UNIVERSIDADE DA CORUÑA

Facultade de Economía e Empresa

Trabajo de fin de máster

# Aplicación de técnicas de aprendizaje automático a la probabilidad de default en la gestión de riesgo de crédito

Óscar Paul Sánchez Riveros

Tutor: Xosé Manuel Martínez Filgueira

Máster Universitario en Banca y Finanzas

Curso académico 2023/24

Trabajo de Fin de Máster presentado en la Facultad de Economía y Empresa de la Universidad de la Coruña para la obtención del Máster Universitario en Banca e Finanzas.

# Resumen

En el presente estudio se analizan los beneficios de aplicar modelos de aprendizaje automático para determinar la probabilidad de defaut en la concesión de créditos. La mayor capacídad de predicción supone mejoras en identificar clientes buenos y malos, así como generar ahorros de capital regulatorio. Se realizó una revisión del proceso de construcción y validación de los modelos y una revisión de la literartura pertinente. Para ilustrar la discusión, se llevó a cabo un caso práctico con una base de datos de préstamos de libre acceso, la cual fue ajustada, validada y comparada en tres modelos de aprendizaje automático: Árbol de desición, Random Forest y Gradient Boosting, además del modelo tradicional Logit. Al analizar las métricas de rendimiento, se concluyó que los modelos de aprendizaje automático superan al modelo tradicional en la clasificación de clientes. Para traducir estas mejoras en beneficios económicos, se estimaron los ahorros de capital usando un enfoque basado en calificaciones internas (IRB).Los resultados mostraron que implementar modelos de aprendizaje automático como Gradient Boosting podría generar ahorros del 8 % en capital regulatorio.

Palabras claves: Aprendizaje automático, Probabilidad de default, Riesgo de crédito, Capital regulatorio, Sistema IRB.

# Abstract

This study analyzes the benefits of applying machine learning models to determine the probability of default in the granting of credits. The greater prediction capacity means improvements in identifying good and bad clients, as well as generating regulatory capital savings. A review of the model construction and validation process and a review of the relevant literature was carried out. To illustrate the discussion, a practical case was carried out with an open access loan database, which was adjusted, validated and compared in three machine learning models: Decision Tree, Random Forest and Gradient Boosting, in addition to the traditional Logit model. By analyzing the performance metrics, it was concluded that the machine learning models outperform the traditional model in customer classification. To translate these improvements into economic benefits, capital savings were estimated using an internal ratings-based (IRB) approach. The results showed that implementing machine learning models such as Gradient Boosting could generate savings of 8% in regulatory capital.

Keywords: Machine learning, Probability of default, Credit risk, Regulatory capital, IRB system.

# Índice

## Contenido

# Índice de figuras

# Índice de tablas

# 1. Introducción

Desde la segunda mitad del siglo XX, con la irrupción de los ordenadores e Internet, se han logrado avances significativos en las Tecnologías de la Información y la Comunicación (TIC), dando lugar a lo que hoy se conoce como la tercera revolución industrial. Este período ha presenciado el desarrollo de redes de banda ancha, masificación de dispositivos móviles inteligentes, servicios de computación en la nube y la capacidad para procesar grandes cantidades de datos, transformando radicalmente tanto la industria como la sociedad en general. Este cambio se ha denominado comúnmente como transformación digital (Villar Mir, 2020), marcando una era en la que la informática se ha generalizado en las interacciones sociales e incluido en los procesos de producción y comercialización de casi todos los sectores de la economía.

El sector financiero no ha permanecido al margen de la transformación digital. La presencia de consumidores más habituados a las tecnologías actuales ha forzado a las instituciones financieras a redefinir sus procesos y a establecer nuevos canales de atención al cliente. Esta interacción creciente entre empresas y consumidores ha desencadenado una avalancha de datos disponibles y las instituciones han visto en ello nuevas oportunidades para conocer las características de sus clientes y clientes potenciales reforzando e ideando nuevos planes estratégicos, el camino no ha sido fácil, procesar enormes cantidades de datos con diferentes características requiere de gran inversión en tecnología e innovación y además en capacitación al personal. Es por eso que las instituciones han encontrado un aliado en la inteligencia artificial (IA) y modelos de aprendizaje automático, Según la encuesta Global de McKinsey (2023), más del 20% de las empresas líderes en rendimiento informaron que al menos el 20% de sus ganancias antes de intereses e impuestos en 2022 se debieron al uso de IA.

La integración del aprendizaje automático en las instituciones financieras, han sido utilizados para la detección de fraudes en tiempo real, la predicción de riesgos crediticios, la personalización de recomendaciones a los clientes y la segmentación de clientes para mejorar la experiencia del usuario (Ginzo Technologies S.L. ,2024). En este estudio se abordará con particular relevancia los modelos destinados a determinar la probabilidad de default en la gestión de riesgos crediticios, es decir en realizar una clasificación binaria para saber si un cliente cumplirá o no con el pago de su crédito, para tal objetivo los modelos realizan una clasificación precisa y segmentan poblaciones de clientes considerando una amplia gama de variables o características. El financiamiento en la modalidad de crédito es la principal operación activa de los bancos, representa por tal la mayor fuente de sus ingresos, en ese sentido es crucial evaluar la verdadera capacidad predictiva y discriminativa de los modelos para revisar el impacto económico que le generé en su organización.

En el contexto descrito, el presente estudio tiene como objetivo principal, investigar y analizar el impacto económico de la aplicación de técnicas de aprendizaje automático en la evaluación y predicción de la probabilidad de default en el riesgo de crédito. Se analizan los posibles beneficios de utilizar estas técnicas en comparación con los métodos tradicionales, así como la identificación de posibles problemas a los que los agentes podrían enfrentarse al implementarlas. Aunque existe literatura sobre la aplicación de modelos de aprendizaje automático al cálculo de la probabilidad de default, gran parte de esta se limita a comparar los modelos mediante métricas de rendimiento, sin abordar si los rendimientos se traducen en beneficios económicos tangibles.

Para alcanzar este objetivo, se realizará una comparación de los rendimientos de modelos de aprendizaje automático evaluados en un caso práctico. Los resultados obtenidos se utilizarán para evaluar los beneficios asociados al uso de estos modelos.

El estudio cobra una particular relevancia en el actual contexto descrito inicialmente, donde la comprensión de temas relacionados con la inteligencia artificial es fundamental.

Estos aspectos están redefiniendo no solo la forma en que vivimos, trabajamos y nos relacionamos, sino también la manera en que se gestionan los riesgos financieros y se toman decisiones en el ámbito crediticio.

El estudio se estructura en seis bloques fundamentales. En primer lugar, la introducción establece el contexto actual del estudio, define los objetivos y justifica su relevancia, además de proporcionar un resumen de la estructura del trabajo. El segundo bloque, el marco teórico, aborda los conceptos principales de las técnicas de aprendizaje automático y su aplicación en el análisis de riesgo de crédito. Asimismo, se detallan los modelos que se utilizarán (regresión logística, árbol de decisión, random forest y gradient boosting) y las técnicas estadísticas asociadas, además de realizar una revisión de estudios previos sobre el tema. El tercer bloque se centra en la metodología aplicada al caso práctico, describiendo el caso, el tratamiento de la base de datos, el análisis exploratorio, preprocesamiento de variables, la ejecución de los modelos y sus resultados individuales. En el caso de estudio, se incluirá el análisis de un conjunto de créditos para evaluar la aplicación práctica de las técnicas estudiadas en el ámbito empresarial, destacando la técnica más efectiva para determinar la probabilidad de incumplimiento en función de las características crediticias. Los resultados comparativos se presentan en el cuarto bloque, donde se describen y analizan comparativamente los distintos modelos empleados. En el quinto bloque se exponen las conclusiones, evaluando si se alcanzaron los objetivos planteados. Finalmente, el sexto bloque detalla la bibliografía utilizada en la investigación.

# 2. Marco Teórico

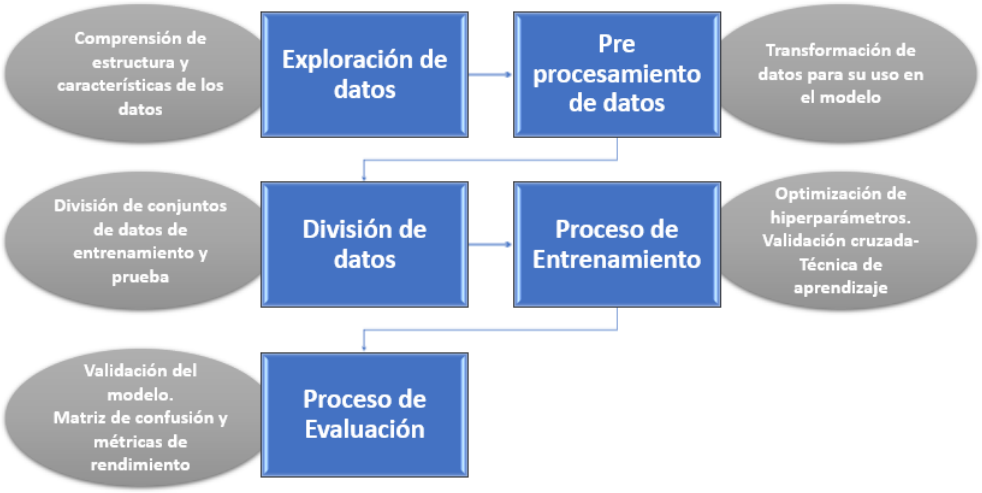
## 2.1 Aprendizaje Automático.

El aprendizaje automático se define como una disciplina científica que se ocupa del diseño y desarrollo de algoritmos que permiten a las computadoras aprender a partir de los datos. Bishop, C. M. (2006). La definición subraya la naturaleza rigurosa y sistemática del aprendizaje automático que se basa en métodos experimentales para avanzar en su entendimiento y aplicación, no es solo un conjunto de herramientas tecnológicas sino un campo de estudio que impulsa el progreso y la innovación digital.

### 2.1.1. Construcción y validación de modelos de aprendizaje automático.

Generalmente aceptado por distintos autores el proceso de construcción y establecimiento de un modelo implican las etapas de exploración de datos, pre procesamiento de datos, división del conjunto de datos, entrenamiento del modelo y evaluación.

***Figura 1: Proceso de validación y construcción de un modelo***




*Fuente: Elaboración propia*

**Exploración de datos**: En esta etapa, se analizan y examinan los datos disponibles para comprender su estructura, características y relaciones entre variables. Esto incluye visualizar datos, detectar anomalías, desigualdades de clases e identificar patrones. El objetivo es obtener una compresión profunda de los datos.

**Preprocesamiento de datos:** En esta etapa, se preparan los datos para su uso en el modelo. Esto incluye la limpieza de datos (eliminación de valores nulos o duplicados), normalización, estandarización, codificación de variables categóricas y la creación de nuevas variables a partir de las existentes. El objetivo es transformar los datos crudos en un formato adecuado y limpio que mejore la eficiencia y precisión del modelo predictivo.

**División del conjunto de datos:** El conjunto de datos disponibles, es dividido en dos subconjuntos de datos. Datos de entrenamiento y datos de prueba. Los datos de entrenamiento consisten en una parte significativa del conjunto de datos (alrededor del 70% o 80%) y sirven para entrenar el modelo, por otra parte, los datos de prueba consisten en una parte pequeña de los datos (alrededor del 30% o 20%), se utilizan para verificar como el modelo generaliza a datos no vistos después de que haya sido entrenado.

**Entrenamiento:** Durante esta etapa, se utiliza el conjunto de datos de entrenamiento para que el algoritmo ajuste los parámetros del modelo y aprenda las relaciones entre las variables explicativas y variable objetivo, esto a fin de minimizar su error de predicción. La duración o profundidad del entrenamiento dependen de dos procesos fundamentales: la optimización de hiperparámetros y la validación cruzada.

Los hiperparámetros son parámetros que controlan aspectos específicos del proceso de aprendizaje y la arquitectura del modelo. Estos incluyen la magnitud del ajuste que el modelo hace a sus parámetros en cada iteración, la selección del número de variables a utilizar, parámetros de regularización para evitar sobreajuste, etc. La selección y ajuste de hiperparámetros adecuados pueden mejorar significativamente el rendimiento del modelo,

una mala elección de los mismos puede llevar a que el modelo no generalice bien o que no converja adecuadamente.

Otro aspecto crucial en el entrenamiento es el proceso de validación cruzada, es una técnica que permite evaluar el rendimiento de un modelo utilizando solo los datos de entrenamiento y asegurar que se generalice bien a datos no vistos. La técnica consiste en dividir el conjunto de datos de entrenamiento en K pliegues (folds) de igual tamaño. En cada iteración, se utiliza un pliegue como conjunto de validación y los K – 1 pliegues restantes como conjunto de entrenamiento, el proceso se repite K veces, utilizando un pliegue diferente como conjunto de validación en cada iteración, el rendimiento se calcula, en caso de ser un problema de clasificación, usando la votación mayoritaria para determinar la clasificación final. Este proceso permite una mejor estimación del rendimiento y reducción de la varianza.

Aunque las técnicas de entrenamiento empleadas por cada algoritmo son específicas y dependen del modelo utilizado, es posible agrupar estos modelos según enfoques comunes de aprendizaje. A continuación, se detallarán los dos principales:

***Tabla 1: Enfoques de Aprendizaje***

| ***Tipo de aprendizaje*** | ***Modo de aprendizaje*** | ***Problemas que aborda*** |
|---|---|---|
| **Aprendizaje Supervisado** | Se entrena al algoritmo proporcionándole información sobre datos etiquetados, en los cuales se incluye las características y la variable objetivo. | ***Regresión.*** Cuando el objetivo de aprendizaje supervisado es un resultado numérico y continuo.<br>***Clasificación***. Cuando el objetivo de aprendizaje supervisado es predecir un resultado categórico (binomial - multimodal) |
| **Aprendizaje no supervisado** | Se entrena al algoritmo proporcionándole información sobre características sin darle una variable objetivo. | ***Agrupar.*** Cuando el objetivo del aprendizaje no supervisado es segmentar las observaciones en grupos similares según las variables observadas |

*Fuente: Elaboración propia basado en Universidad Europea. (2022)*

Se pueden mencionar también los enfoques de aprendizaje semi supervisado, donde el algoritmo se entrena con solo algunas de las observaciones de la variable objetivo. Otro enfoque conocido es la técnica de refuerzo, donde el algoritmo recibe un premio en función de los resultados obtenidos.

