/when-were-credit-scores-invented/#gref

Ganuza Fernández, J.J. (2020). *Economía digital en tiempos de pandemia (II). El crédito social (social scoring).* Funcas. https://blog.funcas.es/economia-digital-en-tiempos-de-pandemia-ii-el-credito-social-social-scoring/

García Montalvo, J. (2014). *El impacto del big data en los servicios financieros*. https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_PEE/EXT0114art06.pdf

Gastesi, N. (2022). *Qué son las Hipotech y cuáles son sus ventajas al comprar una casa nueva*. La información. https://www.lainformacion.com/vivienda-inmobiliario/que-son-hipotech-ventajas-comprar-casa-nueva/2867509/

González Cuervo, F. (2021). *El gran futuro del sector de la construcción español en 2022*. EY. https://www.ey.com/es_es/espana-2022-grandes-tendencias-a-corto-plazo/el-gran-futuro-del-sector-de-la-construccion-espanol-en-2022

Guide To Business. (2020). *El sistema financiero español*. https://www.guidetobusinessinspain.com/wp-content/uploads/2020/09/ANEXO_II_ESP_INVESTINSPAIN.pdf

Guidi, L. R. (2021). *Quantitative Easing: ¿Por qué no causó inflación antes y por qué podría causarla ahora?* LinkedIn. https://www.linkedin.com/pulse/quantitative-easing-por-qu%C3%A9-caus%C3%B3-inflaci%C3%B3n-antes-y-podr%C3%ADa-guidi/?originalSubdomain=es

Gutiérrez Girault, M. (2007). *Credit scoring models: what, how, when and for what purposes*. https://core.ac.uk/download/pdf/6538561.pdf

Helloteca. (2022). *HIPOTECH: LA NUEVA FORMA DE CONTRATAR TU HIPOTECA*. Helloteca. https://helloteca.com/hipotech-contratar-hipoteca/

IBM. (2023). *Árbol de clasificación*. IBM. https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-classification-tree

IBM. (s.f.). *Algoritmo de k vecinos más cercanos*. IBM. https://www.ibm.com/es-es/topics/knn

Iranzo, S. (2008). *Introducción al riesgo-país.* https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/08/Fic/do0802.pdf

Jiménez, G., & Saurina, J. (2006). *Ciclo crediticio, riesgo de crédito y regulación prudencial*. https://www.bde.es/f/webbde/Secciones/Publicaciones/InformesBoletinesRevistas/InformesEstabilidadFinancera/06/Fic/01_Ciclo_regulacion.pdf

Kagan, J. (2020). *Default Risk: Definition, Types, and Ways to Measure*. Investopedia. https://www.investopedia.com/terms/d/defaultrisk.asp

Karamelikli, H. (2016). *La relación entre el sector inmobiliario y crecimiento económico*. TRT español. https://www.trt.net.tr/espanol/programas/2016/02/10/la-relacion-entre-el-sector-inmobiliario-y-crecimiento-economico-430122

KPMG. (2021). *Banca y digitalización: un salto de gigante.* KPMG Tendencias. https://www.tendencias.kpmg.es/2021/06/banca-y-digitalizacion-un-salto-de-gigante/

KPMG. (2022). *Financiación alternativa*. KPMG Tendencias. https://www.tendencias.kpmg.es/2021/07/financiacion-alternativa-abre-paso-sector-inmobiliario/

López Millán, M. (2022). *Economía real y financiera: principales diferencias*. IMF. https://blogs.imf-formacion.com/blog/mba/economia-real-financiera-diferencias/

López, M. (2005). *La vivienda como colateral: política monetaria, precios de la vivienda y consumo en Colombia*. https://www.banrep.gov.co/docum/ftp/borra299.pdf

Manterola, C., & Pineda, V. (2008). *El valor de" p" y la" significación estadística": Aspectos generales y su valor en la práctica clínica.* https://www.scielo.cl/pdf/rchcir/v60n1/art18.pdf

Marrero, D. (2022). *A más habitantes, casas más caras: la evolución del precio donde más ha crecido la población desde 2011*. Idealista. https://www.idealista.com/news/inmobiliario/vivienda/2022/12/22/800731-a-mas-habitantes-casas-mas-caras-la-evolucion-del-precio-donde-mas-poblacion-ha

MathWorks. (2023). *Support Vector Machine (SVM)*. MathWorks. https://es.mathworks.com/discovery/support-vector-machine.html

Maudos, J. (2022). *Cambios en la estructura de ingresos de la banca española*. https://www.funcas.es/wp-content/uploads/2022/07/Cie-289-Joaqu%C3%ADn-Maudos.pdf

Mendez, J. (2019). *Stepwise Regresión*. RPubs. https://rpubs.com/jorge_mendez/609253#:~:text=La%20regresi%C3%B3n%20paso%20a%20paso,modelos%20con%20cientos%20de%20variables

Montoriol Garriga, J. (2022). *¿Cómo puede impactar el aumento de los tipos de interés en el esfuerzo hipotecario?* https://www.caixabankresearch.com/sites/default/files/content/file/2022/05/09/34454/im05_22-07-ee-focus-3-es.pdf

Nanney, R. (2017). *Consideraciones de la disrupción digital*. https://www2.deloitte.com/content/dam/Deloitte/mx/Documents/risk/Gobierno-Corporativo/2017/Disrupcion-digital-GobCorp.pdf

Nguyen, J. (2021). *4 Key Factors That Drive the Real Estate Market*. Investopedia. https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp

Otero, M., & Blanco, J. (2014). *El sector inmobiliario en España*. https://www.ieemadrid.es/wp-content/uploads/El-sector-inmobiliario-en-Espa%C3%B1a.pdf

Paul, T. (2023). *When did credit scores start? A brief look at the long history behind credit reporting.* CNBC. https://www.cnbc.com/select/when-did-credit-scores-start/

Porras Castaño, J. (2018). *Big Data Analytics en el Sector Financiero*. LinkedIn. https://www.linkedin.com/pulse/big-data-analytics-en-el-sector-financiero-javier-porras-casta%C3%B1o/?originalSubdomain=es

Prieto, B. (2022). *Data Driven: la importancia de los datos en la nueva realidad.* KPMG Tendencias. https://www.tendencias.kpmg.es/2022/03/data-driven-importancia-datos-nueva-realidad/

Puertas Medina, R. M., & Martí Selva, M. L. (2013). *Análisis del credit scoring*. https://www.redalyc.org/pdf/1551/155127485011.pdf

Pulido, E., Díaz, A., Santana, I., & Giménez, J. (2023). *La inflación, la otra cara de la guerra de Ucrania*. RTVC. https://rtvc.es/la-inflacion-la-otra-cara-de-la-guerra-de-ucrania/

PwC. (2021). La pandemia tensa la gestión de riesgos de la banca y apremia su digitalización. PwC. https://www.pwc.es/es/sala-prensa/notas-prensa/2021/gestion-riesgos-banca-digitalizacion-informe-union-bancaria.html

Real Academia Española (2022). Real Academia Española. https://dle.rae.es/inmobiliario

REALIA. (2018). *¿Cuál es la diferencia entre crédito y préstamo hipotecario?* REALIA. https://www.realia.es/diferencia-entre-credito-prestamo-hipotecario

Rey, L. (2022). *Valores cíclicos vs defensivos: ¿en qué momento invertir en ellos?* FundsPeople España. https://fundspeople.com/es/glosario/valores-ciclicos-vs-defensivos-en-que-momento-invertir-en-ellos/

Riera, M. (2022). *Previsiones para 2023: ¿cómo cambiarán las hipotecas el año que viene?* HelpMyCash. https://www.helpmycash.com/blog/previsiones-para-2023-como-cambiaran-las-hipotecas-el-ano-que-viene/

San Millán, M. A., & Viejo, F. (2008). *Economía Real versus Economía Financiera Mercados y Activos Financieros*. http://www.emp.uva.es/miguelsanmillan/FinanzasB%E1sicas2008TEMA1.pdf

Santaella, J. (2022). *Economía real y Economía financiera: ¿Cuáles son sus diferencias?* Economia3. https://economia3.com/diferencias-economia-real-financiera/

Schneeweiss, Z. (2022). *ECB's Lagarde Says Policy Rates Must Be Higher to Curb Inflation*. Bloomberg. https://www.bloomberg.com/tosv2.html?vid=&uuid=c50ab029-8b63-11ed-b360-565a6d6e4e47&url=L25ld3MvYXJ0aWNsZXMvMjAyMi0xMi0zMS9lY2Itcy1sYWdhcmRlLXNheXMtcG9saWN5LXJhdGVzLW11c3QtYmUtaGlnaGVyLXRvLWN1cmItaW5mbGF0aW9u

Toloba, C., & del Río, J. M. (2020). *La perspectiva de la digitalización de la banca española: riesgos y oportunidades*. https://repositorio.bde.es/bitstream/123456789/13547/1/Digitalizacion.pdf

Uría Menéndez. (2022). *Residencias de estudiantes implantación en España.* Uría Menéndez. https://www.uria.com/documentos/circulares/1581/documento/12997/Residencias_de_estudiantes.pdf?id=12997&forceDownload=true

Utrera, E. (2019). *Las 'hipotech' llegan al mercado español*. Expansión. https://www.expansion.com/ahorro/2019/06/10/5cfa69f0468aeb362f8b4623.html

Valls, A. (2022). *El sector de la promoción inmobiliaria residencial y los mercados de capitales*. Deloitte. https://www2.deloitte.com/es/es/pages/finance/articles/El-

sector-de-la-promocion-inmobiliaria-residencial-y-los-mercados-de-capitales.html

Zárate, J. (2018). *¡Cuidado! El exceso de datos puede dar indigestión*. LinkedIn. https://www.linkedin.com/pulse/cuidado-el-exceso-de-datos-puede-dar-indigesti%C3%B3n-zarate-sousa/?originalSubdomain=es

Zhu, M. (2014). *Los mercados inmobiliarios, la estabilidad financiera y la economía*. Fondo Monetario Internacional. https://www.imf.org/es/News/Articles/2015/09/28/04/53/sp060514

