uc3m | Universidad Carlos III de Madrid

MÁSTER UNIVERSITARIO EN CIENCIAS ACTUARIALES Y FINANCIERAS
2022/2024

"EARLY WARNING SYSTEM (EWS): MODELOS DE INTELIGENCIA ARTIFICIAL PARA ALERTA TEMPRANA DE RIESGO DE DEFAULT BANCARIO"

TRABAJO DE FIN DE MÁSTER

Autor:
Raúl Alonso Cancino Reyes

Tutora:
Raquel Pérez Calderón

Madrid, 13 de junio de 2024

Resumen

El presente trabajo final de máster (TFM) se centra en el desarrollo de un Sistema de Alerta Temprana (EWS) para la evaluación del riesgo crediticio, específicamente enfocado en la predicción de la Probabilidad de Default (PD) mediante diversos modelos de Inteligencia Artificial (IA) en el ámbito de crédito. Utilizando Python como plataforma de programación, se presenta un análisis exhaustivo que abarca desde la estructura de la base de datos sintética empleada hasta la implementación y evaluación de modelos predictivos.

La metodología empleada se enfoca en técnicas avanzadas, comenzando por la transformación y selección de variables relevantes, seguido por la aplicación de modelos de aprendizaje automático supervisado y no supervisado. Los resultados muestran que los modelos de boosting, en particular CatBoost, superan a otras técnicas en términos de rendimiento predictivo, destacando su capacidad para capturar patrones complejos en los datos.

El análisis de variables revela correlaciones significativas con el riesgo de default, subrayando la importancia de considerar no solo las variables en su totalidad, sino también sus rangos específicos. A pesar de la alta capacidad predictiva del modelo, se identifican ciertos desafíos para mejorar la precisión del modelo y la calibración de los umbrales de clasificación.

Este estudio confirma la viabilidad y la importancia de los sistemas avanzados de alerta temprana en la gestión del riesgo crediticio, proporcionando una herramienta viable y valiosa para la toma de decisiones financieras informadas y contribuyendo a la estabilidad de las instituciones bancarias.

Palabras clave: Sistema de alerta temprana, Probabilidades de default, Basilea III, Banca.

Abstract

This master's thesis focuses on the development of an Early Warning System (EWS) for credit risk assessment, specifically focused on the prediction of the Probability of Default (PD) using various Artificial Intelligence (AI) models in the credit field. Using Python as a programming platform, a comprehensive analysis is presented, ranging from the structure of the synthetic database used to the implementation and evaluation of predictive models.

The methodology employed focuses on advanced techniques, starting with the transformation and selection of relevant variables, followed by the application of supervised and unsupervised machine learning models. The results show that boosting models, particularly CatBoost, outperform other techniques in terms of predictive performance, highlighting their ability to capture complex patterns in the data.

The analysis of variables reveals significant correlations with default risk, highlighting the importance of considering not only the variables, but also their specific ranges. Despite the high predictive power of the model, some challenges are identified to improve the accuracy of the model and the calibration of the classification thresholds.

This study confirms the feasibility and importance of advanced early warning systems in credit risk management, providing a viable and valuable tool for informed financial decision-making and contributing to the stability of banking institutions.

Keywords: Early warning system, Default probabilities, Basel III.

Agradecimientos

A mi madre, que vaya donde vaya estará conmigo,
A mi padre, por haberme transmitido la curiosidad por aprender,
A mi hermana, mi primera amiga,
A mi abuelo, por siempre acompañarme,
A Amy, por que estando tan lejos me hace sentir en casa,
Y a mis amigos que día a día me muestran su aprecio.

Índice

Introducción

La evolución de los créditos ha sido una de las actividades más dinámicas y cruciales en el sector bancario, impulsando el crecimiento económico y facilitando el desarrollo empresarial y el poder adquisitivo de los consumidores. No obstante, las crisis financieras a lo largo de la historia han demostrado la necesidad de una regulación estricta para asegurar la estabilidad del sistema financiero. Esto llevó a la creación del marco regulatorio de Basilea, esencial para una gestión centralizada y efectiva del riesgo crediticio.

Para abordar estos desafíos, los bancos han desarrollado diversas estrategias de gestión del riesgo crediticio, entre las cuales destacan los sistemas de alerta temprana (EWS). Estos sistemas permiten a las instituciones anticiparse a posibles incumplimientos y tomar medidas preventivas.

La motivación para estudiar la aplicación de un EWS en este Trabajo de fin de Máster (TFM) radica en la interrogante de cómo mejorar la solidez y estabilidad de las instituciones financieras a través de una gestión proactiva del riesgo crediticio. En un contexto donde los bancos enfrentan crecientes desafíos debido a la volatilidad económica y la globalización, implementar un EWS puede prevenir potenciales pérdidas económicas para cualquier entidad financiera. Además, el avance de la inteligencia artificial ofrece técnicas para mejorar la precisión y eficiencia en la detección de riesgos financieros, contribuyendo así al fortalecimiento de los sistemas de alerta temprana del sector bancario. Este TFM busca explorar y aprovechar estas tecnologías para proporcionar soluciones prácticas y avanzadas que pueden ser potencialmente adoptadas por instituciones financieras para prevenir el riesgo de default.

En este TFM, se empleará una base de datos sintética de libre acceso obtenida en Kaggle, correspondiente a un banco mediano que ofrece préstamos de consumo y avances de efectivo. Este TFM se centra en el desarrollo de un Sistema de Alerta Temprana basado en los parámetros de Probabilidad de Incumplimiento (PD). Para alcanzar este objetivo, se utilizarán técnicas de inteligencia artificial, una herramienta poderosa que permite analizar grandes volúmenes de datos de manera eficiente y precisa. Mediante el uso de técnicas de aprendizaje automático, tanto supervisado como no supervisado, se buscará agrupar a los individuos según su nivel de riesgo, proporcionando así una solución avanzada y automatizada para la gestión proactiva del riesgo crediticio. Esto mejorará la solidez y la estabilidad financiera de la institución.

La investigación está estructurada en varios capítulos. La primera parte presenta una introducción al tema propuesto y al marco regulatorio. La siguiente parte del documento expone los fundamentos teóricos y los modelos matemático-estadísticos, incorporando técnicas de inteligencia artificial aplicables al modelado del riesgo crediticio. La sección final ilustra los resultados obtenidos mediante los diferentes métodos propuestos, seguida de una discusión detallada.

1. Breve Historia Bancaria: Inicios de la banca moderna al contexto de crédito contemporáneo.

1.1 Orígenes.

La historia de la banca, tal como se conoce hoy en día, surge principalmente de las relaciones entre mercaderes, prestamistas y comerciantes, quienes, como parte de sus actividades, comenzaron a diversificar servicios como el resguardo de dinero y operaciones financieras básicas (Sánchez Marcos, M. 2021).

La consolidación de la banca como un sistema moderno se inicia en el Renacimiento, con hitos importantes como la creación del Banco de Venecia, que contribuyó al desarrollo económico y comercial de la región mediterránea occidental (Igual Luis, D. 1998. Valencia e Italia en el siglo XV: Rutas, mercados y hombres de negocios en el espacio económico del Mediterráneo occidental). Su modelo de operación, que incluía el mantenimiento de reservas completas y la facilitación de préstamos, se convirtió en un estándar para otras instituciones en Europa. Con el tiempo y la expansión mercantil, surgieron más bancos en Europa, consolidando un nuevo estilo de institución moderna similar a lo que conocemos hoy.

Sin embargo, el contexto bancario nunca ha estado -ni probablemente estará- ajeno a crisis. En un contexto más contemporáneo, la Gran Depresión de 1929, con epicentro en los Estados Unidos y gigantes repercusiones globales, marcó uno de los acontecimientos más devastadores en la historia económica moderna, afectando profundamente el sistema financiero y la confianza en la economía. En menos de una semana, la Bolsa de Nueva York cayó, y una percepción de vulnerabilidad invadió el lugar que en aquellos años se consideraba una tierra de oportunidades, debilitando el denominado "Sueño Americano" (Moncayo, J. 2019. 1929: el mayor apocalipsis financiero. Diario La Vanguardia).

Los avances tecnológicos impulsados por la Revolución Industrial en décadas anteriores trajeron consigo un crecimiento económico y de consumo masivo en múltiples industrias, siendo pioneras la automotriz y la de electrodomésticos. Sin embargo, a pesar de que en septiembre de 1929 se alcanzaron máximos históricos, solo un mes más tarde se vivió el famoso "Jueves Negro" y posteriormente el "Martes Negro", lo que desencadenó millonarias pérdidas económicas, devastando reputaciones y quebrantando la confianza del consumidor común.

Aunque inicialmente muchos expertos consideraron la situación un traspié, rápidamente se convirtió en una profunda y duradera recesión -una depresión desencadenada, como se le conoce-, cuya magnitud marcó un antes y después. El sistema económico entró en una nueva dinámica, mostrando su fragilidad intrínseca y creando nuevas interrogantes frente a los desafíos que la creciente globalización traería consigo (Fuller, E. W. 2019. La banca del 100% y sus defensores: una breve historia. Mises Wire).

Como consecuencia de la Gran Depresión y la necesidad de una reconstrucción económico-social, así como la búsqueda de estabilidad financiera después de la Segunda Guerra Mundial, se creó el Banco Mundial en 1944 durante la Conferencia de Bretton Woods. Esta institución fue fundada con la intención de establecer mecanismos financieros globales y transversales para prevenir futuras crisis económicas y financieras (World Bank History, World Bank 2010). Inicialmente, el Banco Mundial se centró en proporcionar financiamiento y asistencia para impulsar el crecimiento económico y reducir la pobreza, aunque en la actualidad abarca áreas como la educación, la sanidad y la infraestructura.

Otro punto relevante en el contexto de la Gran Depresión y el período de posguerra fue la creación del Fondo Monetario Internacional (FMI). También fundado durante la Conferencia de Bretton Woods, a diferencia del Banco Mundial, el objetivo principal del FMI era fomentar la cooperación internacional en términos monetarios para estimular la estabilidad financiera entre los estados miembros. Sus políticas se orientaban a la asistencia financiera en términos de balanzas de pago, ajustes estructurales, estabilización de tipos de cambio y la promoción de prácticas alineadas con una buena salud financiera (IMF. 2022).

Tras la creación del Banco Mundial y el FMI, como parte de los esfuerzos de recuperación de la Gran Depresión y el período de posguerra, se desarrollaron nuevas formas de financiamiento para impulsar el consumo y el crecimiento económico. Una de las más innovadoras fue la invención de las tarjetas de crédito (Forbes Magazine, Frankel, R. S. 2021). Aunque surgieron en la década de 1930, su uso masificado comenzó con la "Diners Club Card" en 1950, inicialmente destinada a gastos de viaje y entretenimiento, pero posteriormente expandida a múltiples áreas del consumo (National Museum of American History, 2020).

El uso de las tarjetas de crédito se expandió rápidamente a múltiples sectores, experimentando un auge sin precedentes. Este crecimiento se debió a la innovación que representaba la posibilidad de realizar compras sin la necesidad de efectivo y financiar compras a corto, mediano y largo plazo, ofreciendo flexibilidad financiera. Las instituciones bancarias y financieras, que emitían estas tarjetas, vieron una oportunidad para expandir sus fuentes de ingresos a través de intereses y tarifas.

Con el aumento del uso de las tarjetas de crédito, también se produjo un incremento sostenido en la competencia entre entidades financieras. Entre los años 1980 y principios de 1990, especialmente en Estados Unidos y Reino Unido, se produjo un fuerte movimiento de desregulación bancaria. Este fenómeno se basó en cambios en la percepción política sobre la influencia del gobierno en la economía, los avances tecnológicos exponenciales y la globalización (Solis-Mullen, J. 2022. Mises Wire).

En Estados Unidos, la desregulación comenzó con la eliminación de restricciones en torno a las tasas de interés que las entidades financieras podían cobrar a quienes solicitaban créditos o préstamos. Este cambio se conoció como la Ley de Depósitos Bancarios (Depository Institutions Deregulation and Monetary Control Act). La implementación de esta ley incrementó la competencia entre las entidades financieras, redujo las barreras de entrada y facilitó las adquisiciones y fusiones entre compañías del sector (Depository Institutions Deregulation and Monetary Control Act of 1980).

Otro evento importante en el contexto de la desregulación bancaria fue la derogación de la Ley Glass-Steagall en 1999 (Dodd, R. 2009. La reforma del sistema financiero en Estados Unidos propone la reforma más radical de la regulación financiera desde el New Deal. Finanzas y Desarrollo, FMI). Esta ley, promulgada durante la Gran Depresión, había separado las actividades bancarias comerciales de las de inversión. Su derogación permitió una mayor integración y desarrollo de un sistema bancario más complejo, extendiéndose a nivel internacional y creando un mercado financiero global y conectado entre diferentes países.

Aunque la desregulación bancaria trajo grandes beneficios, también planteó nuevos desafíos, como una mayor exposición al riesgo sistémico, lo que eventualmente contribuyó a la crisis financiera global de 2008. Conocida también como la Gran Recesión, esta crisis sacudió los cimientos del sistema financiero mundial, donde la desregulación bancaria y las prácticas financieras arriesgadas sentaron las bases para el desastre económico que siguió (Ocampo, J. A. 2010. Impactos de la crisis financiera mundial. CEPAL).

La desregulación bancaria facilitó la proliferación de prácticas financieras arriesgadas, como la concesión de préstamos hipotecarios subprime (Oficina para la Protección Financiera del Consumidor. 2024. ¿Qué es una hipoteca de alto riesgo o subprime?). Estos préstamos, destinados a prestatarios con historiales crediticios deficientes, se empaquetaron en productos financieros complejos, como los activos respaldados por hipotecas(MBS) (Rocket Mortgage. 2020) y se vendieron a inversores en todo el mundo. La búsqueda de mayores ganancias y rendimientos llevó a una relajación de los estándares crediticios y a una explosión de la deuda hipotecaria, alimentando una burbuja inmobiliaria que eventualmente estallaría con consecuencias catastróficas.

La burbuja inmobiliaria, caracterizada por un aumento vertiginoso en los precios de la vivienda, creó una ilusión de riqueza y estabilidad financiera. Sin embargo, esta bonanza se basaba en cimientos frágiles, alimentados por préstamos hipotecarios insostenibles y una confianza excesiva en la capacidad de los mercados inmobiliarios para seguir creciendo indefinidamente. Cuando los precios de la vivienda comenzaron a caer en 2006 (Brainrenews, 2020), los prestatarios se encontraron con hipotecas que superaban el valor de sus propiedades, lo que provocó un aumento en los incumplimientos de pagos y las ejecuciones hipotecarias.

La crisis hipotecaria en Estados Unidos no se limitó a las fronteras del país, sino que tuvo repercusiones globales debido a la interconexión del sistema financiero mundial. Los MBS y otros productos financieros derivados, vendidos y negociados en mercados internacionales, propagaron los riesgos asociados con los activos tóxicos vinculados a las hipotecas subprime. La quiebra de grandes instituciones financieras, como Lehman Brothers, y la congelación de los mercados crediticios desencadenaron una crisis sistémica que se extendió por todo el mundo, sumiendo a las economías en una recesión profunda y prolongada (Pozzi, S. 2019. 10 años de la crisis. Economía. Diario EL PAIS).

La crisis financiera de 2008 dejó a su paso un rastro de devastación económica, con millones de personas perdiendo sus empleos, hogares y ahorros. Reveló las deficiencias y fragilidades del sistema financiero global, así como la necesidad urgente de una regulación más estricta y una supervisión más efectiva para prevenir futuras crisis (Wenjie Chen et al. 2018). Además, resaltó la importancia de una gestión prudente de riesgos y una cultura financiera basada en la transparencia, la responsabilidad y la ética empresarial. En última instancia, la crisis de 2008 sirvió como un recordatorio doloroso pero necesario de los peligros de la complacencia y la codicia desmedida en los mercados financieros,

sugiriendo la necesidad constante de supervisar la calidad crediticia y la concesión de créditos (Ponz, C.S. 2020).

Aunque muchos consideran que la crisis del subprime podría haberse evitado y, por ende, haber prevenido el colapso de los mercados bursátiles en tal magnitud, hay ciertos eventos que son inevitables dada su naturaleza impredecible.

La crisis del COVID-19 emergió como un evento completamente inesperado y sin precedentes en la escena económica mundial, con ramificaciones que afectaron profundamente al riesgo de crédito y a las medidas exógenas de los mercados financieros (OMS, 2020). Esta crisis, desencadenada por una pandemia global, evidenció la vulnerabilidad inherente a la interconexión y la interdependencia de los mercados financieros en un mundo globalizado. La propagación rápida y descontrolada del virus llevó a medidas de confinamiento y cierres generalizados, lo que impactó directamente en la actividad económica (Adrian, T. et al, 2023) y, por ende, en la capacidad de pago de los deudores (Crespo, L. et al, 2023).

Uno de los efectos más inmediatos fue el aumento del riesgo de crédito, ya que muchas empresas y particulares se vieron imposibilitados de cumplir con sus obligaciones financieras debido a la paralización de la actividad económica y la pérdida de ingresos. Esta situación generó una creciente preocupación entre los acreedores sobre la solvencia de sus deudores, lo que se tradujo en un aumento de las tasas de morosidad y en una mayor reticencia de los bancos y otras instituciones financieras para otorgar nuevos préstamos (IMF, “Global Financial Stability Report”, 2021).

Además, las medidas exógenas adoptadas para contener la propagación del virus, como los confinamientos obligatorios y las restricciones a la movilidad, resultaron en un entorno económico altamente impredecible y volátil (Adrian, T. et al, 2020). Estas medidas, aunque necesarias desde una perspectiva de salud pública, tuvieron un impacto significativo en la actividad empresarial, el consumo y la inversión, lo que dificultó aún más la evaluación y gestión del riesgo crediticio.

1.2 Créditos; la evolución del consumo

Como se mencionó en el apartado anterior, los diversos modelos de tarjetas de crédito fueron evolucionando a lo largo de las décadas, con la introducción de nuevas tarjetas de consumo masivo por parte de distintas empresas y bancos. Sin embargo, ya en 1914, algunas tarjetas ofrecían créditos con un nicho específico; por ejemplo, Western Union emitió una tarjeta de crédito, seguida por otras compañías como las compañías petroleras, que lanzaron sus propias versiones para la compra de gasolina (Muy Interesante Magazine, 2024).

Conforme la economía estadounidense creció, el consumo aumentó significativamente, lo que llevó a una expansión exponencial de la demanda de créditos. Esta situación resultó en un auge bancario en la oferta de préstamos personales y créditos, permitiendo a los consumidores financiar sus compras y aumentar su poder adquisitivo (Aguilar Corbacho, G. 2015).

Este auge sentó las bases para el desarrollo de los mercados financieros más avanzados en los años 60 con los bonos. A medida que los consumidores tomaban más crédito, las instituciones financieras buscaban nuevas formas de financiar esos préstamos. Los bonos se convirtieron en una herramienta clave para canalizar fondos hacia el mercado de crédito al consumidor, permitiendo a los prestamistas obtener liquidez al vender sus préstamos en forma de títulos (Alvargonzález, V. 2023).

Esto impulsó el crecimiento de los mercados de bonos y proporcionó a los inversores una nueva clase de activos en la que invertir. Además, la diversificación de los productos financieros derivados del crédito al consumidor, respaldados por bonos, ayudó a mejorar la eficiencia y la liquidez en los mercados financieros (Nash, M. 2023).

Con el paso del tiempo, se innovaron nuevas formas de financiamiento hipotecario, como los préstamos de tasa ajustable y los préstamos con garantía colateralizada. Estas innovaciones permitieron una mayor accesibilidad al crédito hipotecario y aumentaron la liquidez en el mercado de vivienda (Oficina para la Protección del Consumidor, 2018). Esto llevó, en la década de los 80, a la creciente titulización. La titulización implica agrupar préstamos en paquetes negociables y se convirtió en una forma eficiente de distribuir el riesgo y aumentar la disponibilidad de fondos para el crédito al consumidor e hipotecario. Esto permitió a los bancos liberar capital para nuevos préstamos y facilitó la inversión en una variedad de activos respaldados por deudas (Bankinter, 2018).

El notable avance del sector financiero y bancario trajo consigo, considerando las crisis anteriores mencionadas, la necesidad de regular el sector. Así nació la regulación bancaria de Basilea, establecida por el Comité de Basilea en 1988 (BCBS, 2015), la cual introdujo estándares internacionales para la gestión del riesgo y la solvencia de los bancos. Su objetivo es fortalecer la estabilidad financiera y prevenir crisis bancarias. Las regulaciones incluyen requisitos de capital mínimo y directrices sobre la evaluación del riesgo crediticio.

En los años 2010, tras la crisis de 2008, se implementaron regulaciones más estrictas. Estas medidas, conocidas como regulaciones post-crisis, incluyeron la revisión y fortalecimiento de los estándares de Basilea, como el Acuerdo de Basilea III, que aumentó los requisitos de capital y estableció nuevos estándares para la gestión de liquidez y el riesgo sistémico (BIS, 2010).

Además, se introdujeron regulaciones específicas para controlar el riesgo en ciertas áreas, como la Ley Dodd-Frank en Estados Unidos, que supervisa los derivados financieros y refuerza la protección al consumidor. Estas regulaciones post-crisis buscaron mitigar los riesgos sistémicos y mejorar la resiliencia del sistema financiero frente a futuras crisis (Comisión Nacional del Mercado de Valores, 2024).

2. Regulación en la industria Bancaria.

El Comité de Basilea, desde su formulación, ha establecido una serie de normas internacionales para regular la actividad bancaria (Ciby, J., 2013), destacando sus acuerdos sobre adecuación del capital para adaptarse a la evolución de la industria, las innovaciones y las mejoras en el sistema (Basel Committee on Banking Supervision, 2013), tales como Basilea I, II y III.

La adecuación del capital se refiere a los niveles mínimos de capital que una institución financiera debe tener para ser operativamente solvente y poder operar con normalidad. Cuando se establecieron las bases de la supervisión global de los bancos internacionales (o aquellos con operaciones en mercados internacionales), la adecuación del capital se convirtió en el núcleo del comité, lo que trajo consigo un sólido consenso sobre el enfoque ponderado para la medición del riesgo (BIS, 2013).

2.1 Basilea I

Basilea I es el primer acuerdo internacional sobre regulación bancaria establecido por el Comité de Supervisión Bancaria de Basilea en 1988. Su objetivo principal era fortalecer la estabilidad del sistema financiero global mediante la implementación de estándares mínimos de capital que los bancos debían mantener. El acuerdo surgió en respuesta a la crisis de deuda de América Latina y a las inestabilidades financieras de la década de 1980 (BIS, 2013).

El pilar fundamental de Basilea I es el requisito de capital mínimo, que establece que los bancos deben mantener un capital mínimo equivalente al 8% de sus activos ponderados por riesgo (APR). Estos APR son los activos de un banco ajustados según el riesgo que presentan, con diferentes clases de activos asignados a categorías de riesgo específicas.

Basilea I introduce dos tipos principales de capital: el capital básico (Tier 1), que incluye el capital social y las reservas declaradas, y el capital complementario (Tier 2), que abarca otras formas de capital menos líquidas. La combinación de estos tipos de capital debe cumplir con el requisito del 8%.

El APR establece una escala ponderada de riesgos (0%, 20%, 50%, 100%). Dependiendo de la naturaleza del banco, se le asigna un porcentaje. Los bancos centrales tienen una ponderación del 0%, a diferencia de los privados, que tienen un 100%. En cuanto al nivel mínimo de capital, este se determina según la naturaleza de riesgo de un banco (que incluye el riesgo de crédito). Por ejemplo, los sustitutos directos de crédito se ponderan en un 100%, mientras que los compromisos libremente cancelables tienen un riesgo del 0% (BIS, 2013).

Aunque Basilea I representó un avance significativo en la regulación bancaria global, también fue objeto de críticas. Posterior a su implementación, se consideró demasiado simplista y no adecuadamente sensible a los diferentes niveles de riesgo de los activos. No lograba capturar otros riesgos primordiales, como el de mercado, operativo y de liquidez para el cálculo de la ponderación del riesgo, dado que su enfoque principal estaba en el riesgo crediticio.

2.2 Basilea II

Basilea II, publicado en junio de 2004, representa una evolución del marco regulatorio establecido por Basilea I. Su principal objetivo fue fortalecer la regulación, supervisión y gestión de riesgos en el sector bancario, abordando las limitaciones identificadas en Basilea I. Para lograr este objetivo, Basilea II se estructura en tres pilares fundamentales:

2.2.1 Pilar 1: Requisitos de Capital Mínimo

Este pilar mantiene el enfoque de Basilea I en los requisitos de capital mínimo, pero introduce un cálculo más avanzado y sensible al riesgo. Además de los riesgos de crédito, se consideran los riesgos operativos y de mercado. Para los riesgos de crédito, Basilea II permite dos enfoques: el método estándar y el método basado en calificaciones internas (IRB) (BIS. An explanatory note on the Basel II IRB risk weight functions, 2005).

Método Estándar: Similar a Basilea I, utiliza calificaciones externas de agencias de rating para determinar los ponderadores de riesgo de los activos. Las evaluaciones surgidas a partir de la calificación externa de riesgo comienzan desde AAA hasta las más altas (inferior a B-), llevando también la ponderación de riesgo entre el 20% y 150%.

Método IRB (Internal Ratings-Based): Permite a los bancos utilizar sus propios modelos internos para evaluar el riesgo de crédito, sujeto a la previa aprobación de los supervisores. Hay dos variantes: el IRB básico, donde los bancos estiman solo la probabilidad de incumplimiento (PD) y los supervisores proporcionan otros parámetros; y el IRB avanzado, donde los bancos estiman todos los parámetros del riesgo, incluyendo la pérdida dada la falta (LGD) y la exposición en caso de incumplimiento (EAD).

Ya sea que se use el enfoque básico o avanzado, se usa la siguiente fórmula para las perdidas esperadas (EL):

$$Pérdidas\ Esperadas\ =\ PD\ *\ LGD\ *\ EAD$$

(EL) es la estimación media de las pérdidas crediticias que un banco puede esperar dentro de marcos razonables en el próximo período (usualmente un año), y se considera como parte del costo.

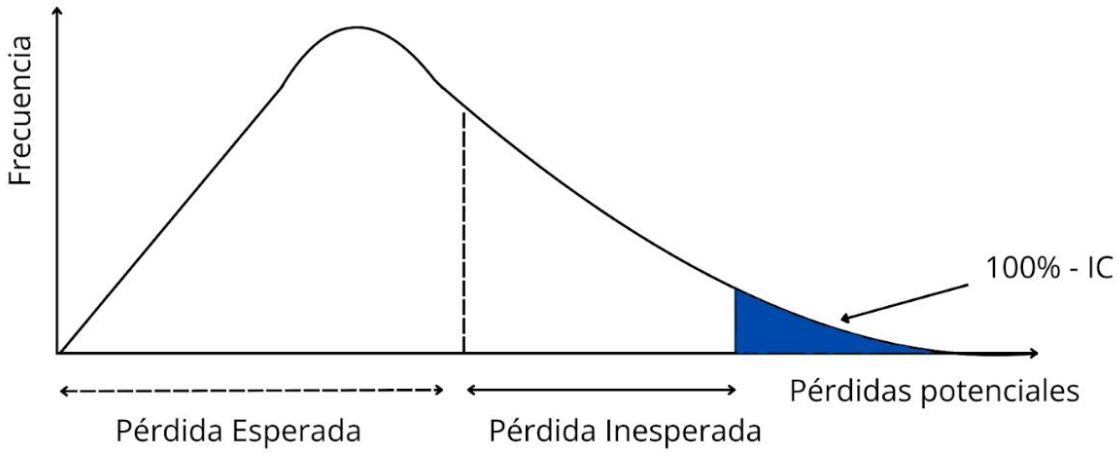
Si las pérdidas son mayores a las esperadas hablaremos de Perdidas Inesperadas (UL). El reto está en que no se sabe a ciencia cierta cuándo podrá pasar esto y predecir con exactitud asoma con el principal desafío. Es por esto que una institución financiera tiene un capital mínimo para enfrentar situaciones adversas.

