BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MACHINE LEARNING PARA EL ANÁLISIS DE RIESGO CREDITICIO.
COMPARACIÓN DE MODELOS Y PROPUESTA DE APLICACIÓN AL CRÉDITO AUTOMOTRIZ EN MÉXICO

T E S I S

PRESENTADA PARA OBTENER EL GRADO DE:
LIC. EN ACTUARÍA

P R E S E N T A:

SILVIA GUADALUPE GUTIÉRREZ HERNÁNDEZ

A S E S O R:
MTRO. DAVID NEXICAPAN CORTES

PUEBLA, PUE. FEBRERO 2026

Agradecimientos.

Quiero expresar mi más profunda gratitud a todas las personas que me acompañaron y apoyaron durante mi trayecto universitario.

En primer lugar, agradezco a mi director de tesis, el Mtro. David Nexicapan Cortes, por su orientación, guía y acompañamiento a lo largo del desarrollo de este trabajo. Sus conocimientos y experiencia en el campo de la Actuaría fueron fundamentales para consolidar esta investigación.

A mis padres, Silvia Hernández y Javier Gutiérrez, gracias por su apoyo incondicional, su confianza y su amor. Su motivación me permitió superar obstáculos y llegar hasta esta meta.

A mi pareja, Milton Sánchez, gracias por llegar a mi vida y por ser un apoyo clave durante mi proceso de titulación. Gracias por tu motivación constante y por confiar en mí.

A mis amigos y compañeros, y de manera muy especial a Ana Mixcoatl, Alejandra González, Wendy Hernández y Anahí Pérez, les agradezco su compañía, su motivación y su amistad. Su apoyo fue esencial en los momentos más difíciles; las considero mi segunda familia.

A mi madrina, gracias por acompañarme a lo largo de mi carrera universitaria; tu apoyo y confianza me ayudaron a llegar hasta aquí. A mi padrino, gracias por tus sabias palabras, tu apoyo y la confianza depositada en mí, que también fue un impulso importante en este camino.

A los profesores que formaron parte de mi formación académica, les agradezco su enseñanza y su tiempo. En especial, al Act. Gerardo Castillo, por su sabiduría y por creer en mí cuando yo misma llegué a dudar.

A la Comunidad BUAP, gracias por el apoyo y la colaboración brindados en mi emprendimiento, lo cual me permitió financiar mi universidad.

A mis perritos, gracias por su compañía y por el confort emocional en los momentos difíciles. Aunque no leerán estas líneas, quiero dejar plasmado mi amor y gratitud hacia ellos. En especial, a mi perrita Mimi, por verme crecer y cuidarme; un abrazo hasta el cielo.

Agradezco también a los miembros del jurado por aceptar formar parte de este proceso y por sus observaciones y contribuciones a este trabajo.

A todos ustedes, muchas gracias.

Índice

Resumen.

La presente investigación propone una aproximación metodológica basada en Machine Learning para el análisis de riesgo crediticio, con énfasis en su posible aplicación al crédito automotriz en México como alternativa para mejorar la evaluación de solicitantes. Para ello se utilizó el conjunto de datos German Credit, el cual contiene características financieras y socioeconómicas de clientes, así como su clasificación como buen o mal pagador. Se realizó un análisis exploratorio de datos y el preprocesamiento necesario (limpieza y adecuación de variables) para su uso en modelos predictivos. Posteriormente, se construyeron modelos de clasificación supervisada —regresión logística, árboles de decisión y bosque aleatorio— con el objetivo de estimar la probabilidad de incumplimiento y clasificar a los solicitantes. El desempeño se evaluó mediante matriz de confusión, exactitud, sensibilidad/especificidad, curva ROC y AUC, y en el caso del bosque aleatorio, mediante el error Out-of-Bag (OOB) e importancia de variables. Los resultados evidencian que el bosque aleatorio alcanza el mejor desempeño global, mientras que la regresión logística ofrece una ventaja relevante en interpretabilidad a través de odds ratios, y los árboles aportan reglas de decisión claras, pero con desempeño moderado. En conjunto, la comparación permite sustentar una propuesta de aplicación metodológica al contexto mexicano, orientada a fortalecer la toma de decisiones y reducir errores en la selección de clientes.

Introducción.

El riesgo crediticio representa uno de los principales retos para instituciones financieras, SOFOMES y entidades que otorgan financiamiento, ya que un proceso de originación deficiente puede traducirse en incrementos de morosidad, deterioro de cartera y pérdidas económicas.

En particular, el crédito automotriz es un producto de alta relevancia por su relación con el consumo, la movilidad y el dinamismo del mercado, además de involucrar montos y plazos que requieren una evaluación adecuada de la capacidad y comportamiento de pago del solicitante. En este contexto, disponer de metodologías objetivas para estimar la probabilidad de incumplimiento y distinguir entre solicitantes de bajo y alto riesgo se vuelve fundamental para fortalecer la toma de decisiones.

Tradicionalmente, el análisis de solicitudes de crédito se apoya en criterios financieros, reglas internas y validación documental; sin embargo, la disponibilidad de datos y el avance de técnicas de analítica permiten complementar dichos enfoques mediante modelos predictivos. El Machine Learning, a través de algoritmos de clasificación supervisada, ofrece herramientas capaces de identificar patrones complejos en información histórica y generar estimaciones consistentes del riesgo de incumplimiento.

Además de mejorar la precisión de clasificación, estos modelos pueden aportar evidencia cuantitativa para justificar decisiones, priorizar variables relevantes y fortalecer el proceso de evaluación de clientes.

Bajo esta perspectiva, la presente investigación propone una aproximación metodológica basada en Machine Learning para el análisis de riesgo crediticio, utilizando como caso de estudio el conjunto de datos German Credit, el cual contiene características socioeconómicas y financieras de solicitantes, así como su resultado crediticio (buen o mal pagador). A partir de este conjunto, se realiza un análisis exploratorio y preprocesamiento de variables, seguido de la construcción de modelos de clasificación —regresión logística, árboles de decisión y bosque aleatorio— con el fin de estimar la probabilidad de incumplimiento y comparar su desempeño mediante métricas como matriz de confusión, exactitud, sensibilidad/especificidad, curva ROC y AUC, así como el error Out-of-Bag en el caso del bosque aleatorio.

Finalmente, el trabajo busca aportar evidencia para sustentar una propuesta de aplicación al crédito automotriz en México, planteando el uso de estos métodos como una alternativa para fortalecer la evaluación de posibles clientes. A través de la comparación entre modelos, se identifican ventajas

y limitaciones en términos de precisión predictiva e interpretabilidad, con el objetivo de orientar una implementación que contribuya a reducir errores de aprobación o rechazo y mejorar la sostenibilidad de la cartera.