7. ANEXOS

ANEXO I: Variables de la base de datos de partida

Nombre	Descripción (inglés)
SK_ID_CURR	*ID of loan in our sample*
TARGET	*Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)*
NAME_CONTRACT_TYPE	*Identification if loan is cash or revolving*
CODE_GENDER	*Gender of the client*
FLAG_OWN_CAR	*Flag if the client owns a car*
FLAG_OWN_REALTY	*Flag if client owns a house or flat*
CNT_CHILDREN	*Number of children the client has*
AMT_INCOME_TOTAL	*Income of the client*
AMT_CREDIT	*Credit amount of the loan*
AMT_ANNUITY	*Loan annuity*
AMT_GOODS_PRICE	*For consumer loans it is the price of the goods for which the loan is given*
NAME_TYPE_SUITE	*Who was accompanying client when he was applying for the loan*
NAME_INCOME_TYPE	*Clients income type (businessman, working, maternity leave,…)*
NAME_EDUCATION_TYPE	*Level of highest education the client achieved*
NAME_FAMILY_STATUS	*Family status of the client*
NAME_HOUSING_TYPE	*What is the housing situation of the client (renting, living with parents, …)*
REGION_POPULATION_RELATIVE	*Normalized population of region where client lives (higher number means the client lives in more populated region)*
DAYS_BIRTH	*Client's age in days at the time of application*
DAYS_EMPLOYED	*How many days before the application the person started current employment*
DAYS_REGISTRATION	*How many days before the application did client change his registration*
DAYS_ID_PUBLISH	*How many days before the application did client change the identity document with which he applied for the loan*
OWN_CAR_AGE	*Age of client's car*
FLAG_MOBIL	*Did client provide mobile phone (1=YES, 0=NO)*
FLAG_EMP_PHONE	*Did client provide work phone (1=YES, 0=NO)*
FLAG_WORK_PHONE	*Did client provide home phone (1=YES, 0=NO)*
FLAG_CONT_MOBILE	*Was mobile phone reachable (1=YES, 0=NO)*
FLAG_PHONE	*Did client provide home phone (1=YES, 0=NO)*

FLAG_EMAIL	*Did client provide email (1=YES, 0=NO)*
OCCUPATION_TYPE	*What kind of occupation does the client have*
CNT_FAM_MEMBERS	*How many family members does client have*
REGION_RATING_CLIENT	*Our rating of the region where client lives (1,2,3)*
REGION_RATING_CLIENT_W_CITY	*Our rating of the region where client lives with taking city into account (1,2,3)*
WEEKDAY_APPR_PROCESS_START	*On which day of the week did the client apply for the loan*
HOUR_APPR_PROCESS_START	*Approximately at what hour did the client apply for the loan*
REG_REGION_NOT_LIVE_REGION	*Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)*
REG_REGION_NOT_WORK_REGION	*Flag if client's permanent address does not match work address (1=different, 0=same, at region level)*
LIVE_REGION_NOT_WORK_REGION	*Flag if client's contact address does not match work address (1=different, 0=same, at region level)*
REG_CITY_NOT_LIVE_CITY	*Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)*
REG_CITY_NOT_WORK_CITY	*Flag if client's permanent address does not match work address (1=different, 0=same, at city level)*
LIVE_CITY_NOT_WORK_CITY	*Flag if client's contact address does not match work address (1=different, 0=same, at city level)*
ORGANIZATION_TYPE	*Type of organization where client works*
EXT_SOURCE_1	*Normalized score from external data source*
EXT_SOURCE_2	*Normalized score from external data source*
EXT_SOURCE_3	*Normalized score from external data source*
APARTMENTS_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
BASEMENTAREA_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
YEARS_BEGINEXPLUATATION_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
YEARS_BUILD_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
COMMONAREA_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
ELEVATORS_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
ENTRANCES_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FLOORSMAX_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FLOORSMIN_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LANDAREA_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LIVINGAPARTMENTS_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LIVINGAREA_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
NONLIVINGAPARTMENTS_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
NONLIVINGAREA_AVG	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
APARTMENTS_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
BASEMENTAREA_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
YEARS_BEGINEXPLUATATION_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
YEARS_BUILD_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
COMMONAREA_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
ELEVATORS_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
ENTRANCES_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FLOORSMAX_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FLOORSMIN_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LANDAREA_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LIVINGAPARTMENTS_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LIVINGAREA_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
NONLIVINGAPARTMENTS_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
NONLIVINGAREA_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
APARTMENTS_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*

BASEMENTAREA_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
YEARS_BEGINEXPLUATATION_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
YEARS_BUILD_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
COMMONAREA_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
ELEVATORS_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
ENTRANCES_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FLOORSMAX_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FLOORSMIN_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LANDAREA_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LIVINGAPARTMENTS_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
LIVINGAREA_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
NONLIVINGAPARTMENTS_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
NONLIVINGAREA_MEDI	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
FONDKAPREMONT_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
HOUSETYPE_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
TOTALAREA_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
WALLSMATERIAL_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
EMERGENCYSTATE_MODE	*Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of*
OBS_30_CNT_SOCIAL_CIRCLE	*How many observation of client's social surroundings with observable 30 DPD (days past due) default*
DEF_30_CNT_SOCIAL_CIRCLE	*How many observation of client's social surroundings defaulted on 30 DPD (days past due)*
OBS_60_CNT_SOCIAL_CIRCLE	*How many observation of client's social surroundings with observable 60 DPD (days past due) default*
DEF_60_CNT_SOCIAL_CIRCLE	*How many observation of client's social surroundings defaulted on 60 (days past due) DPD*
DAYS_LAST_PHONE_CHANGE	*How many days before application did client change phone*
FLAG_DOCUMENT_2	*Did client provide document 2*
FLAG_DOCUMENT_3	*Did client provide document 3*
FLAG_DOCUMENT_4	*Did client provide document 4*
FLAG_DOCUMENT_5	*Did client provide document 5*
FLAG_DOCUMENT_6	*Did client provide document 6*
FLAG_DOCUMENT_7	*Did client provide document 7*
FLAG_DOCUMENT_8	*Did client provide document 8*
FLAG_DOCUMENT_9	*Did client provide document 9*
FLAG_DOCUMENT_10	*Did client provide document 10*
FLAG_DOCUMENT_11	*Did client provide document 11*
FLAG_DOCUMENT_12	*Did client provide document 12*
FLAG_DOCUMENT_13	*Did client provide document 13*
FLAG_DOCUMENT_14	*Did client provide document 14*
FLAG_DOCUMENT_15	*Did client provide document 15*
FLAG_DOCUMENT_16	*Did client provide document 16*
FLAG_DOCUMENT_17	*Did client provide document 17*
FLAG_DOCUMENT_18	*Did client provide document 18*
FLAG_DOCUMENT_19	*Did client provide document 19*
FLAG_DOCUMENT_20	*Did client provide document 20*
FLAG_DOCUMENT_21	*Did client provide document 21*
AMT_REQ_CREDIT_BUREAU_HOUR	*Number of enquiries to Credit Bureau about the client one hour before application*
AMT_REQ_CREDIT_BUREAU_DAY	*Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)*
AMT_REQ_CREDIT_BUREAU_WEEK	*Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)*
AMT_REQ_CREDIT_BUREAU_MON	*Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)*
AMT_REQ_CREDIT_BUREAU_QRT	*Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)*
AMT_REQ_CREDIT_BUREAU_YEAR	*Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)*