Para poder determinar que:

$$Pr(Unexpected\ Losses \geq Expected\ Losses)$$

Se fija un nivel de confianza y el valor en riesgo (VaR) a dicho nivel de confianza se le denomina umbral. Por ende, la probabilidad de que el banco pueda seguir siendo solvente en el siguiente periodo es igual a dicho nivel de confianza (BIS, 2005).

Imagen 1 - VaR



Fuente: Bank for international settlements. An explanatory note on the basel ii irb risk weight functions. 2005.

El enfoque IRB (Internal Ratings-Based) de Basilea II considera factores adicionales como la correlación y la madurez en la evaluación del riesgo de crédito. En el caso de la pérdida inesperada (UL), este método asume que todos los prestatarios están vinculados por un único factor de riesgo sistemático; el estado general de la economía en general. El enfoque IRB es más sensible al riesgo que el método estándar, ya que utiliza una gama más diversa de ponderaciones de riesgo.

2.2.2 Pilar 2: Revisión Supervisora

Este pilar introdujo un proceso de revisión supervisora más robusto, incentivando a los bancos a evaluar adecuadamente sus necesidades de capital en relación con su perfil de riesgo. Los supervisores, de acuerdo a la normativa, tienen la obligación de evaluar la adecuación del capital y la estrategia de gestión de riesgos de cada banco pudiendo exigir a las entidades mantener capital adicional más allá del mínimo regulatorio si lo consideran necesario. Este proceso se conoce como Proceso de Revisión Supervisora (SREP, por sus siglas en inglés).

2.2.3 Pilar 3: Disciplina del Mercado

El tercer pilar se centra en la transparencia y la disciplina del mercado. Los bancos deben divulgar información suficiente sobre su exposición a riesgos, sus procesos de gestión de riesgos y su capital, permitiendo a los actores del mercado evaluar la solidez de la entidad. Esta transparencia tiene como objetivo fomentar un comportamiento más prudente por parte de los bancos y permitir una mejor toma de decisiones por parte de inversores y otros interesados.

Durante la crisis del subprime, Basilea II fue ampliamente criticado. Un problema significativo fue el supuesto de que la correlación disminuye con la probabilidad de incumplimiento (PD). Sin embargo, estudios demostraron que esto no siempre es cierto (Ciby, Joseph, 2013).

Además, Basilea II al asignar gran importancia a los ratings creo una alta demanda de activos con una buena calificación crediticia. Lo que pudo haber incentivado una calificación fuera de lugar con el objetivo de tener menos requisitos de capital.

Otra crítica a Basilea II es que, aunque introdujo la consideración de riesgos operativos y de mercado, no abordó adecuadamente el riesgo de liquidez, un factor crucial revelado por la crisis financiera. La falta de atención al riesgo de liquidez dejó a muchos bancos expuestos a problemas de financiación a corto plazo, exacerbando las dificultades durante la crisis.

2.3 Basilea III

Basilea III es un conjunto de reformas regulatorias desarrolladas por el Comité de Supervisión Bancaria de Basilea en respuesta a las deficiencias reveladas durante la crisis financiera de 2008. Publicado en 2010 y con implementación gradual hasta 2019, Basilea III tiene como objetivo fortalecer la regulación, supervisión y gestión de riesgos en el sector bancario global.

Uno de los principales enfoques de esta normativa es aumentar los requisitos de capital de los bancos. A diferencia de su anterior versión, se introduce un aumento en la calidad y cantidad del capital, con un énfasis en el capital de nivel 1 (Tier 1), que incluye el capital común y las reservas retenidas. El ratio de capital mínimo de nivel 1 se incrementó del 4% al 6%, y se introdujo un nuevo requerimiento de capital común de nivel 1 del 4.5%.

Además, Basilea III establece colchones de capital adicionales para absorber pérdidas durante periodos de estrés financiero. El colchón de conservación de capital (2.5%) y el colchón contracíclico (hasta 2.5%) están diseñados para asegurar que los bancos acumulen capital en tiempos de bonanza que puedan usar en tiempos de crisis.

Para determinar los requisitos de capital según Basilea III, los bancos emplearán datos estresados (stress tests) con el fin de eliminar factores cíclicos que puedan afectar el cálculo. Por lo tanto, la PD calculada en situaciones de estrés, considerando los peores escenarios posibles, será la utilizada para determinar el capital mínimo requerido.

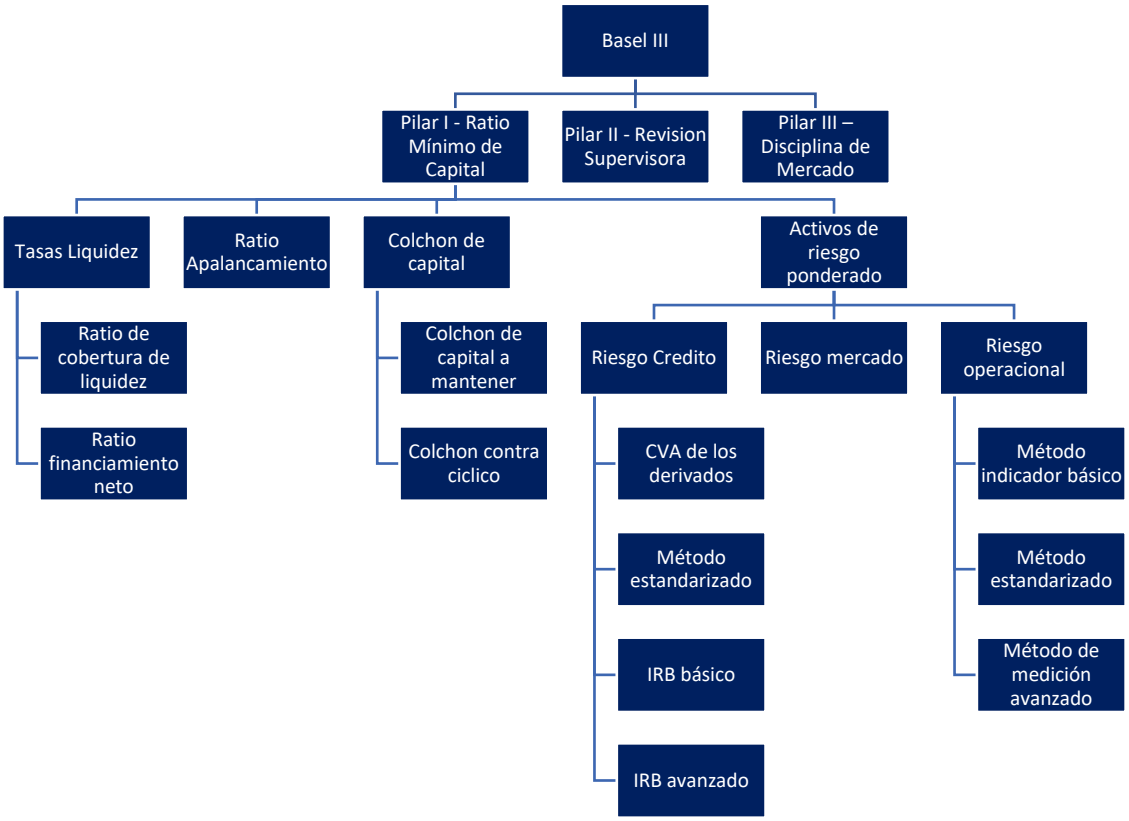
Una innovación clave en la actualización de esta normativa es la introducción de requisitos de liquidez, los cuales no estuvieron presentes en sus dos versiones anteriores.

- Ratio de cobertura de liquidez (LCR) asegura que los bancos mantengan suficientes activos líquidos de alta calidad para cubrir las salidas netas de efectivo durante un periodo de 30 días de estrés financiero.
- El Ratio de financiación estable neto (NSFR) requiere que los bancos mantengan una proporción estable de fuentes de financiación a largo plazo en relación con sus activos y actividades fuera del balance a lo largo de un horizonte de un año.

Basilea III también introduce un límite de apalancamiento para evitar la excesiva toma de riesgos y reducir el apalancamiento excesivo. Este ratio no ponderado por riesgo establece un mínimo del 3%, asegurando que los bancos mantengan un nivel mínimo de capital en relación con su exposición total.

Además, se mejoraron las medidas para gestionar el riesgo sistémico y el riesgo de contraparte. Basilea III impone mayores requerimientos de capital para las exposiciones a entidades financieras y para las transacciones derivadas, reflejando mejor los riesgos asociados (Anexo Imágenes 2).

Imagen 2 - Basilea III



Fuente: Joseph, C. (2013). Advanced Credit Risk Analysis and Management.

2.3.1 El core de Basilea; Probabilidad de Default (PD)

La dificultad de estimar con precisión, confiabilidad y en el momento adecuado la ocurrencia de estos eventos representa un reto significativo para las instituciones financieras.

La probabilidad de default (PD) es una métrica clave en la gestión del riesgo de crédito bancario. En un esfuerzo por reducir la variabilidad en los requisitos de capital, la Autoridad Bancaria Europea ha emitido directrices para estandarizar y mejorar la precisión de estos modelos. La calibración de los modelos de PD es crucial para asegurar resultados precisos y confiables que consideren diversas condiciones económicas. El default se define generalmente como la situación en la que un deudor no ha realizado pagos en más de 90 días. Las PD suelen calcularse para un horizonte de un año, proporcionando una estimación de la probabilidad de que un deudor incumpla sus obligaciones dentro de ese período (EBA, 2017).

Para desarrollar un modelo predictivo de PD efectivo, es esencial la identificación adecuada de variables relevantes y la correcta clasificación de los deudores. Estas variables pueden incluir datos cuantitativos como ingresos, patrimonio y comportamiento de pago, así como datos cualitativos como el tipo de empleo y garantías. Un grupo de riesgo homogéneo, basado en características similares de los deudores, es fundamental para evitar el solapamiento que podría distorsionar el análisis estadístico. Aunque el análisis estadístico es vital para la selección de variables, el criterio experto también juega un papel crucial. Este criterio ayuda a refinar el modelo al incorporar información adicional que puede no ser capturada completamente por los datos cuantitativos. Las tasas de default observadas en un año dependen en gran medida de la asignación experta de los ratings. Por ejemplo, si un banco utiliza un proceso de rating exclusivamente ligado a factores externos que tengan relación directa a fluctuaciones de la economía, los ratings pueden cambiar significativamente.

La calibración del modelo implica ajustar los parámetros para que las PD reflejen con precisión las probabilidades de default observadas. Esto requiere un equilibrio entre la sensibilidad del modelo a diferentes factores y la estabilidad de las PD a lo largo del tiempo. Un modelo bien calibrado permite a las entidades bancarias agrupar a los deudores en grupos de riesgo homogéneos, lo cual es crucial para realizar un análisis estadístico preciso y para predecir de manera confiable si habrá un default.

Es fundamental también considerar el impacto de factores macroeconómicos en las probabilidades de default. La inclusión de variables como la tasa de desempleo, la inflación y el crecimiento del PIB puede mejorar significativamente la precisión del modelo. Además, las condiciones económicas globales pueden influir en el comportamiento del crédito, haciendo necesario que los modelos sean adaptativos y puedan ajustarse a cambios en el entorno económico.

La integración de sistemas de alerta temprana en el marco regulatorio existente, como Basilea III, puede proporcionar una ventaja adicional. Estos sistemas permiten a los bancos identificar potenciales señales tempranas de deterioro en la calidad del crédito y tomar medidas preventivas. La implementación de tecnologías avanzadas como el machine learning y la inteligencia artificial en estos modelos predictivos puede mejorar la capacidad de los bancos para anticipar defaults y gestionar proactivamente el riesgo de crédito.

3. Modelamiento Matemático - Estadístico

En este capítulo, se exponen las diversas técnicas de inteligencia artificial que se emplearán para calcular la probabilidad de default. Además, se presentarán a nivel matemático las ecuaciones correspondientes a cada algoritmo.

Inteligencia Artificial

La inteligencia artificial (IA) constituye un campo de la informática dedicado a la creación de sistemas capaces de realizar tareas que habitualmente requieren inteligencia humana. Entre estas tareas se incluyen el reconocimiento de voz, el aprendizaje, la planificación y la resolución de problemas. Los sistemas de IA se valen de algoritmos y modelos matemáticos para procesar datos y tomar decisiones fundamentadas en dichos datos (Banda, H. 2014).

Cuanto más elaborado sea el razonamiento de la IA, mayor será la cantidad de datos y la capacidad de procesamiento requerida. El propósito fundamental de la IA es desarrollar sistemas capaces de razonar, comprender, planificar, aprender y adaptarse a problemas específicos.

Los sistemas de inteligencia artificial pueden emplear una amplia gama de técnicas, entre las que se incluyen el aprendizaje automático (Machine Learning - ML) y el aprendizaje profundo (Deep Learning - DP), así como la lógica simbólica y otros enfoques.

Aprendizaje automático/Machine Learning

El aprendizaje automático (ML) es una subdisciplina de la inteligencia artificial centrada en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y mejorar a partir de la experiencia. En lugar de seguir instrucciones programadas explícitamente, los sistemas de ML analizan datos y reconocen patrones para realizar predicciones o tomar decisiones (Banda, H. 2014).

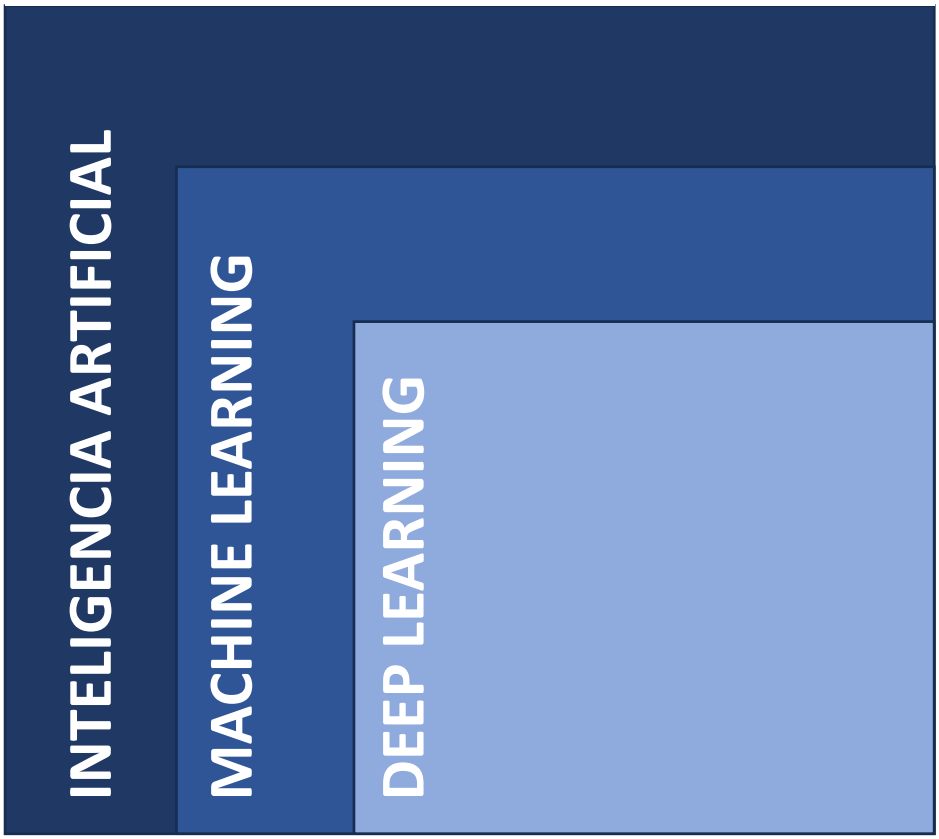
Existen varios tipos de aprendizaje automático, incluyendo el aprendizaje supervisado, donde el modelo se entrena con datos etiquetados, y el aprendizaje no supervisado, donde el modelo descubre estructuras ocultas en datos sin etiquetas. Otro tipo es el aprendizaje por refuerzo, donde un agente aprende a tomar decisiones mediante ensayo y error, recibiendo "recompensas o castigos" según su desempeño. Las aplicaciones de ML son diversas y van desde la detección de fraudes y el análisis predictivo en finanzas hasta la personalización de contenido en plataformas de streaming, contribuyendo a la innovación y eficiencia en diversas industrias.

Aprendizaje profundo/Deep Learning

El aprendizaje profundo (Deep Learning) es una rama del aprendizaje automático que utiliza redes neuronales artificiales con múltiples capas para modelar y comprender patrones complejos en grandes volúmenes de datos. Estas redes neuronales, inspiradas en la estructura del cerebro humano, están compuestas por nodos (neuronas) interconectados que procesan información a través de pesos ajustables. Estas redes pueden aprender representaciones jerárquicas de los datos, lo que permite una mayor precisión en tareas complejas como el reconocimiento de imágenes y el procesamiento del lenguaje natural (Banda, H. 2014).

El aprendizaje profundo es especialmente efectivo en el procesamiento de datos no estructurados, como imágenes, texto y audio. Se ha utilizado con éxito en aplicaciones como el reconocimiento de voz, la visión por computadora, la traducción automática y los vehículos

Imagen 3 - AI - ML - DL



Fuente: Elaboración propia

3.1 La lógica detrás de la clasificación y regresión.

Como se mencionó anteriormente, el marco de la inteligencia artificial se fundamenta en modelos matemático-estadísticos para construir modelos basados en datos y poder realizar predicciones (Nabanita, D. et al, 2018).

En este TFM, se comienza realizando un conjunto de técnicas de IA basadas en un problema de clasificación. Posteriormente, se prueba los diferentes modelos propuestos para determinar cuál entrega mejores predicciones en lo que respecta a probabilidades de default. Luego, se seleccionan las variables más importantes, validadas mediante algoritmos y se utiliza el modelo seleccionado con una lógica de problema de regresión para construir el Sistema de Alerta Temprana (EWS, por sus siglas en inglés). De esta manera, se compara el default real entregado por la base de datos con el potencial default entregado por el EWS predicho ahora mediante el problema de regresión.

¿Cuál es la diferencia entre un problema de regresión y clasificación?

La regresión implica predecir valores continuos basados en los datos de entrada. Por ejemplo, un algoritmo de regresión podría predecir el rating crediticio de un cliente basado en diferentes características financieras, como historial de crédito, ingresos, deudas, etc. Por otro lado, en un problema de clasificación, el objetivo es asignar una etiqueta o categoría a un conjunto de datos. Por ejemplo, determinar si un cliente es de alto riesgo o bajo riesgo de default en función de sus características financieras (Deisenroth, M. P., et al 2020).

El resultado será modelado por una variable aleatoria comprendida entre los reales; es decir $Y \epsilon \mathbb{R}$, donde la variable aleatoria Y_i corresponderá a la predicción del individuo i-ésimo del conjunto de datos.
La clasificación es un proceso en el que un modelo separa los datos en categorías discretas. Se puede usar para predecir la probabilidad de default de un cliente. Esta predicción se podría expresar como una etiqueta binaria, donde "0" representa que el cliente no incumplirá y "1" representa que sí lo hará. El resultado será modelado por una variable aleatoria comprendida entre los valores 0 o 1 (Deisenroth, M. P., et al 2020).; es decir $Y \epsilon \mathbb{R}$, donde la variable aleatoria Y_i corresponderá a la predicción del individuo i-ésimo del conjunto de datos.

Para ambos problemas se tiene que:

- Las variables explicativas o predictivas, denotadas por x serán las utilizadas como inputs en el modelo. Donde $X_i \epsilon \mathbb{R}^P$ con valores observables del conjunto de datos definidos como $x = [x_{i1}, x_{i2}, x_{i3}, \ldots\ldots, x_{in}]$ donde n será el número total de variables utilizadas en la muestra.
- Conjunto de datos representado por $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), , \ldots, (x_n, y_n)\}$

3.2 Modelos y estadística

El Modelo estadístico; siguiendo lo anteriormente explicado, (donde ε_i representa el error aleatorio) será representado como (Fahrmeir, Ludwing et al 2021):

$$Y_i = f(X_i) + \varepsilon_i \quad \textbf{(1)}$$

El modelo estadístico proviene del aprendizaje supervisado. En este enfoque, se intenta aprender una función f a partir de un conjunto de observaciones de entrenamiento (training set). Estas consisten en valores de entrada x_i que producen salidas $f(x_i)$ como respuesta.

El algoritmo de aprendizaje ajusta su relación input/output en función de las diferencias entre la salida original y la generada. El objetivo del entrenamiento es lograr con una alta precisión generar outputs lo más parecidos a los originales (Hastie, T et al 2008).

Para el desarrollo de este estudio se utilizan los siguientes modelos:

Tabla 1 - Modelos a emplear

MODELO	AI ALGORITHM	ON-BASED
Regresión Logística	Machine Learning	Supervised Learning
Decision Tree	Machine Learning	Supervised Learning
XGBoost	Machine Learning	Ensamble learning - Boosting
LightGBM	Machine Learning	Gradient Boosting Machine
CatBoost	Machine Learning	Gradient Boosting Machine
Random Forest	Machine Learning	Ensamble learning
ANN	Deep Learning	Artificial Neural Network
Naive Bayes	Machine Learning	Supervised Learning
Adaboost	Machine Learning	Ensamble learning - Boosting
Gradient Boost Classifier	Machine Learning	Ensamble learning - Boosting

Fuente: Elaboración propia

3.2.1 Regresion Logistica

Este modelo estadístico es comúnmente utilizado para modelar una variable dependiente a través de una función logística (Turner, H, 2008). En la literatura, a veces también se le conoce como función sigmoide, y se representa matemáticamente como:

$$F(x) = \frac{e^x}{e^x+1} \quad (2)$$

Esta función tiene como objetivo principal reducir la dimensión de los valores evaluados, suponiendo que van de (-y, y) a (0,1). Aunque este tipo de modelos se utiliza generalmente para problemas binarios, también puede ser útil en problemas de regresión. Esto se logra mediante los "log-odds". Técnicamente, es una regresión lineal pero con las probabilidades utilizadas a través de logaritmos, que, mediante la función sigmoide, entregará un valor entre 0 y 1.

Si comenzamos asumiendo que $p(x)$ es una función lineal y para llevar los resultados entre 0 y 1 usaremos:

$$log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_{x1} + \ldots + \beta_{xn} \quad (3)$$

Resolviendo y despejando p(x): (por simpleza trabajaremos solo hasta β_{x1})

$$p(x) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \quad (4)$$

Posterior a que la Regresión Logística realice sus predicciones, podemos ajustar esta mediante Likelihood, donde para cada punto, en cada variable i-ésima tendría una variable predicha y-ésima.

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i} * (1 - p(x_i)^{1-y_i}) \quad (5)$$

Aplicando logaritmos tendremos que:

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} y_i \, log\, p(x_i) + (1 - y_i) log(1 - p(x_i)) \quad (6)$$

Re ordenando:

$$l(\beta_0, \beta_1) = \sum_{i=0}^{n} log 1 - p(x_i) + \sum_{i=0}^{n} y_i \, log\frac{p(x_i)}{1-p(x_i)} \quad (7)$$

Por lo que el valor de p(x) quedará definido por

$$l(\beta_0, \beta_1) = \sum_{i=0}^{n} -log 1 + e^{\beta_0+\beta} + \sum_{i=0}^{n} y_i \, (\beta_0 + \beta_{x1}) \quad (8)$$

El siguiente paso es tomar el máximo de la función definida en la ecuación anterior. Esta técnica se le conoce como Maximun Likelihood Estimation. Este método se utiliza para estimar los parámetros de la distribución de la probabilidad. La cual matemáticamente está definida como (Fahrmeir, Ludwing et al 2021):

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=0}^{n}(y_i - p(y_i\,;\beta_0,\beta))y_{ij} \quad (9)$$

3.2.2 Decision Tree

Los árboles de decisión (DT) son uno de los algoritmos de ML más populares y frecuentemente utilizados debido a su simplicidad y alta capacidad adaptativa. Pertenecen a la familia del aprendizaje supervisado, siendo aplicables tanto a problemas de regresión como de clasificación (Schmidt-Thieme, L. 2007).

Los DT se dividen el conjunto de datos analizados en grupos más pequeños y homogéneos usando reglas condicionales de programación. Cada subconjunto de características formado por la separación de los grupos es conocido como nodo. El algoritmo detrás del DT decide qué características usar y en qué punto dividir cada nodo.

En cada nodo m, que representa el subconjunto de características dentro del espacio S_m con N_m observaciones se denotará que;

$$p_{mk} = \frac{1}{N_m}\sum_{x_i \in S_m} I(y_i = k) \quad (10)$$

Es la parte que corresponde a la proporción de la clase k dentro del espacio S_m siendo $I(.)$ el indicador de la función.

DT usa el criterio de impureza de Gini para la decisión de la división del conjunto de datos en otros subgrupos. Donde, suponiendo que k toma valores del conjunto $(1,2,\ldots,z)$ y ocupando la ecuación anterior;

$$I_g(f) = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) = \sum_{k=1}^{K}(p_{mk} - {p_{mk}}^2) = \sum_{k=1}^{K}(p_{mk}) - \sum_{k=1}^{K}({p_{mk}}^2) = 1 - \sum_{k=1}^{K}({p_{mk}}^2) \quad (11)$$

Este proceso de partición de datos es iterativo hasta llegar a una profundidad predefinida o una máxima encontrada por el algoritmo.

Posteriormente, para cada nodo m del DT se trabajará en relación si corresponde a un problema de regresión o clasificación. Para ambos casos se analizan los simples analizados pertenecientes al conjunto $\mathrm{Z} = \{(x_i, y_i)\}_{i=1}^{N}$ En el caso de la clasificación el input en caso de superar un umbral se le asignará 1, en caso contrario 0. Para la regresión devuelve la media de cada predicción.

3.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) es otro algoritmo popular y altamente efectivo en técnicas de machine learning, ampliamente utilizado debido a su capacidad para producir modelos de alta precisión. Se trata de un algoritmo con técnicas de boosting pertenecientes a la familia de algoritmos de aprendizaje supervisado, aplicables tanto a problemas de regresión como de clasificación (Malik, S, et al 2020).

En XGBoost, los modelos se construyen secuencialmente a partir de árboles de decisión conocidos como débiles, y se utiliza un proceso de boosting para mejorar su precisión. Cada árbol de decisión débil se entrena de manera que se centra en los errores de predicción del modelo anterior, lo que significa que los árboles posteriores se enfocan en corregir los errores del modelo anterior.

A diferencia de un árbol de decisión convencional, donde se desarrolla completamente antes de pasar al siguiente árbol, XGBoost desarrolla árboles secuenciales, uno tras otro, minimizando una función de pérdida específica. Otra característica interesante es que este algoritmo posee la capacidad para sintonizarse con la hiperparametrización, lo que significa que no solo construye modelos precisos, sino que también es capaz de calcular cuáles son las características con mayor importancia y así potenciar el rendimiento del modelo.

El objetivo de este algoritmo es minimizar una función de pérdida, que varía según el tipo de problema. Para la regresión, se utiliza una función de pérdida cuadrática, mientras que para la clasificación, se utiliza una función de pérdida logística.

Como se mencionó anteriormente, XGBoost a partir de la combinación de los árboles de decisión débiles se construyen modelos secuenciales, con el objetivo de minimizar la función de pérdida, y este proceso se realiza de manera similar a los algoritmos de la familia de gradient boosting (Freund, Y., & Schapire, R. E. 1999).