Capítulo 1.

1.1 Problemática.

El crédito automotriz se ha consolidado como uno de los principales motores del crédito al consume en México. En los últimos años, las instituciones bancarias y las SOFOMES (Sociedades Financieras de Objeto Múltiple) que son entidades en México que otorgan créditos y financiamientos, han incrementado su oferta significativamente debido a la demanda sostenida de vehículos y a la integración de esquemas de financiamiento en agencias y plataformas digitales de venta. (Banco de México, 2024)

En los Reportes de Indicadores Básicos de Crédito Automotriz, Banco de México señala que este producto es objeto de seguimiento específico porque involucra montos relevantes, plazos multianuales y condiciones financieras variables entre instituciones, lo que refleja su peso en el sistema financiero (Banco de México, 2025).

A pesar de esta expansión, el mercado crediticio automotriz enfrenta tensiones estructurales y un entorno de riesgo cambiante. En términos macroeconómicos, la elevación y el posterior ajuste de las tasas de interés entre 2022 y 2024 han encarecido el financiamiento. Este incremento, en el costo total de adquisición del automóvil presiona la capacidad de pago de los hogares (Banco de México, 2024), lo que acentúa la necesidad de un análisis de riesgo más riguroso. Un mayor servicio de deuda incrementa la probabilidad de incumplimiento ante choques económicos, laborales o inflacionarios.

En la dinámica propia del crédito automotriz existe un elemento que lo distingue de otros productos de consumo: el automóvil se utiliza como colateral o garantía. Los reportes oficiales muestran que, gracias a dicha garantía, el índice de morosidad del crédito automotriz suele ser el más bajo dentro del crédito al consumo bancario (Banco de México, 2024).

Cada incumplimiento genera costos operativos, pérdidas por depreciación del vehículo recuperado y presión sobre la rentabilidad de la cartera. Además, el crecimiento acelerado del producto exige controles más finos para evitar deterioro futuro de calidad crediticia (Banco de México, 2024).

A nivel microeconómico, la problemática se expresa en la heterogeneidad de perfiles solicitantes. Los bancos mexicanos evalúan múltiples criterios al otorgar crédito automotriz: edad del solicitante, historial en Buró de Crédito, ingreso comprobable, antigüedad laboral, relación deuda–ingreso, enganche, plazo y valor del vehículo. En materiales informativos de bancos líderes, se

subraya que contar con “buen historial crediticio” y estabilidad de ingresos es requisito central (Banco de México, 2025).

La aplicación conjunta de tales filtros genera un reto particular para la población joven. De acuerdo con información del mercado laboral mexicano, los jóvenes presentan mayor rotación ocupacional, trayectorias laborales más cortas y, en muchos casos, inserción en empleos informales. Esto se traduce en historial crediticio escaso y menor antigüedad laboral, dos variables claves en los procesos tradicionales de aprobación. En consecuencia, aun cuando un joven cuente con capacidad de pago y un enganche adecuado, la probabilidad de rechazo tiende a ser mayor frente a solicitantes de más edad con historial financiero más amplio. Este fenómeno se vincula con un problema de inclusión financiera: el segmento que más demanda movilidad financiada es también el que enfrenta mayores barreras de acceso.

Otra arista de la problemática es la asimetría de información entre solicitante e institución. El banco observa variables documentales (ingreso, empleo, historial), pero no puede conocer con precisión la probabilidad futura de incumplimiento ni la estabilidad real del ingreso. Por ello, en la práctica se utilizan aproximaciones basadas en reglas generales por edad y trayectoria, que pueden generar errores de tipo I (aprobar a un cliente riesgoso) o tipo II (rechazar a un cliente solvente). La pérdida asociada al primer error es financiera; la del segundo es de negocio y de inclusión.

Adicionalmente, existen señales de que el entorno de riesgo crediticio en México se ha vuelto más complejo. Informes de agencias y prensa especializada documentan que el crédito al consumo, especialmente tarjetas de crédito, ha mostrado repuntes en cartera vencida del crédito al consumo, especialmente en tarjetas de crédito, lo que refuerza la necesidad de fortalecer modelos de riesgo en todos los productos de consumo, incluido el automotriz (EL CEO, 2024) .

En revisiones académicas sobre evaluación de riesgo en México se observa que la mayoría de los modelos cuantitativos nacionales se orientan a crédito al consumo agregado o tarjetas, mientras que los modelos específicos para crédito automotriz son escasos (José Carlos Trejo-García).

En síntesis, el crédito automotriz en México combina alta relevancia económica con un contexto de riesgo cambiante: expansión de cartera, costos financieros fluctuantes, perfiles solicitantes diversos y brechas de acceso particularmente visibles en jóvenes. Este escenario exige herramientas analíticas que permitan discriminar con mayor precisión entre clientes de bajo y alto riesgo, reduciendo errores de aprobación o rechazo y fortaleciendo la sostenibilidad de la cartera. (Reyes Morales, 2022)

1.2 Pregunta general de investigación.

¿En qué medida los modelos de *Machine Learning* de clasificación aplicados a un conjunto de datos permiten predecir correctamente si los solicitantes de crédito serán buenos o malos pagadores?

1.3 Objetivos

Objetivo general de investigación.

Evaluar modelos de *Machine Learning* de clasificación aplicados al conjunto de datos German Credit, con el fin de determinar la capacidad de cada modelo para predecir si los solicitantes de crédito serán buenos o malos pagadores, aportando evidencia para su posible adaptación al contexto del crédito automotriz en México.

Objetivos particulares.

- Analizar la base de datos German Credit, identificando su estructura, variables relevantes y características asociadas al riesgo crediticio, así como realizando el preprocesamiento necesario (limpieza, codificación y adecuación de variables) para su uso en modelos predictivos.
- Construir modelos de *Machine Learning* de clasificación (regresión logística, árboles de decisión y bosque aleatorio) para estimar la probabilidad de incumplimiento de los solicitantes y determinar si serán buenos o malos pagadores, comparando su desempeño mediante métricas de evaluación.
- Comparar los modelos de clasificación desarrollados, evaluando sus ventajas y limitaciones en términos de precisión predictiva e interpretabilidad con el fin de identificar el enfoque más adecuado como propuesta de aplicación al análisis del riesgo en crédito automotriz en México.

Preguntas particulares de investigación.

1. ¿Cuáles son las características principales y variables relevantes del conjunto German Credit que se asocian con el riesgo crediticio, y cómo deben prepararse para su análisis predictivo?
2. ¿Qué tan efectivos son los modelos de clasificación (regresión logística, árboles de decisión y bosque aleatorio) para estimar la probabilidad de incumplimiento y clasificar a los solicitantes como buenos o malos pagadores?