Fuente: Kaggle, Elaboración propia

ANEXO II: Modelo predictivo del suceso de impago: código RStudio

```
###LIBRERÍAS
#remove.packages("rlang")
#install.packages("rlang")
library(rlang)
#install.packages("corrplot")
library(corrplot)
#install.packages("ggplot2")
library(ggplot2)
#install.packages("caret")
library(caret)
#install.packages("glmnet")
library(glmnet)
#install.packages("boot")
library(boot)
#install.packages("MASS")
library(MASS)
#install.packages("tidyverse")
library(tidyverse)
library(Matrix)
#install.packages("rpart")
#install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
#install.packages("kernlab")
library(kernlab)
#install.packages("ROCR")
library(ROCR)
#install.packages("caretEnsemble")
library(caretEnsemble)
#install.packages("Rcpp")
library(Rcpp)
#install.packages("e1071")
library(e1071)
#install.packages("scales")
library(scales)
library(dplyr)
#install.packages("recipes")
library(recipes)
#install.packages("tidyr")
library(tidyr)
```

```
#install.packages("tidymodels")
library(tidymodels)
library(pROC)
###CARGA DE DATOS
 ## En primer lugar, fijamos directorio de trabajo
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
getwd()
 ##En segundo lugar, cargamos el conjunto de datos
library(readxl)
data <- read_excel("Home Credit - TFG data.xlsx")
View(data)
###ANALISIS EXPLORATORIO
summary(data)
str(data)
dim(data)
 ##Analizamos la variable target, creada anteriormente
str(data$TARGET)
tabla1<-table(data$TARGET)
tabla1/nrow(data)
barplot(tabla1, xlab='TARGET',
        ylab='Frecuencia absoluta', col = c("lightblue", "lightgreen"))
 ##Analisis de correlaciones
 #Para analizar la matriz de correlaciones, es necesario contener únicamente datos numéricos.
 #Para ello, creamos un dataset solo con variables numéricas
datanum<- data[,c(6,7,8,9,13,14,15,16,18,19)]
View(datanum)
round(cor(datanum),2)
 ##Como podemos ver, si se ejecuta el comando "cor" tal y como hemos hecho, no funciona, debido a que incluimos los
NA
correlacion<- round(cor(datanum, use='complete.obs'),2)
View(correlacion)
corrplot(correlacion, method="pie", type="upper")
### ADECUACION DE LA BASE DE DATOS Y SELECCION DE PREDICTORES
 ##Tratamiento de NAs
   #Primero miramos si hay algún valor perdido en el conjunto de datos
any(is.na(data))
   #Como vemos que si lo hay, miramos que variables contienen NA y cuantos hay
sapply(data, anyNA)
sum(is.na(data))
sum(is.na(data$AMT_GOODS_PRICE))
```

```
sum(is.na(data$EXT_SOURCE_2))

  # Mirando los datos, podemos ver como la mayoría de valores perdidos se encuentran concentrados en las mismas observaciones

  # Esto, sumado al gran numero de observaciones con las que contamos (no hay necesidad de imputar valores por falta de datos)

  # Y a la distribución de los valores perdidos (los NA se encuentran en muchas variables, algo que se hace muy difícil de tratar)

  # Nos lleva a decidir ELIMINAR LAS OBSERVACIONES CON VALORES PERDIDOS del dataset

data <- na.omit(data)

  # Comprobamos que el dataset no contiene valores vacíos (NA)

any(is.na(data))

dim(data)


 ##TRATAMIENTO DE OUTLIERS

  # Primero dibujaremos un boxplot de los outliers de la variable Businesses_100m

boxplot(data$AMT_CREDIT,

      horizontal = FALSE,

      col = rgb(0, 0.4, 0.8, alpha = 0.4), # Color red/green/blue  # Etiqueta eje X

      ylab = "Amount",  # Etiqueta eje Y

      main = "Antes", # Titulo

      border = "black",  # Color del borde del boxplot

      whiskcol = "black", # Color de los bigotes

      whisklty = 3,     # Tipo de linea para los bigotes

      lty = 1) # Tipo de linea (caja y mediana)

  #Remplazamiento de outliers

impute_outliers <- function(x, removeNA = TRUE){

 quantiles <-  c(quantile(x,prob=0.25)-1.5*IQR(x), quantile(x,prob=0.75)+1.5*IQR(x))

 x[x<quantiles[1]] <- mean(x)

 x[x>quantiles[2]] <- median(x)

 x

}

  #Suistituimos los outliers en nuestro dataset

data$AMT_CREDIT<- impute_outliers(data$AMT_CREDIT)

data$AMT_INCOME_TOTAL<- impute_outliers(data$AMT_INCOME_TOTAL)

data$AMT_ANNUITY<- impute_outliers(data$AMT_ANNUITY)

data$AGE<- impute_outliers(data$AGE)