3.2.3.1 Matemática detrás de los modelos de ensamble- gradient boosting.

En el caso de la *regresión*; la función de perdida cuadrática se define como (Freund, Yoav, et al 1996):

$$L(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2 \quad (12)$$

Donde:

- y_i es el valor real en la i-esima variable.
- $\hat{y}_i$ es la predicción realizada para la i-esima variable.

La predicción realizada se calcula como la suma de las predicciones de todos los árboles y luego se le aplica una transformación para corregir los errores. Expresado matemáticamente como (Freund, Yoav, et al 1996):

$$\hat{y}_i = \sum_{j=1}^{N} f_j(x_i) \quad (13)$$

Donde:

- N como el número total de árboles del modelo
- $f_j(x_i)$ es la predicción del j-esimo árbol para la i-esima observación.

Para calcular $f_j(x_i)$ se sigue el árbol desde la "raíz" hasta una "hoja" y se obtiene la predicción correspondiente a dicha hoja. En XGBoost, a diferencia de un DT convencional esta predicción no es directamente el valor de la hoja, sino un valor ponderado que se ajusta en cada iteración del algoritmo (Freund, Y., & Schapire, R. E. 1999).

En el caso de la *clasificación*; la función de perdida cuadrática se define como:

$$L(y_i, \hat{p}_i) = -[y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)] \quad (14)$$

Donde:

- y_i es la etiqueta real (0 o 1) en la i-esima variable.
- $\hat{p}_i$ es la probabilidad predicha de que y_i sea igual a 1.

Para calcular $\hat{p}_i$, se aplica una función de transformación logística a y_i, la cual corresponde a la suma de las predicciones de cada árbol. Denotado por:

$$\hat{p}_i = \frac{1}{1 + e^{-y_i}} \quad (15)$$

3.2.4 LightGBM

LightGBM (Light Gradient Boosting Machine) es un algoritmo de aprendizaje automático basado en árboles de decisión, conocido por su eficiencia y alta velocidad de entrenamiento. Se clasifica dentro de la familia de algoritmos de gradient boosting, una técnica de ensamble de árboles (Ke G. et al 2017).

Aunque matemáticamente las ecuaciones de cálculo son similares a las de XGBoost (al ser de la familia de Gradient boosting), LightGBM presenta diferencias significativas en su implementación y rendimiento.

Una de las principales diferencias radica en que LightGBM utiliza una estructura de árbol vertical en lugar de una estructura de árbol horizontal utilizada por XGBoost. En un árbol vertical, cada nodo representa una característica, y las divisiones se realizan de manera recursiva a lo largo de esa característica, lo que permite un acceso más rápido a las características y una menor cantidad de datos que se deben recorrer durante el proceso de división.

El método de crecimiento de árboles predeterminado en LightGBM se basa en la diferencia de ganancia (Gain), lo que contribuye a su eficiencia computacional. Este algoritmo está diseñado para ser altamente eficiente en términos de memoria y tiempo de cálculo. Su estructura de árbol vertical y su algoritmo de crecimiento permiten un entrenamiento más rápido y un uso más eficiente de los recursos computacionales, lo que lo hace especialmente útil para conjuntos de datos grandes (Ke G. et al 2017).

Además, LightGBM maneja automáticamente las características categóricas sin necesidad de preprocesamiento adicional, convirtiéndolas internamente en números enteros. También ofrece opciones de regularización, como la “poda” de árboles y la limitación del número de hojas, para evitar el sobreajuste.

3.2.5 CatBoost

CatBoost, al igual que los árboles de decisión, se utiliza para problemas de clasificación y regresión en el aprendizaje supervisado y forma parte de la familia de gradient boosting. En este algoritmo, se divide el conjunto de datos en subconjuntos más pequeños y homogéneos mediante reglas condicionales. Cada subconjunto de características formado por esta división se conoce como nodo (Prokhorenkova, L, et al 2018).

Para entender cómo funciona CatBoost, es útil conocer algunos conceptos clave. En primer lugar, el gradient boosting, en el que en cada iteración se ajusta un nuevo árbol de decisión a los residuos del modelo anterior, intentando minimizar la función de pérdida. Además, este modelo tiene la capacidad de manejar automáticamente las variables categóricas sin necesidad de codificarlas previamente, utilizando una técnica de codificación llamada target encoding para transformar estas variables en características numéricas. También emplea técnicas de regularización para evitar el sobreajuste, lo que mejora la generalización del modelo en datos no vistos. Finalmente, el proceso de entrenamiento de CatBoost consiste en entrenar árboles de decisión de forma secuencial, utilizando gradientes calculados en cada iteración para mejorar la precisión del modelo (Prokhorenkova, L, et al 2018).

Donde la lógica y fórmulas relevantes para entender CatBoost:

Función de Pérdida; ya definida en en la ecuación (1) para regresión y en (3) para clasificación.

Regularización: CatBoost utiliza regularización L1 y L2 para evitar el sobreajuste. Teniendo:

$$Regularización = \lambda \sum_{j=1}^{J} \left\| w_j \right\|^2 \qquad (16)$$

Donde:

- λ es el parámetro de regularización.
- w_j son los pesos de los nodos del árbol j.

Codificación de variables categóricas:

$$\hat{u}_c = \frac{\sum_{i=1}^{N} I(x_i=c) y_i}{\sum_{i=1}^{N} I(x_i=c)} \qquad (17)$$

Donde:

- $\hat{u}_c$ es el valor codificado para la categoría c.
- N es el numero total de muestras.
- x_i es el valor de la variable categórica de la variable i-esima.
- y_i es la etiqueta en la clasificación o el valor en la regresión de la variable i-esima.
- $I(.)$ es la función indicadora

3.2.6 Gradient Boost

Gradient Boosting (GB) es una técnica de ensamble utilizada en aprendizaje supervisado perteneciente a la familia de los algoritmos que llevan el mismo nombre. La idea principal, que lo diferencia de sus "familiares", se sustenta en ajustar cada árbol para predecir los residuos del modelo anterior, en lugar de predecir directamente la variable objetivo (Friedman, J. 2001). Para entender cómo funciona sustancialmente GB, es útil conocer algunos conceptos clave. Primero, la regularización. Esta se utiliza para controlar la complejidad del modelo y así evitar el sobreajuste. Las técnicas más comunes son la profundidad máxima o la tasa de aprendizaje. Segundo, ya habiendo definido que el objetivo es la predicción de los residuos, también se ocupa la función de pérdida (1) para regresión y (3) para clasificación.

Para el caso de los residuos en un problema de clasificación:

$$r_i = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \quad (18)$$

$$r_i = \frac{y_i - F(x_i)}{F(x_i)\,(1-F(x_i))} \quad (19)$$

Donde:

- L es la función de perdida (15).
- $F(x_i)$ es la predicción del modelo para la variable i-esima.

Para el caso de regresión:

$$r_i = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \quad (20)$$

$$r_i = y_i - F(x_i) \quad (21)$$

Donde:

- L es la función de perdida (12).
- $F(x_i)$ es la predicción del modelo para la variable i-esima.

Para la actualización del modelo, considerando el aprendizaje a partir de las ecuaciones de los residuos anteriores. Sin embargo, este aprendizaje está asociado a una tasa, la cual es un hiperparametro que controla la contribución de cada árbol al modelo final. Una tasa más baja requiere más árboles para así alcanzar un rendimiento similar y viceversa.

La actualización del modelo está definida como:

$$F^{(t)}(x) = F^{(t-1)}(x) + v * h^{(t)}(x) \quad (22)$$

Donde:

- $F^{(t)}(x)$ es la predicción del modelo en la iteración t-esima.
- $F^{(t-1)}(x)$ es la predicción del modelo en la iteración anterior a la t-esima.
- v es la tasa de aprendizaje
- $h^{(t)}(x)$ es el árbol de decisión, con el ajuste pertinente de en la iteración t-esima.

3.2.7 Naive Bayes

Este es un modelo, que tiene sus raíces en la teoría de probabilidad desarrollada en el siglo XVIII, sustentado en el Teorema de Bayes, desarrollado por el científico a quien Thomas Bayes. Este teorema describe la probabilidad de un evento, la que está basada en el conocimiento de condiciones previas relacionadas con este. Muchos expertos sustentan este modelo como el padre del machine learning.

Naive Bayes, debido a que es un modelo simple y efectivo es ampliamente utilizado, lo cual inclusive a grandes avances en otros modelos sigue presente por su alta capacidad predictiva en problemas precisos. La principal suposición detrás de este modelo es que todas las características son independientes entre sí (Mitchel, T.M, 2017).

Naive bayes, como se menciona, está sustentado en el Teorema de Bayes, el cual se define como:

$$P(y \mid x_1, x_2, x_3, \ldots, x_n) = \frac{P(y)\, P(x_1 \mid y)\, P(x_2 \mid y) \ldots\ldots P(x_n \mid y)}{P(x_1)\, P(x_2) \ldots\ldots P(x_n)} \qquad (23)$$

Donde:

- $P(y \mid x_1, x_2, x_3, \ldots, x_n)$ es la probabilidad de que el evento pertenezca a la clase y considerando las características $x_1, x_2, x_3, \ldots, x_n$
- $P(y)$ es la probabilidad a priori de la clase y.
- $P(x_i \mid y)$ es la probabilidad de la característica x_i dado que el evento pertenece a la clase y.
- $P(x_i)$ es la probabilidad de la característica x_i en el conjunto de datos.

Ahora, se determina la Función de probabilidad de Clase:

$$P(y) = \frac{Número\ de\ eventos\ de\ la\ clase\ y}{Número\ total\ de\ eventos\ en\ el\ conjunto\ de\ datos} \qquad (24)$$

Ahora, la Función de probabilidad de característica:

$$P(x_i \mid y) = \frac{Número\ de\ eventos\ con\ la\ caracteristica\ x_i\ y\ pertenecientes\ a\ la\ la\ clase\ y}{Número\ total\ de\ eventos\ en\ el\ conjunto\ de\ la\ clase\ y} \qquad (25)$$

Para así aplicar la regla de decisión Naive Bayes;

$$\hat{y} = argmax_y P(y)\ \prod_{i=1}^{n} P(x_i \mid y) \quad (26)$$

Donde:

- $\hat{y}$ es la clase predicha para el evento.
- $argmax_y$ es el valor de y que maximiza la expresión.
- n es el número de características.

¿Pero que pasa para los problemas de regresión?

En estos contextos existe una nueva versión de Naive Bayes, llamada “Naive Bayes Gaussiano”, el cual asume que las características se distribuyen normalmente. Para este caso se asume que las caracteristicas continuas siguen una distribución normal.

Se estandarizan las características, mediante z-score:

$$z_i = \frac{x_i - u_i}{\sigma_i} \quad (27)$$

Donde:

- z_i es el valor estanzarizado de la característica x_i.
- x_i es el valor original de la característica i.
- u_i es la media de la característica i en el conjunto de entrenamiento.
- σ_i es la desviación estándar de la caraceristica i en el conjunto de entrenamiento.

Ahora se calcula la Función de Probabilidad Gaussiana (Yan, L., & Cain, J. 2020):

$$P(xi \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(\frac{(xi - \mu iy)^2}{2\sigma_y^2}\right) \quad (28)$$

- $P(xi \mid y)$ es la probabilidad de la característica x_i dado que la instancia pertenece a la clase y
- μ_{iy} es la media de la característica i en la clase y.
- σ_y^2 es la varianza de la característica i en la clase y.

Así ya que se ha obtenido esto, se procede finalmente a la Regla de decisión Naive Bayes Gaussiano:

$$\hat{y} = argmax_y P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (29)$$

Donde:

- $\hat{y}$ es la clase predicha para el evento.
- $argmax_y$ es el valor de y que maximiza la expresión.
- n es el número de características.

3.2.8 Random Forest

Esta es una de las técnicas más populares para desarrollar modelos predictivos. Su simplicidad, estabilidad y adaptabilidad en tareas de regresión y clasificación la hacen atractiva para cualquier Data Scientist. Random Forest construye múltiples Decision Trees (de ahí su nombre de bosque aleatorio). Las predicciones individuales de cada árbol se agrupan para formar una predicción final. Para los problemas de clasificación, se utiliza la moda de los datos predichos, mientras que para los de regresión se utiliza la media de las predicciones. Cada árbol se entrena de manera independiente mediante una muestra aleatoria de datos y una selección aleatoria de las variables, lo que ayuda a reducir el sobreajuste estadísticamente (Cutler A. et al 2011).

A manera de detalle el algoritmo detrás de Random Forest se encuentra:

Para la construcción de los árboles del modelo:

- Selecciona de manera aleatoria el subconjunto de características para cada árbol.

La importancia de Gini en Random Forest es crucial ya que se utiliza como criterio para decidir cómo dividir los conjuntos de datos en subgrupos en la construcción de los árboles de decisión.

El índice de Gini es una medida de impureza en un conjunto de datos. En el contexto de un nodo de un árbol de decisión, Gini mide qué tan impuros son los datos en ese nodo. Un nodo se considera puro si todos los ejemplos pertenecen a la misma clase. Cuanto menor sea el índice de Gini, más puro será el nodo (Farris, F. 2010).

La formula de Gini para un nodo m con K clases es:

$$I_g(m) = 1 - \sum_{k=1}^{K}({p_{mk}}^2) \quad (30)$$

Donde:

- p_{mk} es la proporción de la clase k en el nodo m.

Para cada nodo en el árbol de decisión, Random Forest considera todas las características y todas las posibles divisiones para calcular la ganancia de información utilizando Gini. La división que produce la mayor ganancia de información se elige la división óptima para ese nodo. El proceso de partición se repite en cada nodo del árbol hasta que se alcanza un criterio de parada, como una profundidad máxima o una pureza mínima en los nodos.

- Usa Bootstrap para entrenar cada árbol, a partir de:

Creación de conjuntos de datos de entrenamiento; se generan B muestras, del mismo tamaño que el conjunto de datos de entrenamiento original D. Para luego realizar una muestra Bootstrap D_b tomando aleatoriamente observaciones del conjunto de datos.

Para cada muestra D_b se crea un DT T_b utilizando solo esa muestra y un subconjunto aleatorio de características.

- Luego usa un árbol de decisión con estos datos y características seleccionadas.
- Se procede a realizar la predicción

Como se habló anteriormente, para problema de clasificación, la predicción final se determinará por medio de la moda.

Donde sí se tienen T arboles del bosque y K clases posibles, la predicción $\hat{y}$ para la muestra x se calculará como;

$$\hat{y} = argmax_y(\sum_{t=1}^{T} I(y_t = k) \quad (31)$$

Siendo $I(.)$ la función indicadora que devolverá 1 si la condición es verdadera y 0 en caso contrario.

En el caso de los problemas de regresión, como también fue mencionado la predicción final será simplemente el promedio de las predicciones de todos los árboles.

Donde si y_t es la predicción del árbol t, entonces la predicción final $\hat{y}$ se calculará como:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \quad (32)$$

3.2.9 Artificial Neural Network

Esta técnica, Artificial Neural Network (AAN) correspondiente al tipo de aprendizaje profundo (Deep learning), tiene a capacidad de emular el funcionamiento de las redes neuronales humanas. Su alta capacidad de resolver problemas complejos con es una de las características más interesantes de este algoritmo. Dentro de cada conjunto neuronal, se visualizan una señal de entrada, un nodo y la salida, la cual va hacia otra neurona. En general, dependiendo de la complejidad del problema, la red será mayor o menormente profunda, lo que también evidentemente determina al número de neuronas de la red. Tal como se realiza en un cerebro humano, en una ANN se realiza sinapsis. Esta transferencia de información se multiplica por un valor peso. La lógica detrás de estos pesos es que en la conexión estos pueden aumentar o inhibir la activación de la neurona posterior. También, dado que existe este mecanismo que puede filtrar la relación entre neuronas, pero tambien existe una funcion de umbral, la cual el valor debe ser mayor a este para poder continuar a la siguiente neurona de ser el caso para luego pasar la funcion de activación (Kabbay, H, 2022).

La funcion de activación es aquella por medio se transfiere la información formada por la combinación lineal de las respectivas entradas y los pesos hacia la salida. Esta información puede tener las siguientes características según el tipo de función o problema requerido a resolver:

- Funcion identidad: no posee modificaciones la información.
- Funcion Escalón
- Función Sigmoidal.
- Etc.

Las entradas, corresponden a variables independientes, donde la totalidad de los valores son multiplicadas por sus respectivos pesos (multiplicación de vectores). Lo que algebraicamente se ve como (Lichtner-Bajjaoui, A. 2020):

$$X * W^t = (x_1, x_2, x_3, \dots., x_p) * \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_p \end{pmatrix} = \sum_{i=1} x_1 * w_i \qquad (33)$$

Donde:

- X vector de entrada correspondiente a los $(x_1, x_2, x_3, \dots., x_p)$ valores.
- W^t vector traspuesto de los $\begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_p \end{pmatrix}$ pesos.

Y tambien algebraicamente una funcion de activación donde:

$$\phi(\sum_{i=1} x_1 * w_i) \qquad (34)$$

Donde:

- ϕ corresponde a la funcion de activación.

El valor de la salida, corresponde a una nueva entrada para otra neurona creando asi la Neural Network. Cabe señalar que puede ser ocupada para un problema de regresión como para de clasificación.

3.2.9.1 Problema de Clasificación:

- Clasificación binaria

$$\hat{y} = \sigma(\sum_{i=1}^{n} x_1 * w_i + \varepsilon) \quad (35)$$

  - $\hat{y}$ es la predicción.
  - x_i son las entradas.
  - w_i son los pesos asociados a cada entrada.
  - ε es el sesgo (bias).

  - Donde σ corresponde a la función sigmoidal.

$$\sigma = \frac{1}{1+e^{-x}} \quad (36)$$

- Clasificación Multiclase

$$\hat{y}_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad (37)$$

  - Donde:
    - $\hat{y}_j$ es la probabilidad de la clase j.
    - e^{z_j} es la entrada para la clase j.
    - K es el número de clases.

3.2.9.2 Problemas de Regresión

$$\hat{y} = f(\sum_{i=1}^{n} x_1 * w_i + \varepsilon) \quad (38)$$

  - Donde:
    - $\hat{y}$ es la predicción.
    - f es la funcion de activación.
    - x_i son las entradas.
    - w_i son los pesos asociados a cada entrada.
    - ε es el sesgo (bias).

3.3 Model Performance

Después de haber explorado los diversos modelos predictivos de Inteligencia Artificial en el apartado anterior, es fundamental comprender cómo evaluar la efectividad de estos modelos en la detección de potenciales probabilidades de default para definir el EWS. En este nuevo apartado, se ofrece al lector los conocimientos necesarios para la necesaria comprensión de la evaluación del rendimiento de los modelos, ofreciendo una guía clara y didáctica para entender cómo medir su eficacia.

Para empezar, se discuten algunas técnicas de trabajo clave que son fundamentales para preparar los datos y optimizar el rendimiento de los modelos. Desde la reducción de la dimensionalidad con PCA[1] hasta el proceso de ingeniería de características, explorando cómo estas prácticas pueden mejorar la capacidad predictiva de cualquier modelo en general.

Posteriormente, se estudia la evaluación de resultados, donde se analizan métricas cruciales para medir el rendimiento de los modelos. Desde métricas clásicas como la precisión (Accuracy) y la puntuación F1 (F1-Score) hasta técnicas más avanzadas como la validación cruzada (Cross-Validation) además de métodos como MAE, MSE y RMSE explorando cómo estas métricas proporcionan una visión detallada de la eficacia de los modelos.

Al comprender y aplicar estas técnicas, se poseen las herramientas para así aplicar los modelos y poder evaluar de manera precisa y eficiente, permitiendo una toma de decisiones más informada y efectiva en lo que respecta al problema propuesto.

3.3.1 Tecnicas de trabajo

Este apartado está centrado en las técnicas de trabajo fundamentales que desempeñan un papel crucial en cualquier proyecto predictivo. Estas técnicas son esenciales para garantizar la calidad y precisión de nuestros modelos, así como para optimizar su rendimiento en entornos del mundo real (Hastie, Trevor et al, 2009).

El Análisis de Componentes Principales (PCA), la Ingeniería de Características (Feature Engineering) y SMOTE (Synthetic Minority Over-sampling Technique) son pilares en la construcción y mejora de modelos predictivos. Más allá de su aplicación específica en el problema propuesto en torno a las probabilidades de default, estas técnicas tienen una importancia generalizada en cualquier problema predictivo.

PCA permite reducir la dimensionalidad de los datos, lo que resulta fundamental para lidiar con conjuntos complejos y grandes. Esta reducción no solo facilita la visualización de los datos, sino que también mejora la eficiencia de los algoritmos de Inteligencia Artificial al reducir el ruido y la redundancia que pudiese existir (Deisenroth, M. P., et al 2020)..

Por otro lado, la Ingeniería de Características es esencial para identificar y crear nuevas variables que capturen la información relevante en los datos. Esto incluye la selección de variables, la creación de variables con características polinómicas, la normalización y más, con el objetivo de mejorar la capacidad predictiva de los modelos y hacerlos más robustos frente a datos nuevos.

[1] Principal Analysis Component – (Análisis de componentes principales en español).

Por último, SMOTE aborda el desequilibrio de clases, un problema común en conjuntos de datos donde una clase es mucho más frecuente que la otra. Esta técnica es crucial para evitar sesgos en los modelos y así garantizar que sean capaces de detectar correctamente las clases minoritarias (Deisenroth, M. P., et al 2020)..

3.3.1.1 PCA

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos mientras se conserva la mayor cantidad de información posible. Esta técnica transforma un conjunto de variables posiblemente correlacionadas en un conjunto de variables no correlacionadas, denominadas componentes principales. Cada componente principal es una combinación lineal de las variables originales y está ordenado de manera que el primer componente tenga la mayor varianza posible, el segundo componente tiene la segunda mayor varianza, y así sucesivamente (Abdi, H., & Williams, L. J. 2010).

Gráficamente, PCA permite visualizar los datos en un espacio de menor dimensión (a menudo 2D), facilitando la identificación de patrones, tendencias y agrupaciones que no serían evidentes en un espacio de mayor dimensión.

Asumiendo que se tiene un determinado número de datos, cada uno k variables, PCA, como técnica avanzada, se centra en encontrar un numero de factores, el cual se denominará como v, que logre tener la misma explicabilidad que las k variables originales con una alta aproximación, para reducir la dimensionalidad para realizar un análisis, es decir la información contenida en v aunque es menor que k la información entregada hace que $k \approx v$. Siendo estas v variables llamadas componentes principales (Amat, J. 2017).

El cálculo de PCA, algebraicamente se obtiene como una combinación lineal de las variables originales. Donde la n-esima componente principal del grupo de variables $(X_1, X_2, X_3, \dots, X_P)$ es la combinación lineal normalizada de las variables con mayor varianza (Jolliffe, I. T. 2002), donde:

$$Z_n = \emptyset_{n1}X_1 + \emptyset_{n1}X_2 + \emptyset_{n1}X_3 + \cdots + \emptyset_{p1}X_p \quad (39)$$

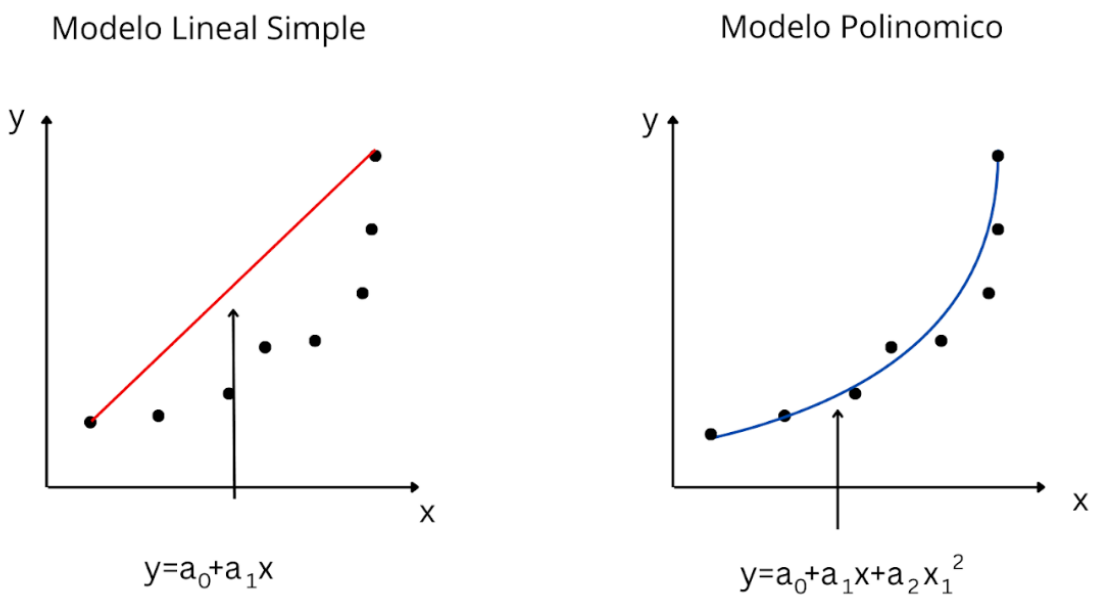
Donde:

- $\emptyset_{j1}$ corresponde al peso que tiene cada variable en cada componente.

3.3.1.2 Polynomial Features – Engineering

En el ámbito de la Inteligencia Artificial y el modelado estadístico, el análisis de regresión desempeña un papel vital en la predicción y comprensión de las relaciones entre variables. Sin embargo, en muchos escenarios, si no es que la mayoría, esta relación no es estrictamente lineal. Es aquí donde entra en juego el concepto de regresión lineal polinómica, permitiendo un modelado más con mayor flexibilidad y adaptación al problema en cuestión (Florescu, D., & England, M. 2019).

Imagen 4 - Modelo Lineal Simple vs Polinómico



Fuente: Elaboración propia

La transformación polinómica implica elevar las características existentes a potencias. Por ejemplo, si para cada dato tenemos las variables [X,Y], podemos añadir una nuevas características (columnas) cuyos valores serán los polinomios de los valores de [X,Y] denotados como X^2, Y^2, YX^2, Y^2X, XY. Este proceso, en el caso de existir más variables que los valores de ejemplo, se repite para cada característica de entrada, creando así una versión transformada de cada una (Manorathna, R. 2020).

Esta técnica de ingeniería de características añade nuevas características basadas en las existentes. El grado del polinomio determina cuántas características se añaden. Usualmente, se elige un grado bajo, como 2 o 3, para evitar que la curva polinómica se vuelva demasiado flexible y distorsione los datos provocando un sobreajuste al conjunto de entrenamiento del modelo.

3.3.1.3 SMOTE

SMOTE, abreviación para Synthetic Minority Over-Sampling Technique, es una técnica de preprocesamiento para abordar el desequilibrio de clases del conjunto de datos analizados (Fernandez, A. et al, 2013).

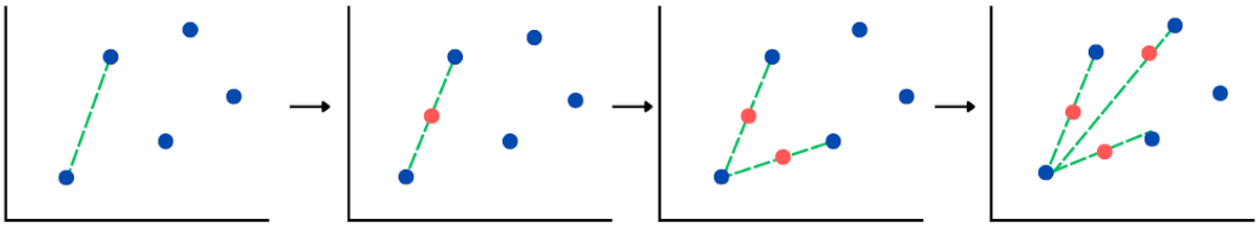
En la práctica, con frecuencia, cuando se entrena un modelo se posee un conjunto de datos con un desbalance en sus variables. Es decir, dentro del data set se cuenta con muy pocos ejemplos de una clase determinada (casos raros, poco frecuentes, outlyers, etc.) lo que lleva a un mal performance del modelo.