3. ¿Cuál de los modelos evaluados presenta el mejor desempeño predictivo según métricas como exactitud, matriz de confusión, AUC y error OOB, y qué variables resultan más influyentes en la clasificación?

Capítulo 2. Justificación

2.1 Justificación teórica y metodológica

El desarrollo de un modelo de evaluación de riesgo crediticio para crédito automotriz mediante técnicas de Machine Learning se justifica por razones teóricas, operativas y sociales. En el plano metodológico, los algoritmos de aprendizaje automático han mostrado ventajas consistentes frente a métodos tradicionales en problemas de clasificación financiera, ya que capturan relaciones no lineales e interacciones entre variables sin requerir supuestos paramétricos estrictos (Dialnet, 2023)

2.2 Justificación operativa y de mercado

El crédito automotriz es un producto de elevada competitividad. Los Reportes de Indicadores Básicos de Banco de México evidencian diferencias en tasas, plazos y estructuras de pago entre instituciones, lo cual incentiva a mejorar procesos internos de originación para conservar rentabilidad y participación de mercado(Banco de México, 2024).

2.3 Justificación en gestión del riesgo

Asimismo, herramientas predictivas robustas contribuyen a gestionar de forma preventiva el riesgo de incumplimiento. Aunque la morosidad del crédito automotriz es relativamente baja, cada incumplimiento representa costos significativos por recuperación, depreciación del vehículo y capital regulatorio (Banco de México, 2024) .

2.4 Justificación social y aporte del estudio

Por último, ante la limitada documentación pública de modelos específicos para crédito automotriz en México, esta propuesta contribuye al conocimiento aplicado y ofrece una metodología replicable para instituciones bancarias y no bancarias. Al construir un marco conceptual adaptable al contexto mexicano y validarlo posteriormente con técnicas de Machine Learning, se aporta evidencia útil para la práctica financiera (Reyes Morales, 2022).

Capítulo 3. Marco Teórico

3.1. Crédito automotriz en México: definición, estructura y características

El crédito automotriz es un instrumento de financiamiento destinado a la adquisición de un vehículo nuevo o usado, mediante el cual una institución financiera entrega al solicitante un monto de dinero que deberá ser reembolsado en pagos periódicos durante un plazo determinado. Este tipo de financiamiento se clasifica dentro del crédito al consumo y se caracteriza por estar directamente vinculado a un bien durable cuya compra se realiza de forma inmediata, mientras el pago se distribuye en el tiempo (Banco de México, 2025).

Una particularidad central del crédito automotriz es que el propio vehículo adquirido funciona como colateral o garantía prendaria. Esto significa que, si el acreditado incumple con el pago, la institución puede reclamar el automóvil y venderlo para recuperar parcial o totalmente el saldo adeudado. En México, esta figura suele formalizarse mediante un contrato de crédito simple con garantía prendaria: el vehículo permanece en posesión del deudor, pero la factura queda endosada a favor de la institución hasta que se liquide la obligación

(Banco de México, 2025).

3.1.1 Elementos que integran un crédito automotriz

En el sistema financiero mexicano, los créditos automotrices se definen por un conjunto de condiciones que determinan su costo y su riesgo. Entre los elementos principales se encuentran el monto del crédito, el plazo de pago, el enganche, la tasa de interés, el CAT (Costo Anual Total) y los seguros asociados.

El monto del crédito corresponde a la cantidad financiada por la institución para cubrir total o parcialmente el valor del vehículo. El plazo o duración se refiere al número de meses o años en los que se pagará el financiamiento. Esta variable es relevante porque afecta el nivel de riesgo: mientras más largo sea el plazo, mayor exposición enfrenta el banco ante cambios económicos o laborales del cliente (Banco de México, 2025).

El enganche es el pago inicial realizado por el solicitante. Un enganche mayor reduce el monto financiado y, por lo tanto, disminuye el riesgo para la institución. El Simulador de Crédito Automotriz de CONDUSEF considera el enganche como variable esencial para estimar el crédito al que puede acceder un cliente (CONDUSEF, 2025).

La tasa de interés representa el costo financiero expresado en porcentaje anual sobre el saldo financiado. Complementariamente, el CAT incorpora intereses, comisiones, seguros y gastos asociados al crédito, por lo que permite comparar el costo real entre distintas ofertas automotrices y no únicamente la tasa nominal (Banco de México, 2025). Finalmente, los seguros asociados —principalmente seguro de auto y, en ocasiones, seguros de vida o desempleo— suelen ser obligatorios y forman parte del costo total del crédito (Banco de México, 2025).

3.1.2 Importancia del crédito automotriz dentro del crédito al consumo.

A nivel agregado, el crédito automotriz es uno de los segmentos más relevantes del crédito al consumo en México. Banco de México elabora reportes periódicos de Indicadores Básicos de Crédito Automotriz para transparentar tasas, comisiones, costos y comportamiento del mercado, lo que da cuenta de su importancia sistémica (Banco de México, 2025). Estos reportes muestran que el financiamiento automotriz está presente tanto en banca comercial como en otras entidades reguladas, y constituye un canal clave para la adquisición de vehículos para fines particulares (Banco de México, 2025). Dado el valor elevado de los automóviles respecto al ingreso promedio, el acceso a crédito automotriz permite ampliar la capacidad de compra de los hogares y sostiene la dinámica del mercado vehicular.

3.1.3 Riesgo y morosidad del crédito automotriz

Aunque el crédito automotriz pertenece al crédito al consumo, su riesgo relativo suele ser menor que otros productos del mismo segmento debido a su garantía prendaria. De acuerdo con Banco de México y datos de la CNBV, el índice de morosidad del crédito automotriz es el más bajo entre los créditos al consumo bancario, precisamente porque el vehículo respalda la operación (México, 2023). Sin embargo, una morosidad menor no implica ausencia de riesgo. Cada incumplimiento genera costos administrativos, depreciación del vehículo recuperado y pérdidas para la institución. Además, el mercado automotriz crediticio ha mostrado un crecimiento acelerado en años recientes, lo que incrementa la necesidad de controles sólidos y modelos de evaluación más precisos para evitar deterioro futuro de cartera (EL CEO, 2024).

3.1.4 Síntesis del apartado

En México, el crédito automotriz puede definirse como un financiamiento al consumo destinado a adquirir un vehículo, cuya característica central es que el automóvil funciona como garantía

prendaria. Sus condiciones fundamentales incluyen monto, enganche, plazo, tasa y CAT, elementos que determinan el costo total y el riesgo de incumplimiento. Aunque presenta niveles de morosidad bajos frente a otros créditos al consumo debido al colateral, su expansión y los costos operativos del incumplimiento justifican la necesidad de analizarlo bajo enfoques cuantitativos y predictivos dentro de las instituciones financieras mexicanas.