data$AMT_GOODS_PRICE<- impute_outliers(data$AMT_GOODS_PRICE)

data$YEARS_ID_PUBLISH<- impute_outliers(data$YEARS_ID_PUBLISH)

data$CNT_FAM_MEMBERS<- impute_outliers(data$CNT_FAM_MEMBERS)

data$HOUR_APPR_PROCESS_START<- impute_outliers(data$HOUR_APPR_PROCESS_START)

data$EXT_SOURCE_2<- impute_outliers(data$EXT_SOURCE_2)
```

```
   #Mostramos el nuevo Boxplot con outliers tratados
boxplot(data$AMT_CREDIT,
        horizontal = FALSE,
        col = rgb(0, 0.4, 0.8, alpha = 0.4), # Color red/green/blue  # Etiqueta eje X
        ylab = "Amount",  # Etiqueta eje Y
        main = "Despues", # Titulo
        border = "black",  # Color del borde del boxplot
        whiskcol = "black", # Color de los bigotes
        whisklty = 3,      # Tipo de linea para los bigotes
        lty = 1) # Tipo de linea (caja y mediana)


  ##TRANSFORMACION DE VARIABLES
   #DICOTOMIZAMOS LA VARIABLE NAME_CONTRACT_TYPE
   #Para ello, asignamos el valor 0 cuando la variable tome el valor Cash loan y 1 cuando tome el valor Revolving loan
data$NAME_CONTRACT_TYPE<-as.factor(ifelse(data$NAME_CONTRACT_TYPE=="Cash loans",0,1))
   #DICOTOMIZAMOS LA VARIABLE CODE_GENDER
   #Para ello, asignamos el valor 0 cuando la variable tome el valor M y 1 cuando tome el valor F
data$CODE_GENDER<-as.factor(ifelse(data$CODE_GENDER=="M",0,1))
   #DICOTOMIZAMOS LA VARIABLE FLAG_OWN_CAR
   #Para ello, asignamos el valor 0 cuando la variable tome el valor N y 1 cuando tome el valor Y
data$FLAG_OWN_CAR<-as.factor(ifelse(data$FLAG_OWN_CAR=="N",0,1))
   #DICOTOMIZAMOS LA VARIABLE FLAG_OWN_REALTY
   #Para ello, asignamos el valor 0 cuando la variable tome el valor N y 1 cuando tome el valor Y
data$FLAG_OWN_REALTY<-as.factor(ifelse(data$FLAG_OWN_REALTY=="N",0,1))
   #DICOTOMIZAMOS LA VARIABLE NAME_INCOME_TYPE
   #Al haber muchos posibles valores, dividimos en dos grupos principales: Employed (1) y Not Employed (0)
data$NAME_INCOME_TYPE<-as.factor(ifelse(data$NAME_INCOME_TYPE=="Unemployed" |
data$NAME_INCOME_TYPE=="Student" | data$NAME_INCOME_TYPE=="Pensioner" |
data$NAME_INCOME_TYPE=="Maternity leave",0,1))
   #DICOTOMIZAMOS LA VARIABLE NAME_EDUCATION_TYPE
   #Al haber muchos posibles valores, dividimos en dos grupos principales: Secondary (1) y Higher (0)
data$NAME_EDUCATION_TYPE<-as.factor(ifelse(data$NAME_EDUCATION_TYPE=="Secondary / secondary
special" | data$NAME_EDUCATION_TYPE=="Lower secondary",1,0))
   #DICOTOMIZAMOS LA VARIABLE WEEKDAY_APPR_PROCESS_START
   #Al haber muchos posibles valores, dividimos en dos grupos principales: Weekend (1) y Weekday (0)
data$WEEKDAY_APPR_PROCESS_START<-
as.factor(ifelse(data$WEEKDAY_APPR_PROCESS_START=="SATURDAY" |
data$WEEKDAY_APPR_PROCESS_START=="SUNDAY"|
data$WEEKDAY_APPR_PROCESS_START=="FRIDAY",1,0))
   #DICOTOMIZAMOS LA VARIABLE NAME_FAMILY_STATUS
   #Dividimos los posibles valores de la variable en  dos grupos principales: Individual(0) y Couple (1)
data$NAME_FAMILY_STATUS<-as.factor(ifelse(data$NAME_FAMILY_STATUS=="Married" |
```

```
data$NAME_FAMILY_STATUS=="Civil marriage", 1,0))
  #GRAFICOS INTERESANTES
tabla3<-table(data$TARGET, data$NAME_CONTRACT_TYPE)
tabla4<-table(data$TARGET, data$CODE_GENDER)
tabla5<-table(data$TARGET, data$FLAG_OWN_CAR)
tabla6<-table(data$TARGET, data$FLAG_OWN_REALTY)
tabla7<-table(data$TARGET, data$NAME_INCOME_TYPE)
tabla8<-table(data$TARGET, data$NAME_EDUCATION_TYPE)
tabla9<-table(data$TARGET, data$NAME_FAMILY_STATUS)
tabla10<-table(data$TARGET, data$WEEKDAY_APPR_PROCESS_START)
Conf4x2 = matrix(c(1:8), nrow=2, byrow=TRUE)
layout(Conf4x2)
layout.show(8)
barplot(tabla3, main="TARGET/CONTRATO", xlab="TIPO DE CONTRATO", ylab="OBSERVACIONES",
      legend = rownames(tabla3), col=c("lightblue", "lightgreen"))
barplot(tabla4, main="TARGET/GENERO", xlab="HOMBRE/MUJER", ylab="OBSERVACIONES",
      legend = rownames(tabla4), col=c("lightblue", "lightgreen"))
barplot(tabla5, main="TARGET/VEHICULO", xlab="CON/SIN VEHICULO", ylab="OBSERVACIONES",
      legend = rownames(tabla5), col=c("lightblue", "lightgreen"))
barplot(tabla6, main="TARGET/VIVIENDA", xlab="CON/SIN VIVIENDA", ylab="OBSERVACIONES",