En este estudio, nos centraremos en modelos con un enfoque de aprendizaje supervisado para abordar problemas de clasificación.

**Evaluación:** En esta etapa, se evalúa el rendimiento del modelo entrenado comparando sus predicciones con los valores reales del conjunto de datos de prueba. el objetivo en esta etapa es determinar la capacidad del modelo para generalizar a nuevos datos y asegurar que cumple con los criterios de desempeño requeridos antes de su implementación en un entorno de producción.

Al aplicar modelos de clasificación, se suele utilizar la matriz de confusión para evaluar medidas de desempeño, en esta se compara eventos categóricos reales con los eventos categóricos predichos. Cuando se predice el evento correcto se refiere a esto como un verdadero positivo (TP). Si se predice un evento que no sucede esto se denomina falso positivo (FP). Alternativamente cuando no se predice un evento y sucede, esto se denomina falso negativo (FN), por último, si predice un evento como que no va suceder y no sucede se le denomina verdadero negativo (TN).

***Figura 2: Matriz de confusión***


*Fuente: Extraído de* Boehmke y Greenwell (2020)

De la matriz de confusión se extraen medidas de rendimiento de clasificadores binarios para medir la efectividad del modelo:

***Tabla 2 : Medidas de rendimiento***

| *Medida* | *Descripción* |
|---|---|
| **Exactitud** | Porcentaje de predicciones correctas realizadas por el modelo en comparación con total de predicciones.<br>Mide la efectividad general del modelo<br>$Exactitud = \frac{TP + TN}{Total\ de\ predicciones}$ |
| **Sensibilidad** | Proporción de verdaderos positivos (casos correctamente identificados) sobre el total de verdaderos positivos y falsos negativos.<br>Evalúa la capacidad del modelo para identificar correctamente todas las instancias de la clase positiva.<br>$Sensibilidad = \frac{TP}{TP + FN}$ |
| **Especificidad** | Proporción de verdaderos negativos entre todas las instancias que realmente son negativas.<br>Evalúa la capacidad del modelo para identificar correctamente todas las instancias de la clase negativa.<br>$Especificidad = \frac{TN}{TN + FP}$ |
| **Precisión** | Proporción de verdaderos positivos entre todos los positivos predichos en el modelo.<br>$Precisión = \frac{TP}{TP + FP}$ |
| **F1 - Score** | Media armónica de la precisión y la sensibilidad.<br>Útil para obtener una sola métrica que balance a tanto precisión como sensibilidad, especial cuando las clases están desequilibradas.<br>$F1 - score = 2\ x \frac{Precision\ x\ Sensibilidad}{Precision + Sensibilidad}$ |
| **AUC – ROC (Área bajo la curva – Característica operativa del receptor)** | La curva ROC es la gráfica que muestra la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos a varios umbrales de clasificación.<br>El área bajo la curva proporciona una medida agregada del rendimiento del modelo en todos los posibles umbrales de clasificación. AUC cercano a 1 indica un buen rendimiento. |

*Fuente: Elaboración propia basado en Barrios, J. (2019)*

La aplicación de cada uno de los procesos de construcción y validación de un modelo son esenciales para desarrollar modelos de aprendizaje automático robustos y precisos.

### 2.1.2. Aplicación de aprendizaje automático en riesgo de crédito.

El Comité de Basilea define el riesgo de crédito como "la posibilidad de que un prestatario o contraparte incumpla sus obligaciones contractuales en los términos acordados" (Comité de Basilea de Supervisión Bancaria, 2017). Ante este riesgo el mismo comité recomienda a las entidades financieras implementar modelos que permitan reducirlo de manera efectiva.

El estudio se concentrará en dos tipos de beneficios ligados al uso de modelos de aprendizaje automático en la determinación de la probabilidad de incumplimiento en la gestión de riesgo de crédito: (i) la identificación eficiente de clientes buenos y malos, (ii) la optimización de los requerimientos de capital.

La identificación eficiente de clientes buenos y malos, está referido a la capacidad de clasificación que posee un modelo, pudiendo ordenar correctamente a los prestatarios según su riesgo de incumplimiento, la ventaja que los modelos de aprendizaje automático poseen es que permiten procesar y analizar grandes volúmenes de datos pudiendo analizar patrones complejos y relaciones no lineales que los métodos tradicionales pueden obviar. Esto resultaría en decisiones de admisión crediticia más informadas y eficientes, permitiendo a las entidades financieras minimizar el riesgo de crédito al seleccionar con mayor precisión a los prestatarios más solventes y rechazar a aquellos con mayor probabilidad de incumplimiento.

La optimización de los requerimientos de capital, esta referido a la cantidad de capital propio que las normas y regulaciones obligan a las instituciones financieras a mantener en relación con sus activos ponderados por riesgo, esto garantiza que las instituciones financieras tengan suficiente capital para absorber pérdidas inesperadas y proteger a los depositantes y al sistema financiero en general. El cálculo de los requerimientos de capital se basa en los activos ponderados por riesgo (RWA), es decir que cada activo que tenga la institución financiera se pondera de acuerdo a su riesgo asociado, por ejemplo, un préstamo a un prestatario con alta probabilidad de incumplimiento tendrá una mayor ponderación de riesgo en comparación con un préstamo a un prestatario de alta calidad crediticia.

Para tal efecto el Comité de Supervisión Bancaria de Basilea (2017) establece dos métodos para el cálculo del requerimiento de capital, el método estándar y el método de calificaciones internas.

El método estándar consiste en una ponderación fija, establecida por el ente regulador para medir la exposición al riesgo de un activo. La ponderación fija el requerimiento de capital necesario para cubrir posibles pérdidas. Este es un enfoque simple y uniforme donde se aplica las mismas ponderaciones a todos los activos de una misma categoría. En cambio, con el método basado en calificaciones internas (IRB, por siglas en inglés) la ponderación al riesgo es determinada por un modelo interno de la entidad financiera y está en función, entre otras variables, de la probabilidad de incumplimiento de cada activo. Esta característica ofrece una ventaja significativa al ajustar el requerimiento de capital según el riesgo específico de cada activo, en lugar de aplicar una ponderación general como el método estándar, esto tendría como consecuencia una asignación más eficiente del capital.

La categoría del activo que se evaluará en este estudio corresponde a las exposiciones minoristas[1], esto porque en el caso de estudio evaluaremos un conjunto de datos de créditos minoristas, en Comité de Supervisión Bancaria de Basilea (2017) se establece como parte del método estándar que para exposiciones minoristas se usará la ponderación al 75%, por otro lado el método basado en las calificaciones internas (IRB), el requerimiento de capital se calcula como se indica en la Eq(1).

$$K = LGD * N\left(\sqrt{\frac{1-R}{R}} * G(PD) + \sqrt{\frac{R}{1-R}} * G(0.999)\right) - PD * LGD \quad (1)$$

[1] El comité de Basilea establece diferentes ponderaciones, fórmulas de requerimiento de capital y correlaciones según las categorías de los activos, esto tanto en sus métodos estándar como en los métodos IRB.

Donde K es el requerimiento de capital, PD es la probabilidad de incumplimiento, LGD es la probabilidad dada el incumplimiento, R es la correlación, G es la función inversa de la distribución normal estándar y N es la función de distribución normal acumulativa. La fórmula para la correlación R se indica en la Eq(2).

$$R = 0.03 * \frac{(1 - e^{-35*PD})}{(1 - e^{-35})} + 0.16 * (1 - \frac{(1 - e^{-35*PD})}{(1 - e^{-35})} \quad (2)$$

Luego la cantidad de activos ponderados por riesgo se puede calcular como se indica en la Eq (3).

$$RWA = k * 12.5 * EAD \quad (3)$$

Donde EAD es la exposición en caso de incumplimiento y 12.5 convierte el requerimiento de capital en términos porcentuales a una base anual, esto en concordancia con Basilea III para asegurar que los requerimientos de capital estén expresados en términos anuales, cumpliendo con las normativas regulatorias.

Los activos ponderados por riesgo (RWA por sus siglas en inglés) aseguran que los bancos tengan suficiente capital reservado para cubrir perdidas potenciales en caso de que los prestatarios incumplan sus obligaciones. Este RWA al ser ajustado por la PD adecuaría mejor los requerimientos de capital de manera más precisa a los riesgos reales asociado a los activos. Estas fórmulas se usarán en el caso práctico para evaluar a los modelos en razón de estos beneficios que puedan otorgar su aplicación.

En el siguiente apartado se abordará la descripción de una serie de modelos utilizados en la gestión de riesgo de crédito.

## 2.2. Descripción de los modelos

Las técnicas de aprendizaje automático se han utilizado predominantemente para detectar las relaciones entre las características de un prestatario y posibles escenarios de

incumplimiento. En este estudio, los modelos que abordarán el riesgo crediticio se basarán tanto en métodos estadísticos, como la regresión logística (modelo tradicional), como en modelos de aprendizaje automático, tales como Árboles de Decisión, Random Forest y Gradient Boosting (GBoost), esto con el fin de comparar sus diferencias en la evaluación del riesgo.

Aunque los modelos seleccionados abordan tanto problemas de regresión como de clasificación, este estudio se centrará en los problemas de clasificación. Esto se debe a que la probabilidad de incumplimiento es, por naturaleza, un problema de clasificación. Por lo tanto, se prestará especial atención a cómo cada modelo maneja este tipo de problema para determinar la probabilidad de default.

### 2.2.1. Regresión Lineal y Regresión Logit

La regresión lineal y logística son técnicas ampliamente utilizadas en diversas aplicaciones prácticas. El modelo Logit fue presentando por primera vez por Cox, D.R. (1958). Ambas han servido como pilares para el desarrollo de métodos más avanzados en el aprendizaje automático. Incluir estas técnicas en una comparación con métodos más avanzados permitirá evaluar las mejoras en términos de precisión y manejo de datos complejos.

En una regresión lineal, el objetivo es encontrar los coeficientes que minimicen la suma de los errores al cuadrado entre las predicciones y los valores reales de los datos. La ecuación de minimización, de forma generalizada para una regresión lineal múltiple donde hay más de una variable independiente, se puede expresar en su notación matricial como se indica en la Eq(4).

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

Donde $\boldsymbol{y}$ es el vector de valores dependientes, $\boldsymbol{X}$ es la matriz de variables independientes, $\boldsymbol{\beta}$ es el vector de coeficientes y $\boldsymbol{\varepsilon}$ es el vector de errores.

La función objetivo de los modelos lineales se puede expresar mediante la Eq(5).

$$\min \boldsymbol{e}^T \boldsymbol{e} = \min\ (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{5}$$

El vector $\boldsymbol{\beta}$ de coeficientes que minimiza la suma de los errores al cuadrado se determina mediante la Eq(6).

$$\boldsymbol{\beta} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \tag{6}$$

Los modelos de regresión lineal pueden utilizar variables independientes con valores continuos, dicretos o categoricos (por ejemplo, promedio de ingresos, promedio de gastos, número de hijos, sexo, etc.) para estimar una variable dependiente continua (por ejemplo, probabilidad de incumplimiento crediticio). Esto puede ser útil para pronosticar si un prestatario pagará o no un crédito. Si la probabilidad de incumplimiento es muy alta, es posible que el cliente presente un incumplimiento. Sin embargo, esto requiere la intervención de un experto humano para decidir cuál es el umbral adecuado para un cliente en particular, es decir, ¿qué tan alta debe ser la probabilidad de incumplimiento para declarar que un cliente presentará un default?

En lugar de depender del criterio de un experto humano, se puede entrenar a una computadora para que establezca el umbral automáticamente, proporcionándole al agente artificial muchas etiquetas binarias (por ejemplo, 0 si el cliente no presenta default y 1 si lo hace) como variable dependiente. En este caso, se debe modificar el modelo de regresión lineal para adaptarlo a una salida discreta, utilizando un modelo de regresión logística.

La regresión logística se usa para problemas de clasificación binaria (donde los resultados son categorias como 0 o 1. Para transformar los valores predichos en rango [0,1] se usa la funcion Sigmoide, la cual transforma cualquier valor real en un valor entre 0 y 1, esta se indica en la Eq(7).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

Aplicando la funcion sigmoide al predictor lineal $\boldsymbol{X\beta}$ **:**

$$\widehat{\mathbf{P}} = \sigma(\boldsymbol{X\beta}) = \frac{1}{1 + e^{-\boldsymbol{X\beta}}} \tag{8}$$

Donde $\widehat{\mathbf{P}}$ es el vector de probabilidades predichas.

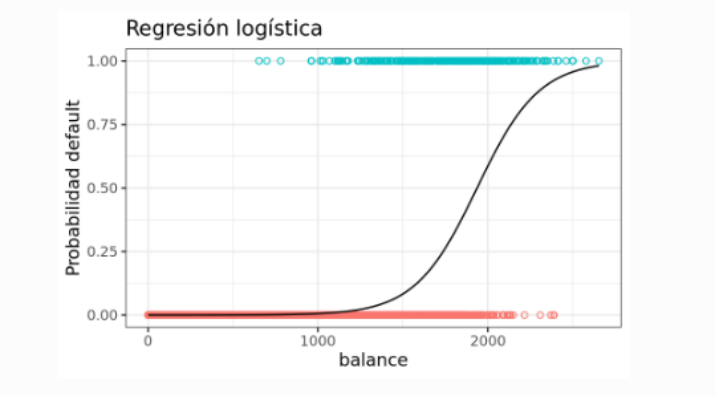
La regresión logística primero calcula las probabilidades de que ocurra un evento para diferentes niveles de la variable independiente. Luego, toma el logaritmo de estas probabilidades para crear un criterio continuo e ilimitado. El logaritmo de la probabilidad, conocido como logit, se modela como una combinación lineal de las variables independientes, lo que resulta más manejable en un contexto de regresión. De esta manera, la función objetivo en una regresión logística se puede expresar como en la Eq(9).

$$\min J(\boldsymbol{\beta}) = \boldsymbol{min} - \frac{1}{n}[\boldsymbol{y^T log}(\widehat{\mathbf{P}}) + (1 + \boldsymbol{y})^T \boldsymbol{log}(\mathbf{1} - \widehat{\mathbf{P}})] \tag{9}$$

Donde $\boldsymbol{y}$ es el vector de valores observados (0 o 1) de tamaño $n\ x\ 1$

Para encontrar los coeficientes $\boldsymbol{\beta}$ que minimizan la funcion objetivo se utiliza un metodo iterativo de optimizacion, como el descenso de gradiante.