Puesto que en existen estos pocos frecuentes casos, no es realista buscar más casos para complementar la base de datos. Una forma de resolver este problema mediante el submuestreo de la clase mayoritaria. Es decir, excluyendo las filas correspondientes a la clase mayoritaria de forma que haya aproximadamente la misma cantidad de filas para las clases mayoritaria y minoritaria. Sin embargo, el hacer esto implica perdida de información que puede afectar negativamente a la etapa de entrenamiento perdiendo precisión predictiva. Algunas técnicas es el sobre muestreo de la clase minoritaria, es decir, aumentar aleatoriamente estas observaciones. El problema de este enfoque es que implica un sobreajuste de los datos (Chawla, et al. (2002).

Se podría continuar argumentando técnicas, sin embargo, son todas “arcaicas” o deficientes. Es en este punto donde se desarrolla el algoritmo de SMOTE, el cual:

- Toma la diferencia entre la muestra y su vecino más cercano.
- Esta diferencia se multiplica por un numero aleatorio que está entre 0 y 1.
- El resultado se agrega a la muestra generando un nuevo ejemplo sintético.
- El proceso iterativo continua al siguiente vecino, hasta un numero optimo encontrado
- por el algoritmo o predefinido por el usuario.

Imagen 5 - SMOTE



Fuente: Elaboración propia

3.3.2 Evaluación de resultados

En este apartado, se profundiza en la evaluación de los resultados generados, una etapa crucial en el desarrollo y validación de modelos predictivos. La importancia de esta fase radica en la necesidad de comprender cómo se desempeñan los modelos en la práctica y cuán confiables son en la toma de decisiones, considerando el contexto de cualquier problema que se enfrente.

Evaluar los resultados de los modelos permite determinar su capacidad para generalizar patrones y hacer predicciones precisas sobre datos nuevos y no estudiados (Deisenroth, M. P., et al 2020).. Esto es esencial para garantizar que los modelos a continuación de la etapa de entrenamiento sean efectivos y útiles en aplicaciones del mundo real.

Mediante diversas métricas y técnicas de evaluación, se puede entender la precisión, la robustez y la estabilidad de los modelos. Esto brinda información valiosa para ajustar y mejorar los modelos, así como para tomar decisiones informadas sobre su implementación (Botchkarev, A. 2018).

Se explorarán diferentes métricas de evaluación, desde medidas clásicas como la precisión (Accuracy) y el error cuadrático medio (MSE), hasta técnicas más avanzadas como el Shap Value. Cada una de estas métricas y técnicas proporcionará una perspectiva única sobre el rendimiento de los modelos, permitiendo tomar decisiones informadas y mejorar continuamente su eficacia (Zubair Khalid, 2021).

3.3.2.1 Accuracy

El Accuracy, o precisión, es una de las métricas más simples y fundamentales para evaluar el rendimiento de un modelo de clasificación. En términos simples, el Accuracy mide la proporción de predicciones correctas que realiza el modelo sobre el total de predicciones.

Matemáticamente, se calcula dividiendo el número de predicciones correctas entre el número total de predicciones realizadas por el modelo (Casal, R. et al, 2021). Es decir:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (40)$$

Donde:

- TP es el número de Verdaderos Positivos (muestras positivas correctamente clasificadas).
- TN es el número de Verdaderos Negativos (muestras negativas correctamente clasificadas).
- FP es el número de Falsos Positivos (muestras negativas clasificadas incorrectamente como positivas).
- FN es el número de Falsos Negativos (muestras positivas clasificadas incorrectamente como negativas).

Es importante tener en cuenta que el Accuracy puede ser engañoso en ciertos contextos, especialmente cuando se trata de conjuntos de datos desbalanceados, donde una clase puede ser mucho más prevalente que la otra. En estos casos, un modelo puede lograr un alto Accuracy simplemente prediciendo siempre la clase mayoritaria.

El Accuracy es útil como una primera medida para evaluar el rendimiento general de un modelo, pero no siempre proporciona una imagen completa, especialmente cuando las clases no están equilibradas.

Por ejemplo, en el contexto de un problema de detección de fraudes bancarios, donde solo el 1% de las transacciones son fraudulentas, un modelo que predice que todas las transacciones son legítimas tendría un Accuracy del 99%, pero sería inútil para detectar fraudes.

Por lo tanto, es importante considerar otras métricas además del Accuracy, como el F1-Score, especialmente en problemas de clasificación desbalanceada. Sin embargo, en problemas de clasificación binaria equilibrada, el Accuracy sigue siendo una métrica útil y fácil de interpretar (Zubair Khalid, 2021).

3.3.2.2 F1-Score

El F1-Score es una métrica que combina la precisión y el recall (sensibilidad) de un modelo en una sola medida. Se utiliza comúnmente en problemas de clasificación binaria para evaluar su rendimiento.

Recordando que:

- Precisión: La precisión mide la proporción de verdaderos positivos (TP) sobre todos los elementos clasificados como positivos (TP + FP). Es decir, cuántos de los casos que el modelo clasifica como positivos realmente lo son.
- Recall: Mide la proporción de verdaderos positivos (TP) sobre todos los casos que son realmente positivos (TP + FN). Es decir, cuántos de los casos positivos reales el modelo es capaz de detectar.

Matemáticamente, el F1-Score se calcula como la media armónica de precisión y recall. Esto significa que da más peso a los valores más bajos, lo que lo hace útil cuando hay un desequilibrio entre las clases (Casal, R. et al, 2021). Donde F1-Score se define como:

$$F1\ Score = 2 * \frac{Precisión * Recall}{Precisión + Recall} \qquad (41)$$

Donde la precisión se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos positivos, y el recall se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos. El F1-Score varía entre 0 y 1, donde 1 representa la mejor puntuación posible y 0 la peor.

Es particularmente útil cuando queremos encontrar un equilibrio entre la precisión y el recall. Por ejemplo, en la detección de fraudes, queremos minimizar tanto los falsos negativos como los falsos positivos.

El F1-Score es sensible a las clases desequilibradas, lo que lo hace útil en problemas donde las clases tienen diferentes proporciones.

3.3.2.3 Cross-Validation

Cross-validation (validación cruzada) es una técnica utilizada para evaluar el rendimiento de un modelo de aprendizaje automático de una manera más robusta y precisa. Se utiliza para estimar cómo se comportará un modelo en un conjunto de datos no visto.

La forma más común de cross-validation es la validación cruzada k-fold, donde el conjunto de datos se divide en k subconjuntos (folds) de igual tamaño. Asi, el modelo se entrena k veces, utilizando k-1 folds como datos de entrenamiento y el fold restante como datos de validación (Friedl, H., & Stampfer, E. 2001).

Matematicamente la ecuación usada para el cálculo del k-fold del cross-validation se define como (Casal, R. et al, 2021):

$$CV = \frac{1}{K}\sum_{i=1}^{k} R_i \qquad (42)$$

Donde R representa el rendimiento del modelo analizado en el fold i.

Después de realizar k entrenamientos y validaciones, se calcula un promedio de las métricas de evaluación (por ejemplo, Accuracy, F1-Score) obtenidas en cada fold. Cross-validation proporciona una estimación más confiable del rendimiento del modelo, ya que utiliza todo el conjunto de datos para entrenar y evaluar el modelo, evitando así la dependencia de una única división de datos. Además, esta técnica ayuda a detectar si un modelo está sobreajustando (overfitting) o subajustando (underfitting) los datos, ya que permite evaluar su rendimiento en diferentes particiones de datos.

3.3.2.4 MAE

El MAE (Mean Absolute Error) es una métrica comúnmente utilizada para evaluar la precisión de un modelo de regresión. Mide la magnitud promedio de los errores en las predicciones del modelo, sin considerar su dirección (Zubair Khalid, 2021).

Matemáticamente, el MAE se calcula como la media de las diferencias absolutas entre las predicciones del modelo $\hat{y}_i$ con los valores reales y_i (Casal, R. et al, 2021).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \qquad (43)$$

Donde:

- n es el numero total de la muestra.
- $\hat{y}_i$ predicción del modelo para la muestra i-esima.
- y_i es el valor real de la muestra i.

El MAE es fácil de interpretar, ya que representa el error promedio en la misma unidad que la variable de destino. Por ejemplo, si se está prediciendo el rating crediticio de un cliente (rating que va entre 1-50), un MAE de 4 significa que, en promedio, nuestras predicciones tienen un error absoluto de 4.

El MAE es menos sensible a valores atípicos en los datos que otras métricas de error, como el MSE (Mean Squared Error), lo que lo hace útil en conjuntos de datos con valores extremos.

3.3.2.5 MSE

El MSE (Mean Squared Error) es una métrica comúnmente utilizada en problemas de regresión para evaluar la calidad de las predicciones de un modelo. Calcula el promedio de los cuadrados de las diferencias entre las predicciones del modelo $\hat{y}_i$ con los valores reales y_i (Zubair Khalid, 2021).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \qquad (44)$$

Donde:

- n es el número total de la muestra.
- $\hat{y}_i$ predicción del modelo para la muestra i-esima.
- y_i es el valor real de la muestra i.

El MSE penaliza más fuertemente los errores grandes que los pequeños debido al cuadrado en la ecuación. Esto significa que los errores más grandes tienen un impacto proporcionalmente mayor en el resultado final.

El MSE tiene la ventaja de que es continuo y diferenciable, lo que lo hace útil en problemas de optimización. Sin embargo, puede ser sensible a valores atípicos en los datos, ya que eleva al cuadrado las diferencias.

Por ejemplo, si seguimos prediciendo el rating crediticio de un cliente, un MSE de 2 significa que, en promedio, nuestras predicciones tienen un error cuadrático medio de 4. Sin embargo, si fuera de 6, sería de 36, 9 veces más que el anterior.

3.3.2.6 RMSE

El RMSE (Root Mean Squared Error) es una métrica ampliamente utilizada en problemas de regresión que representa la raíz cuadrada del MSE, lo que proporciona una medida de la dispersión de los errores en las predicciones del modelo en la misma escala que la variable objetivo (Casal, R. et al, 2021).

Matemáticamente, el RMSE se calcula como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (45)$$

Donde:

- n es el número total de la muestra.
- $\hat{y}_i$ predicción del modelo para la muestra i-esima.
- y_i es el valor real de la muestra i.

El RMSE es útil porque proporciona una medida de dispersión que es fácilmente interpretable en la misma unidad que la variable objetivo, a diferencia del MSE.

Comparado con el MSE, el RMSE tiene la ventaja de que sus valores están en la misma escala que la variable objetivo, lo que lo hace más intuitivo de interpretar. Siguiendo el mismo ejemplo, si estamos prediciendo el rating crediticio, un RMSE de 9 significa que, en promedio, nuestras predicciones tienen un error de 9.

Además, el RMSE penaliza los errores más grandes de manera más significativa que los errores pequeños, debido a la raíz cuadrada. Esto lo hace más sensible a los errores grandes que el MSE, lo que puede ser útil para detectar predicciones particularmente incorrectas.

3.3.2.7 Varianza Explicada

La varianza explicada es una métrica que indica la proporción de la varianza total en los datos que es explicada por el modelo. Es una medida importante para comprender qué tan bien el modelo captura la variabilidad de los datos (Casal, R. et al, 2021).

Matemáticamente, la varianza explicada se calcula como:

$$Varianza\ explicada = 1 - \frac{Varianza\ Residual}{Varianza\ Total} \quad (46)$$

Donde:

- Varianza residual es la varianza de los residuos, es decir, la diferencia entre los valores reales y las predicciones del modelo.
- Varianza total es la varianza total de los valores reales.

Un valor de varianza explicada cercano a 1 indica que el modelo explica la mayoría de la varianza en los datos, mientras que un valor cercano a 0 indica que el modelo no explica mucha varianza y puede que no se ajuste bien a los datos (Rosenthal & Rosenthal, 2011).

La varianza explicada es una métrica útil para evaluar la capacidad predictiva de un modelo y comprender cuánta información está capturando sobre los datos. Es especialmente importante en problemas donde se busca explicar la variabilidad de una variable objetivo.

3.3.2.8 Shap Value

SHAP (SHapley Additive exPlanations) es una técnica utilizada para explicar las predicciones de modelos de aprendizaje automático de manera individual y global. Su lógica se basa en asignar un valor a cada característica (feature) que contribuye al resultado de la predicción, permitiendo entender cómo cada característica influye en la salida del modelo.

Matemáticamente, el valor SHAP se calcula utilizando la teoría de juegos, específicamente utilizando los valores de Shapley. Estos valores representan la contribución promedio de una característica específica a través de todas las posibles combinaciones de características (Lundberg, S. M., & Lee, S.-I. 2017).

La ecuación para calcular el valor SHAP para una característica particular j en una predicción i es:

$$\emptyset_j{}^i = \sum S \subseteq F \setminus \{j\} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left(f\left(x_i{}^{S\cup\{j\}}\right) - f(x_i{}^S)\right) \qquad (47)$$

Donde:

- $\emptyset_j{}^i$ es el SHAP value de la característica j para la predicción i.
- F es el conjunto de todas las características.
- $x_i{}^S$ es el conjunto de características en la observación i que están presentes en el conjunto S.
- $f(x_i{}^S)$ es la predicción del modelo para la observación i con las características en S.
- $|S|$ es el número de características S.

3.3.2.9 Clasificaciones, Error tipo I y Error tipo II.

El Error de Tipo I y el Error de Tipo II son dos conceptos fundamentales en la evaluación de modelos predictivos desde la perspectiva de los resultados de clasificación, es decir, los Verdaderos Positivos (TP), los Verdaderos Negativos (TN), los Falsos Positivos (FP) y los Falsos Negativos (FN) (Kim, H.-Y. 2015).

3.3.2.9.1 Error de Tipo I (α)

El Error de Tipo I ocurre cuando se clasifica incorrectamente como positivo (1) lo que en realidad es negativo (0) (Casal, R. et al, 2021).

Matemáticamente, se calcula como:

$$Error\ tipo\ I = \frac{FP}{FP+TN} \qquad (48)$$

Es decir, es la proporción de falsos positivos (clasificaciones incorrectas como positivos) sobre el total de negativos reales.

Por ejemplo, si clasificamos como potenciales individuos que realizarán default (1) a clientes en el sistema de alerta temprana (EWS), el Error de Tipo I sería la proporción de clientes que finalmente pagaron su deuda, siendo que fuero clasificadas incorrectamente como potenciales individuos que realizarían default.

3.3.2.9.2 Error de Tipo II (β):

El Error de Tipo II ocurre cuando se clasifica incorrectamente como negativo (0) lo que en realidad es positivo (1) (Casal, R. et al, 2021).

Matemáticamente, se calcula como:

$$Error\ tipo\ II = \frac{FN}{FN+TP} \qquad (49)$$

Es decir, es la proporción de falsos negativos (clasificaciones incorrectas como negativos) sobre el total de positivos reales.

Por ejemplo, si en el EWS clasificamos como clientes que potencialmente **no** harán default (0) a individuos que finalmente terminan realizando, el Error de Tipo II sería la proporción de individuos que realizaron default clasificados incorrectamente como aquellos que no lo harían.

4. Early Warning System (EWS)

En esta sección del TFM, se aborda el proceso completo de desarrollo del Sistema de Alerta Temprana (EWS) enfocado en los parámetros de Probabilidad de Default (PD) en el ámbito crediticio, utilizando el software de programación de uso libre Python.

Se comienza explicando la estructura y composición de la base de datos sintética empleada, la cual contiene información detallada sobre el comportamiento crediticio de los clientes. En el siguiente paso se describe la metodología utilizada para el procesamiento de los datos, destacando técnicas de limpieza, transformación y selección de variables relevantes.

Posteriormente, se aborda la implementación de modelos predictivos utilizando técnicas de aprendizaje automático supervisado y no supervisado. Estos modelos son entrenados y evaluados para determinar su capacidad predictiva en cuanto a la probabilidad de que los clientes realicen default, con el fin de seleccionar el mejor modelo y las variables correspondientes. Además, a estas variables se les aplican técnicas de tramificación para la creación de un rating basado en criterio experto.

Finalmente, se presentan los resultados obtenidos a partir de la aplicación de estos modelos, evaluando su precisión en la clasificación del riesgo crediticio y su capacidad para anticipar incumplimientos basados en este rating, confeccionando así el EWS. Este análisis permite determinar el grado de eficacia del Sistema de Alerta Temprana propuesto, así como su potencial impacto en la gestión proactiva del riesgo crediticio en instituciones financieras

4.1 Predicción del Riesgo de Default

A continuación, se procede a explicar el proceso de selección del modelo que mejor predice la probabilidad de default, para su posterior uso en el Sistema de Alerta Temprana (EWS).

La base de datos sintética utilizada proviene de Kaggle (Ana Montoya, et al., 2018) y ejemplifica un banco pequeño el cual contiene información sobre clientes que han solicitado préstamos personales y avances en efectivo, también conocidos como créditos de liquidez o de consumo en algunos lugares.

4.1.1 Análisis exploratorio y descriptivo

4.1.1.1 Librería y Data shape

Para este trabajo, mediante el software de uso libre Python se utilizaron las librerías adecuadas para el análisis y procesamiento de datos (ver Anexo 1).

La base de datos sintética empleada se compone de un total de 307,511 registros de clientes independientes, cada uno con 122 características/variables (ver Anexo 2).

Tabla 2 - Descripción data set

DATA SHAPE	
Número de filas del data set	307.511
Número de columnas del data set	122
Número total de datos	37.516.342

Cuando se trabaja con una base de datos de gran tamaño, es fundamental utilizar técnicas de estadística descriptiva para comprender la naturaleza de los datos, identificar posibles valores atípicos (outliers) y detectar datos incorrectos. En Python, esto se facilita mediante la función **describe()** de las librerías de análisis de datos.

La función **describe()** proporciona un resumen estadístico de los datos, que incluye medidas como la media, la desviación estándar, los cuartiles, y los valores mínimo y máximo. A continuación, se presenta un ejemplo reducido del recuadro generado por la función **describe()** para una mejor visualización y ejemplificación (ver Anexo 3).

Tabla 3 - Estadistica descriptiva

Index	TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	...	AMT_GOODS_PRICE	CNT_FAM_MEMBERS
count	307511	307511	307511	...	307511	307511
mean	0,08	168.797	599.025	...	538.396	2
std	0,27	237.123	402.490	...	369.446	0,91
min	0	25650	45000	...	40500	1
25%	0	112500	270000	...	238500	2
50%	0	147150	513531	...	450000	2
75%	0	202500	808650	...	679500	3
max	1	117000000	4050000	...	4050000	20

Fuente: Elaboración propia

4.1.2 Trabajo categórico y variables

Posterior al estudio de las variables mediante la estadística descriptiva, se observó que los datos de DAYS_EMPLOYED requieren modificación, dado que contienen un valor erróneo debido a la naturaleza de esta variable. Esta variable representa los días trabajados al momento de solicitar el crédito y debería contener valores negativos. Por ejemplo, si un cliente tiene un valor de -1500 en DAYS_EMPLOYED, esto significa que el cliente ha estado empleado durante 1500 días laborales al momento de la solicitud de crédito.

Sin embargo, se identificó un valor positivo extremadamente alto que altera la muestra. Este valor se reemplazó con un valor nulo (ver Anexo 4). A continuación, se presenta un resumen estadístico de DAYS_EMPLOYED antes y después de la modificación.

Tabla 4 - Estadistica descriptiva DAYS_EMPLOYED

DAYS_EMPLOYED	
Count	307.511
Mean	63.815
Std	141.275,76651
Min	-17.912
25%	-2.760
50%	-1.213
75%	-289
max	365.243

Fuente: Elaboración propia

Posterior a la modificación:

Tabla 5 - Estadistica descriptiva DAYS_EMPLOYED modificada

DAYS_EMPLOYED	
Count	252137
Mean	-2384
Std	2338,360
Min	-17.912
25%	-3.175
50%	-1.648
75%	-767
max	0

Fuente: Elaboración propia

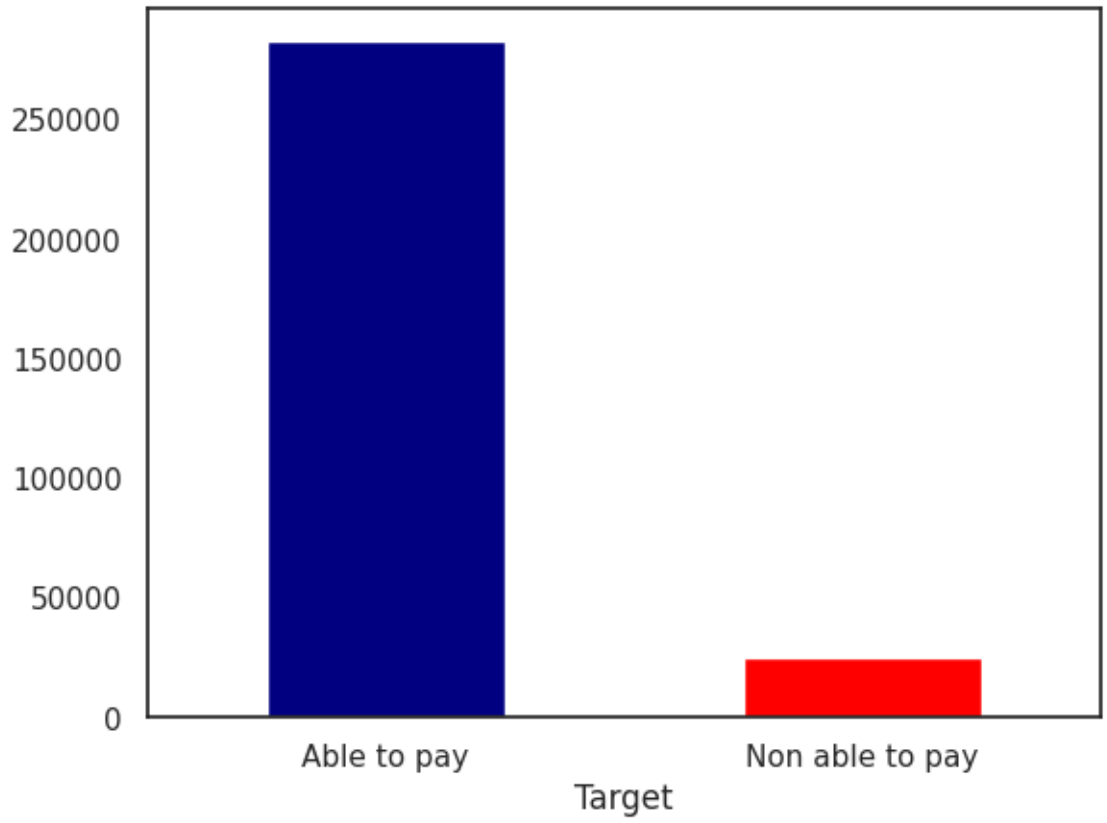
Ahora que se ha trabajado en la anomalía encontrada, se procede a analizar los casos de default presentes en la base de datos sintética en la variable TARGET. Estos datos ya están proporcionados y la base trabajada representa un historial de los tipos de préstamos solicitados y si hubo o no default, lo que permite realizar el modelo con mayor robustez.

Para una mejor comprensión, se define el default en la base de datos como:

$$Clasificación\ de\ los\ individuos = \begin{cases} Default - Non\ ablet\ to\ pay, & x = 1 \\ Non\ Default - Able\ to\ pay, & x = 0 \end{cases} \quad (50)$$

De los cuales 282.686 corresponden a “Non Default – Able to pay” y 24.825 a “Default – Non able to pay” lo que gráficamente, se observa como

Imagen 6 - Comparación clientes



Fuente: Elaboración propia

4.1.2.1 Nuevas variables

En la mayoría de los modelos predictivos, es crucial trabajar con la base de datos antes de aplicar cualquier modelo para eliminar impurezas en las variables o agregar otras que se consideren relevantes. A continuación, se describen las variables adicionales agregadas y el proceso para su creación (ver Anexo 6).

En este caso, se proceden a agregar las variables (Anexo 6):

i. risk_rate que representa una tasa de interés en función de los ingresos de cada individuo.

La creación de esta variable se construye a partir de los deciles de los salarios de la base de datos, de los cuales se tiene que:

Tabla 6 - Deciles salarios

DECÍL	MONTO
1	81000
2	99000
3	112500
4	135000
5	147150
6	162000
7	180000
8	225000
9	270000

Fuente: Elaboración Propia

A mayor salario, se asigna una mayor tasa de interés, distribuida en tramos aleatorios con un rango de 0,02 en cada tramo. Este método permite una mayor reproducibilidad y variabilidad en las tasas de interés asignadas a los clientes.

ii. IR-US, que es la tasa de política monetaria de la FED.

Esta será una tasa fija, que se dejará para evaluar su impacto en el modelo predictivo (de ser el caso). El valor a mayo 2024 corresponde a 3,8%

iii. Trabajo_aval, variable que muestra un rating anexo a cada cliente en proporción a si posee trabajo y/o aval bancario.

En este caso, es sumamente importante a juicio experto agregar el impacto de poseer un aval bancario, puesto que juega un rol crucial en cada evaluación de crédito.

Para poder crear esta variable se tiene como criterio

- NAME_INCOME_TYPE: si el individuo recibe sus ingresos mediante un trabajo estable.
- Bank_guarantee: Si el individuo tiene o no un aval bancario

Tabla 7 - Trabajo aval

NAME_INCOME_TYPE="Working"	Bank_guarantee	Rating para Trabajo_aval
1	1	1
1	0	2
0	1	3
0	0	4

Fuente: Elaboración propia

iv. Credit_limit, representa el monto máximo que potencialmente el cliente podría solicitar, previo a una revisión. Este monto esta concatenado con el sueldo y patrimonio declarado a la entidad bancaria.

Para poder construir esta variable utilizaremos otras 3;

-FLAG_OWN_CAR= Hace referencia si se posee ("Y") o no ("N) uno o más automóviles a nombre del individuo.

-FLAG_OWN_REALTY= Hace referencia si se posee ("Y") o no ("N) uno o más bienes inmuebles a nombre del individuo.

-AMT_INCOME_TOTAL= Sueldo anual declarado por el individuo al banco.

Donde las combinaciones nos darán:

Tabla 8 - Limite de crédito

FLAG_OWN_CAR	FLAG_OWN_REALTY	Credit_limit
N	N	AMT_INCOME_TOTAL*0,4
Y	N	AMT_INCOME_TOTAL*0,5
N	Y	AMT_INCOME_TOTAL*0,55
Y	Y	AMT_INCOME_TOTAL*0,65

Fuente: Elaboración propia

4.1.2.1.1 Verificación de missing values

Después de haber trabajado con la anomalía en la variable DAYS_EMPLOYED y haber agregado nuevas variables, es crucial verificar que no existan datos vacíos ni variables incompletas en el dataset (ver Anexo 7). El trabajo de verificación se divide en dos partes: una para las variables numéricas y otra para las categóricas.

Variables numéricas: Los valores faltantes se rellenan con la media de los datos de su variable respectiva.

Variables categóricas: Los valores faltantes se rellenan con el valor más frecuente (la moda).

Este proceso garantiza que ninguna variable tenga datos faltantes ni vacíos.

4.1.2.1.2 One-hot encoding

Esta técnica de codificación es fundamental en modelos predictivos, particularmente cuando se trabajan con variables categóricas. Después de haber detallado la selección y procesamiento de variables, incluyendo la identificación de las que se agregaron y eliminaron, así como la verificación y tratamiento de valores faltantes mediante un imputador, es esencial comprender el papel del One Hot Encoding (Anexo 8).