      legend = rownames(tabla6), col=c("lightblue", "lightgreen"))
barplot(tabla7, main="TARGET/INGRESO", xlab="TIPO DE INGRESO", ylab="OBSERVACIONES",
      legend = rownames(tabla7), col=c("lightblue", "lightgreen"))
barplot(tabla8, main="TARGET/EDUCACION", xlab="TIPO DE EDUCACION", ylab="OBSERVACIONES",
      legend = rownames(tabla8), col=c("lightblue", "lightgreen"))
barplot(tabla9, main="TARGET/ESTADO FAMILIAR", xlab="ESTADO FAMILIAR", ylab="OBSERVACIONES",
      legend = rownames(tabla9), col=c("lightblue", "lightgreen"))
barplot(tabla10, main="TARGET/DIA DE APLICACION", xlab="DIA DE APLICACION",
ylab="OBSERVACIONES",
      legend = rownames(tabla10), col=c("lightblue", "lightgreen"))
  #Volvemos a fijar los gráficos para que solo aparezca uno
Conf1x1 = matrix(c(1:1), nrow=1, byrow=TRUE)
  #Normalizamos las variables numericas
data[,c(6,7,8,9,13,14,15,16,18,19)]<-scale(data[,c(6,7,8,9,13,14,15,16,18,19)])
View(data)
  #Creamos una base de datos sin las variables AMT_CREDIT y AMT_GOOD_PRICE, muy correlacionadas entre si
data1<-data[,c(1,2,3,4,5,6,8,10,11,12,13,14,15,16,17,18,19)]
View(data1)
 ###PARTICIÓN TRAINING SET (70%) Y TEST SET (80%)
RNGkind("Super", "Inversion", "Rounding")
```

```
   #Establecemos la semilla para tener la misma partición
set.seed(1357)
   #Emplearemos la función create Data Partition, que asegura que hay la misma proporción de 1 y 0 en la muestra de entrenamiento y en el test set
index<-createDataPartition(data$TARGET, p = 0.7, list = FALSE )
train<- data[index,]
test<-data[-index,]
   #Hacemos lo mismo sobre la base de datos "data1"
index<-createDataPartition(data1$TARGET, p = 0.7, list = FALSE )
train1<- data1[index,]
test1<-data1[-index,]
   #Visualizamos como queda distribuida la variable target
sum(train$TARGET)
sum(test$TARGET)
str(train$TARGET)
tabla11<-table(train$TARGET)
tabla11/nrow(train)
tabla12<-table(test$TARGET)
tabla12/nrow(test)
View(train)
View(test)
 ##SELECCION DE PREDICTORES-Modelo Explicativo
   #Usamos la función glm (General Linear Model) y creamos nuestro full model
modelo<-glm((TARGET)~.,family = binomial(logit), data = data)
summary(modelo)
#Hacemos lo mismo sobre la base de datos "data1"
modelo<-glm((TARGET)~.,family = binomial(logit), data = data1)
summary(modelo)
###ENTRENAMIENTO Y TESTING DE MODELOS
  ##MODELO PREDICTIVO LOGIT
   #Llevaremos a cabo la predicción mediante un modelo stepwise
   #Utilizaremos el siguiente modelo
modelo<-glm((TARGET)~.,family = binomial(logit), data = data)
   #Hacemos numérica la variable target para poder hacer calculos
test$TARGET<-as.numeric(test$TARGET)
   #Establecemos el cutoff del 8%(ya que es la probabilidad de que TARGET=1 en la base de datos de entrada)
cutoff=0.35
   #Modelo
stepmodelo1<-stepAIC(modelo, direction="both", trace=1) #both: stepwise, trace=1 muestra el proceso
summary(stepmodelo1)
```

```
summary(stepmodelo1)
 #Prediciones
test$predictionStep1<-as.numeric(predict(modelo, newdata=test, type = "response"))
test$TARGETStep1<-ifelse(test$predictionStep1 >cutoff,1,0)
str(test$TARGETStep1)
sum(test$TARGETStep1)
 # Matriz de confusión
test[1:30,c("TARGET","TARGETStep1")]
matrizConfusion1<-table(PREDICTED= test$TARGETStep1, REAL= test$TARGET)
matrizConfusion1
accuracy1<-(matrizConfusion1[1,1]+matrizConfusion1[2,2])/sum(matrizConfusion1)*100
accuracy1
Sensibilidad1<-(matrizConfusion1[2,2])/((matrizConfusion1[2,2])+matrizConfusion1[1,2])*100
Sensibilidad1
Especificidad1<-(matrizConfusion1[1,1])/((matrizConfusion1[1,1])+matrizConfusion1[2,1])*100
Especificidad1
 #Pintamos la curva ROC
pred1 = prediction(test$predictionStep1, test$TARGET)
 # logit.reg.prediction son las predicciones que hemos calculado anteriormente
perf1 <- performance(pred1,"tpr","fpr")
 # tpr es True positive rate, y fpr es False positive rate
plot(perf1,col="red") # Dibujamos la curva ROC
abline(a=0, b= 1)
AUC <- performance( pred1, measure = "auc")
AUCaltura <- AUC@y.values
plot(perf1,
    avg= "threshold",
    colorize=TRUE,
    lwd= 3,
    main= "ROC curve Step1")
plot(perf1,
    lty=3,
    col="grey78",
    add=TRUE)
abline(a=0, b= 1, col="lightblue")