***Figura 3: Ajuste de modelo logit***



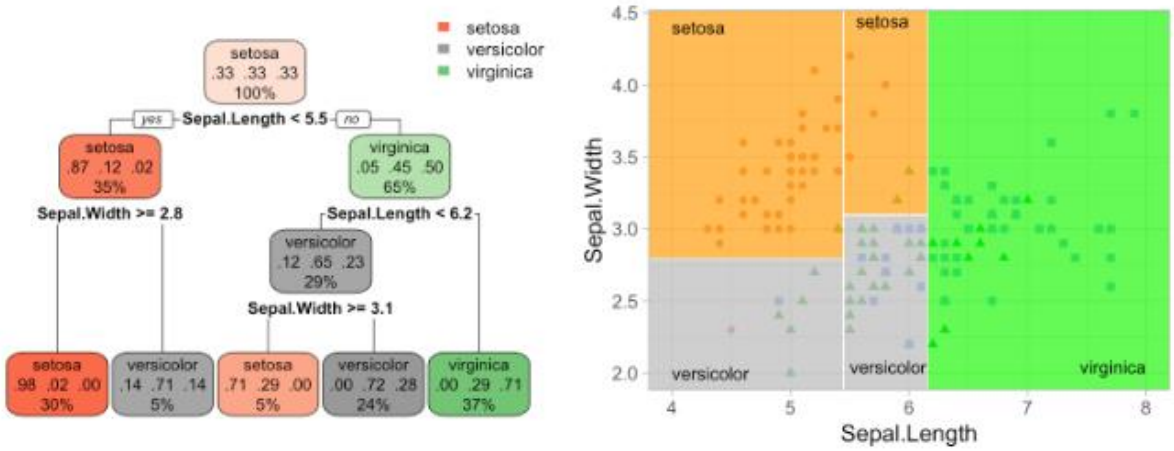

Fuente: Extraído de Amat, J. (2016)

Se ha establecido el modelo logit como un enfoque fundamental para abordar problemas de clasificación. Dado su estatus de modelo pilar, será comparado con modelos de aprendizaje automático en este estudio.

## 2.2.2. Árbol de decisión

Un árbol de decisión es un algoritmo de aprendizaje automático supervisado presentado por Breiman, Friedman, Olshenque y Stone (1984), se emplea tanto en problemas de clasificación como de regresión. Este algoritmo segmenta repetidamente el conjunto de datos en subconjuntos más pequeños basándose en características específicas, logrando que cada división resulte en subconjuntos más homogéneos en términos de la variable objetivo. Su estructura se asemeja a un árbol binario y su lógica de toma de decisiones es fácil de interpretar.

***Figura 4: Árbol de decisión para problema de clasificación***




Fuente: Extraido de Boehmke y Greenwell (2020)

### 2.2.2.1. Funcionamiento del algoritmo árbol de decisión

El algoritmo de árbol de decisión comienza con la división del nodo raíz, el cual contiene la totalidad de los datos de entrenamiento. Esta división se realiza en función de la característica que mejor segmenta los datos en subconjuntos más homogéneos. Para medir dicha homogeneidad en problemas de clasificación, se utiliza principalmente el índice de Gini,

el cual evalúa la impureza o heterogeneidad de los datos en cada subconjunto resultante de la división de datos.

- Cálculo del índice de Gini

Para un conjunto de datos con K clases, el índice de Gini se calcula como se indica en la Eq(10).

$$Gini = 1 - \sum_{i=1}^{K} p_i^2 \tag{10}$$

Donde $p_i$ es la proporción de elementos en la clase $i$ en el nodo.

Si el índice de Gini es igual a 0, indica que el nodo es puro, es decir, todas las observaciones pertenecen a una misma clase. Valores cercanos a 0 indican baja impureza, lo que significa que la mayoría de las observaciones en el nodo pertenecen a una sola clase. En contraste, si el índice de Gini toma un valor cercano a 0.5, indicaría alta impureza, sugiriendo que las observaciones en el nodo están distribuidas de manera uniforme entre las diferentes clases.

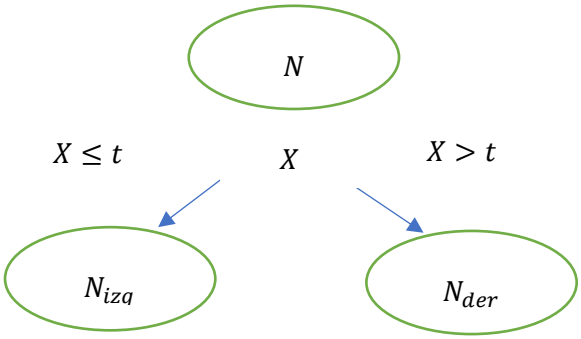
- Función objetivo – Minimización del índice de Gini ponderado

Para cada par de subconjuntos formados a partir de la división de un nodo, se calculará su índice Gini. Utilizando estos índices, se determinará el índice Gini ponderado, que combina las impurezas de ambos subconjuntos.

La función objetivo que emplea el algoritmo en cada nodo para decidir la característica y el punto de división se basa en la minimización de la impureza de Gini ponderada de los subconjuntos resultantes de una división. Este proceso puede expresarse de la siguiente manera:

Sea una característica $X$ y un punto de división $t$ , dividimos el nodo $N$ en dos nodos $N_{izq}$ y $N_{der}$. El $N_{izq}$ contiene las observaciones que cumplen $X \leq t$ $N_{der}$ contiene las observaciones que cumplen $X > t$.

***Figura 5: División de nodos***




Fuente: Elaboración propia

La función objetivo es encontrar la característica $X$ y el punto de división $t$ que minimice la impureza de Gini ponderada, esto se expresa en la Eq(11).

$$\min_{X,t} Gini_{ponderada} = \min_{X,t}\left(\frac{|N_{izq}|}{N}.Gini(N_{izq}) + \frac{|N_{der}|}{N}.Gini(N_{der})\right) \quad (11)$$

Una vez dividido el nodo inicial, el proceso de selección y división se repite en cada subconjunto, formando un proceso de ramificación. Este procedimiento continuará hasta alcanzar un criterio de parada, como un número mínimo de observaciones en un nodo o una profundidad máxima del árbol. El proceso también puede detenerse y realizar una poda para simplificar el árbol, recortando y fusionando nodos menos informativos que no contribuyen significativamente a la capacidad de generalización del modelo. Estos criterios mencionados forman parte de los hiperparámetros para determinar la complejidad del modelo.

- Consideraciones Adicionales

Una vez realizado el proceso de clasificación de una nueva observación, el usuario puede rastrear todos los nodos recorridos que determinaron su clasificación en una determinada

clase. A diferencia de otras técnicas de aprendizaje automático, esto proporciona una explicación fácil de entender del razonamiento y la decisión del modelo.

Determinadas características pueden ser utilizadas más de una vez como criterio de decisión en diferentes nodos. La frecuencia con la que una característica es seleccionada puede indicar su importancia relativa para la clasificación.

Un proceso de ramificación muy extenso en un árbol de decisión puede llevar a problemas de sobreajuste, capturando tanto las relaciones subyacentes como el ruido específico del conjunto de datos. Esto puede resultar en un rendimiento excelente para los datos de entrenamiento, pero deficiente para datos nuevos.

### 2.2.3. Random Forest

Random Forest es un modelo de aprendizaje automático presentado por Breiman, L. (2001), utiliza una técnica de ensamblaje combinando múltiples modelos. En este enfoque cada modelo genera una predicción diferente y, a partir de estas, se obtiene una única predicción agregada. En el caso de Random Forest, los modelos utilizados son árboles de decisión, cuyo funcionamiento se describió en el apartado anterior.

#### 2.2.3.1. Funcionamiento del algoritmo Random Forest

El algoritmo Random Forest construye una colección de árboles de decisión, donde cada árbol se entrena utilizando subconjuntos diferente del conjunto de datos original, los subconjuntos se forman a partir del método Bootstrap (muestreo aleatorio con reemplazo).

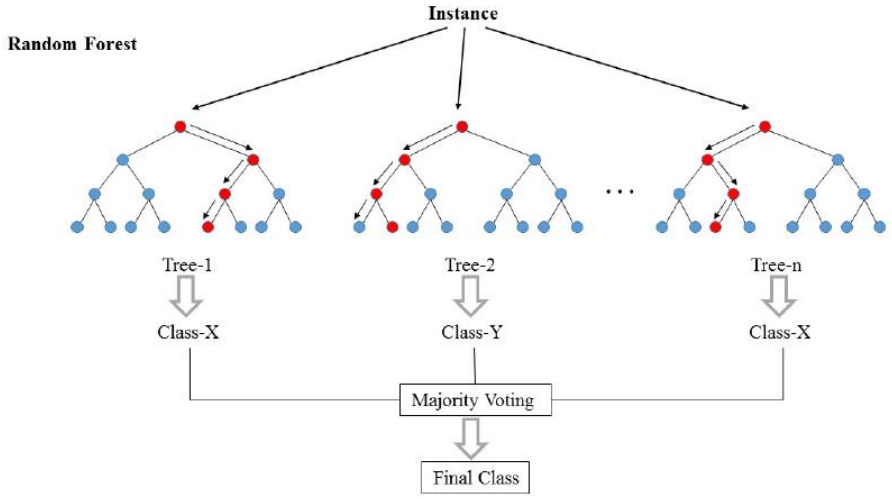
El método Bootstrap funciona de la siguiente manera: a partir de una muestra original de tamaño $n$, se generan múltiples nuevas muestras, también de tamaño $n$. Cada una de estas muestras se forma seleccionando observaciones de la muestra original con reemplazo, lo que significa que una misma observación puede aparecer varias veces en una sola muestra. Cada uno de estos conjuntos de muestras se utiliza para entrenar un árbol de decisión independiente.

Además, el algoritmo introduce un componente adicional de aleatoriedad en la selección de características. En cada nodo de cada árbol de decisión, se selecciona aleatoriamente un subconjunto de características del total disponible, utilizando normalmente$\sqrt{X}$ características para problemas de clasificación, donde $X$ es el número total de caracteriticas.

A partir de este punto, el algoritmo sigue el mismo proceso que en los árboles de decisión para cada árbol independiente, evaluando recursivamente cada característica y su posible punto de división, eligiendo aquel que minimice la impureza (índice Gini ponderado) de los nodos resultantes. Este proceso de selección de características y división se repetirá hasta que se cumpla un criterio de parada establecidos previamente en sus hiperparámetros que optimicen el modelo.

Una vez que todos los árboles de decisión en el bosque han sido entrenados, se utilizan para hacer predicciones. Cada árbol realiza una predicción diferente para una nueva observación. La clase con la mayoría de los votos de todos los árboles se convierte en la predicción final del modelo.

***Figura 6: Método de clasificación - Random Forest***




Fuente: Extraído de Sanchez (2021)

- Consideraciones adicionales

Random Forest ayuda a reducir la correlación entre los árboles inyectando más aleatoriedad en el proceso de estructuración de los árboles individuales (Boehmke y Greenwell ,2020). Esta aleatoriedad se introduce al dividir los subconjuntos de muestras y al seleccionar aleatoriamente un subconjunto de características sobre las cuales se evalúa la impureza de los nodos.

En cuanto a su interpretabilidad, aunque los árboles de decisión individuales son fáciles de interpretar, el conjunto de árboles en un Random Forest es más complejo. En este nivel de algoritmo, se empieza a manifestar el fenómeno conocido como "caja negra" en los algoritmos de aprendizaje automático, donde el proceso de selección del algoritmo se vuelve más difícil de discernir.

### 2.2.4. Gradient Boosting

El funcionamiento del algoritmo Gradient Boosting presentado por Friedman, J.H. (2001), consiste en una serie de pasos secuenciales donde cada modelo mejora la predicción del modelo anterior, centrándose en los errores cometidos por el modelo predecesor.

#### 2.2.4.1. Funcionamiento del algoritmo Gradient Boosting

La secuencia comienza con un modelo débil, cuya tasa de error es solo ligeramente mejor que el azar. Un árbol de decisión poco profundo representa un aprendiz débil. Después de construir el primer árbol, se calcularán los residuos para cada observación, los cuales son la diferencia entre los valores reales y las predicciones del modelo actual.

El nuevo árbol se entrenará utilizando todas las observaciones del conjunto de datos original, pero el objetivo se centrará en predecir los residuos del modelo anterior, enfocándose en corregir los errores que el modelo anterior cometió. Una vez que el nuevo árbol está entrenado para predecir los residuos, sus predicciones se agregarán a las predicciones del modelo existente para mejorar la precisión global del modelo.

- Función objetivo – minimización de residuos

Está secuencia se representa de la siguiente forma (Boehmke y Greenwell ,2020):

1.Ajustar el árbol de decisión al conjunto de datos: $F_1(x) = y$

2. Ajustar el siguiente árbol de decisión a los residuos: $h_1(x) = y - F_1(X)$.

3.Agregar este nuevo árbol al algortimo anterior: $F_2(X) = F_1(X) + h_1(X)$,

4.Ajustar el siguiente árbol de decisión a los residuos $F_2 : h_2(x) = y - F_2(X)$.

5.Agregar este nuevo árbol al algortimo: $F_3(X) = F_2(X) + h_2(X)$,

6.Este proceso continuara hasta que algún mecanismo indique la detención.

El modelo final es un modelo aditivo por etapas de $b$ arboles individuales:

$$F(X) = \sum_{b=1}^{B} F^b(X) \tag{12}$$

- Consideraciones adicionales

El algoritmo GBoost es una de las técnicas más utilizadas y preferidas para problemas de regresión y clasificación. A partir del enfoque básico de GBoost, se han desarrollado modelos más avanzados. Sin embargo, al igual que otros modelos, el aumento de la complejidad puede incrementar los costos computacionales y dificultar la explicación de sus decisiones.