3.2. Machine Learning: aprendizaje supervisado y modelos de clasificación

El *Machine Learning* (ML) es una rama de la inteligencia artificial enfocada en el desarrollo de algoritmos capaces de aprender patrones a partir de datos históricos para realizar predicciones o clasificaciones sin ser programados de forma explícita para cada caso. En términos generales, el ML busca aproximar una función que relacione variables de entrada (*features*) con una variable objetivo (*target*), con la finalidad de generalizar el aprendizaje a casos futuros (Caparrós, 2023).
Dentro del ML, el enfoque más utilizado en problemas de riesgo crediticio es el aprendizaje supervisado, debido a que las observaciones históricas cuentan con una etiqueta conocida (por ejemplo, "buen pagador" o "mal pagador"), lo cual permite entrenar modelos que aprendan las relaciones entre características del solicitante y su comportamiento real de pago (Kotsiantis, 2007).

3.2.1 Aprendizaje supervisado: definición formal y lógica general

En el aprendizaje supervisado se trabaja con un conjunto de datos compuesto por pares (X_i, Y_i), donde X_i representa el vector de variables explicativas de la observación i y Y_i corresponde al valor observado de la variable objetivo. El objetivo es estimar una función $f(\cdot)$ capaz de predecir Y a partir de X :

$$\textit{Ecuación (1)} \quad \hat{Y}_i = f(X_i)$$

El algoritmo se ajusta minimizando un error de entrenamiento y buscando que el error permanezca bajo también en datos no vistos (capacidad de generalización). Conceptualmente, el error de generalización se expresa como:

$$\textit{Ecuación (2)} \quad E_{gen} = \mathbb{E}\left[\, \mathrm{L}\big(\mathrm{Y}, \mathrm{f}(\mathrm{X})\big) \,\right]$$

Donde $L(\cdot)$ es una función de pérdida o error, y $\mathbb{E}[\cdot]$ representa la esperanza matemática sobre datos futuros (Caparrós, 2023). Este enfoque es adecuado para riesgo crediticio porque el historial de pagos permite observar Yen el pasado y predecirlo en nuevos solicitantes.

3.2.2 Modelos de clasificación en Machine Learning.

En clasificación, la variable objetivo es categórica. De forma general:

$$\textit{Ecuación (3)} \quad Y \in \{1, 2, \dots, K\}$$

donde Kes el número de clases. En riesgo crediticio suele utilizarse clasificación binaria:

$$\textit{Ecuación (4)} \quad Y \in \{0, 1\}$$

Muchos clasificadores estiman primero la probabilidad condicional de pertenencia a la clase positiva:

$$\text{Ecuación (5) Probabilidad estimada} \quad \hat{p}(X) = P(Y = 1|X)$$

y luego asignan la clase final mediante un umbral c(comúnmente 0.5):

$$\text{Ecuación (6)} \qquad \hat{Y} = \begin{cases} 1, si\ \hat{p}(X) \geq c \\ 0, si\ \ \hat{p}(X) < c \end{cases}$$

Esto permite clasificar solicitantes en categorías de riesgo de manera directa.

3.2.3 Proceso general de construcción de un clasificador supervisado.

Un modelo supervisado de clasificación suele seguir estas etapas: preparación y limpieza de datos; selección de variables predictoras; entrenamiento del modelo con datos etiquetados; validación y prueba con datos no utilizados en el ajuste; y evaluación por métricas de desempeño (Caparrós, 2023). Estas fases aseguran que el modelo no sólo aprenda patrones históricos, sino que sea capaz de generalizar adecuadamente.

3.2.4 Métricas para evaluar modelos de clasificación.

La evaluación de clasificadores parte de la matriz de confusión, de la cual se derivan métricas clave. La exactitud (*accuracy*) se define como:

$$\text{Ecuación (7)} \quad Accuracy = \frac{(VP + VN)}{(VP + FP + FN + VN}$$

La sensibilidad (*recall* o TPR) se calcula como:

$$\text{Ecuación (8)} \quad Recall = \frac{VP}{VP + FN}$$

La especificidad (TNR) se define como:

$$\text{Ecuación (9)} \quad Specificity = \frac{VN}{VN + FP}$$

En adición, la curva ROC grafica la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR), donde:

Ecuación (10) $FPR = \frac{FP}{FP + VN}$

El área bajo la curva ROC (AUC) resume la capacidad discriminatoria del modelo y toma valores entre 0.5 y 1; valores más cercanos a 1 indican mejor separación entre clases (Press, Credit scoring using machine learning and deep learning-based models, 2024).

3.2.5 Machine Learning aplicado al riesgo crediticio.

La literatura de *credit scoring* destaca que los algoritmos supervisados permiten mejorar la capacidad predictiva frente a enfoques tradicionales al capturar relaciones no lineales e interacciones entre variables financieras y sociodemográficas (Press, Credit scoring using machine learning and deep learning-based models, 2024). Además, estudios comparativos reportan que métodos basados en árboles y ensambles, como los bosques aleatorios, pueden superar en desempeño a modelos lineales en clasificación crediticia, manteniendo una estructura adecuada para evaluación de riesgo (Lou, 2021).

3.3 Modelos de clasificación aplicados al riesgo crediticio.

En evaluación de riesgo crediticio, los modelos de clasificación supervisada permiten estimar la probabilidad de incumplimiento de un solicitante a partir de variables demográficas, financieras y del propio crédito. De manera general, estos modelos construyen una regla de decisión que asigna a cada individuo una clase de riesgo, por ejemplo: buen pagador (0) o mal pagador (1). La literatura destaca que tanto los modelos estadísticos clásicos como los algoritmos de ML basados en árboles son herramientas estándar para diferenciar perfiles de riesgo y optimizar la asignación de crédito (Press, Credit scoring using machine learning and deep learning-based models, 2024).
En esta investigación se emplean tres clasificadores ampliamente utilizados en riesgo crediticio: regresión logística, árboles de decisión y bosques aleatorios, los cuales combinan interpretabilidad y desempeño predictivo en problemas financieros (Lou, 2021).