Esta técnica convierte las variables categóricas en un formato numérico adecuado para el análisis predictivo. Por ejemplo, teniendo en cuenta la variable categórica 'NAME_INCOME_TYPE' con categorías como Working, Pension, el One Hot Encoding crea nuevas columnas binarias para cada categoría ('NAME_INCOME_TYPE_Working', 'NAME_INCOME_TYPE_Pension') con los valores 1/True si aplica dicho color o 0/False en caso contrario. Esto asegura que cada categoría tenga su propia representación sin ambigüedades y permite a los modelos capturar mejor las relaciones entre las variables. Además, al evitar sesgos y mantener la independencia entre las categorías, se mejora la precisión de las predicciones.

4.1.2.2 Matriz de correlación

Habiendo realizado el trabajo de variables procedemos a estudiar la correlaciòn. Este paso es esencial en el estudio, ya que proporciona una comprensión detallada de cómo cada variable se relaciona con la variable objetivo en esta etapa del análisis (en esta primera parte, 'TARGET'), lo que arroja luz sobre su relevancia para el problema de predicción de crédito.

La correlación es una medida estadística que indica la fuerza y la dirección de la relación entre dos variables. En este contexto, al ordenar las correlaciones con respecto a la variable objetivo, se puede observar qué variables están más fuertemente relacionadas positivamente y cuáles negativamente. Las correlaciones positivas indican que a medida que el valor de la variable aumenta, es más probable que el cliente incumpla el crédito (TARGET=1). Por otro lado, las correlaciones negativas sugieren que a medida que el valor de la variable aumenta, es menos probable el incumplimiento del crédito (TARGET=0).

Las variables altamente correlacionadas positiva o negativamente con la variable objetivo son aquellas que probablemente tengan mayor importancia en el modelo predictivo. Esto permite priorizar qué características son más informativas y pueden ser útiles para predecir el incumplimiento del crédito. A través de estas correlaciones, se pueden capturar relaciones complejas entre las variables predictoras y la variable objetivo. Por ejemplo, una correlación positiva alta de una variable podría indicar que cierto comportamiento crediticio está asociado con un mayor riesgo de incumplimiento, como uno alto negativo el caso contratrio.

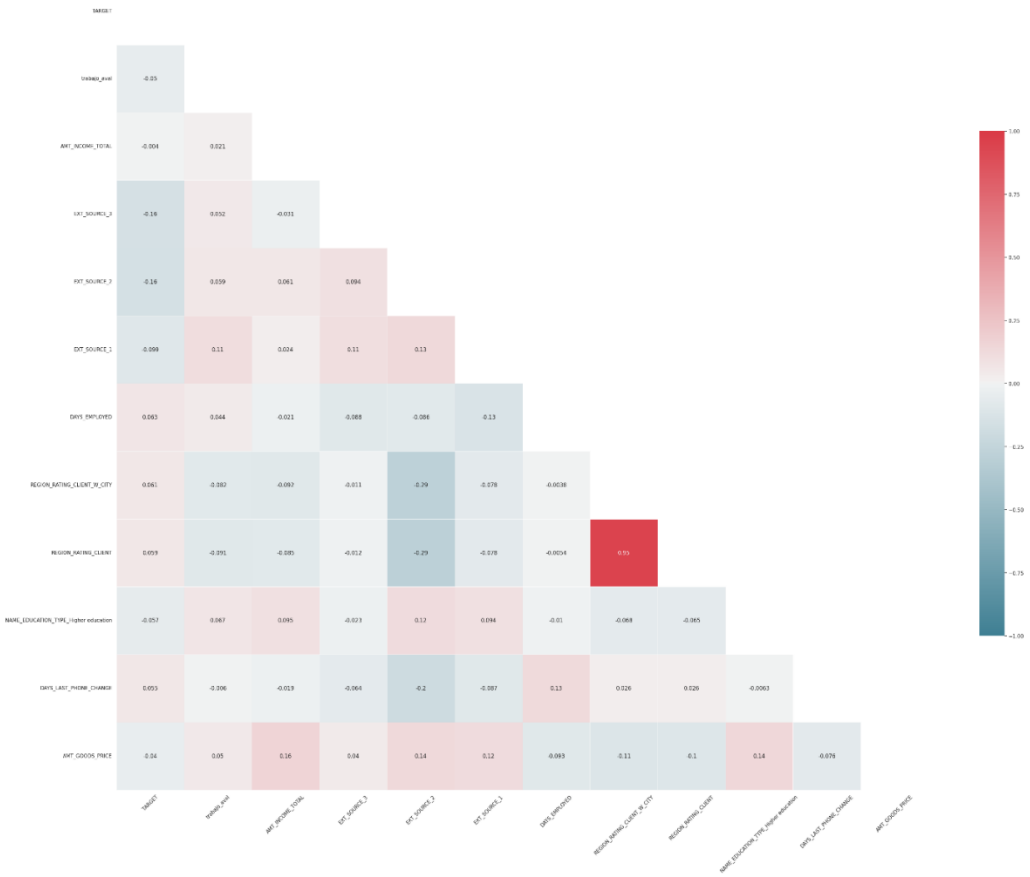
Utilizando estas correlaciones, se pueden seleccionar las variables más relevantes para el modelo predictivo, lo que ayuda a construir un modelo más preciso y eficiente. Además, este análisis puede sugerir la necesidad de incluir nuevas variables o de eliminar aquellas que no aportan información significativa.

Habiendo aplicado el codigo (Anexo 9), se observa que las variables con los valores de correlacion más alto respecto a la variable TARGET corresponden a:

- 'trabajo_aval': rating interno en torno a si posee un aval y trabajo estable.
- 'AMT_INCOME_TOTAL': Sueldo anual declarado por el cliente.
- 'EXT_SOURCE_1': Fuente extra formal de ingresos por arriendos de inmuebles.
- 'EXT_SOURCE_2': Fuente extra formal de ingresos por negocios no vinculados con ocupación que genera el sueldo anual declarado por el cliente.
- 'EXT_SOURCE_3': Fuentes extra de dinero declaradas por el cliente por medio de cartolas bancarias pasadas.
- 'DAYS_EMPLOYED': Dias que el individuo lleva empleado al momento de realizar la solicitud crediticia.
- 'REGION_RATING_CLIENT_W_CITY': Variable generada mediante el one-hot-encoding la cual dice si el cliente que se le ha generado un rating vive o no en un centro urbano.
- 'REGION_RATING_CLIENT': Rating del cliente en proporción a la región geografica donde vive.
- 'NAME_EDUCATION_TYPE_Higher education': Variable creada mediante one-hot-encoding que menciona si el individuo posee estudios superiores.
- 'DAYS_LAST_PHONE_CHANGE': Cantidad de dias transcurridos donde el cliente realizó su cambio de número de movil al momento de solicitar el credito.
- 'AMT_GOODS_PRICE': Patrimonio total declarado por el cliente.

4.1.2.2.1 Gráfico correlación

Imagen 7 - Grafica de correlación



Fuente: Elaboración propia

4.1.2.2.2 Histogramas Variables

Imagen 8 - Histograma variables

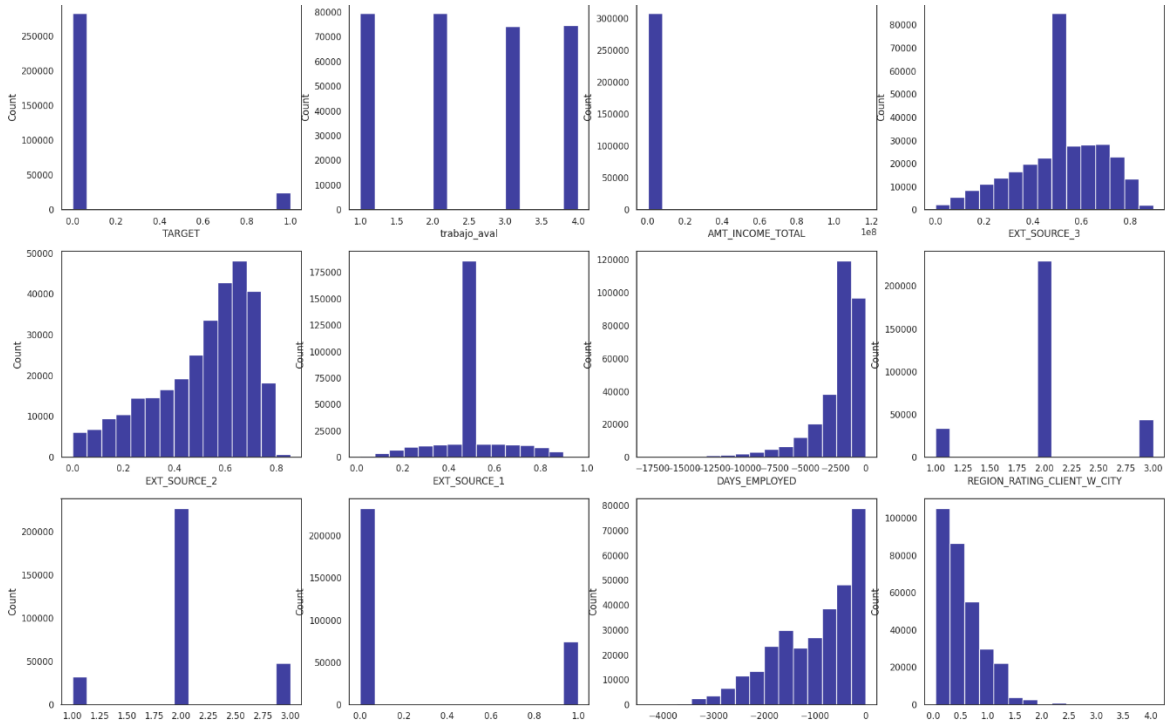


Ilustración 1 - Fuente: Elaboración propia

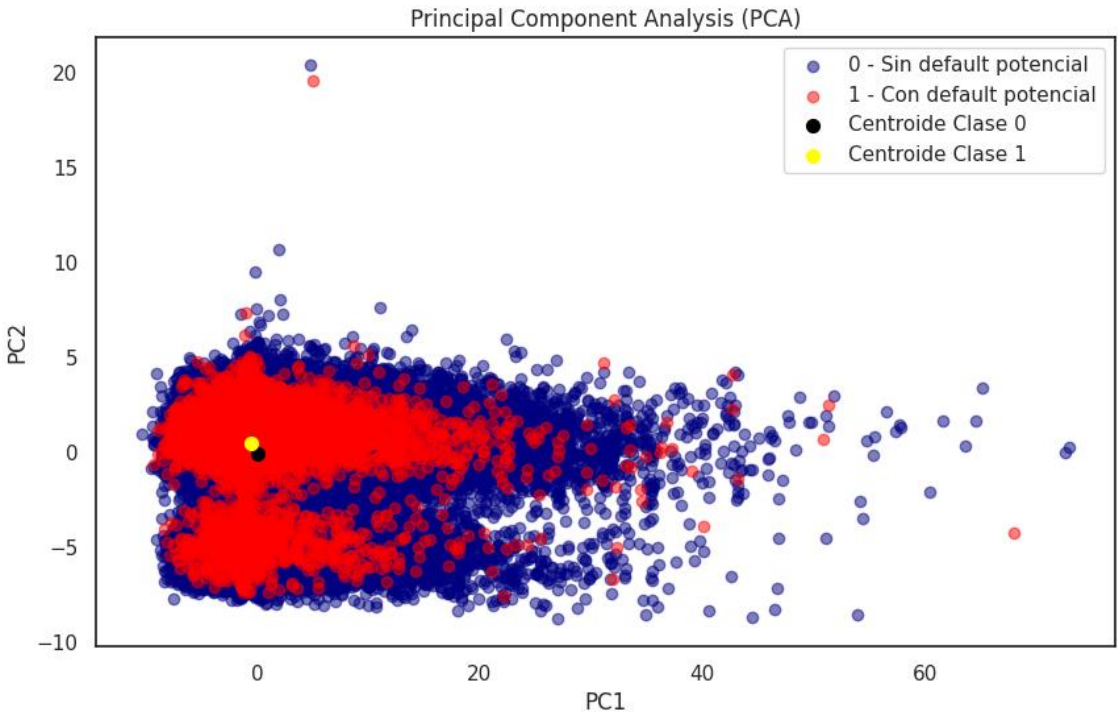
4.1.3 Reducción de la dimensionalidad

4.1.3.1 PCA

En el análisis de datos, como en el de detección de default bancario, el uso de técnicas como el Análisis de Componentes Principales (PCA) como se menciona en los capítulos anteriores es fundamental para comprender la complejidad de los datos y encontrar patrones significativos. En el estudio, se comienza el análisis (posterior aplicación del one-hot encoding) con un conjunto de datos que contenía una gran cantidad de características (244 en total) y como objetivo se tiene reducir esta dimensionalidad para poder visualizar los datos de manera más efectiva (Anexo 10).

Al aplicar PCA, se proyectan los datos en un espacio bidimensional, lo que permitió crear un gráfico que muestra la distribución de los clientes bancarios en función de dos componentes principales para realizar un análisis en torno al comportamiento de las dos clases de cliente. En este gráfico, pude observar claramente dos grupos distintos: los clientes sin default potencial (clase 0) y los clientes con default potencial (clase 1).

Imagen 9 - Analisis de los componentes principales



Fuente: Elaboración propia

Lo más interesante de este análisis es que se puede ver una clara separación entre las dos clases. Los clientes con default potencial tienden a agruparse en una región diferente de los clientes sin default potencial, aunque con superposiciones en ciertas regiones del gráfico. Esta separación es crucial, ya que indica que existen diferencias significativas en las características entre los dos grupos de clientes.

Al calcular los centroides de las dos clases, obtuvimos información adicional sobre la distribución de los datos. Los centroides representan el punto medio de cada clase en el espacio de dos dimensiones definido por los componentes principales. Donde:

Centroide de la clase 0 (Sin default potencial), con coordenadas:

$$PC1 = 0.046556 \quad PC2 = -0.045322$$

Este centroide al estar ubicado cerca del origen del sistema de coordenadas PCA indica que los clientes sin default potencial tienen valores moderados en ambas componentes principales. Sugiere que, en promedio, los clientes de esta clase tienen características financieras típicas, sin extremos significativos en ninguna dirección.

Centroide de la clase 1 (Con default potencial), con coordenadas:

$$PC1 = -0.533721 \qquad PC2 = 0.519575$$

Este centroide está alejado del origen y se encuentra en la parte inferior izquierda del gráfico lo que indica que los clientes con default potencial tienden a tener valores más extremos en las componentes principales, también sugiere que los clientes de esta clase muestran características más polarizadas, con valores significativamente bajos en PC1 y valores altos en PC2.

Al interpretar estos resultados, podemos afirmar que la separación observada en el gráfico PCA se debe en gran medida a las diferencias entre las características financieras de los dos grupos de clientes.

Sin embargo, también de forma gráfica se observa cierta superposición entre las dos clases, lo que indica que algunos clientes sin default potencial comparten características con los clientes con default potencial. Estos casos son particularmente interesantes ya que representan desafíos para la clasificación precisa.

4.1.4 Técnicas de trabajo

A continuación, se presentan las técnicas utilizadas con la intención de mejorar la capacidad predictiva de los modelos propuestos para el estudio.

4.1.4.1 Polynomial Features

La aplicación de Polynomial Features en el contexto del sistema de alerta temprana bancaria implica la expansión de las características originales del conjunto de datos mediante la generación de combinaciones polinómicas de las variables existentes. Esta técnica, como se menciona en el capítulo de Model Performance es fundamental para capturar relaciones no lineales entre las variables y el objetivo de detectar default.

Al aplicar Polynomial Features con un grado de 2 (Anexo 11), se generaron todas las combinaciones posibles de pares de características originales, junto con las características cuadráticas de cada variable individual. Este proceso aumentó significativamente el número de características, pasando de 13 variables originales a 91 características derivadas.

El propósito de esta expansión fue enriquecer la información disponible para el modelo de aprendizaje automático. Cada una de estas nuevas características representa una interacción polinómica entre las variables originales, lo que permite al modelo capturar relaciones más complejas y sutiles entre los datos.

El objetivo es mejorar la capacidad predictiva del modelo y su capacidad para identificar tempranamente el riesgo de incumplimiento. Sin embargo, es importante reconocer que este enriquecimiento de características también puede aumentar el riesgo de sobreajuste. Por lo tanto, se llevará a cabo una validación cuidadosa del modelo para garantizar su capacidad de generalización efectiva a nuevos datos y así fortalecer el sistema de alerta temprana bancaria.

4.1.4.1.2 Nuevos datos, Imputer y matriz de correlación

Al haber evaluado las nuevas variables creadas se presenta el nuevo grafico de correlación (Anexo 12). Además, se aplica la técnica para evitar datos vacíos tanto para las variables categóricas como para las numéricas del nuevo conjunto de datos generado por Polynomial Features, que recapitulando generará combinaciones polinómicas entre las variables. (Anexo 13).

4.1.4.2 Training and Testing Set

Teniendo ya la base de datos trabajada para poder realizar predicciones, la división del conjunto de datos en conjuntos de entrenamiento y prueba es un paso crucial en el desarrollo de los modelos predictivos, y su realización es fundamental para garantizar la efectividad y la evaluación adecuada de los resultados.

Al dividir el conjunto de datos en conjuntos de entrenamiento y prueba, se sigue una lógica clara: el conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba se reserva para evaluar el rendimiento del modelo en datos no vistos.

La importancia de este paso radica en varias razones. Al utilizar un conjunto de datos de prueba independiente, se puede evaluar de manera más objetiva el rendimiento del modelo. Esto proporciona una estimación más precisa de cómo el modelo generaliza a nuevos datos. También se tiene que al entrenar un modelo, existe el riesgo de que este se ajuste demasiado a los datos de

entrenamiento y no pueda generalizar bien a nuevos datos. Al reservar un conjunto de datos de prueba, se puede verificar si el modelo está sobre ajustando al comparar su rendimiento en los conjuntos de entrenamiento y prueba.

Al evaluar el rendimiento del modelo en datos de prueba, se pueden tomar decisiones informadas sobre su capacidad para predecir datos nuevos y no vistos. Esto es crucial en un contexto bancario, donde la precisión en la predicción de defaults es esencial para la toma de decisiones financieras.

La función train_test_split divide aleatoriamente el conjunto de datos en conjuntos de entrenamiento y prueba, lo que permite controlar el tamaño de cada conjunto. Al especificar un valor para test_size, se determina la proporción del conjunto de datos que se utilizará como conjunto de prueba, que en este estudio será 80% training y 20% testing (Anexo 14).

Tabla 9 - Training & Testing set

CATEGORÍA	CLIENTES	VARIABLES
Training	246008	91
Testing	61503	91

Fuente: Elaboración propia

El uso de una semilla (random_state) garantiza reproducibilidad en la división de los datos. Esto significa que, si se utiliza la misma semilla en diferentes ejecuciones del código, se obtendrá la misma división de datos, lo que facilita la comparación de resultados y la reproducibilidad del análisis.

Finalmente se verifica nuevamente que no exista ningún dato vacio en los conjuntos de datos generados (Anexo 15).

4.1.4.3 SMOTE

Finalmente, se aplica SMOTE para el nuevo conjunto de datos y asi asegurar una muestra equilibrada (Anexo 16). Por lo que finalmente para iniciar los modelos predictivos tenemos una muestra de training:

Tabla 10 - Variables con SMOTE

CATEGORÍA	CLIENTES	VARIABLES
Training	452396	91

Fuente: Elaboración propia

4.1.4.4 Naturaleza del problema

La naturaleza del problema a predecir es del tipo clasificación, ya explicado en el capítulo de Modelamiento Matemático-Estadístico.

4.2 Resultados

En esta sección se presentarán los resultados obtenidos por los modelos desarrollados para predecir la probabilidad de default en el sistema de alerta temprana bancaria. Se analizarán varios aspectos de rendimiento de cada modelo, incluida la matriz de confusión, la curva ROC (Receiver Operating Characteristic) y las métricas de evaluación como el accuracy, f1-score y cross-validation (Anexo 17).

La matriz de confusión es una herramienta fundamental en la evaluación del rendimiento de un modelo de clasificación. Esta matriz muestra la relación entre las predicciones del modelo y las clases reales de los datos. En este caso, donde 0 representa que no hubo default y 1 indica la presencia de default, la matriz de confusión se divide en cuatro cuadrantes: true positives (TP), true negatives (TN), false positives (FP) y false negatives (FN). Estos cuadrantes permiten evaluar cómo el modelo está clasificando correctamente o incorrectamente las instancias.

Por otro lado, la curva ROC es una representación gráfica que ilustra la capacidad de un modelo de clasificación para distinguir entre clases. Esta curva muestra la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) para varios umbrales de clasificación. El área bajo la curva ROC (AUC) proporciona una medida de la capacidad de discriminación del modelo: cuanto mayor sea el AUC, mejor será el rendimiento del modelo en la clasificación (Tilmann, G, 2013).

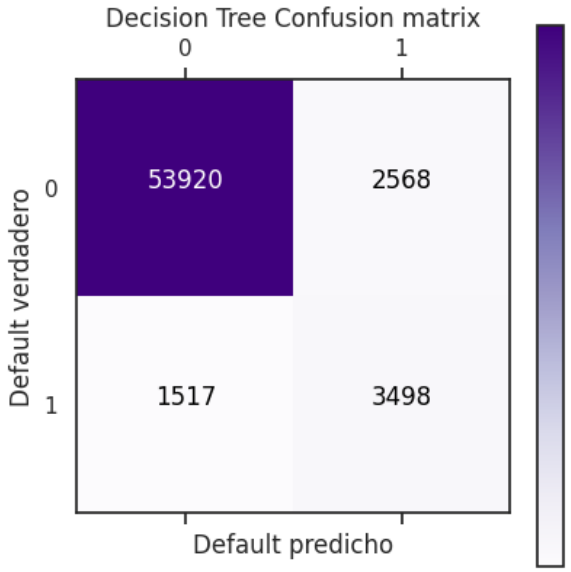
Por tanto, los resultados presentados en esta sección proporcionarán una visión integral del rendimiento de cada modelo en términos de su capacidad para predecir el default bancario. Esto permitirá una evaluación comparativa entre los diferentes modelos y facilitará la selección del modelo más adecuado para el sistema de alerta temprana.

4.2.1 Decision Tree

4.2.1.1 Matriz de Confusión y ROC

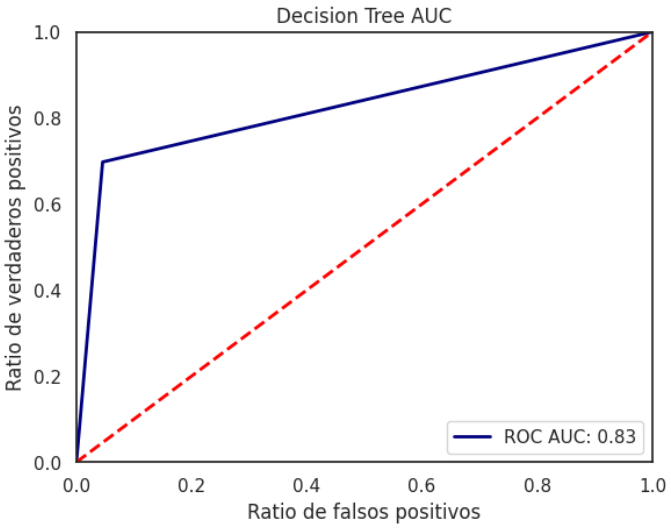
El modelo Decision Tree muestra una buena capacidad de predicción, con un ROC de 0.83 y una precisión del 93.30%. La matriz de confusión revela una tasa razonable de falsos positivos y negativos, con 2568 y 1517 respectivamente, reflejando un equilibrio aceptable entre ambos tipos de errores.

Imagen 11 - Matriz confusión DT



Fuente: Elaboración propia

Imagen 10 - AUC DT



Fuente: Elaboración propia

4.2.1.2 Performance

Tabla 11 - Performance DT

MÉTRICA	RESULTADO
CROSS-VALIDATION	94,7%
ACCURACY	93,3%
F1-SCORE	63,13%

Fuente: Elaboración propia

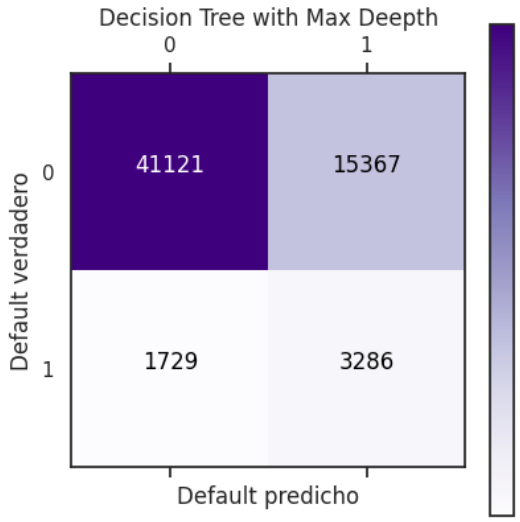
Decision Tree muestra un excelente rendimiento en términos de validación cruzada con un 94.70%, lo que indica una alta estabilidad y consistencia. Su precisión del 93.30% confirma su capacidad para clasificar correctamente la mayoría de las instancias, mientras que un F1-score de 63.13% sugiere que, aunque efectivo, aún hay margen de mejora en el balance entre precisión y exhaustividad

4.2.2 Decision Tree con máxima profundidad

4.2.2.1 Matriz de Confusión y ROC

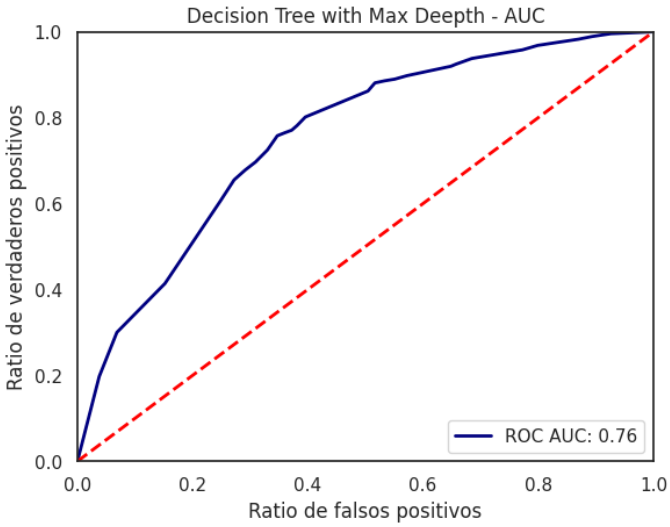
Este modelo presenta una disminución en el rendimiento comparado con el Decision Tree básico, con un ROC de 0.76 y una precisión del 72.20%. La matriz de confusión muestra un aumento significativo en los falsos positivos (15367), indicando que el modelo tiende a sobreajustar los datos.

Imagen 13 - Matriz confusión DT con profundidad máxima



Fuente: Elaboración propia

Imagen 12 - AUC DT MD



Fuente: Elaboración propia

4.2.2.2 Performance

Tabla 12 - Performance DT MD

MÉTRICA	RESULTADO
CROSS-VALIDATION	78,64%
ACCURACY	72,2%
F1-SCORE	27,76%

Fuente: Elaboración propia

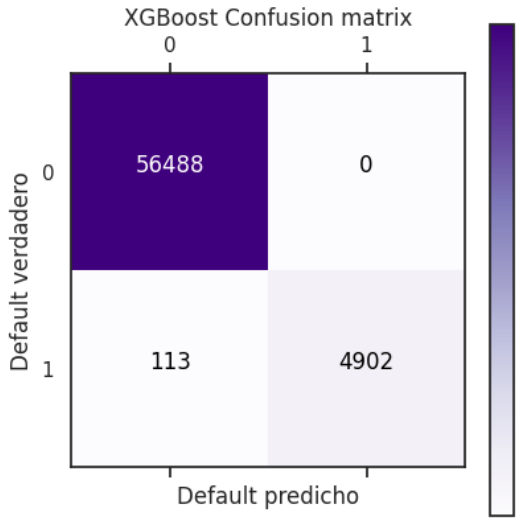
Este modelo exhibe una validación cruzada del 78.64%, señalando una menor estabilidad en comparación con el modelo básico. La precisión del 72.20% refleja una capacidad predictiva más baja, y un F1-score de 27.76% indica que el modelo tiene dificultades significativas para equilibrar correctamente las predicciones positivas y negativas.