## 2.2 Revisión de la literatura

El uso de técnicas de aprendizaje automático en la evaluación del riesgo crediticio ha suscitado un creciente interés entre los investigadores debido al notable avance de los modelos en la última década (Assef y Steiner, 2020). Diversos estudios han explorado las relaciones y avances en este campo. Uno de los estudios más destacados es el de Lessmann, Baesens, Seow y Thomas (2015), titulado "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research", en el cual se analizan 48

investigaciones relacionadas con la creación de algoritmos de clasificación enfocados en el riesgo crediticio, realizadas entre 2003 y 2014.

- Hallazgos clave del Estudio de Lessmann et al. (2015)

Entre los hallazgos más relevantes de este estudio se destaca que la mayoría de las investigaciones utilizan un número reducido de conjuntos de datos, en promedio 1.9. Los autores argumentan que esto es inadecuado, dado que los conjuntos de datos del mundo real suelen ser grandes y de alta dimensión. Además, se señala que la mayoría de las investigaciones se basan en una única medida de desempeño, lo cual también resulta inapropiado debido a que diferentes indicadores incorporan nociones distintas del desempeño del modelo.

En este mismo estudio se lleva a cabo un análisis comparativo del rendimiento en precisión de 41 modelos, utilizando el área bajo la curva (AUC) y otros indicadores de desempeño. El mejor clasificador individual resultó ser las Redes Neuronales Artificiales (ANN); sin embargo, se evidencia que no todos los métodos sofisticados mejoran la precisión. Por ejemplo, los clasificadores extendidos como las Extreme Learning Machines (ELMs) y Rotation Forest (RotFor) no superan a sus predecesores, y algoritmos de selección dinámica de conjuntos, aunque más complejos, no predicen tan eficazmente como alternativas más simples, como la regresión logística.

- Estudios recientes y avances

Otros estudios más recientes también aportan información relevante. Por ejemplo, Li, Y.; Chen, W. (2020). comparan 10 modelos de ensamble, concluyendo que el aprendizaje en conjunto ha mejorado la precisión en términos generales, aunque técnicas como las Redes Neuronales (NN) y las Máquinas de Vectores de Soporte (SVM) requieren un proceso de entrenamiento prolongado, especialmente para conjuntos de datos grandes.

Alonso-Robisco y Carbó (2020) discuten las oportunidades y riesgos asociados con el uso del aprendizaje automático en la concesión de créditos, destacando la importancia de la

interpretabilidad de las decisiones crediticias y los problemas de sesgos discriminatorios relacionados con los modelos de aprendizaje automático. En su estudio, evaluaron varios modelos, concluyendo que el modelo XGBoost mostró la mejor precisión.

Alonso-Robisco y Carbó (2022) investigan si los modelos de aprendizaje automático pueden ayudar a los bancos a ahorrar capital, utilizando evidencia de una cartera de crédito española. Se evaluó que el modelo de aprendizaje automático XGBoost logra un ahorro del 17% con respecto a un método tradicional logit.

En resumen, el uso de técnicas de aprendizaje automático para la evaluación del riesgo crediticio ha mostrado avances significativos, aunque también presenta desafíos importantes. La adecuada selección y validación de modelos, el uso de múltiples conjuntos de datos y medidas de desempeño, así como la interpretabilidad y los sesgos en las decisiones crediticias. Incluir también la necesidad de explorar en las investigaciones mayores beneficios que conlleva una buena predicción de la probabilidad de incumplimiento en la gestión de riesgo de crédito.

# 3. Metodología

En esta parte del estudio, se abordará un caso práctico utilizando una base de datos de préstamos. Se seguirá el proceso de construcción y validación de modelos de aprendizaje automático discutidos en el apartado anterior. El objetivo es aplicar la base de datos a los modelos para evaluar sus capacidades y beneficios en la determinación de la probabilidad de incumplimiento.

## 3.1. Descripción de la base de datos

Se usará la base de datos "Give Me Some Credit" de la plataforma Kaggle.com. Esta base de datos es parte de los concursos de predicción de probabilidad de incumplimiento que se organizan en la plataforma. El conjunto de datos incluye un total de 150,000 préstamos etiquetados con una variable objetivo de incumplimiento y diez variables explicativas asociadas a cada préstamo. Los detalles de las variables se indican en la Tabla 3.

***Tabla 3: Variables del modelo***

| *Nombre de Variable* | *Descripción* | *Tipo de variable* |
|---|---|---|
| **Default**[2] | Indica si el prestatario experimentó una morosidad de 90 días o más en los últimos dos años, si sucede se consideró Default. | Binaria (Si=1, No=0) |
| **Revolving** | Deuda total en tarjetas de crédito y líneas de créditos personales, excepto bienes raíces y deuda a plazos, dividido por la suma de los límites de crédito. | Porcentaje (Se expresa como decimales) |
| **DebtRatio** | La relación entre la suma de los pagos mensuales de deudas, pensión alimenticia y costos de vida entre el ingreso bruto mensual. | Porcentaje (Se expresa como decimales) |
| **MonthlyIncome** | Ingreso mensual del prestatario. | Real (incluye decimales) |
| **CreditLines** | Números de préstamos y líneas de crédito abiertas | Entero |
| **RealEstate** | Número de préstamos hipotecarios y de bienes raíces, incluidas las líneas de crédito sobre el valor de la vivienda. | Entero |
| **Dependents** | Número de dependientes en la familia, excluyendo al propio prestatario. | Entero |
| **Age** | Edad del prestatario en años | Entero |
| **30-59Days** | Número de veces que el prestatario ha tenido un retraso de 30-59 días, pero no más, en los últimos dos años | Entero |
| **60-89Days** | Número de veces que el prestatario ha tenido un retraso de 60-89 días, pero no más, en los últimos dos años. | Entero |

[2] Default = 1 - clase positiva, No Default = 0 – clase negativa

| | | |
|---|---|---|
| **90Days** | Número de veces que el prestatario ha tenido un retraso de 90 días o más. | Entero |

Fuente: Elaboración propia basado en Kaggle Competition. (2011)

## 3.2. Análisis Exploratorio y Preprocesamiento de datos

En este apartado se analizarán los datos según sus características originales y se describirá el tratamiento aplicado para lograr una base de datos coherente y lógica. Las principales métricas de las variables se presentan en la tabla 4.

***Tabla 4: Descripción de variables***

| *Variable* | *n* | *mean* | *median* | *sd* | *min* | *max* | *skew* | *kurtosis* |
|---|---|---|---|---|---|---|---|---|
| ***Default*** | 150000 | 0.067 | 0.000 | 0.250 | 0.000 | 1.000 | 3.469 | 10.033 |
| ***Revolving*** | 150000 | 6.048 | 0.154 | 249.755 | 0.000 | 50708.000 | 97.630 | 14544.035 |
| ***Age*** | 150000 | 52.295 | 52.000 | 14.772 | 0.000 | 109.000 | 0.189 | -0.495 |
| ***30-59Days*** | 150000 | 0.421 | 0.000 | 4.193 | 0.000 | 98.000 | 22.597 | 522.352 |
| ***DebtRatio*** | 150000 | 353.005 | 0.367 | 2037.819 | 0.000 | 329664.000 | 95.156 | 13733.648 |
| ***MonthlyIncome*** | 120269 | 6670.221 | 5400.000 | 14384.674 | 0.000 | 3008750.000 | 114.037 | 19503.570 |
| ***CreditLines*** | 150000 | 8.453 | 8.000 | 5.146 | 0.000 | 58.000 | 1.215 | 3.091 |
| ***90Days*** | 150000 | 0.266 | 0.000 | 4.169 | 0.000 | 98.000 | 23.087 | 537.714 |
| ***RealEstate*** | 150000 | 1.018 | 1.000 | 1.130 | 0.000 | 54.000 | 3.482 | 60.474 |
| ***60-89Days*** | 150000 | 0.240 | 0.000 | 4.155 | 0.000 | 98.000 | 23.331 | 545.657 |
| ***Dependents*** | 146076 | 0.757 | 0.000 | 1.115 | 0.000 | 20.000 | 1.588 | 3.001 |

Fuente: Elaboración propia

La variable "Default" muestra una media y mediana reducidas, indicando que la mayoría de los prestatarios en el conjunto de datos no han incumplido sus pagos. En efecto, la proporción es del 93.3% para "No Default" (0) y del 6.7% para "Default" (1), evidenciando un desbalance significativo en los datos.

*Tratamiento: Se aplicará la técnica SMOTE (Synthetic Minority Over-samplig Technique) permite generar ejemplos sintéticos de la clase minoritaria a partir de identificar observaciones con características muy similares. (vecinos cercanos)*

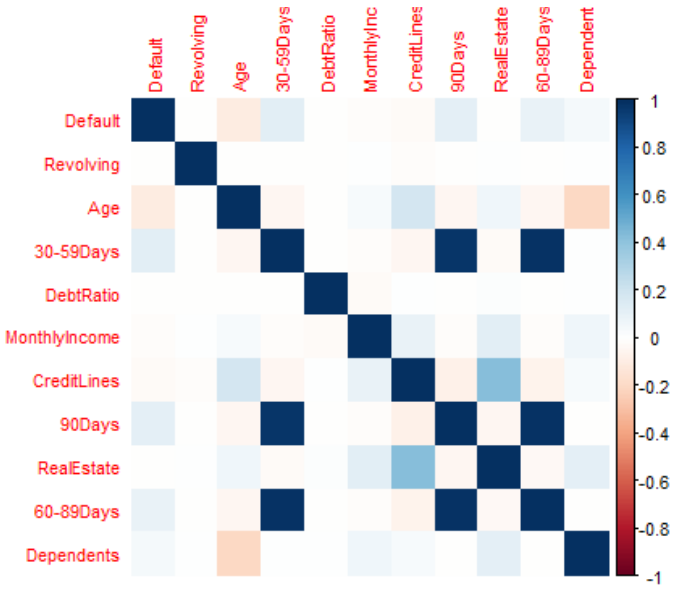
Las variables "Revolving" y "DebtRatio" presentan valores atípicos extremos. Los valores máximos en estas variables sugieren la presencia de datos erróneos que difieren

considerablemente del promedio, incluso escapando de la lógica económica (véase tabla 4). Esto se confirma con los altos valores en los indicadores de asimetría y curtosis.

*Tratamiento: Se mantendrá observaciones con valores de DebtRatio <= 100 y Revolving <= 1, datos por encima del mismo se consideran erróneos y se eliminarán al no poder un prestatario gastar más de 100 veces sus ingresos y además no poder gastar más de lo que su límite de línea de crédito le permita.*

Las variables "Days30-59", "Days60-89" y "Days90" también presentan valores atípicos considerando la gran diferencia de su media con respecto a su valor máximo. En estas tres variables se detecta un valor de 98, lo cual sugiere que el prestatario se atrasó 98 veces, lo que es incongruente dado que el análisis abarca solo 2 años (24 meses). Esto indica la presencia de datos erróneos. Por otro lado, estas tres variables presentan alta correlación. (véase figura 7).

***Figura 7: Correlación de variables***




*Fuente: Elaboración propia*

*Tratamiento: Se mantendrá observaciones con valores de esas tres variables menores o iguales a 24, ya que como se menciona, esta solo puede llegar al límite de 24 en 2 años. Por otro lado, con respecto a la correlación de las mismas se creará dos nuevas variables, la*

*primera **Late _N** que considere el número máximo de veces que el prestatario se atrasó en su pago entre las variables "Days30-59", "Days60-89" y "Days90", y segundo se creará la variable "**Late_Dummy**", una variable categórica donde se considere 1 si el cliente presentó atraso y 0 si no lo hizo. Al crear estas variables que recogen el efecto de "Days30-59", "Days60-89" y "Days90", estas se eliminarán del conjunto de datos, corrigiendo el problema de correlación.*

Las variables "Age", "CreditLines" y "RealEstate" presentan indicadores bastante centrados alrededor de la media y la mediana, aunque aún muestran ligeras asimetrías. En el caso de "Age", la edad mínima registrada es cero, lo cual es claramente erróneo.

*Tratamiento: Se eliminará aquella observación donde la variable Age es igual a cero al ser una incongruencia.*

Finalmente, las variables "MonthlyIncome" y "Dependents" tienen menos observaciones (120,269 y 146,076 respectivamente) comparadas con otras variables (150,000), lo que indica la presencia de valores nulos.

*Tratamiento: Se evidenció que en el 90 % de las observaciones donde la variable MonthlyIncome es nula, DebtRatio presentaba valores superiores a 10, llegando hasta 329,000. Esto sugiere sobreendeudamientos extremos y no viables económicamente. Además, se observó que las observaciones con valores nulos en Dependent también presentaban valores nulos en MonthlyIncome. Ante esta situación, se decidió eliminar todas las observaciones donde MonthyIncome presentaba valores nulos para asegurar la calidad y coherencia de los datos.*

### 3.1.1 Uso de técnica Binning y Weigth of Evidende (WoE)

A pesar de la eliminación de inconsistencias en los datos, este aún presenta dos grandes problemas, presencia de valores atípicos extremos y alta dispersión, por lo que es muy conveniente realizar una transformación de los mismos, se usará Bining y WoE. Estas técnicas existen en el mundo de la calificación crediticia y se han usado para explorar datos y filtrar

variables en proyectos de modelado de riesgo crediticio, como la probabilidad de default (Bhalla, D. 2015); consisten en agrupar los datos de la variable en intervalos (bins) en lugar de considerar cada valor individualmente, para cada intervalo, se calcula el WoE usando la siguiente fórmula:

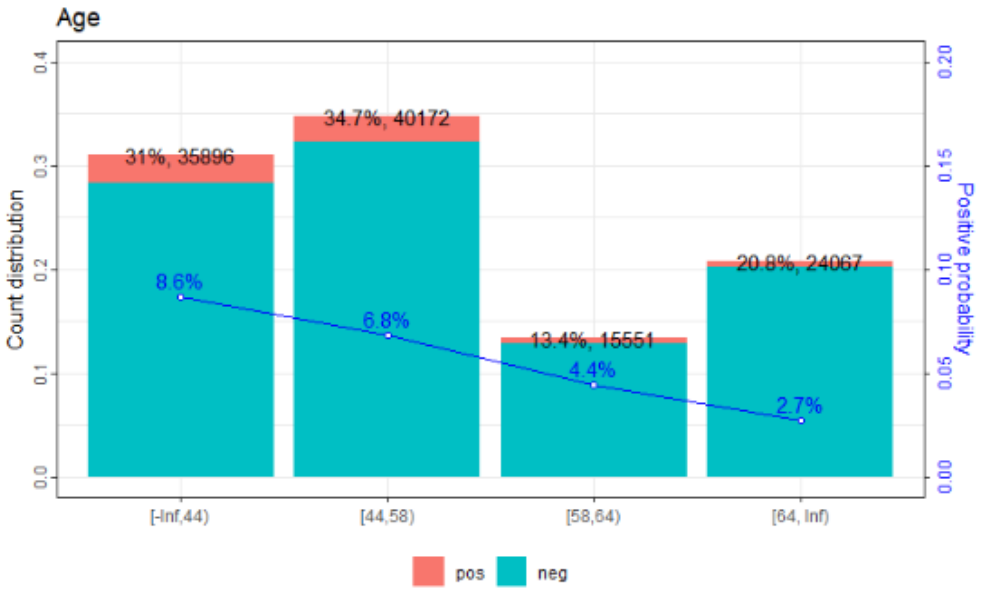
$$WoE = \ln\left(\frac{Distribucion\ de\ no\ incumplimiento}{Distribución\ de\ incumplimiento}\right) \quad (13)$$

Donde la distribución de no incumplimiento es la proporción de observaciones no incumplidas en el intervalo, y la distribución de incumplimiento es la proporción de observaciones incumplidas en el intervalo.