3.3.1 Regresión logística

La regresión logística es un modelo estadístico clásico usado para clasificación binaria. Su objetivo es modelar la probabilidad de que ocurra un evento $Y = 1$(por ejemplo, incumplimiento) en función de un conjunto de variables predictoras X. El modelo se basa en la función logística:

Ecuación (11) $P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}}$

Equivalentemente, se expresa mediante log-odds:

$$\text{Ecuación (12)} \quad log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Cada coeficiente β_j representa el cambio en el log-odds asociado a un incremento unitario en X_j. Su interpretación operativa suele hacerse con odds ratios (OR):

$$\text{Ecuación (13)} \quad OR_j = e^{\beta_j}$$

Si $OR_j > 1$, la variable incrementa la probabilidad del evento (mayor riesgo); si $OR_j < 1$, reduce dicha probabilidad. Esta interpretación es especialmente útil en riesgo crediticio porque permite justificar la dirección y magnitud del efecto de cada variable. La regresión logística es ampliamente utilizada en *scorecards* por su alta interpretabilidad y su capacidad de asignar probabilidades de incumplimiento; sin embargo, su desempeño puede verse limitado cuando existen relaciones no lineales complejas en los datos financieros (Press, Credit scoring using machine learning and deep learning-based models, 2024).

3.3.2 Árboles de decisión.

Los árboles de decisión son modelos no paramétricos que clasifican observaciones mediante reglas secuenciales del tipo "si–entonces", generando una estructura jerárquica de nodos. El algoritmo selecciona particiones del espacio de variables con el objetivo de maximizar la pureza de cada nodo.

Uno de los criterios de partición más utilizados es la impureza de Gini, definida para un nodo t:

$$\text{Ecuación (14)} \quad Gini(t) = 1 - \sum_{k=1}^{K} p_k^2$$

donde $p_k$es la proporción de observaciones de la clase k en el nodo. Cada partición busca maximizar la reducción de impureza:

$$\text{Ecuación (15)} \quad \Delta Gini = Gini(t) - \left[\frac{n_L}{n_t} Gini(t_L) + \frac{n_R}{n_t} Gini(t_R)\right]$$

Este procedimiento permite separar de forma progresiva observaciones asociadas a buen y mal comportamiento de pago. La ventaja principal del árbol es su interpretabilidad, pues las reglas finales pueden leerse como decisiones similares a las usadas por analistas financieros. Sin embargo, los árboles individuales pueden ser inestables y propensos al sobreajuste si no se realiza poda o control de complejidad (QuantInsti., 2022).

3.3.3 Bosques aleatorios (Random Forest)

El bosque aleatorio es un modelo de ensamble que combina múltiples árboles de decisión para mejorar precisión y estabilidad predictiva. Cada árbol se entrena con muestras bootstrap del conjunto original y con un subconjunto aleatorio de variables en cada partición (*mtry*). La predicción final se determina mediante el voto mayoritario:

Ecuación (16) $\hat{Y} = mode(\hat{Y}^{(1)}, \hat{Y}^{(2)}, \ldots, \hat{Y}^{(B)})$

Donde Bes el número total de árboles.

Una característica clave es la validación Out-of-Bag (OOB). Aproximadamente el 36.8% de los datos no entra en la muestra bootstrap (remuestreo con reemplazo) de cada árbol, por lo que resulta útil para medir error interno. El error OOB se estima como:

Ecuación (17) $Err_{OBB} = \frac{1}{n} \sum_{i=1}^{n} I(\hat{Y}_{OBB,i} \neq Y_i$

Donde $I(\cdot)$es función indicadora. Este método permite evaluar el modelo sin necesidad de una partición externa, reduciendo costo computacional sin sacrificar exactitud (AIMS Press, 2024).

En cuanto a interpretabilidad, el bosque aleatorio calcula el Mean Decrease Gini, que mide la contribución promedio de cada variable a la reducción de impureza. Valores mayores indican mayor relevancia predictiva. Diversos estudios concluyen que los bosques aleatorios suelen superar el desempeño de árboles individuales y modelos lineales en tareas de *credit scoring*, especialmente cuando las relaciones entre variables son no lineales, aunque a costa de menor interpretabilidad directa (Lou, 2021).

3.3.4 Síntesis del apartado

Los tres modelos descritos representan enfoques complementarios para clasificación crediticia. La regresión logística ofrece alta interpretabilidad mediante odds ratios; los árboles de decisión permiten reglas claras derivadas de particiones jerárquicas; y los bosques aleatorios mejoran precisión y estabilidad combinando múltiples árboles con validación OOB. Esta diversidad metodológica permite comparar desempeño predictivo y relevancia de variables en la identificación de buenos y malos pagadores.

3.4. Variables clave en riesgo crediticio y su relevancia para el crédito automotriz.

La evaluación del riesgo crediticio consiste en identificar la probabilidad de que un solicitante incumpla sus obligaciones de pago. Para ello, las instituciones financieras utilizan un conjunto de

variables que describen la capacidad y la disposición de pago del cliente, así como las condiciones del crédito solicitado. Estas variables se integran en modelos de *credit scoring* con el propósito de diferenciar entre clientes de bajo y alto riesgo y apoyar decisiones de otorgamiento (Press, Credit scoring using machine learning and deep learning-based models, 2024).

En crédito automotriz, este proceso es particularmente importante porque involucra montos relativamente altos dentro del crédito al consumo, con plazos largos y la presencia de un colateral (el vehículo). Banco de México reporta que, en México, la cartera automotriz se analiza con énfasis en características como monto del crédito, valor del auto, plazo y tasa, por su impacto directo en costo y riesgo de la operación (Banco de México, 2025).

3.4.1 Características del crédito: monto, plazo y carga de pago.

Las variables propias del crédito solicitado son determinantes porque definen el tamaño del compromiso financiero adquirido por el cliente.

Monto del crédito (Amount). El monto representa la exposición directa de la institución frente al cliente. A mayor monto financiado, mayor es el riesgo potencial de pérdida en caso de incumplimiento. Banco de México incluye el monto como una de las características centrales de análisis del crédito automotriz en México (Banco de México, 2025).

Plazo o duración (Duration). El plazo determina el tiempo de exposición. Plazos más largos incrementan la incertidumbre asociada a cambios en empleo, ingresos o condiciones macroeconómicas, lo cual puede elevar la probabilidad de morosidad. Por ello, el plazo es una característica estructural comparada en los indicadores básicos automotrices de Banxico (Banco de México, 2025).

Carga financiera o tasa de pago (Installment Rate). Las instituciones consideran qué proporción del ingreso mensual se destina al pago del préstamo, porque una carga excesiva reduce la capacidad de pago y eleva el riesgo. La CONDUSEF recomienda ajustar el crédito a la capacidad de pago real del cliente considerando ingreso, plazo y monto, para evitar sobreendeudamiento (CONDUSEF, 2025).

3.4.2 Características del solicitante: edad, estabilidad e indicadores de activos

Las variables sociodemográficas no determinan por sí solas el riesgo, pero funcionan como aproximaciones a estabilidad económica y ciclo de vida.