4.2.3 XGBoost

4.2.3.1 Matriz de Confusión y ROC

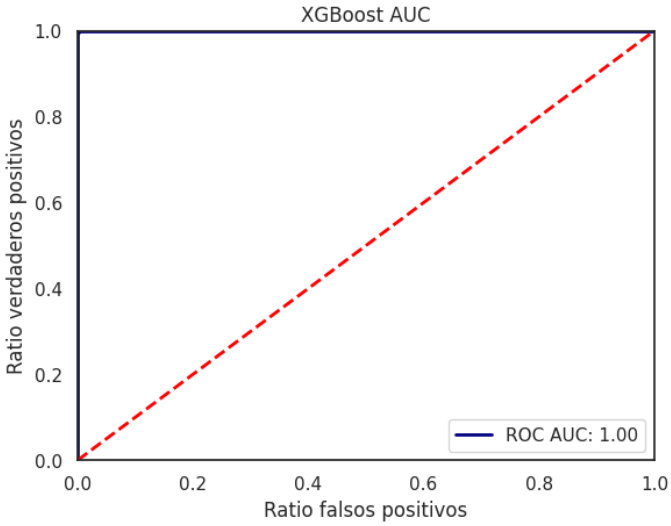
XGBoost ofrece un rendimiento excelente, con un ROC perfecto de 1 y una precisión del 99.81%. La matriz de confusión indica un número muy bajo de errores, con solo 113 falsos negativos, demostrando su alta capacidad discriminativa y predictiva.

Imagen 15 - Matriz confusión XGBoost



Fuente: Elaboración propia

Imagen 14 - AUC XGBoost



Fuente: Elaboración propia

4.2.3.2 Performance

Tabla 13 - Performance XGBoost

MÉTRICA	RESULTADO
CROSS-VALIDATION	98,87%
ACCURACY	99,81%
F1-SCORE	98,86%

Fuente: Elaboración propia

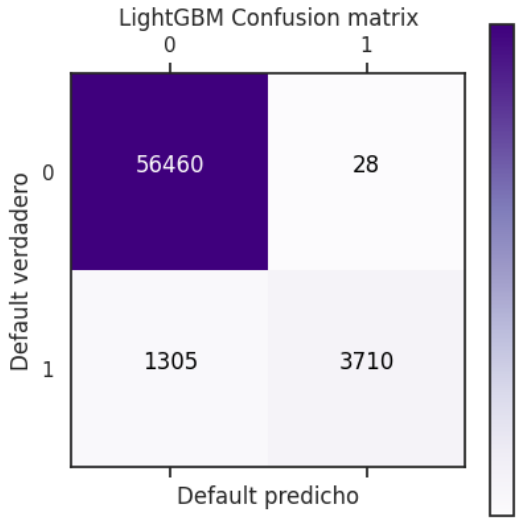
El rendimiento de XGBoost es sobresaliente con una validación cruzada del 98.87%, lo que demuestra su alta fiabilidad. La precisión del 99.81% muestra su habilidad casi perfecta para clasificar las instancias, y un F1-score de 98.86% resalta su excepcional equilibrio entre precisión y exhaustividad.

4.2.4 LightGBM

4.2.4.1 Matriz de Confusión y ROC

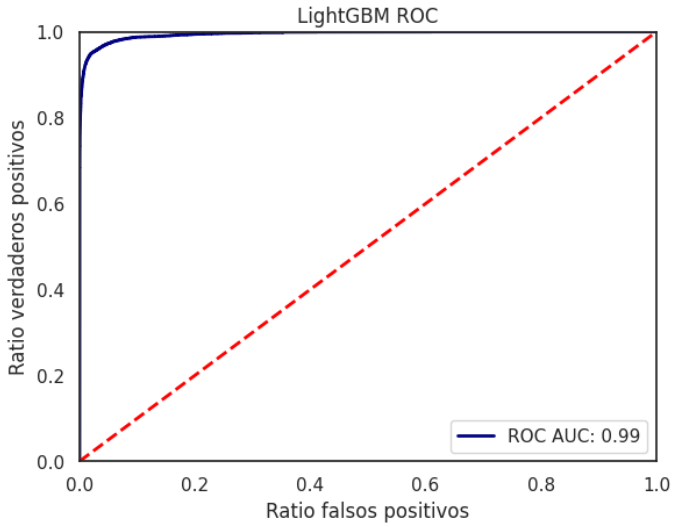
El modelo LightGBM también muestra un rendimiento destacado, con un ROC de 0.99 y una precisión del 97.83%. La matriz de confusión evidencia 28 falsos positivos y 1305 falsos negativos, lo que indica una excelente habilidad para clasificar correctamente los casos de no-default y default.

Imagen 17 - Matriz confusión LightGBM



Fuente: Elaboración propia

Imagen 16 - AUC LightGBM



Fuente: Elaboración propia

4.2.4.2 Performance

Tabla 14 - Performance LightGBM

MÉTRICA	RESULTADO
CROSS-VALIDATION	96,79%
ACCURACY	97,83%
F1-SCORE	84,77%

Fuente: Elaboración propia

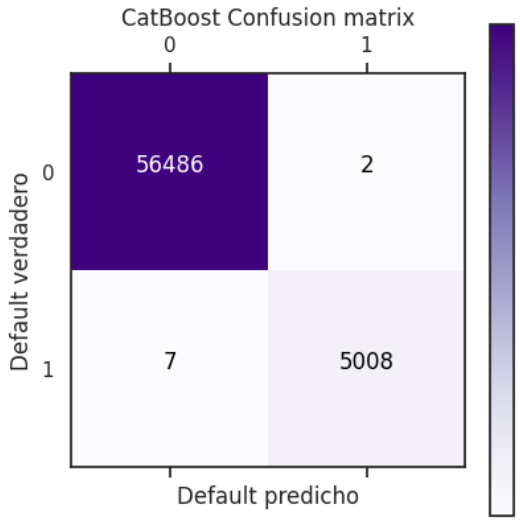
LightGBM presenta una validación cruzada del 96.79%, reflejando su alta estabilidad. Con una precisión del 97.83%, el modelo clasifica correctamente la mayoría de las instancias, y un F1-score de 84.77% indica un buen balance en sus predicciones.

4.2.5 CatBoost

4.2.5.1 Matriz de Confusión y ROC

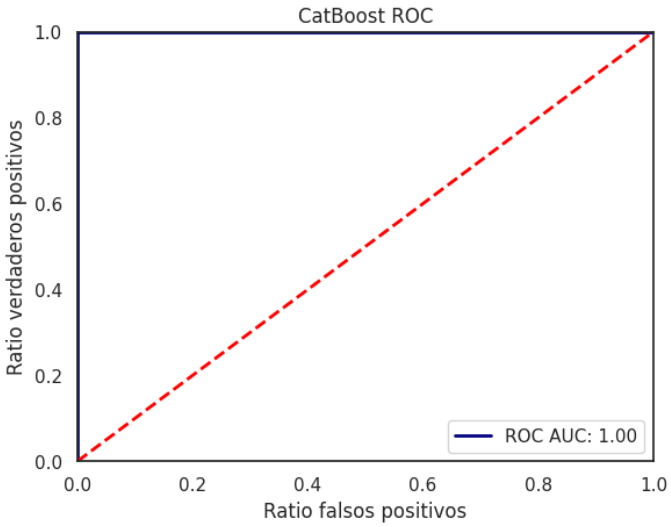
CatBoost sobresale con un ROC perfecto de 1 y una precisión casi perfecta de 99.99%. La matriz de confusión revela mínimos errores, con solo 2 falsos positivos y 7 falsos negativos, destacándose como el modelo más preciso y robusto.

Imagen 19 - Matriz confusión CatBoost



Fuente: Elaboración propia

Imagen 18 - ROC CatBoost



Fuente: Elaboración propia

4.2.5.2 Performance

Tabla 15 - Performance CatBoost

MÉTRICA	RESULTADO
CROSS-VALIDATION	99,42%
ACCURACY	99,985%
F1-SCORE	99,91%

Fuente: Elaboración propia

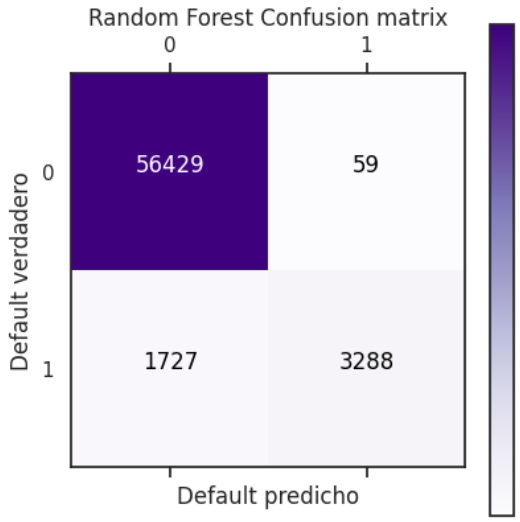
CatBoost se destaca con una validación cruzada del 99.42%, subrayando su notable estabilidad y consistencia. Su precisión casi perfecta del 99.99% y un F1-score de 99.91% indican un rendimiento casi impecable, con un equilibrio excelente entre precisión y exhaustividad.

4.2.6 Random Forest

4.2.6.1 Matriz de Confusión y ROC

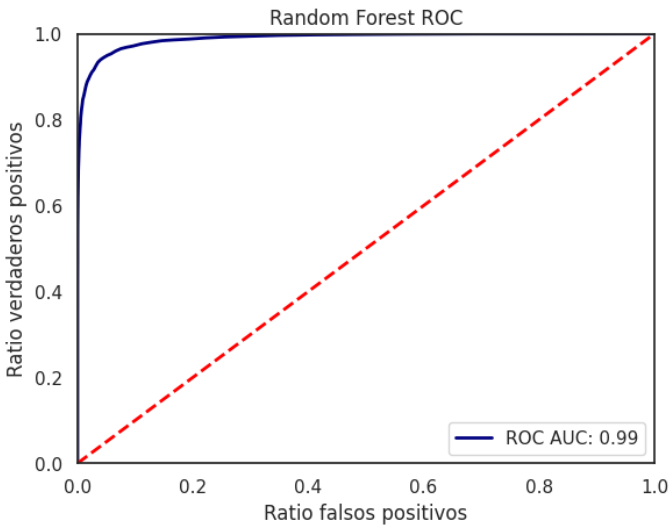
Este modelo ofrece un rendimiento sólido con un ROC de 0.99 y una precisión del 97.10%. La matriz de confusión muestra 59 falsos positivos y 1727 falsos negativos, indicando una alta capacidad predictiva y un buen balance entre los errores.

Imagen 21 -Matriz confusión Random Forest



Fuente: Elaboración propia

Imagen 20 - AUC Random Forest



Fuente: Elaboración propia

4.2.6.2 Performance

Imagen 22 - Performance Random Forest

MÉTRICA	RESULTADO
CROSS-VALIDATION	97,572%
ACCURACY	97,096%
F1-SCORE	78,641%

Fuente: Elaboración propia

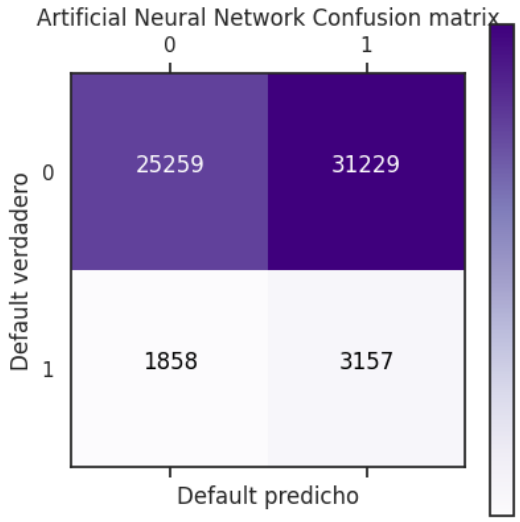
El modelo Random Forest ofrece una validación cruzada del 97.57%, indicando alta fiabilidad. Con una precisión del 97.10%, clasifica correctamente la mayoría de las instancias, y un F1-score de 78.64% muestra un buen equilibrio en las predicciones, aunque con margen de mejora.

4.2.7 Artificial Neural Network

4.2.7.1 Matriz de Confusión y ROC

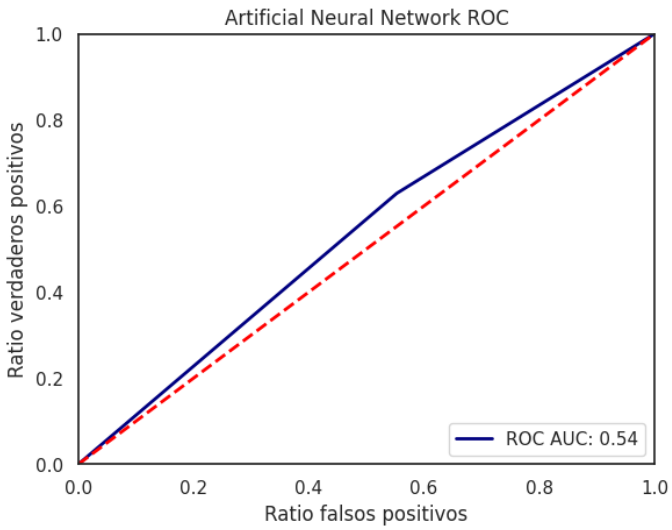
El modelo de red neuronal artificial muestra un rendimiento bajo con un ROC de 0.54 y una precisión del 46.20%. La matriz de confusión revela altos valores de falsos positivos (31229) y falsos negativos (1885), indicando una capacidad predictiva insuficiente.

Imagen 24- Matriz confusión ANN



Fuente: Elaboración propia

Imagen 23 - AUC ANN



Fuente: Elaboración propia

4.2.7.2 Performance

Tabla 16 - Performance ANN

MÉTRICA	RESULTADO
CROSS-VALIDATION	53,99%
ACCURACY	46,202%
F1-SCORE	16,02%

Fuente: Elaboración propia

La red neuronal artificial muestra una validación cruzada baja del 53.99%, sugiriendo poca estabilidad. Su precisión del 46.20% indica un rendimiento insuficiente, y un F1-score de 16.02% resalta su falta de equilibrio en las predicciones, lo que lo hace menos adecuado para este problema.

4.2.8 Naive Bayes

4.2.8.1 Matriz de Confusión y ROC

El modelo Naive Bayes tiene un rendimiento pobre con un ROC de 0.54 y una precisión del 91.58%. La matriz de confusión muestra un alto número de falsos negativos (5011) y 167 falsos positivos, sugiriendo que el modelo no captura adecuadamente la relación entre las variables.

Imagen 26 - Matriz confusión Naive Bayes

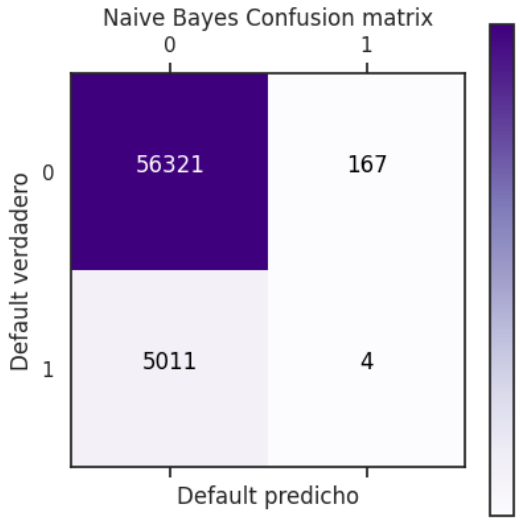


Ilustración 3 - Fuente: Elaboración propia

Imagen 25 – AUC Naive Bayes

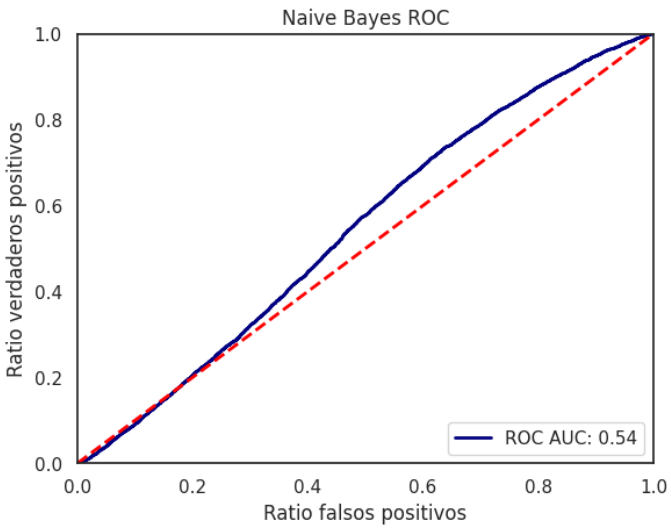


Ilustración 2 - Fuente: Elaboración propia

4.2.8.2 Performance

Tabla 17 - Performance Naive Bayes

MÉTRICA	RESULTADO
CROSS-VALIDATION	49,906%
ACCURACY	91,58%
F1-SCORE	0,1524%

Fuente: Elaboración propia

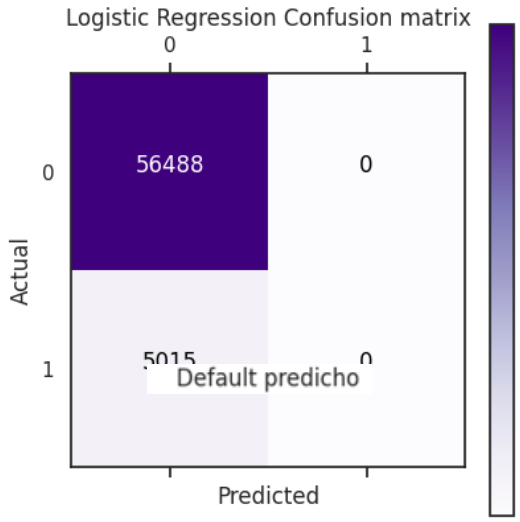
El modelo tiene una validación cruzada baja del 49.91%, reflejando inestabilidad. Aunque su precisión es del 91.58%, el F1-score de 0.15% revela una deficiencia significativa en equilibrar las predicciones positivas y negativas, indicando que el modelo no maneja bien la relación entre las variables.

4.2.9 Regresión Logistica

4.2.9.1 Matriz de Confusión y ROC

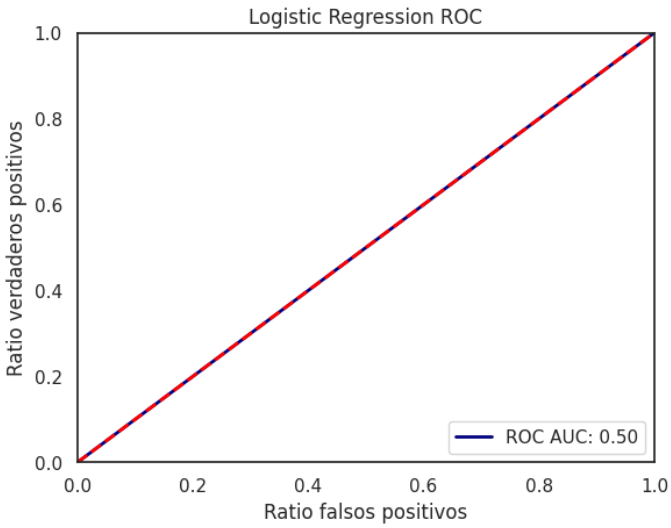
El modelo de regresión logística tiene un rendimiento deficiente con un ROC de 0.5 y una precisión del 91.84%. La matriz de confusión revela que no detecta correctamente los casos de default (5015 falsos negativos), indicando una falta de capacidad discriminativa.

Imagen 28 - Matriz confusión Regresión Logística



Fuente: Elaboración propia

Imagen 27 - AUC Regresión logística



Fuente: Elaboración propia

4.2.9.2 Performance

Tabla 18 - Performance Regresión Logística

MÉTRICA	RESULTADO
CROSS-VALIDATION	50,06%
ACCURACY	91,8445%
F1-SCORE	0

Fuente: Elaboración propia

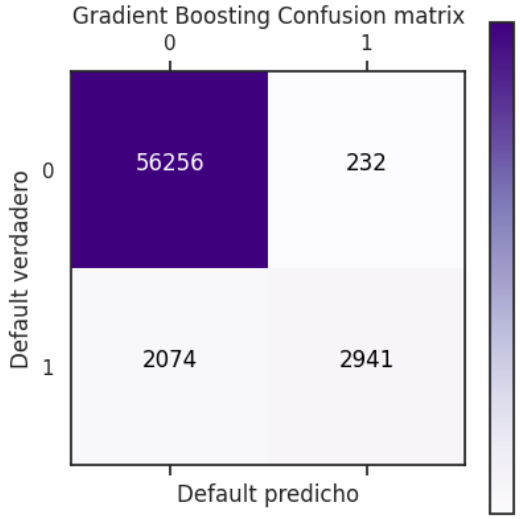
La regresión logística exhibe una validación cruzada del 50.06%, sugiriendo una baja estabilidad. Su precisión del 91.84% es engañosamente alta, ya que el F1-score de 0% indica que no logra equilibrar las predicciones positivas, lo que lo hace ineficaz para detectar casos de default.

4.2.10 GradientBoosting

4.2.10.1Matriz de Confusión y ROC

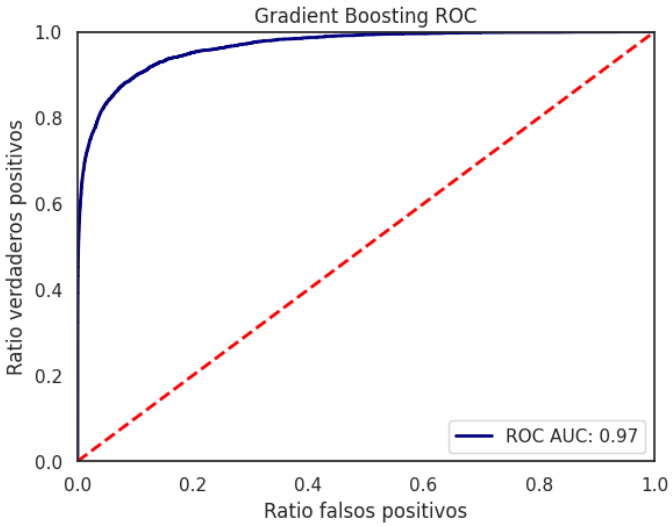
Gradient Boosting muestra un excelente rendimiento con un ROC de 0.97 y una precisión del 96.23%. La matriz de confusión indica 232 falsos positivos y 2074 falsos negativos, reflejando una alta capacidad predictiva y un buen equilibrio entre los errores.

Imagen 30 - Matriz confusión Gradient Boosting



Fuente: Elaboración propia

Imagen 29 - AUC Gradient Boosting



Fuente: Elaboración propia

4.2.10.2 Performance

Tabla 19 - Performance Gradient Boosting

MÉTRICA	RESULTADO
CROSS-VALIDATION	95,92%
ACCURACY	96,23%
F1-SCORE	71,83%

Fuente: Elaboración propia

Este modelo presenta una validación cruzada del 95.92%, señalando alta estabilidad. Su precisión del 96.23% indica una clasificación correcta en la mayoría de los casos, y un F1-score de 71.83% muestra un buen equilibrio entre precisión y exhaustividad, aunque no tan elevado como algunos otros modelos de boosting.

4.2.11 Selección de modelos

A continuación, se presenta un resumen de la evaluación de los resultados de cada modelo.

Tabla 20 - Resumen performance

MODELO	Cross Validation	Test Acurracy	F1 Score
DECISION TREE	94,760797 %	93,358048 %	63,135096 %
DECISION TREE m/p	78,640406 %	72,202982 %	27,767450 %
XGBOOST	98,871352 %	99,816269 %	98,860543 %
LIGHTGBM	96,791774 %	97,832626 %	84,770936 %
CATBOOST	99,420424 %	99,985367 %	99,910224 %
RANDOM FOREST	97,572279 %	97,096077 %	78,641473 %
ARTIFICIAL NEURAL NETWORK	53,884211 %	46,202624 %	16,024974 %
NAIVE BAYES	49,906498 %	91,580898 %	0,154261 %
ADABOOST	96,190298 %	96,975757 %	81,689309 %
REGRESION LOGISTICA	50,060788 %	91,845926 %	0,000000 %
GRADIENTBOOSTING	95,926602 %	96,250589 %	71,836834 %

Fuente: Elaboración propia

Los resultados obtenidos de los modelos muestran una amplia variabilidad en su desempeño en la predicción del default bancario. Entre ellos, destacan positivamente XGBoost, CatBoost y LightGBM, con altos valores de cross-validation, test accuracy y F1-score. Esto sugiere que estos modelos tienen una capacidad significativa para predecir con precisión tanto las instancias positivas como las negativas de default.

Por otro lado, modelos como Naive Bayes, Artificial Neural Network, Logistic Regression y Decision Tree MP presentan un desempeño notablemente inferior en comparación con los anteriores. Naive Bayes y Logistic Regression, en particular, muestran una baja capacidad de predicción, evidenciada por su bajo test accuracy y F1-score. Estos modelos parecen no ser adecuados para capturar la complejidad de los datos y las relaciones entre las variables.

Los modelos Decision Tree y Random Forest muestran un rendimiento decente pero inferior en comparación con XGBoost, CatBoost y LightGBM. Esto sugiere que estos modelos basados en árboles sin técnicas adicionales tienen una capacidad limitada para manejar relaciones más complejas entre las variables.

En términos generales, los modelos basados en boosting, como XGBoost, CatBoost y LightGBM, son los más efectivos para predecir el default bancario, seguidos de cerca por Random Forest. Estos modelos son capaces de aprender patrones más complejos y sutiles en los datos, lo que se refleja en su alto test accuracy y F1-score.

Sin embargo, es importante destacar que, aunque CatBoost y XGBoost tienen un rendimiento ligeramente mejor en comparación con LightGBM, estos últimos todavía muestran una capacidad bastante significativa en la predicción del default. Por lo tanto, aunque ambos predicen bastante bien la elección final entre estos modelos dependerá de consideraciones adicionales que nos enfrentaremos en la construcción del modelo predictivo de alerta temprana como la velocidad de entrenamiento, la interpretabilidad y los recursos computacionales disponibles.

4.2.11.1 Mean Decrease in Gini Importance

La inclusión del Mean Decrease in Gini Importance (MDGI) como un punto adicional de interés en la evaluación de los modelos XGBoost y CatBoost es una estrategia valiosa en la elección del modelo para el sistema de alerta temprana bancaria. El MDGI es una métrica que indica cuánto disminuye la impureza de Gini de un árbol debido a las divisiones realizadas en una variable particular.

Al calcular cómo la pureza de los nodos del árbol disminuye con cada división basada en una variable particular, el MDGI permite identificar qué variables contribuyen más a la separación entre las clases de default y no-default. Esto es crucial para comprender qué características son más influyentes en las decisiones del modelo y cómo afectan las predicciones. Por ejemplo, si una variable tiene un alto MDGI, significa que tiene un gran impacto en la precisión de las predicciones, lo que la convierte en un punto focal para intervenciones o acciones correctivas (Han, H., Guo, X., & Yu, H. August 2016).