El proceso de división de binning intenta garantizar que el WoE de los bins sea monotónico, es decir los que los valores de WoE aumenten o disminuyan de manera ordenada, siendo crucial para la interpretación de variables. Además, se asegura de que cada bin contenga un número mínimo de observaciones para evitar problemas de sobreajuste, por último, se garantiza a través de pruebas de chi- cuadrado para combinar bins hasta que todos los bins resultantes sean estadísticamente diferente entre sí respecto a la variable objetivo. Con esta transformación los datos continuos pasaran a convertirse en intervalos categóricos, garantizando:

- Reducción de la varianza, al agrupar los datos en intervalos en lugar de considerar cada valor individualmente.
- Manejo de Valores Atípicos, al agrupar los valores atípicos dentro de un intervalo específico, reduciendo su impacto.

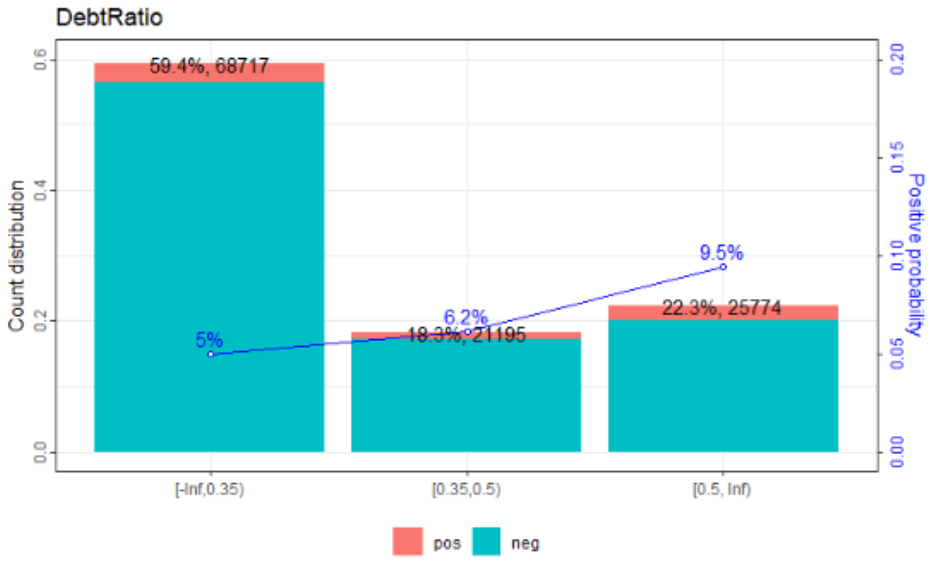
**Figura 8: Binning y WoE en Age**




Fuente: Elaboración propia

La variable Age se ha dividido en cuatro intervalos, se observa que los intervalos de menor edad presentan una mayor probabilidad de incumplimiento. Específicamente clientes más jóvenes muestran tasas de incumplimiento más altas, a medida que la edad de los clientes aumenta a intervalos mayores, la probabilidad de incumplimiento disminuye.

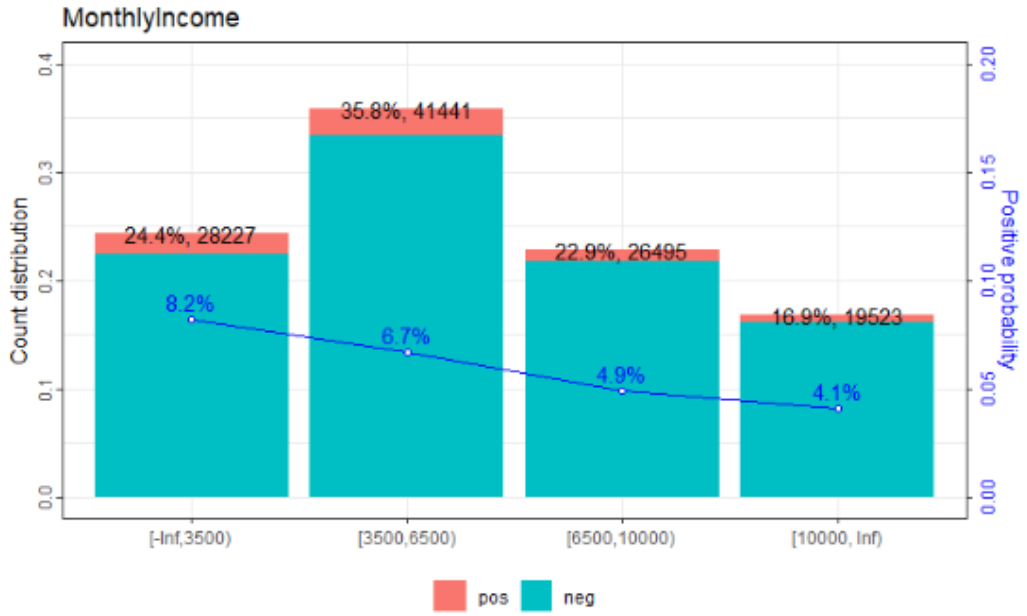
**Figura 9: Binning y WoE en DebtRatio**




Fuente: Elaboración propia

La variable DebtRatio se ha dividido en tres intervalos, se observa que los intervalos donde la ratio es menor hay menos probabilidad de incumpliendo, esto sugiere que los individuos con menor carga de deuda en relación con sus ingresos son menos propensos a incumplir con sus obligaciones, a medida que la ratio aumenta la probabilidad de incumplimiento aumentan significativamente.

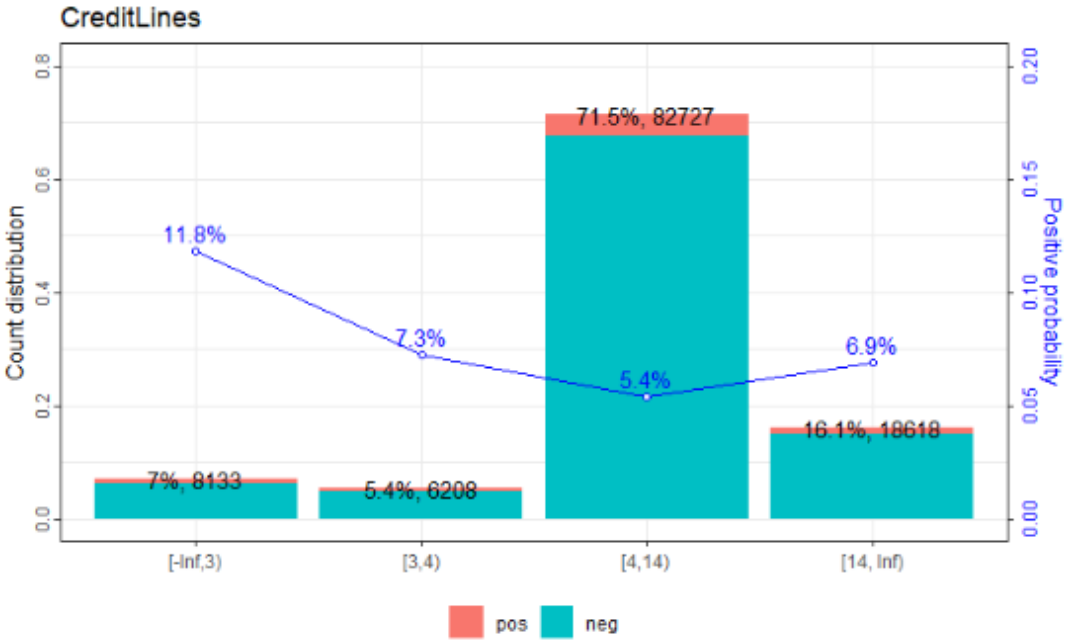
**Figura 10: Binning y Woe en MonthlyIncome**




Fuente: Elaboración propia

La variable MothlyIncome se ha dividido en cuatro intervalos, se observa que los intervalos donde los ingresos mensuales son menores se presenta mayor probabilidad incumplimiento, la probabilidad de incumplimiento va reduciéndose a medida que aumentan los ingresos en intervalos mayores.

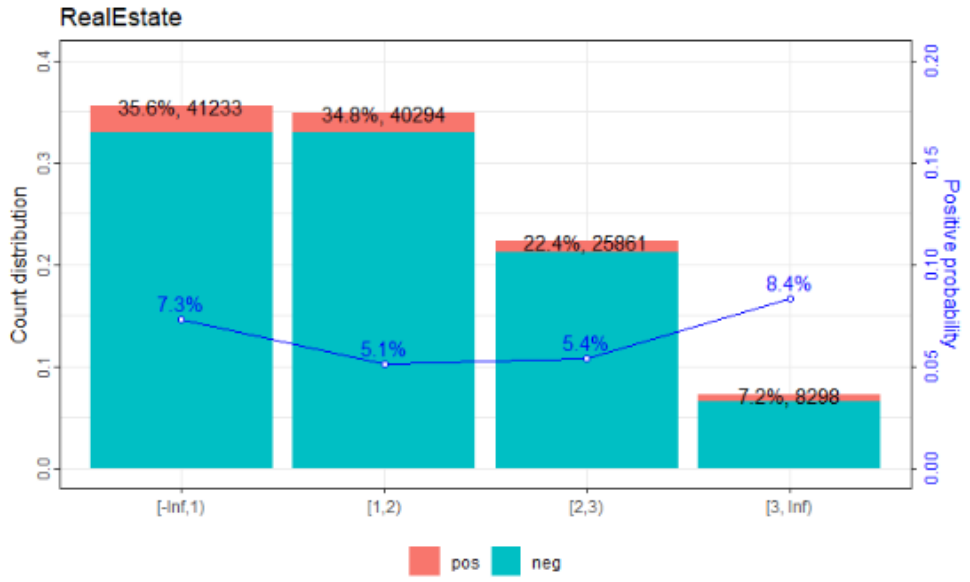
**Figura 11: Binning y Woe en CreditLines**




Fuente: Elaboración propia

La variable CreditLines se ha dividido en cuatro intervalos, sus resultados muestran cierta particularidad pues se observa que a medida que las líneas de crédito que posee un individuo aumentan también se reduce la probabilidad de incumplimiento, lo cual es una particularidad de los datos dado que la lógica económica debería indicar lo contrario, sin embargo el incremento de la probabilidad de incumplimiento se puede observar en los dos últimos intervalos, justamente donde el número de líneas de crédito es mayor.

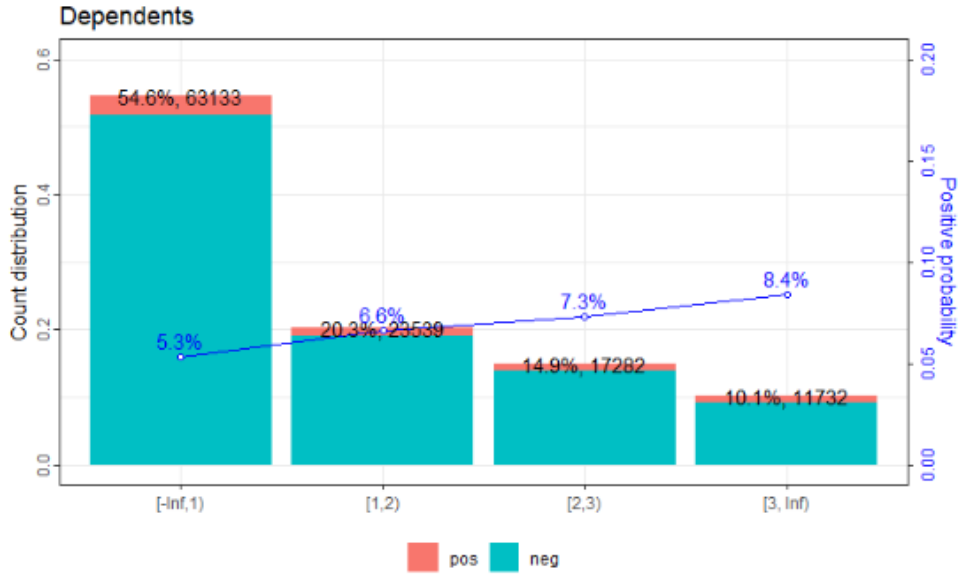
**Figura 12: Binning y WoE en RealEstate**




Fuente: Elaboración propia

La variable RealEstate se ha dividido en cuatro intervalos, se observa que los intervalos donde un individuo presente más préstamos hipotecarios o de bienes raíces presentan también mayor probabilidad de incumplimiento, nuevamente mayor carga de deuda conlleva mayor riesgo de incumplir con sus obligaciones.