Edad (Age). La edad suele actuar como proxy de experiencia laboral, estabilidad de ingreso y madurez financiera. En México, guías de crédito automotriz señalan que la edad es un factor considerado al evaluar capacidad y viabilidad del financiamiento (CONDUSEF, 2025). En etapas tempranas (18–25 años), los solicitantes pueden tener menos historial crediticio o menor estabilidad laboral; mientras que edades medias tienden a asociarse con ingresos más consolidados. Esto no implica causalidad directa, sino que refleja trayectorias económicas típicas por cohorte, razón por la cual aparece frecuentemente en modelos de scoring (Press, Credit scoring using machine learning and deep learning-based models, 2024).

Activos o patrimonio (Property / Housing). La posesión de activos se asocia con mayor respaldo económico del cliente. Aunque el crédito automotriz está garantizado con el automóvil, el patrimonio adicional reduce el riesgo percibido porque indica capacidad financiera acumulada; por ello, las variables de patrimonio son diferenciadores comunes en sistemas de scoring (Kotsiantis, 2007).

3.4.3 Historial crediticio y comportamiento financiero previo.

El bloque más importante en sistemas de scoring suele ser el historial crediticio, porque captura evidencia real de pago.

Historial de pago (Credit History). Las sociedades de información crediticia en México calculan puntajes a partir del comportamiento histórico del cliente: pagos puntuales, atrasos y saldos. Buró de Crédito destaca que su puntaje se basa directamente en el desempeño previo del acreditado y que sirve para separar buenos de malos pagadores (Crédito, 2025).

Estatus financiero / cuentas bancarias (Checking Account Status). Las variables sobre manejo de cuentas, saldos y estabilidad financiera reflejan capacidad de liquidez y disciplina financiera, por lo que son típicas en scoring. En crédito automotriz, CONDUSEF y bancos comerciales destacan que el historial en Buró, la capacidad de pago y el ingreso comprobable son requisitos comunes para acceder a mejores condiciones (CONDUSEF, 2025).

3.4.4 Relevancia de estas variables en el contexto mexicano.

En México, las instituciones financieras enfrentan un mercado donde el crédito automotriz ha crecido de forma acelerada y requiere mecanismos sólidos de selección para evitar deterioro de cartera. Aunque la morosidad automotriz es relativamente baja frente a otros créditos al consumo, sigue siendo sensible a condiciones económicas como inflación o ingreso disponible, lo que

refuerza la necesidad de identificar perfiles con mayor probabilidad de incumplimiento (Banco de México, 2024).

Además, la disponibilidad y uso extendido de sociedades de información crediticia hace que variables como historial de pago sean especialmente determinantes en México, ya que son insumos centrales para la decisión bancaria de otorgar o negar un crédito (Crédito, 2025).

3.5. Pertinencia de modelos de Machine Learning para evaluar crédito automotriz en el contexto mexicano.

En México, el crédito automotriz constituye un segmento relevante del crédito al consumo, tanto por su crecimiento reciente como por el valor económico de las operaciones. Los *Indicadores Básicos de Crédito Automotriz* del Banco de México muestran que este tipo de financiamiento es monitoreado de forma periódica debido a su peso en la cartera de consumo, así como por su relación directa con el mercado interno de vehículos (Banco de México, 2025). Asimismo, reportes recientes señalan que durante 2024 el crédito automotriz bancario presentó un crecimiento anual notable, lo cual confirma su expansión dentro del sistema financiero mexicano (Economista, 2025)Una característica estructural del crédito automotriz es la presencia de un colateral: el automóvil adquirido se utiliza como garantía prendaria. Esto genera un riesgo relativo menor frente a otros productos de consumo. De hecho, Banco de México señala que el índice de morosidad (IMOR) del crédito automotriz ha sido el más bajo en comparación con otros créditos al consumo, precisamente por tratarse de un préstamo respaldado con garantía (México, Indicadores básicos de crédito al consumo., 2023) Sin embargo, que la morosidad sea baja no implica ausencia de riesgo: cada incumplimiento genera costos de recuperación, depreciación acelerada del activo recuperado y pérdidas potenciales para la institución financiera. Además, el crecimiento acelerado del portafolio automotriz incrementa la exposición del sistema ante un posible deterioro de cartera, haciendo indispensable mejorar los mecanismos de selección de clientes (Economista, 2025).

A este contexto se añade que las condiciones macroeconómicas mexicanas —particularmente las variaciones en tasas de interés— influyen de forma directa en el costo del crédito y, por tanto, en la capacidad de pago del acreditado. Cambios recientes en la tasa de referencia de Banco de México y el debate público sobre reducir tasas para facilitar el acceso al crédito reflejan un entorno donde el riesgo asociado al crédito al consumo puede variar con rapidez (Reuters, 2025). En escenarios así, se vuelve más importante anticipar qué perfiles tienen mayor probabilidad de incumplimiento.

En la práctica mexicana, las decisiones de otorgamiento de crédito automotriz se basan en variables de ingreso, estabilidad laboral, carga de pago y, especialmente, historial crediticio. El Buró de Crédito establece que su puntaje (*Mi Score*) se calcula directamente con base en el desempeño previo del cliente en el manejo de sus créditos, lo cual lo convierte en insumo fundamental para distinguir entre buenos y malos pagadores (crédito). Por ello, las instituciones financieras mexicanas dependen en gran medida del historial de pagos pasados para estimar riesgo futuro.

No obstante, la evidencia académica internacional indica que los métodos tradicionales de evaluación (como reglas expertas rígidas o modelos estrictamente lineales) pueden ser insuficientes cuando existen interacciones complejas y relaciones no lineales entre variables. Estudios recientes en *credit scoring* muestran que los algoritmos de *Machine Learning* supervisado incrementan la capacidad predictiva al aprender patrones multivariados más flexibles que los modelos clásicos (Press, Data Science in Finance and Economics, 2024). En particular, los enfoques basados en árboles y ensambles, como los bosques aleatorios, han reportado mejor desempeño que la regresión logística estándar en bases de consumo como German Credit, ya que capturan relaciones complejas entre características del solicitante y condiciones del crédito (al, 2021).

En consecuencia, la aplicación de modelos de ML al crédito automotriz en México es pertinente porque permite mejorar la precisión del *credit scoring* al identificar patrones complejos de riesgo, apoyar a las instituciones financieras en la expansión del crédito sin deteriorar la cartera y optimizar las decisiones de otorgamiento mediante probabilidades de incumplimiento más consistentes para distintos perfiles de solicitantes.