  ##Algoritmo KNN
RNGkind("Super", "Inversion", "Rounding")
#Establecemos la semilla para tener la misma partición
set.seed(1357)
```

```
#Emplearemos la función create Data Partition, que asegura que hay la misma proporción de 1 y 0 en la muestra de entrenamiento y en el test set
index<-createDataPartition(data$TARGET, p = 0.7, list = FALSE )
train<- data[index,]
test<-data[-index,]
 # Setting up train controls
repeats = 2
numbers = 2
tunel = 30     #numero de valores del hiperparametro que se van a probar
RNGkind("Super", "Inversion", "Rounding")
set.seed(1357)
x = trainControl(method = "repeatedcv",
                 number = numbers,                                                    #folds
                 repeats = repeats,
                 classProbs = TRUE,
                 summaryFunction = twoClassSummary) #el resumen es el adecuado para dos clases
train$TARGET<-as.factor(ifelse(train$TARGET==1,"Y","N"))
 #Estimacion del modelo. Funcion TRAIN
modeloknn <- train((TARGET)~. , data = train,
                   method = "knn",
#algoritmo a usar
                   preProcess = c("center","scale"),
                   trControl = x,
                   metric = "ROC",
                   tuneLength = tunel)
modeloknn
plot(modeloknn)
test_class <- predict(modeloknn, newdata=test)          #prediccion clasificada
test_pred <- predict(modeloknn, newdata=test, type="prob")#prediccion probabilidades numericas
view(test)
test$predictionsknn<-test_pred["Y"]
test$TARGETknn1<-ifelse(test$predictionsknn >cutoff,1,0)
test$TARGET<-as.factor(test$TARGET)
matrizConfusion2<-table(PREDICTED= test$TARGETknn1, REAL= test$TARGET)
matrizConfusion2
accuracy2<-(matrizConfusion2[1,1]+matrizConfusion2[2,2])/sum(matrizConfusion2)*100
accuracy2
Sensibilidad2<-(matrizConfusion2[2,2])/((matrizConfusion2[2,2])+matrizConfusion2[1,2])*100
Sensibilidad2
Especificidad2<-(matrizConfusion2[1,1])/((matrizConfusion2[1,1])+matrizConfusion2[2,1])*100
```

```
Especificidad2
 #PLOT ROC KNN
library(ROCR)
test$TARGET<-as.factor(ifelse(test$TARGET==1,"Y","N"))
pred_test <-prediction(test_pred[,2],test$TARGET)
 # Calculating Area under Curve (AUC)
perf_test <- performance(pred_test,"auc")
perf_test1<-performance(pred_test,"tpr", "fpr")
auc=paste0("AUC=",round(perf_test@y.values[[1]],3))
auc
plot(perf_test1,
    avg= "threshold",
    colorize=TRUE,
    lwd= 3,
    main= "ROC curve KNN")
plot(perf_test1,
    lty=3,
    col="grey78",
    add=TRUE)
abline(a=0, b= 1, col="lightblue")
text(auc, x=.1, y=.8)
##MODELO PREDICTIVO ARBOL DE CLASIFICACION
RNGkind("Super", "Inversion", "Rounding")
#Establecemos la semilla para tener la misma partición
set.seed(1357)
#Emplearemos la función create Data Partition, que asegura que hay la misma proporción de 1 y 0 en la muestra de entrenamiento y en el test set
index<-createDataPartition(data$TARGET, p = 0.7, list = FALSE )
train<- data[index,]
test<-data[-index,]
 #En primer lugar, empleamos un árbol que, por defecto, tiene 3 valores del parámetro de complejidad (hiperparámetro)
RNGkind("Super", "Inversion", "Rounding")
set.seed(1357)
modeloarbol1<-train(TARGET~., data=train, method="rpart")
modeloarbol1
plot(modeloarbol1)
 #Pintamos el árbol
rpart.plot(modeloarbol1$finalModel)
rpart.rules(modeloarbol1$finalModel)
 ##Realizamos el mismo proceso, pero esta vez determinamos el valor del hiperparametro mediante cross validation
```

```
numbers <- 10
rep<-4
Grid<-expand.grid(cp=c(seq(0.008,0.9,by=0.002)))  #en este caso los valores a probar del hiperparametro cp van de 0.008 a 0.3 en pasos de 0.002
RNGkind("Super", "Inversion", "Rounding")
set.seed(1357)
Controls<- trainControl(method = "repeatedcv",        # método de remuestreo, en este caso cross validation repetida
                number = numbers,       # número de folds
                repeats = rep,
                classProbs = TRUE,                              # se obtienen las probabilidades de clasificación
                summaryFunction=twoClassSummary,
                verboseIter = TRUE
)
 #Estimación del modelo. Función TRAIN