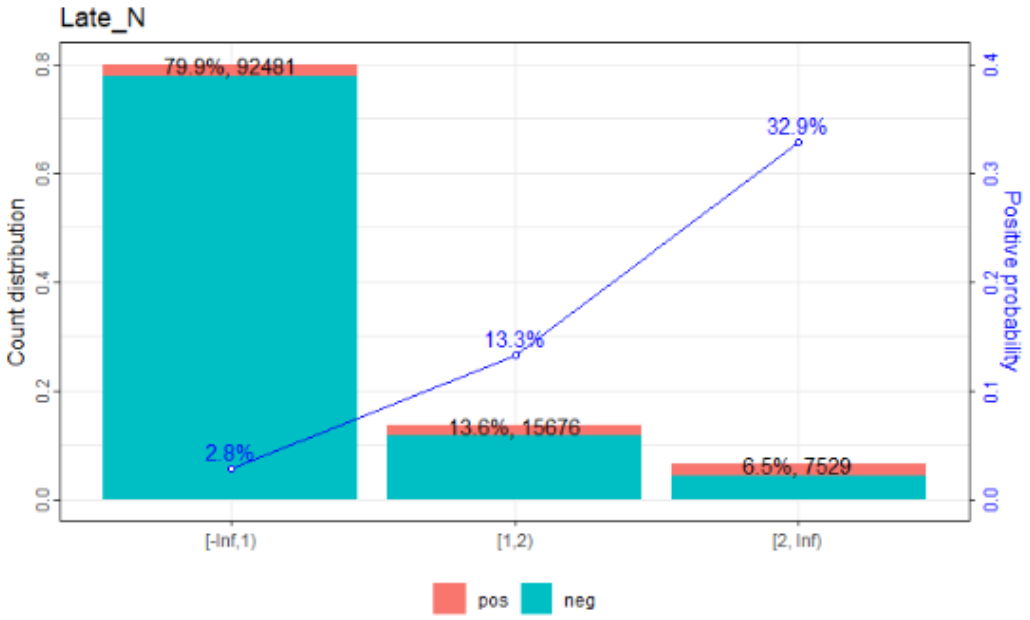
**Figura 13: Binning y Woe en Dependents**




Fuente: Elaboración propia

La variable Dependents se ha dividido en cuatro intervalos, se observa que los intervalos de mayor número de dependientes presentan mayor probabilidad de incumplimiento, esto sugiere que los individuos con mayor carga de personas dependientes presentarían problemas en cumplir con sus obligaciones.

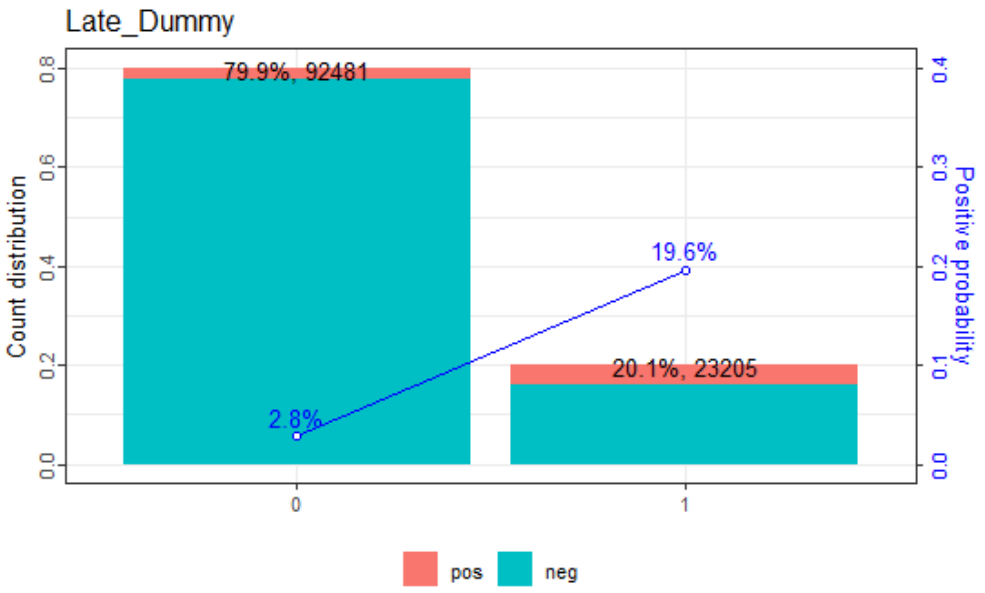
**Figura 14: Binning y WoE en Late_N**




Fuente: Elaboración propia

La variable Late_N se ha dividido en tres intervalos, se observa que los intervalos donde el individuo ha presentado más veces atrasos poseen mayor probabilidad de incumplimiento. La diferencia de probabilidades de incumplimiento entre quienes han tenido atraso y quienes no, es muy significativa.

**Figura 15: Binning y WoE en Late_Dummy**




Fuente: Elaboración propia

La variable Late_Dummy que indica si un individuo se atrasó o no en sus pagos, nos dice que si un individuo se atrasó la probabilidad de incumplimiento es mayor.

- Perfil de individuo con altas probabilidades de incumplimiento.

Luego del análisis de transformación que se realizó a cada variable se puede obtener un perfil del cliente que más probablemente incumpla con sus pagos. Este será un individuo joven que tenga un alto ratio de deuda, bajos ingresos, varios préstamos hipotecarios, muchos dependientes y un historial de retraso en sus pagos.

## 3.3. División de datos

Luego de la limpieza y transformación inicial, el conjunto total de datos restantes de *Give me some credit* (GMC) se divide en conjuntos de entrenamiento y prueba, para el caso de este estudio los datos de entrenamiento será el 80% y los datos de prueba será el 20%.

La división se realiza bajo un proceso aleatorio, sin embargo, se estratifica según la proporción de la variable objetivo Default. En nuestro caso se dividió según la tabla 6.

**Tabla 6: Distribución Entrenamiento - Prueba**

| | |
|---|---|
| Total datos GMC<br>115,686 observaciones | GMC_Entrenamiento<br>92,548 observaciones |
| | GMC_Prueba<br>23,138 observaciones |

Fuente: Elaboración propia

Después de la división de datos, los datos de entrenamiento, que fueron categorizadas en intervalos mediante las técnicas de Binning y WoE, se transformaron en variables numéricas utilizando variables dummy. Además, para abordar el desbalanceo de clases, se aplicó la técnica de sobremuestreo sintético (SMOTE).

## 3.4. Aplicación de modelos

Siguiendo con el proceso de construcción y validación del modelo, en este apartado se mostrará los resultados de forma individual de cada modelo propuesto. Todos los modelos fueron entrenados con los datos de entrenamiento aplicando las técnicas particulares de cada uno expuestos en el marco teórico, además siguiendo con el proceso establecido fueron validados con los datos de prueba para medir sus resultados.

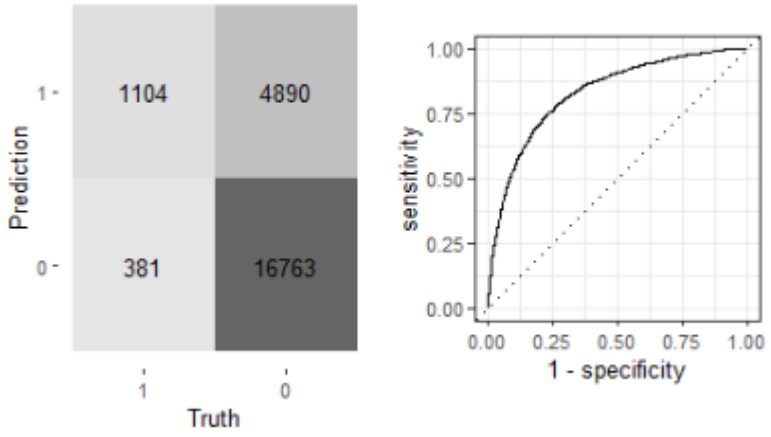
Cada modelo usó un proceso de validación cruzada con tres pliegues(folds) asegurando que las proporciones de la variable objetivo Default se mantengan en cada partición, además se usó un proceso iterativo para hallar los hiperparámetros del modelo que maximicen el ROC-AUC. Todo esto se ejecutó en los modelos de aprendizaje automático (árbol de decisión, random forest y gradient boosting), no se ejecutó en el modelo logit, el cual ha sido ejecutado bajo su estado tradicional para fines comparativos.

### 3.4.1. Aplicación Modelo Logit

El modelo logit se entrenó en su forma tradicional, buscando la curva sigmoide que mejor se ajuste a los datos de entrenamiento mediante las técnicas estadísticas explicadas en el

marco teórico. En este caso, no se incluyó la búsqueda de hiperparámetros óptimos ni se realizó una validación cruzada, dado que el objetivo era evaluar el rendimiento básico del modelo sin optimización adicional. Los resultados del modelo en los datos de prueba, incluyendo métricas de desempeño y curvas ROC, se presentan en la fig. (16)

**Figura 16: Matriz de confusión - Curva Roc – Métricas (Modelo Logit)**



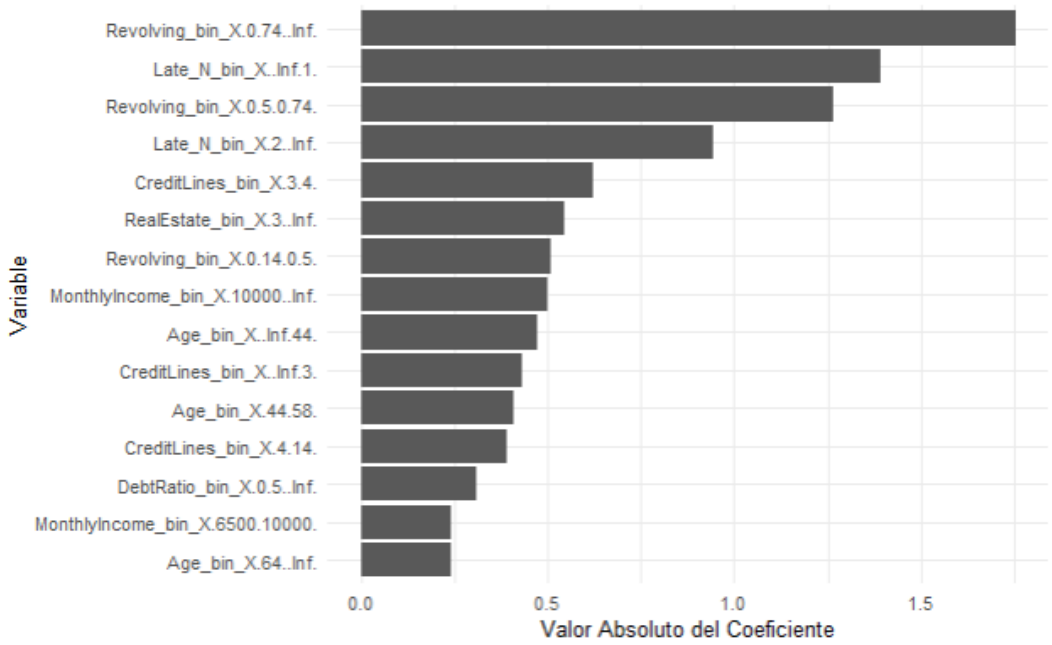

| .metric | .estimator | .estimate | model |
|---|---|---|---|
| accuracy | binary | 0.77219 | Simple LR - Base |
| roc_auc | binary | 0.83014 | Simple LR - Base |
| sensitivity | binary | 0.74343 | Simple LR - Base |
| specificity | binary | 0.77417 | Simple LR - Base |
| f_meas | binary | 0.29523 | Simple LR - Base |

Fuente: Elaboración propia

En la evaluación del desempeño del modelo logit, se destaca que el 77.22% de sus predicciones fueron correctas (exactitud). Además, su capacidad para detectar correctamente los casos de incumplimiento alcanzó el 74.34%(sensibilidad). Finalmente, el modelo mostró una sólida capacidad para distinguir entre clases, logrando una métrica de 0.83 en el área bajo la curva ROC(AUC-ROC). En general, el modelo se benefició significativamente de la transformación de datos utilizando la técnica de binning y WoE.

Las variables más significativas del modelo se han determinado en base a sus coeficientes estimados para cada variable y se muestra en la fig. (17).

**Figura 17: Top 15 Variables por Importancia (Valor Absoluto del Coeficiente)**




Fuente: Elaboración propia

Entre las variables más significativas para explicar las probabilidades de incumplimiento en el modelo logit resaltan las variables Revolving y Late_N. Para la variable Revolving, los intervalos los intervalos [0.5 – 0.74> y [0.74 – Inf> son especialmente importante, lo cual sugiere que niveles de endeudamiento por encima de la mitad de la línea de crédito son indicadores muy significativos de incumplimiento. Por otro lado en Late_N sobresalen los intervalos [Inf – 1> y [2, Inf> , lo cual indica que el récord histórico de pagos es un factor crucial para la determinación de la probabilidad de incumplimiento.

### 3.4.2. Aplicación Modelo árbol de decisión

El proceso de ajuste del modelo de árbol de decisión comenzó con la búsqueda de los hiperparámetros óptimos para mejorar su rendimiento. Se realizó una búsqueda en cuadricula (grid search) sobre los hiperparámetros: Cost Complexity (regula el proceso de poda) en el rango [-5,0], Tree Depth (regula la profundidad del árbol) en el rango [1-15] y min_n (regula el número mínimo de observaciones en un nodo para la determinar la clasificación) en el rango [20-150]. Además, se generaron 3 niveles equidistantes dentro de cada rango, resultando en 27 combinaciones posibles de hiperparámetros.

Los hiperparámetros óptimos que maximizaron la métrica ROC-AUC en el modelo fueron los siguientes:

**Tabla 5: Hiperparámetros - árbol de decisión**

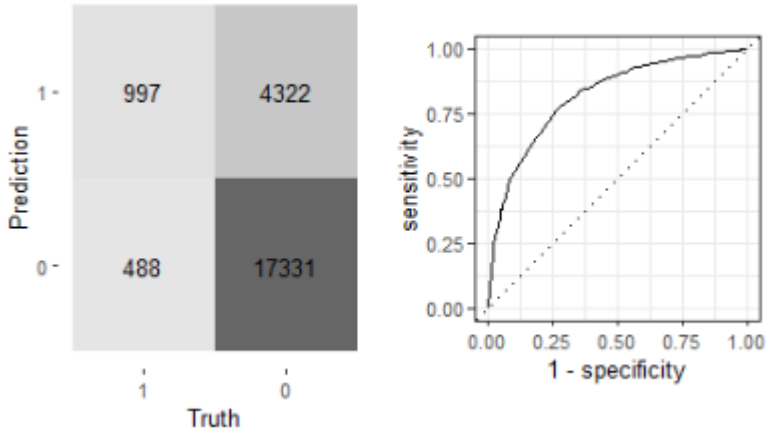
| *Cost_complexity* | *Tree_depth* | *Min_n* | *Mean Roc-Auc* |
|---|---|---|---|
| 0.000010000 | 8 | 150 | 0.815 |

Fuente: Elaboración propia

Con los hiperparámetros óptimos, se procedió a ajustar el modelo de árbol de decisión utilizando el conjunto de datos de entrenamiento y una validación cruzada con 3 folds.