Esta métrica es especialmente interesante en el contexto de los modelos de boosting, como XGBoost y CatBoost, ya que estas técnicas de ensamblaje de árboles tienden a hacer un uso eficiente de las características más relevantes. Por lo tanto, el MDGI puede proporcionar una comprensión detallada de qué variables tienen el mayor impacto en la capacidad predictiva de estos modelos.

Al evaluar los modelos (Anexo 18) XGBoost y CatBoost en función del MDGI, se puede identificar qué características contribuyen más a la precisión de las predicciones. Esto ayuda a comprender mejor la lógica interna de los modelos y a enfocar los esfuerzos en la optimización de las variables más influyentes.

Las características más importantes para cada modelo son:

Tabla 21 - Características más relevantes modelo CatBoost

N°	CatBoost
1	Trabajo_aval
2	Trabajo_aval AMT_INCOME_TOTAL
3	AMT_INCOME_TOTAL
4	REGION_RATING_CLIENT_W_CITY REGION_RATING_CLIENT
5	Trabajo_aval_REGION_RATING_CLIENT
6	AMT_GOODS_PRICE trabajo_aval
7	DAYS_LAST_PHONE_CHANGE
8	EXT_SOURCE_2
9	EXT_SOURCE_3 DAYS_LAST_PHONE_CHANGE
10	DAYS_LAST_PHONE_CHANGE AMT_GOODS_PRICE

Fuente: Elaboración propia

Tabla 22 - Características más relevantes modelo XGBoost

N°	XGBoost
1	DAYS_LAST_PHONE_CHANGE
2	EXT_SOURCE_3 DAYS_EMPLOYED
3	NAME_EDUCATION_TYPE_Higher education
4	DAYS_EMPLOYED DAYS_LAST_PHONE_CHANGE
5	AMT_INCOME_TOTAL
6	REGION_RATING_CLIENT
7	EXT_SOURCE_2 DAYS_EMPLOYED
8	REGION_RATING_CLIENT_W_CITY NAME_EDUCATION_TYPE_Higher education
9	REGION_RATING_CLIENT_W_CITY DAYS_LAST_PHONE_CHANGE
10	AMT_INCOME_TOTAL REGION_RATING_CLIENT_W_CITY

Fuente: Elaboración propia

Para XGBoost, las características más importantes indican que la estabilidad laboral y financiera son cruciales para predecir el default bancario, aunque es curioso que el número de dias que han pasado desde el ultimo cambio de telefono tome un lugar relevante. Variables como la antigüedad en el empleo, el tipo de educación, los ingresos anuales y la región donde vive el cliente son factores clave. Esto sugiere que la consistencia en el empleo, la educación superior y los ingresos más altos tienden a estar asociados con un menor riesgo de default.

Por otro lado, para CatBoost, la presencia de aval y trabajo estable es el factor más influyente, seguido de los ingresos anuales y la ubicación geográfica del cliente. Esto sugiere que la combinación de estabilidad laboral, respaldo financiero y ubicación geográfica son los principales predictores del default bancario, destacando la importancia de tener un respaldo financiero sólido y una situación laboral estable para mitigar el riesgo de default.

4.3 Predicción del Sistema de Alerta Temprana

Habiendo evaluado diversos modelos para la detección temprana de default bancario, se ha llegado a la conclusión de que CatBoost y XGBoost serían potencialmente los más efectivos para esta tarea. Ahora, en una etapa más avanzada del estudio, se plantea un nuevo desafío: utilizar estos modelos para abordar un problema de regresión.

La lógica del sistema de alerta temprana (EWS) se centra en la creación de un rating basado en criterio experto, utilizando las variables más relevantes identificadas por estos modelos. En esta fase, se busca tramificar las variables para dar una mayor amplitud de análisis y capacidad predictiva, cuando sea posible, y crear un registro tipo “semáforo” para clasificar a los clientes en función de su potencial riesgo de default.

La transición hacia un problema de regresión implica un cambio en la metodología, donde el objetivo ya no es clasificar los clientes en categorías de default o no-default, sino predecir la categoría de riesgo donde se encuentra un cliente y asi relacionarla con probabilidad de default. Este enfoque permitirá una mayor precisión en la evaluación del riesgo crediticio de cada cliente, ofreciendo así una herramienta más robusta para la toma de decisiones financieras.

4.3.1 Tramificación de las variables electas

La robustez del modelo en el sistema de alerta temprana (EWS) es esencial para una evaluación precisa del riesgo de default bancario. Si bien las variables en si puede proporcionar información crucial para predecir, su naturaleza diversa requiere una mayor granularidad para capturar la variabilidad en los datos.

Para mejorar la robustez del modelo, se ha optado por la tramificación de las variables (Anexo 19). Esto implica dividir las variables en rangos discretos, como deciles o cuartiles, para asignar un valor binario (1 o 0) a cada dato según el rango en el que se encuentre, similar al enfoque de one-hot encoding. Por ejemplo, en 'AMT_INCOME_TOTAL', los clientes con ingresos en el rango del 10% superior pueden etiquetarse como 1, mientras que el resto como 0 y asi consiguientemente en el tramo que el cliente en cuestión se encuentre. Este enfoque permite considerar la diversidad dentro de cada variable y capturar mejor la relación entre las características y el riesgo de default. La tramificación también aumenta la interpretabilidad del modelo al proporcionar una representación más intuitiva de las variables y sus efectos en la probabilidad de default.

En este contexto, se explorará cómo la tramificación de las variables mejora la capacidad predictiva del modelo en el sistema de alerta temprana bancaria. Este enfoque es fundamental para garantizar una evaluación más precisa y sólida del riesgo de default, lo que es crucial para la toma de decisiones financieras informadas.

4.3.1.1 Metodología de la tramificación

4.3.1.1.1 Deciles

Las siguientes variables, dada su naturaleza, se han tramificado por deciles, donde el numero dentro del paréntesis indica la pertenencia a ese decil:

Tabla 23 - Deciles

VARIABLES	NOMBRE DE VARIABLES TRAMIFICADAS
AMT_INCOME_TOTAL	AIT (1) … AIT (10)
AMT_GOODS_PRICE	AGP (1) … AGP (10)
DAYS_EMPLOYED	DEMP (1) … DEMP (10)
DAYS_LAST_PHONE_CHANGE	PCh (1) …PCh (10)
EXT_SOURCE_1	EX1(1) …EX1(10)
EXT_SOURCE_2	EX2(1) …EX2(10)
EXT_SOURCE_3	EX3(1) …EX3(10)

Fuente: Elaboración propia

4.3.1.1.2 Cuartiles

Donde el numero dentro del paréntesis indica la pertenencia a dicho cuartil:

Tabla 24 - Cuartiles

VARIABLES	NOMBRE DE VARIABLES TRAMIFICADAS
Trabajo_aval	Aval(1), aval(2), aval(3), aval(4)

Fuente: Elaboración propia

4.3.1.1.3 Tercil

Donde el numero dentro del paréntesis indica la pertenencia a dicho tercil:

Tabla 25 - Terciles

VARIABLES	NOMBRE DE VARIABLES TRAMIFICADAS
REGION_RATING_CLIENT	RRC(1), RRC(2), RRC(3)
REGION_RATING_CLIENT_W_CITY	RW(1), RW(2), RW(3)

Fuente: Elaboración propia

Cabe señalar que la variable NAME_EDUCATION_TYPE_Higher_education no se ha tramificado puesto que ya estaba clasificada como 1 o 0 luego de haber aplicado el one-hot-encoding.

4.3.2 Creación de la variable de Rating

La metodología para la creación de la variable "EWS" (Early Warning System) se basa en una tramificación de las variables más relevantes identificadas previamente (Anexo 20). Esta tramificación se realiza asignando valores binarios (1 o 0) a cada dato en función de su pertenencia a determinados rangos discretos definidos para cada variable. Por ejemplo, en 'AMT_INCOME_TOTAL', se han creado 10 rangos (AIT1 a AIT10), donde el rango AIT1 corresponde al 10% más bajo de ingresos y el rango AIT10 al 10% más alto.

La lógica detrás de esta tramificación es asignar un peso decreciente o creciente a cada rango, reflejando su contribución al riesgo de default. Para variables como ingresos, patrimonio y antigüedad laboral, se asignan pesos decrecientes, donde los rangos más altos tienen mayor peso. Mientras que para variables como los rating de región o la cantidad de días desde el último cambio de teléfono, se asignan pesos crecientes, donde los rangos más bajos tienen mayor peso.

Esta metodología permite capturar la variabilidad de las variables de manera más precisa, reflejando su impacto en el riesgo de default. La suma ponderada de los valores asignados a cada variable para cada cliente da como resultado el puntaje de "EWS", que representa su riesgo de default. Cuanto mayor sea el puntaje de EWS, mayor será el riesgo de default del cliente, lo que lo convierte en una herramienta valiosa para el sistema de alerta temprana bancaria.

4.3.3 Tramificación de la variable rating

4.3.3.1 Definición de los tramos de riesgo

Habiendo creado las variables, se han obtenido un nuevo data set compuesto por 84 variables. Posterior a esto se procede a definir los rangos de riesgos (Anexo 21).

La asignación de rangos se basa en una división equitativa de los puntajes de "EWS" para categorizar el riesgo asociado. Primero, se identifica el número de individuos que caen en cada rango definido. En este caso, se establecieron cuatro rangos de riesgo, donde los puntajes más bajos indican menor riesgo y los más altos indican mayor riesgo, lo que se traduce en (donde X es el puntaje para el individuo):

Tabla 26 - Tramos de riesgos

CLASIFICACIÓN	RANGO
Bajo riesgo	X < 4.9
Riesgo moderado	4.9 < X < 6.5
Riesgo considerable	6.5 < X < 8
Alto riesgo	8 < X < 11

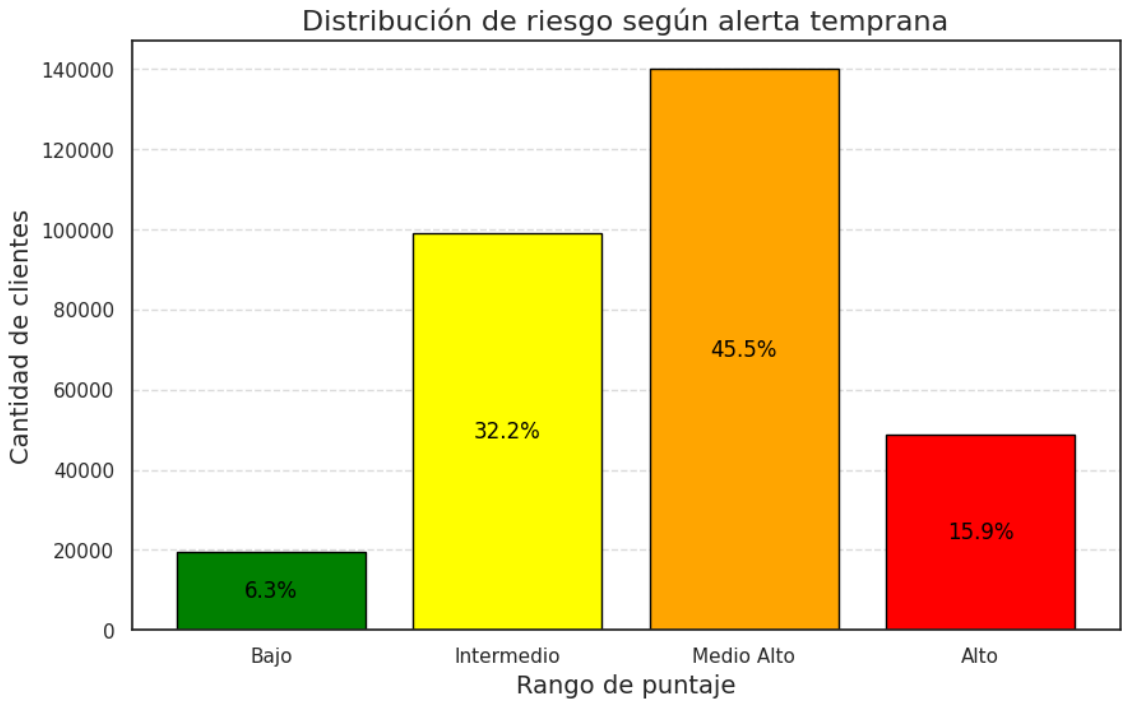
Fuente: Elaboración propia

El primer rango incluye a aquellos con puntajes menores a 4.9, que se consideran de bajo riesgo. El segundo rango va desde 4.9 hasta menos de 6.5, considerado como riesgo moderado. El tercer rango abarca puntajes desde 6.5 hasta menos de 8, indicando un riesgo considerable. Finalmente, el cuarto rango, con puntajes entre 8 y 11, denota un alto riesgo de default.

Esta asignación de rangos permite una clasificación clara y equitativa del riesgo asociado a los puntajes de "EWS", lo que facilita la identificación de los clientes con diferentes niveles de riesgo y orienta las estrategias de mitigación adecuadas dentro del sistema de alerta temprana bancaria.

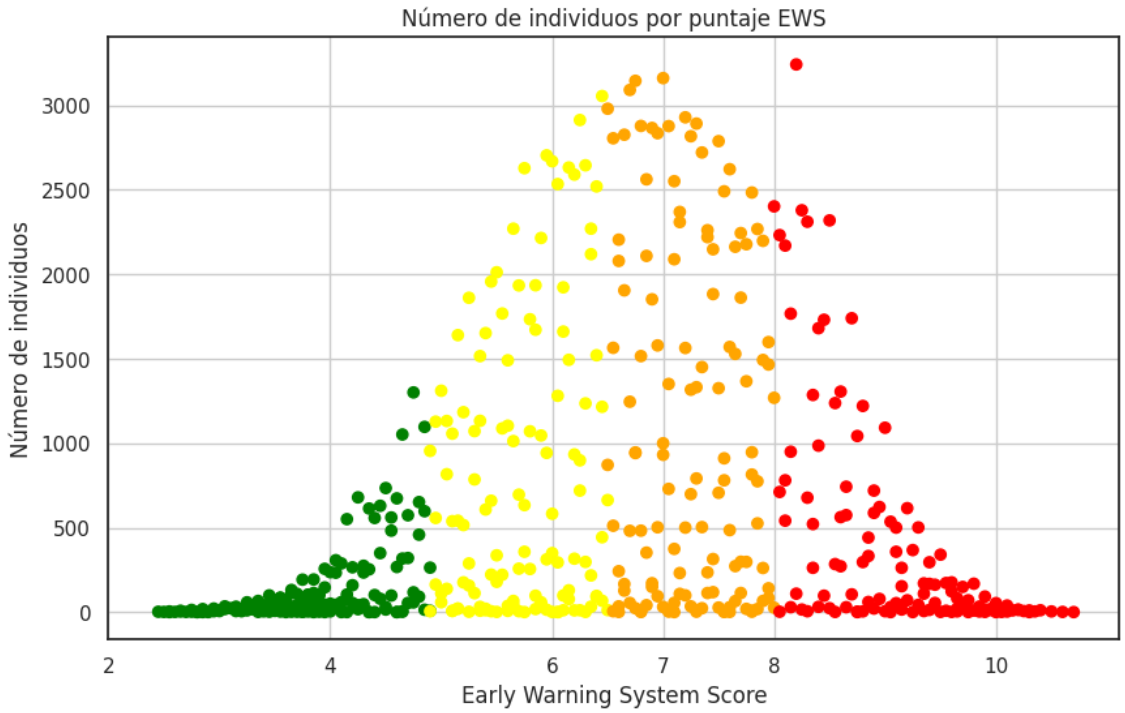
4.3.3.2 Distribución de riesgo

Imagen 31 - Distribución de riesgo



Fuente: Elaboración propia. Anexo 21.1

Imagen 32 - Distribución de individuos

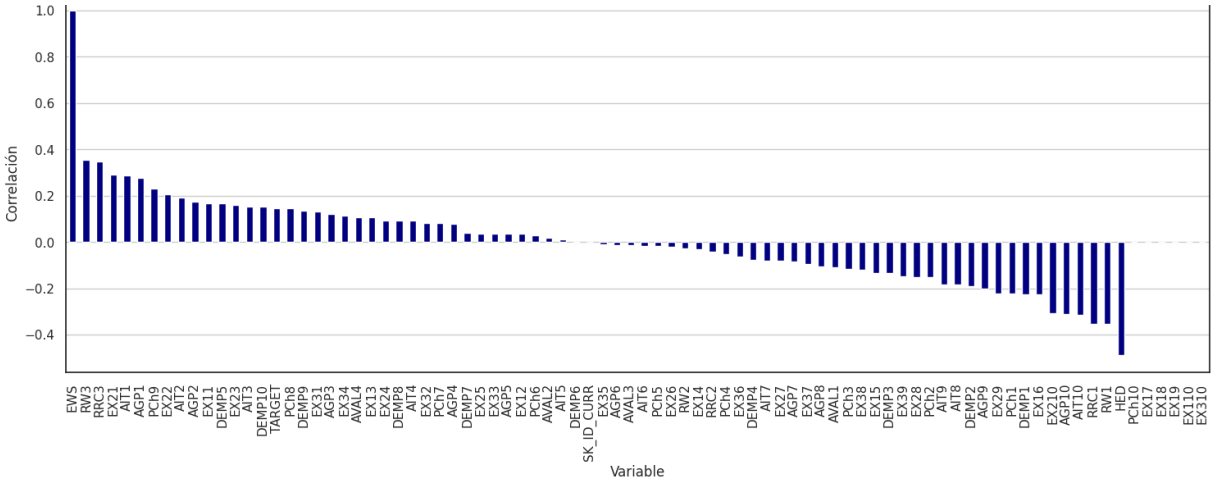


Fuente: Elaboración propia. Anexo 21.2

4.3.4 Correlación de variables

Se realiza un análisis de correlación para identificar las variables más importantes en el proceso de análisis predictivo ocupando como variable a predecir la creada en el punto anterior "EWS". De aquí se usarán las variables que tengan una correlación con la variable sobre 0.1. Esto con la intención de ocupar la máxima cantidad de variables que fueron creadas mediante la tramificación (Anexo 22).

Imagen 33 - Correlación variables EWS



Fuente: Elaboración propia

4.3.5 Técnicas de trabajo

Ya habiendo seleccionado las variables, se procede al trabajo de Imputer, como a esta altura son solo variables numéricas, se llena con la moda de la variable cuando exista un dato vacío. A continuación, se realiza el train_test_split para trabajar con los modelos (Anexo 23). Quedando con un data set final para el trabajo predictivo de:

Tabla 27 - Características variables EWS

CLIENTES	VARIABLES
307511	45

Fuente: Elaboración propia

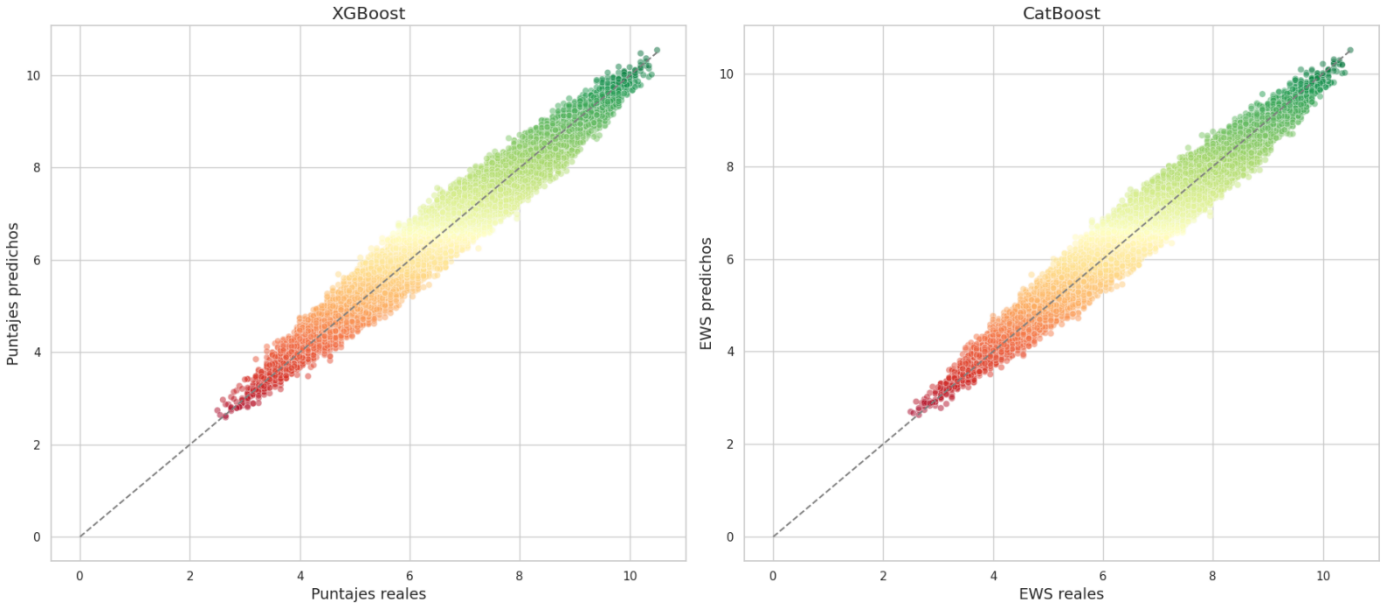
Ahora se procede a trabajar con el data set para los modelos de XGBoost y CatBoost, con la intención de predecir el rating del sistema de alerta temprana.

4.3.6 Resultados

Ahora, se muestran los resultados obtenidos por los modelos XGBoost y CatBoost (Anexo 24). Recapitulando que lo que se está prediciendo es el puntaje del sistema de alerta temprana, donde en el eje X podemos observar los puntajes reales, creados por la asignación de puntajes mediante criterio experto post tramificación de las variables selectas. El eje Y son los puntajes predichos.

4.3.6.1 Gráficos

Imagen 34 - Resultados gráficos EWS



Fuente: Elaboración propia

4.3.6.2 Performance

Dado que gráficamente, los resultados no representan una base solida para la toma de decisión respecto al mejor modelo evaluaremos el modelo con las métricas competentes a un problema de regresión, lo que da:

Tabla 28 - Performance resumen

TÉCNICA	CatBoost	XGBoost
MAE	18,0522 %	18,8535 %
MSE	5,2991 %	5,6938 %
RMSE	23,0198 %	23,8617 %
Varianza	96,2516 %	95,9724 %

Fuente: Elaboración propia

CatBoost muestra un mejor rendimiento al observar los análisis. Primero el MAE, con un valor del 18.05%, en comparación con XGBoost que tiene un MAE ligeramente mayor, que alcanza el 18.85%.

Esta diferencia se refuerza al observar el MSE, donde CatBoost logra un 5.30% y XGBoost un 5.69%. RMSE de CatBoost es de 23.02%, mientras que XGBoost tiene un RMSE del 23.86%. Lo que indica que CatBoost tiene una menor dispersión de errores en sus predicciones.

Además, CatBoost explica un mayor porcentaje de la variabilidad en los datos, con un 96.25% de varianza, en comparación con el 95.97% de XGBoost. Esto sugiere que CatBoost tiene una mejor capacidad para capturar la relación entre las variables y el riesgo de default.

La elección de CatBoost se basa en su rendimiento general superior en todas las métricas evaluadas. Aunque las diferencias entre los modelos son relativamente pequeñas, CatBoost proporciona predicciones más precisas y robustas del riesgo de default, lo que lo convierte en la opción preferida para el sistema de alerta temprana bancaria.

4.3.6.3 Selección de modelos e importancia de variables

Por ende, se decanta por el modelo CatBoost, método innovador en la industria financiera por su alta adaptabilidad en la detección de fraudes. A continuación, se presentan algunos hallazgos interesantes para Shap values como para la importancia relativa (Anexo 54).

4.3.6.3.1 Shap Values

Los resultados para las 10 primeras variables mediante Shap Values son:

Tabla 29 - Shap Valuea

VARIABLE	SHAP VALUE
HED	35,4%
AVAL4	13,3%
AVAL1	12,4%
PCh9	11,4%
RW3	10,9%
AGP1	10,1%
EX11	9,8%
DEMP5	9,7%
AGP10	8,2%
EX21	7,9%

Fuente: Elaboración propia

4.3.6.3.2 Importancia relativa de las variables

Ya realizado el análisis, se muestran las 10 primeras variables con mayor importancia relativa:

Tabla 30 - Importancia de variables

VARIABLE	IMPORTANCIA RELATIVA
DEMP5	3,32 %
EX13	2,73 %
AVAL4	2,36 %
AVAL1	2,33 %
HED	1,95 %
EX39	1,86 %
EX15	1,86 %
EX38	1,72 %
EX31	1,57 %
DEMP10	1,53 %
PCh2	1,49 %

Fuente: Elaboración propia

4.3.6.3.3 Algunos hallazgos interesantes

La variable DAYS_EMPLOYED muestra una importancia significativa en el quinto decil (DEMP5), lo que sugiere que los clientes con una cantidad moderada de días empleados tienen un impacto más fuerte en la probabilidad de default. Esto podría indicar cierta inestabilidad financiera entre los clientes con historiales laborales moderados, donde ni la estabilidad prolongada ni la falta de historial laboral parecieran ser factores determinantes para el riesgo de default.

En el caso de EXT_SOURCE_1, EXT_SOURCE_3 y EXT_SOURCE_2, los rangos EX13 y EX15 emergen como importantes, lo que indica que ciertos valores dentro de estas fuentes de ingresos externos tienen un efecto más pronunciado en el riesgo de default. Específicamente, en EXT_SOURCE_1 podría haber una relación significativa entre los ingresos provenientes de arriendos de inmuebles y el riesgo de default.

Para la variable Trabajo_aval, Aval1 (el no tener un aval y tampoco empleo) y Aval4 (tener aval y empleo) muestran una influencia significativa, sugiriendo que inclusive los clientes con un buen tipo de aval tienen riesgo de default. Esto podría indicar que, aunque los avales pueden proporcionar cierta garantía, en ciertos casos pueden no ser suficientes para mitigar el riesgo de default.

La variable binaria HED (educación superior) también muestra una influencia notable, lo que implica que, a pesar de tener una educación superior, algunos clientes aún enfrentan riesgos financieros considerables. Esto podría deberse a otros factores como niveles de ingresos o historiales crediticios, lo que indica que una educación superior por sí sola no es un indicador definitivo de estabilidad financiera.

Otro aspecto interesante por considerar es el rating de la región con centro urbano (decil 3), variable que según el análisis de shap value tiene un impacto significativo en el modelo de predicción de riesgo, lo que indica que la ubicación geográfica también juega un papel preponderante en el análisis. También la variable de Patrimonio (AGP1), que representa a los clientes con menor patrimonio representa una contribución del 10% a la capacidad predictiva, lo que da a entender que es una variable importante.

En conclusión, este análisis más detallado resalta la importancia de considerar no solo las variables en su totalidad, sino también los rangos específicos dentro de estas variables tramificadas. Proporciona una comprensión más profunda de cómo ciertos segmentos dentro de las variables influyen en la predicción del riesgo de default, lo que es esencial para la toma de decisiones financieras informadas en el sistema de alerta temprana bancaria, aunque siempre será interesante el análisis particular de cada cliente.

4.4 Matriz EWS

Con el modelo ya seleccionado, el haber analizado los resultados de este como las variables y obtenido algunas interesantes conclusiones se procede a comparar la relación entre el default y el rating de los clientes, evaluando así la robustez del sistema de alerta temprana bancario.