En síntesis, aunque el crédito automotriz mexicano presenta niveles relativamente bajos de morosidad, su crecimiento acelerado y la sensibilidad del crédito al consumo ante el entorno macroeconómico justifican el uso de modelos predictivos más robustos. Por ello, comparar clasificadores supervisados como regresión logística, árboles de decisión y bosques aleatorios establece una base metodológica sólida para proponer esquemas de evaluación de riesgo adaptables a instituciones financieras mexicanas.

Capítulo 4. Resultados

El análisis de riesgo crediticio busca determinar la probabilidad de que un cliente incumpla el pago de un crédito. Este estudio utiliza el dataset German Credit, el cual contiene 1,000 observaciones y 62 variables de tipo factor con dos niveles ***Good*** *para Buen pagador y* ***Bad*** *para Mal pagador*,

así como variables explicativas numéricas, incluyendo tanto variables continuas (como Duración, Monto y Edad) como variables binarias (dummy), representadas por valores 0 y 1 que indican la presencia o ausencia de ciertas condiciones que abarcan información financiera, demográfica y laboral de los solicitantes de crédito.

Previo al modelado, se verificó la calidad de los datos, confirmando que el dataset está limpio, ya que: No presenta valores faltantes (NA's), las variables categóricas están correctamente codificadas (sin errores ortográficos ni categorías inconsistentes) y las variables numéricas se encuentran dentro de rangos coherentes para el contexto financiero.

Debido a estas características, el dataset es adecuado para un análisis estadístico sin requerir procesos de imputación o corrección de inconsistencias.

Para la clasificación del riesgo crediticio, se aplicó un modelo de regresión logística, dado que esta técnica es apropiada cuando la variable dependiente es dicotómica, en este caso: *Good (Buen pagador)* vs. *Bad (Mal pagador)*. Además, permite estimar la probabilidad de incumplimiento en función de los predictores e interpretar los efectos de cada variable mediante los Odds Ratios los cuales comparan qué tan probable es que ocurra el evento frente a que no ocurra, lo que resulta útil para la toma de decisiones en el ámbito financiero.

La construcción del modelo permitirá identificar las variables que influyen significativamente en el riesgo crediticio y estimar la probabilidad de incumplimiento de nuevos solicitantes, contribuyendo así a una toma de decisiones más precisa en la evaluación crediticia.

El resumen estadístico muestra que las variables numéricas se encuentran dentro de rangos coherentes: por ejemplo, la duración del crédito varía entre 4 y 72 meses con una media de 20.9 meses, mientras que el monto solicitado oscila entre 250 y 18,424 unidades monetarias, con una media de 3,271. La edad de los clientes va de 19 a 75 años, con una media de 35.55 años.

Respecto a las variables binarias, se observa un comportamiento coherente con su codificación, ya que sus valores se encuentran entre 0 y 1. La variable objetivo "Clase" presenta una distribución de 70% clientes "Buen *pagador*" y 30% "Mal *pagador*", reflejando un leve desbalance de clases, pero aceptable para el modelado.

Además, variables como pagos realizados puntualmente cuentan con una media de 0.53, lo que indica que aproximadamente el 53% de los clientes tiene historial de pagos puntuales.

Tabla 1

Variables seleccionadas

Variable	**Interpretación breve**
Duración del crédito (meses)	Dura entre 4 y 72 meses; la media es 20.9 meses.
Monto del crédito	Créditos entre 250 y 18,424 unidades; promedio de 3,271.
Porcentaje de cuota sobre ingreso	Edad entre 19 y 75 años; media de 35.55 años.
Duración de residencia	Va de 1 a 4; mediana de 3 indica cuotas moderadas.
Edad	La media de 0.963 indica que la mayoría son trabajadores extranjeros.
Número de créditos existentes	Media 0.274 → el 27.4% tiene saldo menor que 0.
Número de personas dependientes	Media 0.53 → 53% ha pagado puntualmente antes.

Nota: Elaboración propia con base en el conjunto de datos German Credit. Las estadísticas descriptivas presentadas corresponden a las variables seleccionadas para el análisis del riesgo crediticio.

De acuerdo con la **Tabla 1** obtenida, podemos observar que, al desagregar las estadísticas por grupo de riesgo, se observa que los clientes clasificados como ‘Mal pagador’ presentan una duración promedio del crédito de 24.9 meses, mientras que en los clientes ‘Buen pagador’ este valor disminuye a 19.2 meses.

De manera similar, el monto promedio solicitado en el grupo ‘Mal pagador’ es de 3,938 unidades, superior al promedio de 2,985 unidades en el grupo ‘Buen pagador’.

Estos resultados sugieren que los créditos de mayor duración y mayor monto podrían estar relacionados con un mayor riesgo de incumplimiento.

Tabla 2.

Medidas descriptivas de las variables numéricas por clase de cliente.

Clase	Duración media	Duración mediana	Desviación estándar	Monto medio	Mediana del monto	Desviación estándar del monto
Mal *pagador*	24.9 meses	24	13.3	3938	2574	3536
Buen *pagador*	19.2 meses	18	11.1	2985	2244	2401

Nota: Elaboración propia con base en el conjunto de datos German Credit. Las medidas descriptivas se presentan por clase de cliente (buen pagador y mal pagador).

Se observa en la **Tabla 2** que el análisis descriptivo muestra en promedio, que los solicitantes clasificados como mal pagadores presentan créditos de mayor duración y mayor monto, lo que sugiere una mayor exposición financiera asociada al riesgo de incumplimiento.

4.1 Variables representativas

Se seleccionó un subconjunto de variables representativas que abarcan dimensiones clave del solicitante, tales como duración y monto del crédito (Duration, Amount), características demográficas (Age), estabilidad financiera (ForeignWorker, CheckingAccountStatus.lt.0, Housing.Own), comportamiento histórico (CreditHistory.PaidDuly), estabilidad laboral (EmploymentDuration.1.to.4) y propósito del crédito (Purpose.NewCar).

Tabla 3.

Tipo de variable representativa.

Variable	Tipo	Dimensión representada
Class	Categórica (objetivo)	Resultado crediticio (Bueno/Malo)
Duration	Numérica	Duración del crédito

Amount	Numérica	Monto solicitado
Age	Numérica	Característica demográfica
ForeignWorker	Binaria	Estatus laboral internacional
CheckingAccountStatus.lt.0	Binaria	Riesgo financiero en cuenta
CreditHistory.PaidDuly	Binaria	Historial de pago
Purpose.NewCar	Binaria	Propósito del crédito
EmploymentDuration.1.to.4	Binaria	Estabilidad laboral
Housing.Own	Binaria	Tenencia de vivienda (capacidad patrimonial)

Nota: Elaboración propia. La clasificación de las variables responde a su función dentro del análisis de riesgo crediticio.