Posteriormente, el modelo ajustado se validó con los datos de prueba, los resultados se muestran en la siguiente fig. (18).

**Figura 18: Matriz de confusión–Curva Roc–Métricas (Modelo árbol de decisión)**




| .metric | .estimator | .estimate | model |
|---|---|---|---|
| accuracy | binary | 0.79212 | Arbol de desicion |
| roc_auc | binary | 0.81862 | Arbol de desicion |
| sensitivity | binary | 0.67138 | Arbol de desicion |
| specificity | binary | 0.80040 | Arbol de desicion |
| f_meas | binary | 0.29306 | Arbol de desicion |

Fuente: Elaboración propia

En la evaluación del modelo de árbol de decisión se obtuvieron las siguientes métricas de desempeño. Destaca la exactitud (accuracy) del modelo la cual es de 0.79212, indicando

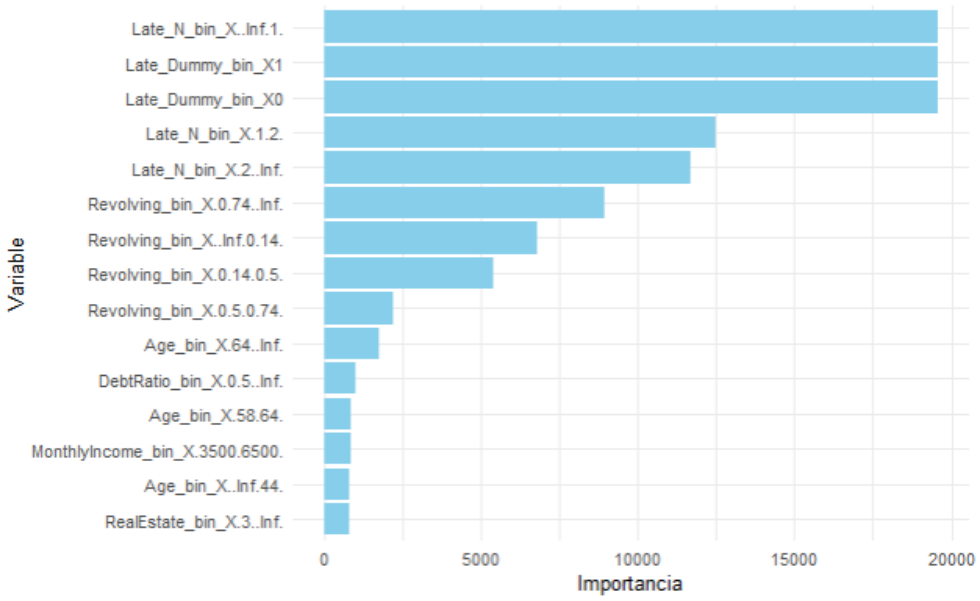
que el modelo clasifica correctamente el 79.21% de las instancias, el área bajo la curva ROC (roc_auc) es de 0.81862, mostrando una buena capacidad para distinguir entre clases, la especificidad (specificity) es de 0.80040 indicando que el 80.04% de los individuos que cumplieron son correctamente identificados.

Por otro lado, se observa debilidades en cuanto a las métricas de sensibilidad y f-score. La sensibilidad (sensibility) es de 0.67138, lo que refleja que el modelo identifica correctamente el 67.14% de las instancias donde el individuo incumple. La medida F (f_meas) es de 0.29306 sugiriendo que le balance entre precisión y sensibilidad no es el óptimo.

En resumen, el modelo de árbol se decisión tiene un desempeño aceptable en precisión y capacidad discriminativa, pero requiere mejoras en sensibilidad y balance general.

Al respecto de las variables más importantes consideradas por el modelo para determinar la clasificación de individuos se puede ver la fig. (19).

**Figura 19: Top 15 variables de importancia - Árbol de decisión**




Fuente: Elaboración propia

La importancia se mide sumando la reducción de impureza (Gini) que la característica aporta en las divisiones en las que participa en el árbol. En algunos casos, una variable puede utilizarse más de una vez, lo que incrementa su aporte y, por lo tanto, su importancia. En este

contexto, el modelo considera que la variable de mayor importancia a Late_N [inf.1.>, es decir, que un cliente no haya tenido atrasos en sus pagos anteriormente se considera que es la característica de mayor importancia para clasificarlo como pagador o no. A continuación, se observa que los cuatros variables siguientes en importancia también son de tipo Late_N y Late_Dummy. Esto resalta el historial de pagos del cliente para determinar la probabilidad de incumplimiento. Además, al igual que el modelo logit, la variable Revolving en sus cuatro intervalos también muestra una significativa relevancia.

### 3.4.3. Aplicación modelo Random Forest

Al igual que el modelo anterior el proceso de ajuste del modelo comenzó con la búsqueda de los hiperparámetros óptimos para mejorar su rendimiento. Se realizó una búsqueda en cuadriculas sobre los hiperparámetros: Mtry (regula el número de variables a considerar en cada árbol) en el rango (3,10), Min_n (regula el número mínimo de observaciones en un nodo para la determinar la clasificación) en el rango (5,200) y Trees (regula el número de árboles en el modelo) en el rango (50,200), Además se generaron 3 niveles equidistantes dentro de cada rango, resultando en 27 combinaciones posibles de hiperparámetros.

Los hiperparámetros óptimos que maximizan la métrica ROC-AUC en el modelo fueron los siguientes:

**Tabla 6: Hiperparámetros - Random Forest**

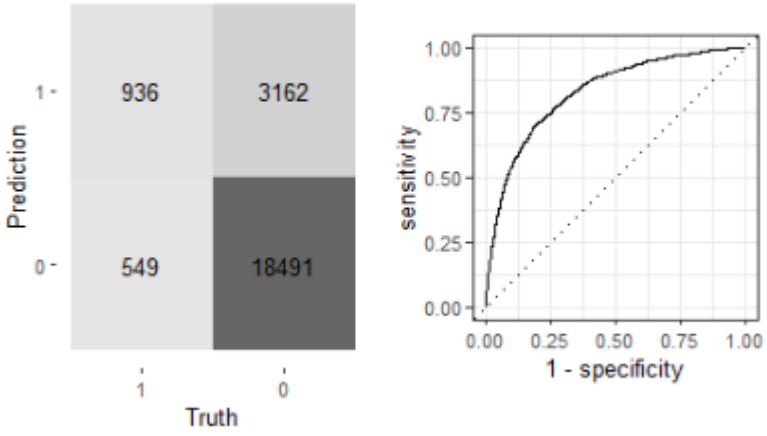
| ***Mtry*** | ***Trees*** | ***Min_n*** | ***Mean Roc-AUC*** |
|---|---|---|---|
| **3** | 200 | 200 | 0.8294 |

Fuente: Elaboración propia

Con los hiperparámetros óptimos, se procedió a ajustar el modelo random forest usando el conjunto de datos de entrenamiento y una validación cruzada con 3 folds.

Posteriormente el modelo ajustado se validó con los datos de prueba, los resultados se muestran en la fig. (20).

**Figura 20: Matriz de confusión - Curva Roc - Métricas (Random Forest)**



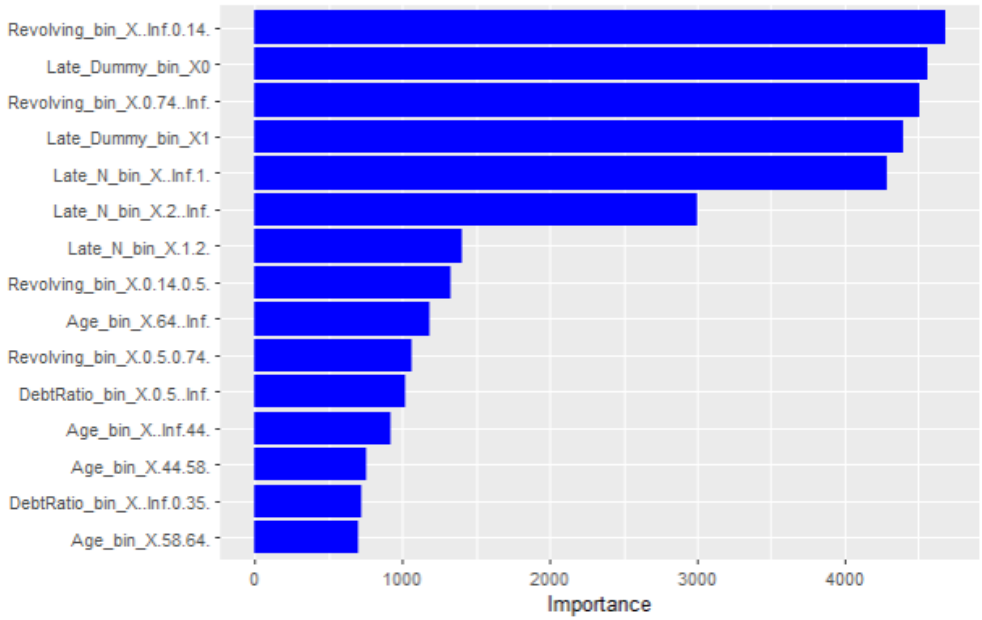

| .metric | .estimator | .estimate | model |
|---|---|---|---|
| accuracy | binary | 0.83961 | Random Forest |
| roc_auc | binary | 0.82841 | Random Forest |
| sensitivity | binary | 0.63030 | Random Forest |
| specificity | binary | 0.85397 | Random Forest |
| f_meas | binary | 0.33530 | Random Forest |

Fuente: Elaboración propia

En la evaluación del modelo de Random Forest se obtuvieron las siguientes métricas de desempeño. Destaca una exactitud (acurracy) del 83.96%, una buena capacidad para distinguir entre clases con un área bajo la curva ROC (roc_auc) de 0.82841, y una notable capacidad para distinguir las clases negativas con una especificidad (specificity) de 85.40%. Aunque supera al modelo de árbol de decisión en estas medidas, aún se requieren mejoras en la sensibilidad y en el balance general del modelo.

Respecto de las variables más importantes consideradas por el modelo para clasificar a los individuos, se puede observar en la fig. (21).

**Figura 21: Top 15 variables de importancia - Random Forest**




Fuente: Elaboración propia

Al igual que en modelo de árbol de decisión, en el modelo Random Forest la importancia se mide sumando la reducción de impureza (Gini) que cada característica aporta en las divisiones de cada árbol de decisión que tiene el modelo random forest. El modelo considera como la variable más importante a Revolving_bin_X [inf. 0.14>, es decir que un bajo nivel de endeudamiento con respecto a la línea de crédito o crédito personales de un individuo es la característica más relevante para determinar si incurrirá de default. En segundo lugar, está la variable Late_Dummy_binX0, indica que, si el individuo no ha tenido atrasos anteriormente, también es crucial para predecir si pagará o no. Al igual que los modelos anteriores, considera de suma importancia Revolving y las variables que miden el atraso como Late_N y Late Dummy.

### 3.4.4. Aplicación modelo Gradient Boosting

El ajuste del modelo comenzó con la búsqueda de los hiperparámetros óptimos para mejorar su rendimiento. Se realizó una búsqueda en cuadrículas sobre los hiperparámetros: Mtry en el rango (3,10), Min_n en el rango (5,100) y Trees en el rango (50,100). Al igual que

en los modelos anteriores se generaron 3 niveles equidistantes dentro de cada rango, resultando en 27 combinaciones posibles de hiperparámetros.

Los hiperparámetros óptimos que maximizaron la métrica ROC-AUC en el modelo fueron los siguientes:

**Tabla 7: Hiperparámetros - Gradient Boosting**

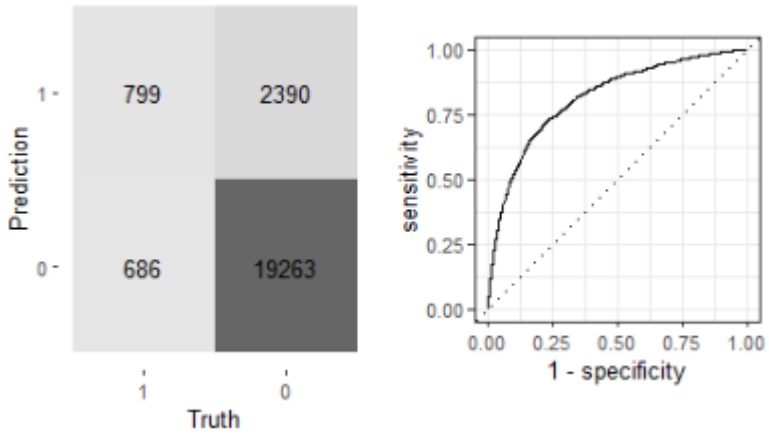
| *Mtry* | *Trees* | *Min_n* | *Mean ROC-AUC* |
|---|---|---|---|
| **3** | 50 | 100 | 0.817 |

Fuente: Elaboración propia

Con los hiperparámetros óptimos, se procedió a ajustar el modelo gradient boosting usando el conjunto de datos de entrenamiento y una validación cruzada con 3 folds.

Posteriormente el modelo ajustado se validó con los datos de prueba, los resultados se muestran en la siguiente fig. (22).