Para comparar esto, recordemos que para realizar esto se procedió en apartados anteriores a clasificar clasificando a los clientes en diferentes niveles de riesgo según su rating. El cálculo que se propone es realizado como una forma de evaluar la capacidad predictiva de un modelo de riesgo de default, y comparando estas clasificaciones con los resultados reales de default (Anexo 26).

Primero, se divide a los clientes en distintos grupos según su rating. Este rating es obtenido mediante un modelo de predicción, y se ha establecido un criterio de clasificación que asigna a cada cliente a uno de cuatro grupos: bajo riesgo, riesgo moderado, riesgo considerable y alto riesgo.

Posteriormente, se contabiliza el número de individuos en cada grupo y se determina cuántos de ellos efectivamente no hicieron default (TARGET = 0), calculando así la tasa de no-default (0) y default (1) para cada categoría de riesgo.

Por ejemplo, para el grupo de bajo riesgo (rating menor a 4.9), se cuenta cuántos individuos están en esa categoría (i) y cuántos de ellos no hicieron default. La tasa de no-default (Bajo_0) se obtiene dividiendo los individuos que realizaron default entre la totalidad de la categoría, y la tasa de default se calcula como la diferencia entre 1 y tasa de no-default, es decir:

- Número de individuos de la categoría j (i_j):

$$i_j = \sum_{i=1}^{n} l_{\ rating\ i\ en\ el\ rango\ j} \qquad (51)$$

Donde n es el número total de individuos, j el grupo de riesgo y l es la función indicadora.

- Número de individuos de la categoría j que no hicieron default:

$$i_{j,0} = \sum_{i=1}^{n} l_{\ (rating\ i\ en\ el\ rango\ j)\cap(TARGETx=0)} \qquad (52)$$

- Tasa de no-default para la categoría $j(Tasa\ {}^{j}_{0})$:

$$Tasa\ {}^{j}_{0} = \frac{i_{j,0}}{i_j} \qquad (53)$$

- Tasa de default para la categoría $j(Tasa\ {}^{j}_{1})$:

$$Tasa\ {}^{j}_{1} = 1 - Tasa\ {}^{j}_{0} \qquad (54)$$

Este procedimiento se repite para cada grupo de riesgo, ajustando los límites correspondientes a cada categoría.

Finalmente, se realiza el mismo cálculo, pero esta vez utilizando el rating real de los clientes (real_EWS), en lugar del rating predicho por el modelo. Esto permite comparar cómo se comportan las tasas de no-default y default en relación con el rating real de los clientes.

Este método proporciona una medida de la capacidad predictiva del modelo de riesgo de default, evaluando cómo se alinean las clasificaciones del modelo con los resultados reales. Si las tasas de no-default y default son consistentes entre el rating predicho y el rating real, esto indicaría que el modelo es capaz de predecir con precisión el riesgo de default de los clientes. Por el contrario, discrepancias significativas entre estas tasas pueden señalar deficiencias en el modelo predictivo.

4.4.1 Default & Non-default rates

Habiendo ya casi finalizado el estudio, se presentan los resultados sustentados por el modelo predictivo CatBoost.

4.4.1.1 Hipótesis

Retomando la hipótesis, ¿El sistema de alerta temprana es congruente con el rating predicho y el default bancario?, revisemos los resultados.

4.4.1.1.1 Default rates:

Tabla 31 - Default rates

CATEGORIA RIESGO	BAJO	MODERADO	CONSIDERABLE	ALTO
Default Rate real EWS	2,36%	4,43%	9,17%	14,8%
Default Rate predicho EWS	2,33%	4,35%	9,24%	15,28%

Fuente: Elaboración propia

4.4.1.1.2 Non default rates:

Tabla 32 - Non default rates

CATEGORIA RIESGO	BAJO	MODERADO	CONSIDERABLE	ALTO
Non-Default Rate real EWS	97,6%	95,577%	90,83%	85,2%
Non-Default Rate predicho EWS	97,7%	95,65%	90,76%	84,72%

Fuente: Elaboración propia

Al observar las tasas de default, se aprecia que el modelo logra capturar con precisión el riesgo de default en cada categoría. Por ejemplo, para la categoría de riesgo "Bajo", la tasa de default predicha es del 2.3%, mientras que la tasa real de default es del 2.36%, lo que indica una correspondencia casi exacta entre la predicción del modelo y la realidad. De manera similar, para el resto de las categorías de riesgo como para las tasas de no-default.

Sin embargo, la más interesante revelación se encuentra al comparar las tasas de no-default. Aunque estas tasas son altas debido a la definición inversa de default (default = 1 - no default), es interesante observar cómo se comportan en relación con las predicciones del modelo.

Un sistema de alerta temprana con un 100% de efectividad debería adelantarse al default con asertividad total, sin embargo, existen múltiples factores que pueden escaparse de cualquier modelo de inteligencia artificial que pueden disminuir la capacidad predictiva de estos.

La mejora constante del sistema de alerta temprana es crucial para las instituciones financieras que empleen esta técnica de mitigación de riesgos, ya que les permite tomar decisiones informadas y proactivas en la gestión del riesgo crediticio, minimizando las pérdidas y optimizando la rentabilidad de sus carteras de clientes.

4.4.1.1.3 Error tipo I

Lo que se espera de un cliente que ha sido clasificado con un riesgo potencial de default, aunque ninguna institución financiera en términos reales lo desea, es que caiga en el impago. Sin embargo, los resultados del modelo, menciona que para el riesgo considerable existen 9 de cada 10 individuos clasificados como potenciales default cuando en realidad son falsos positivos. Lo mismo para aquellos individuos con riesgo alto, que corresponden a 8 de cada 10. Pero ¿es esto un problema?

Tabla 33 - Error tipo I

CATEGORIA	% ERROR TIPO I
Riesgo considerable	90,755%
Riesgo alto	84,716%

Fuente: Elaboración propia

De acuerdo con la consultora PwC en conjunto a la tecnológica británica GALYTIX 8 de cada 10 indicadores de un sistema de alerta temprana son falsos positivos, y aunque existen múltiples esfuerzos para mitigar esto, muchas veces los bancos no concentran la mayor parte de los recursos de mejora del EWS en este aspecto, puesto que en términos reales un falso positivo indica que el cliente que se esperaba que no pagaría lo termina realizando, aunque el fin de cada préstamo es que termine siendo pagado (PwC & GALYTIX, 2024).

4.4.1.1.4 Error tipo II

Pese a que los falsos positivos tienen un costo asociado, (el cual no se explora en este TFM) el real "dolor de cabeza" de un banco es clasificar como clientes no riesgosos a aquellos que terminan haciendo default.

Tabla 34 - Error tipo II

CATEGORIA	% ERROR TIPO II
Riesgo bajo	2,298%
Riesgo moderado	4,349%

Fuente: Elaboración propia

Estos porcentajes representan una pérdida financiera no prevista para el banco, lo que podría afectar su rentabilidad y estabilidad financiera. Esto se debe a que el banco destina recursos y capital a la adquisición y gestión de clientes, proporcionándoles créditos y servicios con la expectativa de obtener un retorno positivo. Sin embargo, cuando estos clientes incumplen sus obligaciones financieras, se generan pérdidas que disminuyen la rentabilidad esperada. Lo resalta la necesidad de mejorar la precisión del modelo de riesgo de default, especialmente en la identificación de clientes con riesgo bajo y moderado. Lo anterior podría lograrse mediante la inclusión de variables adicionales en el modelo, una mejor calibración de los umbrales de clasificación o el uso de técnicas de modelado más avanzadas.

5. Conclusiones y desafíos

5.1 Conclusiones

Este TFM ha demostrado la relevancia y eficacia de los sistemas de alerta temprana (EWS) basados en modelos de inteligencia artificial para la predicción del riesgo de default en clientes bancarios. A lo largo del estudio, se ha comprobado que los modelos de boosting, particularmente CatBoost, presentan un rendimiento superior en la predicción de defaults, superando a otras técnicas de la misma familia como XGBoost y LightGBM. La capacidad de CatBoost para capturar patrones complejos en los datos se refleja en métricas de error reducidas y un alto porcentaje de varianza explicada.

El análisis de variables ha revelado insights importantes: ciertos segmentos dentro de variables muestran una fuerte correlación con el riesgo de default, destacando la necesidad de considerar no solo las variables en su totalidad, sino también sus rangos específicos. Estas observaciones permiten una comprensión más profunda de los factores que influyen en el riesgo de default, proporcionando una base sólida para la toma de decisiones financieras informadas.

La comparación de las tasas de default y no-default entre las predicciones del modelo y los resultados reales muestra que el modelo tiene una alta capacidad predictiva, aunque aún presenta desafíos, como la existencia de falsos positivos y falsos negativos. En particular, los errores de tipo II, que representan clientes clasificados erróneamente como de bajo riesgo pero que terminan en default, subrayan la importancia de mejorar la precisión del modelo. Estos errores tienen un impacto financiero significativo, ya que generan pérdidas no previstas que afectan la rentabilidad del banco.

Para mitigar estos errores, es esencial incorporar variables adicionales y mejorar la calibración de los umbrales de clasificación. Además, la aplicación de técnicas de modelado más avanzadas y el uso de inteligencia artificial generativa pueden ofrecer mejoras sustanciales en la precisión y eficiencia de los sistemas de alerta temprana.

Para finalizar, este estudio confirma la viabilidad y la necesidad de sistemas avanzados de alerta temprana en la gestión del riesgo crediticio mediante un ejemplo acotado y aplicado. La implementación de un EWS basado en inteligencia artificial no solo mejora la capacidad predictiva de los modelos, sino que también proporciona una herramienta invaluable para la gestión proactiva del riesgo, contribuyendo a la estabilidad y solidez financiera de las instituciones bancarias.

5.2 Futuros desafíos

Los sistemas de alerta temprana (EWS) desempeñan un papel crucial en la gestión del riesgo en diversas industrias, incluido el sector financiero. Sin embargo, a pesar de su importancia, estos sistemas enfrentan una serie de problemas y desafíos que pueden limitar su efectividad y precisión.

Uno de los principales problemas radica en la precisión de la asignación de rating a cada cliente. En muchos casos, los modelos se vuelven inexactos debido a la incapacidad para actualizar los puntajes rápidamente o no poder abarcar características difíciles de cuantificar. Esto puede deberse a la falta de datos en tiempo real o a la rigidez de los procesos de actualización. Como resultado, los modelos pueden no reflejar con precisión el riesgo real de los clientes, lo que lleva a decisiones subóptimas en la gestión del crédito.

Además, los enfoques basados en modelos de riesgo de mercado pueden reaccionar exageradamente ante fluctuaciones en los precios y el crédito. Esto significa que los modelos pueden generar señales de alerta excesivas en respuesta a fluctuaciones normales del mercado, lo que conduce a decisiones erróneas.

Los modelos regulatorios también enfrentan desafíos, ya que a menudo se basan en reglas estandarizadas para aumentar los requisitos de capital y liquidez. Estas reglas pueden no tener en cuenta las condiciones o naturaleza propia del banco como también el constante cambio en el mercado, lo que puede resultar en una asignación inadecuada de recursos y una gestión ineficiente del riesgo.

Por otro lado, los indicadores de EWS pueden producir una alta tasa de falsos positivos. Esto significa que los sistemas pueden identificar incorrectamente a los clientes como de alto riesgo cuando en realidad no lo son. Este problema puede ser especialmente relevante en el contexto de clientes que solicitan préstamos medianos, como avances en efectivo o líneas de crédito, donde la precisión en la identificación del riesgo es crucial para tomar decisiones de crédito informadas.

A pesar de estos desafíos, la evolución de la inteligencia artificial (IA) y la inteligencia generativa presenta una oportunidad para mejorar los modelos de EWS. La IA puede aprovechar grandes volúmenes de datos históricos y en tiempo real para mejorar la precisión de los modelos de rating y de riesgo de mercado. Los algoritmos de aprendizaje automático pueden identificar patrones complejos y no lineales en los datos, lo que permite una evaluación más precisa del riesgo de default.

Además, la inteligencia generativa puede ser un pilar aliado en la mejora de los modelos de EWS. Los modelos generativos pueden simular escenarios futuros y generar datos sintéticos que ayuden a calibrar y mejorar los modelos existentes. Esto permite a los analistas de riesgos probar diferentes estrategias y escenarios hipotéticos para evaluar su efectividad y prepararse para posibles eventos de riesgo.

Sin embargo, para aprovechar al máximo el potencial de la IA y la inteligencia generativa, es fundamental abordar los desafíos relacionados con la interpretación y explicabilidad de los modelos. Los modelos de IA a menudo son considerados como cajas negras, lo que dificulta comprender cómo llegan a sus conclusiones. Mejorar la transparencia y la interpretación de los modelos es crucial para generar confianza en su uso en la toma de decisiones financieras.

Por ende, si bien los sistemas de alerta temprana enfrentan una serie de problemas y desafíos, la evolución de la inteligencia artificial y generativa ofrece una oportunidad emocionante para mejorar la precisión y la eficacia de estos sistemas. Al aprovechar estas tecnologías de manera efectiva y abordar los desafíos relacionados, las instituciones financieras pueden fortalecer su capacidad para gestionar el riesgo y tomar decisiones informadas en un entorno financiero cada vez más complejo y volátil.

6. Bibliografía

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, Volume 2, Issue 4 p. 433-459 https://doi.org/10.1002/wics.101

Adrian, T., & Natalucci, F. (2020, 22 de mayo). La COVID-19 empeora vulnerabilidades financieras. *Blog IMF*. [Consulta en abril 2024]. Disponible en: https://www.imf.org/es/Blogs/Articles/2020/05/22/blog-gfsr-covid-19-worsens-pre-existing-financial-vulnerabilities

Aguilar Corbacho, G. (2015). La evolución de la economía estadounidense a partir del crash de 1929 y el debate Keynes vs Hayek (1929-1973). Facultad de Derecho, ICADE. . [Trabajo fin de grado, Pontificia Universidad de Comillas]. E-Archivo https://repositorio.comillas.edu/rest/bitstreams/5358/retrieve

Amat Rodrigo, J. (Junio, 2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) *y t-SNE. Ciencia de Datos*. https://cienciadedatos.net/documentos/35_principal_component_analysis

Alberto Fernández, Salvador Garcia, Mikael Galar, Ronaldo C. Prati, Bartosz Krawczyk, & Francisco Herrera. (2018). Learning from Imbalanced Data Sets. *Springer*.

Alvargonzález, V. (18 octubre, 2023). Bonos USA, historia de una profecía auto cumplida. *ElEconomista.es*

Banco Mundial. (2023). Informe sobre el desarrollo mundial 2022. Los impactos económicos de la COVID-19. *Blog Banco Mundial*. [Consulta en abril 2024]. Disponible en: https://www.bancomundial.org/es/publication/wdr2022/brief/chapter-1-introduction-the-economic-impacts-of-the-covid-19-crisis

Banda, H. (2014). Inteligencia Artificial: Principios y Aplicaciones. *ResearchGate.* [Consulta en abril 2024]. Disponible en: https://www.researchgate.net/publication/262487459_Inteligencia_Artificial_Principios_y_Aplicaciones

Bank for International Settlements (BIS). (n.d.). History of the Basel Committee. *Blog BIS*. [Consulta en abril 2024]. Disponible en: https://www.bis.org/bcbs/history.htm

Bank for International Settlements. (2005). An explanatory note on the Basel II IRB risk weight functions. *Press & Communications BIS*.

Bank for international settlements, BCBS. (2015). *Acuerdos de Basilea.*

Bank for international settlements. (2010). Basilea III: marco regulador internacional para los bancos. *Press & Communications BIS.*

BankInter. (2018). ¿Qué es la titulización hipotecaria? *Blog Bankinter*. [Consulta en abril 2024]. Disponible en: https://www.bankinter.com/banca/preguntas-frecuentes/hipotecas/que-es-la-titulizacion-hipotecaria

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *Interdisciplinary Journal of Information, Knowledge, and Management*, 2019, 14, 45-79 https://doi.org/10.48550/arXiv.1809.03006

Bräuning, M., Malikkidou, D., Scalone, S., & Scricco, G. (2019). A new approach to early warning systems for small European banks. *Working Paper Series*, European Central Bank. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2348~351ba1be4c.en.pdf

Casal, R., Bouzas, J., & Oviedo, M. (2021). Aprendizaje Estadístico. *Apuntes Aprendizaje Estadístico del Máster en Técnicas Estadísticas, Universidad de la Coruña.*

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR*), 16, 321-357. https://doi.org/10.1613/jair.953

Comisión Nacional del Mercado de Valores. (2024). Boletín Internacional de la Comisión Nacional del Mercado de Valores: La Ley Dodd-Frank. *Boletín Internacional, Blog CNMV*. [Consulta en abril 2024]. Disponible en: https://www.boletininternacionalcnmv.es/ficha.php?menu_id=1&jera_id=119&cont_id=90

Crespo, L., Gómez García, M., Jovell, P., Rivera, B., & Villanueva, E. (5 diciembre,2023). Pérdidas de ingresos y de empleo durante la pandemia de COVID-19 y situación financiera de los hogares: evidencia de la EFF. *Boletin Económico 2023/T4 Banco de España.* https://www.bde.es/f/webbe/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEconomico/23/T4/Fich/be2304-art05.pdf

Cutler, A., Cutler, D.R. and Stevens, J.R. (2012) Random Forests. In: Zhang, C. and Ma, Y.Q., Eds., *Ensemble Machine Learning*, Springer, New York, 157-175. http://dx.doi.org/10.1007/978-1-4419-9326-7_5

Depository Institutions Deregulation and Monetary Control Act of 1980. (1980). *Conference report filed in House, H. Rept. 96-842. Title I: Monetary Control Act of 1980.* *https://www.congress.gov/bill/96th-congress/house-bill/4986*

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for Machine Learning. *Cambridge University Press.*

Dodd, R. (2009). La reforma del sistema financiero en Estados Unidos propone la reforma más radical de la regulación financiera desde el New Deal. *Finanzas y Desarrollo, Fondo Monetario Internacional*. [Consulta en abril 2024]. Disponible en: https://www.imf.org/external/pubs/ft/fandd/spa/2009/09/pdf/dodd.pdf

Dutta, N., Subramaniam, U., & Sanjeevikumar, P. (2018, noviembre). Mathematical models of classification algorithm of machine learning. *Hamad bin Khalifa University Press (HBKU Press).* https://doi.org/10.5339/qproc.2019.imat3e2018.3

Fahrmeir, L., et al. (2021). Regression models, methods and applications. *Springer*. pp. 23-84.

Farris, F. A. (2010). The Gini Index and Measures of Inequality. *The American Mathematical Monthly*, 117(10), 851–864. https://doi.org/10.4169/000298910X523344

Florescu, D., & England, M. (2019). Algorithmically generating new algebraic features of polynomial systems for machine learning. *Proceedings of the 4th Workshop on Satisfiability Checking and Symbolic Computation (SC2 '19), 12 pages. CEUR Workshop Proceedings 2460, 2019* https://ceur-ws.org/Vol-2460/paper4.pdf

Freund, Y., & Schapire, R. E. (1999). A short introduction to boosting (N. Abe, Trans.). *Journal of Japanese Society for Artificial Intelligence, 14(5), 771-780*. https://cseweb.ucsd.edu/~yfreund/papers/IntroToBoosting.pdf

Friedl, H., & Stampfer, E. (2001). Cross-validation. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of Environmetrics* (Vol. 1, pp. 452-460). John Wiley & Sons. http://dx.doi.org/10.1002/9780470057339.vac062

Fuller, E. W. (23 octubre, 2019). La banca del 100% y sus defensores: una breve historia. *Mises Wire*.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). *O'Reilly Media, Inc.*

Han, H., Guo, X., & Yu, H. (August 2016). Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. *En 2016 7th IEEE International Conference on Software Engineering and Service Science* (ICSESS) (pp. 1-4). http://dx.doi.org/10.1109/ICSESS.2016.7883053

Hastie, T., et al. (2009). The elements of statistical learning: data mining, inference, and prediction. New York: *Springer.*

Wenjie Chen, Mico Mrkaic, Malhar Nabar.. (2018, 3 de octubre). Efectos perdurables: La recuperación económica mundial a los 10 años de la crisis. *Blog IMF.* [Consulta en abril 2024]. Disponible en: https://www.imf.org/es/Blogs/Articles/2018/10/03/blog-lasting-effects-the-global-economic-recovery-10-years-after-the-crisis

IMF. (2021, octubre). Global Financial Stability Report. *IMF*. [Consulta en abril 2024]. Disponible en: https://www.imf.org/en/Publications/GFSR/Issues/2021/10/12/global-financial-stability-report-october-2021

IMF. (2022). History of IMF and World Bank. *IMF.* [Consulta en abril 2024]. Disponible en: https://www.imf.org/en/About/Factsheets/Sheets/2022/IMF-World-Bank-New

Igual, D. L. (1998). Valencia e Italia en el siglo XV: Rutas, mercados y hombres de negocios en el espacio económico del Mediterráneo occidental. *Bancaixa*.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning in Python. *Springer.*

Jerome, H. F. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, Vol. 29, No. 5, pp. 1189-123*. https://doi.org/10.1214/aos/1013203451

Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition, *Springer.*

Joseph, C. (2013). Advanced Credit Risk Analysis and Management, *Wiley*.

Kabbay, H. (2022). Artificial neural network concepts and examples. [Trabajo fin de máster, University of Missouri, St. Louis]. E-Archivo. https://www.studocu.com/in/document/madras-institute-of-technology-anna-university/artificial-intelligence-and-data-science/artificial-neural-network-concepts-and-examples/84288059

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*. https://dl.acm.org/doi/10.5555/3294996.3295074

Kim HY. Statistical notes for clinical researchers: Type I and type II errors in statistical decision. *Restor Dent Endod*. 2015 Aug;40(3):249-252. https://doi.org/10.5395/rde.2015.40.3.249

Li, W. (2016). Probability of Default and Default Correlations. *J. Risk Financial Manag*. 2016, 9(3), 7 https://doi.org/10.3390/jrfm9030007

Lichtner-Bajjaoui, A. (2020). *A mathematical introduction to neural networks* [Trabajo fin de máster, Universidat de Barcelona]. E-Archivo. https://diposit.ub.edu/dspace/bitstream/2445/180441/2/tfm_lichtner_bajjaoui_aisha.pdf

Malik, S., Harode, R., & Singh, A. (2020). XGBoost: A Deep Dive into Boosting (Introduction Documentation). *SFU Professional Computer Science*. https://doi.org/10.13140/RG.2.2.15243.64803

Manorathna, R. (2020, October 12). Polynomial Regression with a Machine Learning Pipeline. *Data365*. [Consulta en abril 2024]. Disponible en: https://www.researchgate.net/profile/Rukshan-Manorathna/publication/344616388_Polynomial_Regression_with_a_Machine_Learning_Pipeline/links/5f84941ba6fdccfd7b5ca885/Polynomial-Regression-with-a-Machine-Learning-Pipeline.pdf

Mitchell, T. M. (2017). Generative and discriminative classifiers: Chapter 3: Naive Bayes and logistic regression. Machine Learning. *McGraw Hill.*

Moncayo, J. (24 octubre, 2019). 1929: el mayor apocalipsis financiero. *Diario La Vanguardia*.

Montoya, A., Inversion, K., Odintsov, K., & Kotek, M. (2018). Home Credit Default Risk [Dataset]. https://www.kaggle.com/c/home-credit-default-risk/data

Nash, M. (11 abril, 2023). Mercado de bonos: Una historia de 'idas y vueltas'. *ElEconomista.es*

Ocampo, J. A. (2010). Impactos de la crisis financiera mundial. *Revista CEPAL*.

Oficina para la Protección Financiera del Consumidor. (4 septiembre, 2020). ¿Cuál es la diferencia entre un préstamo hipotecario de tasa de interés fija y un préstamo hipotecario de tasa de interés ajustable (ARM, por sus siglas en inglés)? *Consumer Finance.* [Consulta en abril 2024]. Disponible en: *https://www.consumerfinance.gov/es/obtener-respuestas/cual-es-la-diferencia-entre-un-prestamo-hipotecario-de-tasa-de-interes-fija-y-un-prestamo-hipotecario-de-tasa-de-inters-ajustable-arm-es-100/*

Oficina para la Protección Financiera del Consumidor. (2024). ¿Qué es una hipoteca de alto riesgo o subprime? *Consumer Finance.* [Consulta en abril 2024]. Disponible en: https://www.consumerfinance.gov/es/obtener-respuestas/que-es-una-hipoteca-de-alto-riesgo-o-subprime-es-110/#:~:text=Por%20lo%20general%2C%20una%20hipoteca,a%20este%20tipo%20de%20prestatarios

OMS. (2020, 27 de abril). COVID-19: Cronología de la actuación de la OMS. *WHO News*. [Consulta en abril 2024]. Disponible en: https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19

Ponz, C. S. (2020, 4 de mayo). Lecciones clave de la crisis de 2008 para los empresarios. *Sabadell News*.

Pozzi, S. (2019, 10 de septiembre). 10 años de la crisis | La zona cero: Lehman Brothers, el gatillo de la crisis. Economía. *Diario EL PAIS*.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.09516

PwC & GALYTIX. (2024). *Banks must act on their early warning systems or risk ROE downturn*. [Consulta en mayo 2024]. Disponible en:https://www.pwc.co.uk/banking-capital-markets/assets/documents/future-of-ews.pdf

Rocket Mortgage. (2020). ¿Qué son los títulos respaldados por hipotecas (MBS)? *Rocket Mortgage Blog.* [Consulta en abril 2024]. Disponible en: https://www.rocketmortgage.com/es/learn/titulos-valores-respaldados-por-hipotecas

Rosenthal, G., & Rosenthal, J. (2011). Statistics and Data Interpretation for Social Work. *Springer Publishing Company.*

Sánchez Marcos, M. (2021). La Banca. Historia, productos y evolución. [Trabajo fin de grado, Universidad de Valladolid]. E-Archivo https://uvadoc.uva.es/handle/10324/51717

Schmidt-Thieme, L. (2007). *Machine learning: 4. Decision trees*. [Presentación de PowerPoint]. Information Systems and Machine Learning Lab (ISMLL), Institute for Business Economics and Information Systems, & Institute for Computer Science, University of Hildesheim. https://www.ismll.uni-hildesheim.de/lehre/ml-13w/script/ml-04-decisiontrees-2up.pdf

Solis-Mullen, J. (2022, 13 de enero). Dinero y banca en los EE. UU. tras las crisis de los 1970 y 80. *Mises Wire.*

Trevor, H., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning (2nd ed.). *Springer Series in Statistics.*

Turner, H. (2008*). Introduction to generalized linear models*. [Presentación de PowerPoint]. ESRC National Centre for Research Methods, UK, and Department of Statistics, University of Warwick, UK. https://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

Vive Más Vida Magazine. (2022, junio). "¿Sabías que en 1914 Western Unión creó la primera tarjeta de crédito?" *Vive Más, Educación Financiera*. [Consulta en abril 2024]. Disponible en: https://www.vivemasvidas.com/finanzas/educacion/primera-tarjeta-credito#:~:text=Justo%20antes%20de%20la%20Primera,Vaya%2C%20como%20una%20tarjeta%20VIP.

World Bank. (2010). World Bank History. *World Bank Archive*. [Consulta en abril 2024]. Disponible en: https://www.worldbank.org/en/archive/history

Yan, L., & Cain, J. (2020, October 5). *The Normal (Gaussian) Distribution*. [Presentación de PowerPoint]. University of Stanford. https://web.stanford.edu/class/archive/cs/cs109/cs109.1212/lectures/10_normal_gaussian.pdf

Zubair, K. (2021). *Mathematical Foundations for Machine Learning and Data Science Analysis and Evaluation of Classifier's Performance*. [Presentación de PowerPoint]. Lahore University of Management Sciences. https://www.zubairkhalid.org/ee212/2021/notes13.pdf