train$TARGET<-as.factor(ifelse(train$TARGET==1,"Y","N"))
modeloarbol1 <- train(TARGET~. , data = train,
                method = "rpart",                                                                       #algoritmo a usar
                trControl = Controls,      #cómo se realiza el proceso de training
                tuneGrid = Grid,        # la malla de valores de cp (complexity parameter) que se prueban
                metric="ROC      # se selecciona el mejor modelo según el valor de la AUC
 # Resumen del modelo
modeloarbol1
plot(modeloarbol1)
 #Pintamos el arbol
rpart.plot(modeloarbol1$finalModel)
 # reglas del arbol
rpart.rules(modeloarbol1$finalModel)
 #importancia de las variables, está escalado desde 100 la variable que más aporta a las particiones
varImp(modeloarbol1)
 #prediccion
predtree<-predict(modeloarbol1, newdata=test, type="prob")
predictiontree<-as.factor(ifelse(predtree[,2] >0.35,1,0))
matrizConfusiontree<-table(PREDICTED= predictiontree, REAL= test$TARGET)
matrizConfusiontree
accuracytree<-(matrizConfusiontree[1,1]+matrizConfusiontree[2,2])/sum(matrizConfusiontree)*100
accuracytree
Senisibilidadtree<-(matrizConfusiontree[2,2])/((matrizConfusiontree[2,2])+matrizConfusiontree[1,2])*100
Senisibilidadtree
Especificidadtree<-(matrizConfusiontree[1,1])/((matrizConfusiontree[1,1])+matrizConfusiontree[2,1])*100
Especificidadtree
```

```
 #Deshacemos la transformacion de la variable
train$TARGET<-as.factor(ifelse(train$TARGET=="Y",1,0))
 #calculamos el AUC y la curva ROC
predictree=prediction(predtree[,2], test$TARGET)
AUC<-performance(predictree,"auc")
AUCtree <- AUC@y.values
# Calculating Area under Curve (AUC)
roctree <- performance(predictree,"tpr","fpr")
plot(roctree,
    avg= "threshold",
    colorize=TRUE,
    lwd= 3,
    main= "ROC curve Árboles")
plot(roctree,
    lty=3,
    col="grey78",
    add=TRUE)
abline(a=0, b= 1, col="lightblue")
text(AUCtree, x=.1, y=.8)
###MODELO PREDICTIVO SUPPORT VECTOR MACHINE
RNGkind("Super", "Inversion", "Rounding")
#Establecemos la semilla para tener la misma partición
set.seed(1357)
 #Emplearemos la funcion create Data Partition, que asegura que hay la misma proporción de 1 y 0 en la muestra de entrenamiento y en el test set
index<-createDataPartition(data$TARGET, p = 0.7, list = FALSE )
train<- data[index,]
test<-data[-index,]
 #Transformamos la variable train$TARGET
train$TARGET<-as.factor(ifelse(train$TARGET==1,"Y","N"))
 #Establecemos el control, ya que estableceremos el valor de los hiperparametros mediante el empleo de crossvalidation
control<-trainControl(method="repeatedcv", number=2, repeats = 2, classProbs = TRUE, summaryFunction = twoClassSummary, search = "random")
##Introducimos las técnicas que emplearemos
modelLookup("svmRadial")
modelLookup("svmLinear")
modelLookup("svmPoly")
 #Entrenamos el SVM Lineal
RNGkind("Super", "Inversion","Rounding")
set.seed(1357)
```

```
svmlin<-train(TARGET~., data=train, method="svmLinear", trControl=control, metric="ROC")
  #Comprobamos los procesos realizados
 #SVM LINEAL
svmlin
plot(svmlin,pch=19, cex=2.5)
 # Calculamos las predicciones de estos
predsvmlin<-predict(svmlin, newdata=test, type="prob")
predsvmlin1<-as.factor(ifelse(predsvmlin[,2] >0.35,1,0))
predict2=prediction(predsvmlin[,2], test$TARGET)
AUC<-performance(predict2,"auc")
AUClin <- AUC@y.values
matrizConfusionsvm<-table(PREDICTED= predsvmlin1, REAL= test$TARGET)
matrizConfusionsvm
accuracysvm<-(matrizConfusionsvm[1,1]+matrizConfusionsvm[2,2])/sum(matrizConfusionsvm)*100
accuracysvm
Senisibilidadsvm<-(matrizConfusionsvm[2,2])/((matrizConfusionsvm[2,2])+matrizConfusionsvm[1,2])*100
Senisibilidadsvm
Especificidadsvm<-(matrizConfusionsvm[1,1])/((matrizConfusionsvm[1,1])+matrizConfusionsvm[2,1])*100
Especificidadsvm
rocsvm <- performance(predict2,"tpr","fpr")
plot(rocsvm,
    avg= "threshold",
    colorize=TRUE,
    lwd= 3,
    main= "ROC curve Árboles")
plot(rocsvm,
    lty=3,
    col="grey78",
    add=TRUE)
abline(a=0, b= 1, col="lightblue")
text(AUClin, x=.1, y=.8)
 #Deshacemos la transformación de la variable
train$TARGET<-as.factor(ifelse(train$TARGET=="Y",1,0))
```