**Figura 22: Matriz de confusión - Curva Roc – Métricas (Gradient Boosting)**



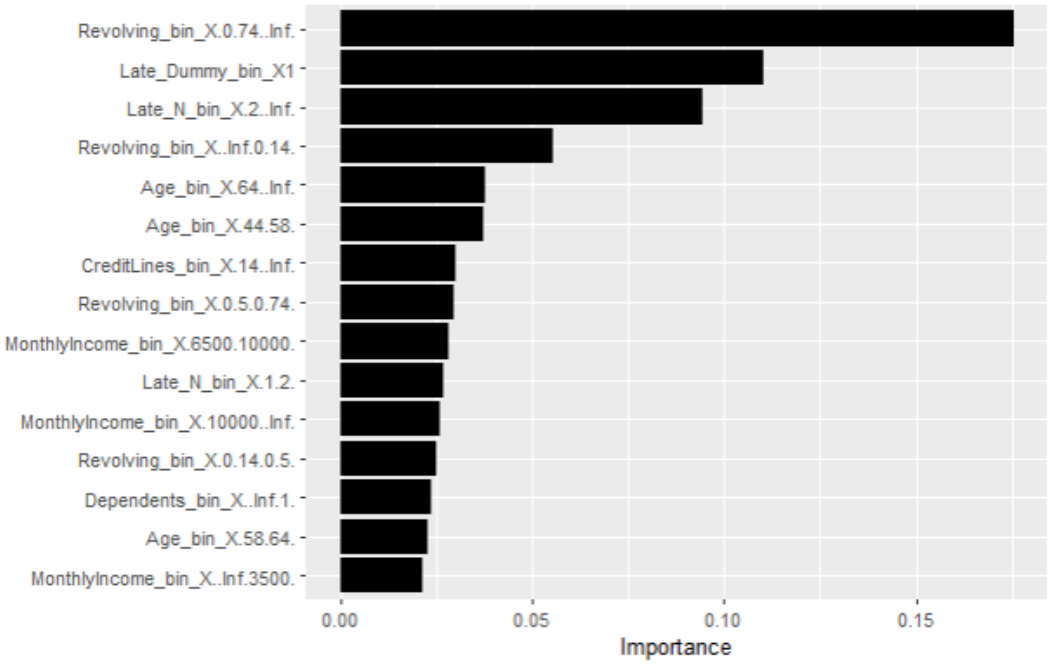

| .metric | .estimator | .estimate | model |
|---|---|---|---|
| accuracy | binary | 0.86706 | Boost tree |
| roc_auc | binary | 0.81557 | Boost tree |
| sensitivity | binary | 0.53805 | Boost tree |
| specificity | binary | 0.88962 | Boost tree |
| f_meas | binary | 0.34189 | Boost tree |

Fuente: Elaboración propia

En la evaluación del modelo gradient boosting se obtuvieron las siguientes métricas. Destaca un buen desempeño general con una exactitud (accuracy) del 86.71%. El área bajo la curva ROC (roc_auc) alcanzó 0.81557, lo que es un buen indicador de su habilidad para discriminar entre clases positivas y negativas, además destaca su especificidad la cual alcanzó el 88.96%. Así mismo se observó una mejora en comparación con otros modelos en cuanto a su balance general, con F-score de 0.34. Sin embargo, aún se requieren mejoras en la sensibilidad.

Respecto de las variables más importantes consideras por el modelo para clasificar a los individuos, se puede observar en la fig. (22).

**Figura 23: Top 15 Variables de importancia - Modelo Gradient Boosting**




Fuente: Elaboración propia

Al igual que en modelo de árbol de decisión, en el modelo de gradient boosting la importancia se mide sumando la reducción de impureza (Gini) que cada característica aporta en las divisiones, esta vez, de cada árbol de decisión que tiene el modelo gradient boosting. En este caso, se considera como la variable más importante Revolving_bin_X[0.74.inf>, lo que indica que un alto nivel de endeudamiento en líneas de tarjeta de crédito o créditos

personales con respecto a sus límites, es importante para discriminar entres clientes que pagan y los que no. En segundo lugar, se encuentra la variable Late Dummy_bin_X1, que indica si el cliente ha tenido atrasos, también considerada de gran importancia. Además, el modelo también considera a la variable Age en los rangos [44-58> y [58, inf>, sugiriendo que la edad mayor a 44 años es un criterio relevante para predecir la probabilidad de incumplimiento.

# 4. Resultados

En esta sección del estudio, se revisará los resultados comparativos de los modelos analizados previamente. Se evaluará su performance en clasificación para determinar la probabilidad de default y se analizarán sus implicaciones en los requerimientos de capital regulatorio.

**Tabla 8: Cuadro comparativo de resultados**

| ***Métrica*** | ***Logit*** | ***Árbol de decisión*** | ***Random forest*** | ***Gradiante boost*** |
|---|---|---|---|---|
| **acurracy** | 0.77 | 0.79 | 0.84 | 0.87 |
| **roc_auc** | 0.83 | 0.82 | 0.83 | 0.82 |
| **sensivility** | 0.74 | 0.67 | 0.63 | 0.54 |
| **specificity** | 0.77 | 0.80 | 0.85 | 0.89 |
| **f-score** | 0.30 | 0.29 | 0.34 | 0.34 |

Fuente: Elaboración propia

El análisis comparativo de métricas en la Tabla 8, nos indica que el modelo gradient boosting presenta la mayor exactitud (accuracy) del 87%. Esto implica que, en general, este modelo clasifica correctamente un mayor número de instancias de defaut y no-default en comparación con los otros modelos. Además, gradient boosting también muestra la mayor especificidad (89%), indicando una excelente capacidad para identificar correctamente las instancias de no-default, reduciendo así los falso positivos.

Por otro lado, aunque logit tiene la mayor sensibilidad (74%) que es crucial para identificar correctamente lo casos de default, su rendimiento global es inferior debido a su menor exactitud (77%) y especificidad (77%). Esto indica que, aunque logit es efectivo para detectar defaults, puede no ser tan confiable en general, ya que también comete más errores al identificar no-defaults.

Random forest se destaca con un buen equilibrio entre precisión y sensibilidad, reflejado en su f-score(34%), pero su especificidad es menor que la de gradient boosting, lo que lo hace menos eficiente en la identificación correcta de no-defaults.

Finalmente si consideramos como objetivo buscar un alto rendimiento global y un equilibrio óptimo entre exactitud y capacidad para discriminar correctamente entre default y no defaults, el modelo de gradient boosting se consolida como la mejor opción, su combinación de alta exactitud y especificidad, junto con un sólido f-score , lo posiciona como el modelo más robusto y confiable para determinar la probabilidad de default en el conjunto de datos de GMC analizado.

Lo demostrado con el caso práctico, confirma la literatura revisada, demostrando que los modelos de aprendizaje automático vienen siendo superiores a las técnicas tradicionales para determinar la probabilidad de default en una cartera de crédito.

Siguiendo con el análisis de resultados, se ha evaluado el impacto de los modelos en la determinación del capital regulatorio. El objetivo es demostrar que los beneficios en

clasificación se traducen en beneficios económicos al reducir el capital regulatorio requerido, lo cual es crucial para las instituciones financieras.

Para este análisis, se calcularon el capital medio aplicando las distribuciones de probabilidad de default de cada modelo en las ecuaciones (Eq .1), (Eq.2) y (Eq.3) de requerimiento de capital nombradas en el marco teórico. Con la finalidad de evaluar únicamente el impacto de las probabilidades de default en los modelos, se utilizó un valor fijo de LGD = 0.45 y EAD = 1, los resultados se presentan en la tabla 9.

**Tabla 9: Capital medio por modelo**

| *Modelo* | *Capital Medio* | *Desviación estándar* | *Capital Máximo* | *Capital Mínimo* |
|---|---|---|---|---|
| **Gradient Boosting** | 0.8564 | 0.2124 | 1.1962 | 0.2317 |
| **Random Forest** | 0.9144 | 0.2063 | 1.1962 | 0.2122 |
| **Desicion Tree** | 0.8792 | 0.2218 | 1.1956 | 0.2750 |
| **Logit** | 0.9377 | 0.2225 | 1.1962 | 0.0782 |

Fuente: Elaboración propia

Los resultados de la tabla muestran que gradient boosting no solo es el modelo más preciso en términos de clasificación, sino que también ofrece el capital medio más bajo (0.8564). Esto implica una reducción significativa en el capital regulatorio requerido, lo que puede traducirse en un beneficio financiero considerable para una institución. En comparación con el modelo logit, gradient boosting logra una reducción del capital medio del 8.66%, demostrado su eficiencia superior.

Otro punto a destacar es que gradient boosting presenta una desviación estándar relativamente baja (0.2124) lo que indica una mayor estabilidad y previsibilidad en las estimaciones de capital. El Modelo de árbol de decisión y random forest también presentan capitales medio más bajos que el logit, logrando reducciones de capital de 6.65% y 2.55% respectivamente.

Los resultados del caso de estudio demuestran que los modelos de aprendizaje automático, específicamente gradient boosting, no solo mejoran la exactitud de la clasificación de incumplimientos, sino que también ofrecen una mayor granularidad en sus predicciones de probabilidad de default. Esto permite obtener beneficios significativos en la reducción del capital regulatorio.

# 5. Conclusiones

La investigación de este estudio ha demostrado la eficacia de aplicar técnicas de aprendizaje automático para determinar la probabilidad de default en un contexto práctico. A través del análisis comparativo de cuatro modelos distintos, se ha evidenciado a través de los resultados obtenidos que las técnicas de aprendizaje automático no solo mejoran la exactitud en la clasificación de incumplimientos, sino que también ofrecen una mayor granularidad en las predicciones de probabilidad de default, lo que permite obtener beneficios significativos en la reducción de capital regulatorio.

Las implicaciones prácticas de los resultados hallados sugieren que la implementación de los modelos como gradient boosting puede llevar a una reducción considerable en los fondos que una institución financiera debe reservar para cubrir posibles pérdidas por incumplimiento. En línea con los objetivos de esta investigación, se ha demostrado que los modelos de aprendizaje automático generan beneficios económicos significativos (8% en ahorro de capital regulatorio) en la gestión de riesgo de crédito, contribuyendo a una gestión más eficiente y efectiva de los recursos financieros.

Otros hallazgos vinculados a la interpretabilidad del modelo pueden contribuir a la literatura existente. La interpretabilidad de los modelos, en el caso práctico, se evaluó mediante el análisis de importancia de las variables, se pudo identificar que variables relacionadas al historial de pago y nivel de endeudamiento son cruciales en la mayoría de los modelos. Esto siguiere que dichas variables deben ser altamente valoradas al momento de decidir el acceso al crédito de un individuo.

Sin embargo, el estudio no está exento de limitaciones. Los resultados dependen en gran medida de la calidad de los datos utilizados, así como el tratamiento que se haga con ellos. Por lo cual en un contexto donde la aplicación de los modelos se realice sobre una base más amplia de datos (millones de observaciones) podría requerir de una infraestructura tecnológica avanzada y personal capacitado, futuras investigaciones podrían explorar el costo de implementar tal infraestructura y compararlo con los beneficios obtenidos.

Futuras investigaciones también podrían explorar otras variables y algoritmos de aprendizaje automáticos que no fueron considerados en este estudio, un elemento agregado podría ser evaluar los modelos a lo largo del tiempo y distintas condiciones económicas.

Finalmente, el estudio no solo cumple con los objetivos propuestos, sino que también puede servir como guía para el desarrollo de modelos de aprendizaje automático, permitiendo que más personas adquieran conocimiento sobre el tema y puedan aplicar sus técnicas en sus propios contextos profesionales.

# 6. Bibliografía

Amat, J (2016). *Regresión logística y simple.* Disponible en: https://rpubs.com/Joaquin_AR/229736

Assef, F. M. & Steiner, M. T. A. (2020). *Ten-year evolution on credit risk research: a systematic literature review approach and discussion*. Ingeniería e Investigación, 40(2), 50–71. Disponible en: https://doi.org/10.15446/ing.investig.v40n2.78649

Barrios, J. (2019). *La matriz de confusión y sus métricas.* Disponible en: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Bhalla, D. (2015). *Weight of evidence (Woe) and information value (IV) Explained*. Disponible en: https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. Disponible en: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Boehmke, B., & Greenwell, B. (2020). *Hands-On Machine Learning with R*. Disponible en: https://bradleyboehmke.github.io/HOML/

Breiman, L. (2001) *Random Forests*. Machine Learning 45, 5–32 Disponible en: https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees (1st ed.).* Chapman and Hall/CRC. Disponible en: https://doi.org/10.1201/9781315139470

Comité de Supervisión de Basilea. (2017). *Basel III: Finalising post – crisis reforms*. Disponible en: https://www.bis.org/bcbs/publ/d424.pdf

Cox, D. R. (1958). "The Regression Analysis of Binary Sequences". Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215-232. Disponible en: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Friedman, J.H. (2001). *Greedy function approximation: A gradient boosting machine*. Ann. Statist. 29 (5) 1189 – 1232. Disponible en: https://doi.org/10.1214/aos/1013203451

Ginzo Technologies S.L. (2024). *Inteligencia Artificial y Machine Learning en el sector Bancario y Financiero.* Disponible en: https://ginzo.tech/inteligencia-artificial-y-machine-learning-en-el-sector-bancario-y-financiero/

Kaggle Competition (2011) . *Give Me Some Credit*.. Disponible en: https://www.kaggle.com/datasets/brycecf/give-me-some-credit-dataset

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). *Benchmarking state-of-the-art classification algorithms for credit scoring*: *An update of research.* European Journal of Operational Research. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0377221715004208

Li, Y.; Chen, W. (2020). *A Comparative Performance Assessment of Ensemble Learning for Credit Scoring*. Mathematics , 8, no. 10: 1756. Disponible en: https://doi.org/10.3390/math8101756

McKinsey & Company. (2023). *El estado de la IA en 2023: El año clave de la IA generativa*. Disponible en: https://www.mckinsey.com/featured-insights/destacados/el-estado-de-la-ia-en-2022-y-el-balance-de-media-decada/es

Robisco, A., & Carbó, J. M. (2022). *Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio*. *International Review of Financial Analysis, 84*, 102372. Disponible en: https://doi.org/10.1016/j.irfa.2022.102372

Robisco, A., & Carbó, J. M. (2023). Capítulo III. *Aprendizaje automático en modelos de concesión de crédito: oportunidades y riesgos*. En Carbó. S, Ganuza. J, Peña. D, Poncela. P(Eds.), Big Data (pp. 79-104). Funcas. Disponible en: https://www.funcas.es/wp-content/uploads/2023/05/Analisis-financiero-y-big-data.pdf.

Sánchez, G. (2021). *Random forest classifier sp500*. Disponible en: https://gsnchez.com/blog/article/Random-forest-classifier-sp500

Universidad Europea. (2022). *Aprendizaje supervisado y no supervisado*. Disponible en: https://universidadeuropea.com/blog/aprendizaje-supervisado-no-supervisado/

Villar Mir, J. (2020). *El impacto en la sociedad de la era digital*. Disponible en: https://www.boe.es/biblioteca_juridica/anuarios_derecho/abrir_pdf.php?id=ANU-M-2020-10016100178