La **Tabla 3** resume las variables seleccionadas y la dimensión del riesgo que representan, agrupando factores del crédito (monto y plazo), características del solicitante (edad), estabilidad y respaldo (empleo y vivienda), así como señales de comportamiento financiero (estatus de cuenta e historial de pago).

El resumen estadístico indica que la mayoría de los solicitantes son considerados 'Buen pagador' (70%), tienen una edad promedio de 35 años y solicitan créditos con una duración promedio de 21 meses por un monto aproximado de 3,271 unidades monetarias. Alrededor del 53% presenta historial de pagos puntuales y el 71% posee vivienda propia, lo cual podría asociarse con menor riesgo. Asimismo, el 27% registra saldo negativo en cuenta, lo cual puede representar una señal de alerta.

En cuanto a las variables numéricas, la duración promedio del crédito (Duration) es de 20.9 meses, lo que sugiere que los préstamos se concentran mayormente en un rango de entre 18 y 24 meses. El monto promedio solicitado (Amount) se sitúa en aproximadamente 3,271 unidades monetarias, con una mayoría de valores ubicados entre 1,366 y 3,972 unidades. La edad media de los solicitantes (Age) es de 35.5 años, reflejando que el perfil predominante corresponde a adultos jóvenes en etapa económicamente activa.

Respecto a las variables binarias, se observa que el 96.3% de los solicitantes son identificados como trabajadores extranjeros (ForeignWorker), lo cual es una característica propia del conjunto de datos. Asimismo, el 27.4% muestra un estado de cuenta corriente con saldo negativo (CheckingAccountStatus.lt.0), lo que puede asociarse a un mayor riesgo financiero. Por otro lado, el 53% de los clientes cuenta con un historial de pagos puntuales (CreditHistory.PaidDuly), lo cual constituye un indicador favorable de comportamiento crediticio.

En términos de propósito del crédito, el 23.4% de los préstamos se destinan a la adquisición de un automóvil nuevo (Purpose.NewCar). En cuanto a la estabilidad laboral, el 33.9% de los clientes presenta una antigüedad en el empleo de entre 1 y 4 años (EmploymentDuration.1.to.4), lo que refleja una situación laboral moderadamente estable. Finalmente, el 71.3% de los solicitantes posee vivienda propia (Housing.Own), lo que podría interpretarse como un indicio de solidez patrimonial.

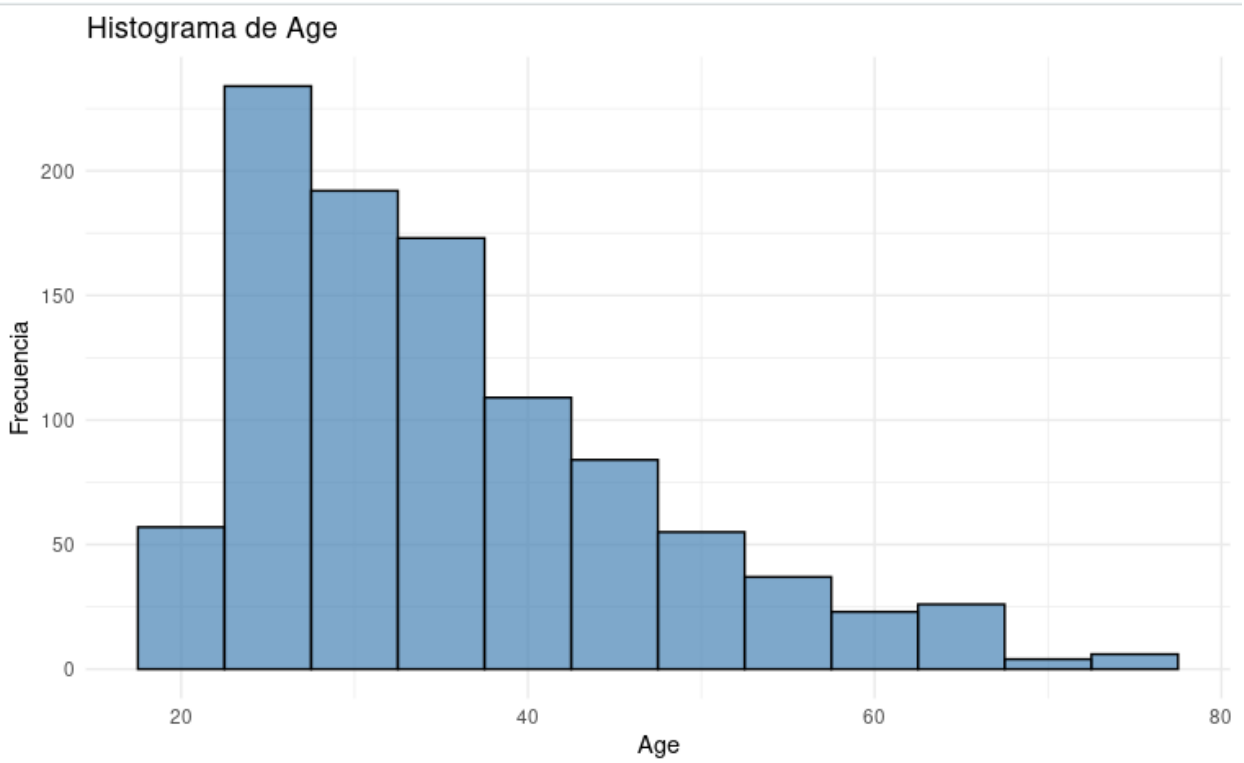
Los resultados descriptivos sugieren que ciertos factores podrían estar vinculados al riesgo crediticio. Por ejemplo, el hecho de que una proporción considerable de clientes con historial de pago puntual y vivienda propia predominen en la muestra podría asociarse a un menor riesgo de incumplimiento. En contraste, la presencia de un saldo negativo en la cuenta corriente y montos crediticios más elevados podrían representar indicadores de alerta. Asimismo, la estabilidad laboral moderada observada en una parte de los solicitantes podría influir en la capacidad de pago.

Estos patrones preliminares permiten anticipar que variables como la duración del crédito, el monto solicitado, el historial de pagos, la tenencia de vivienda y el estado de la cuenta bancaria podrían desempeñar un papel relevante en la clasificación del riesgo crediticio. No obstante, será a través del modelo de regresión logística donde se determinará con mayor precisión el efecto estadístico y la significancia de cada uno de estos factores sobre la probabilidad de incumplimiento.

Los gráficos de a continuación se realizaron sobre el conjunto completo de 1,000 observaciones con el fin de identificar patrones generales en las variables representativas.

Figura 1

Variable edad.

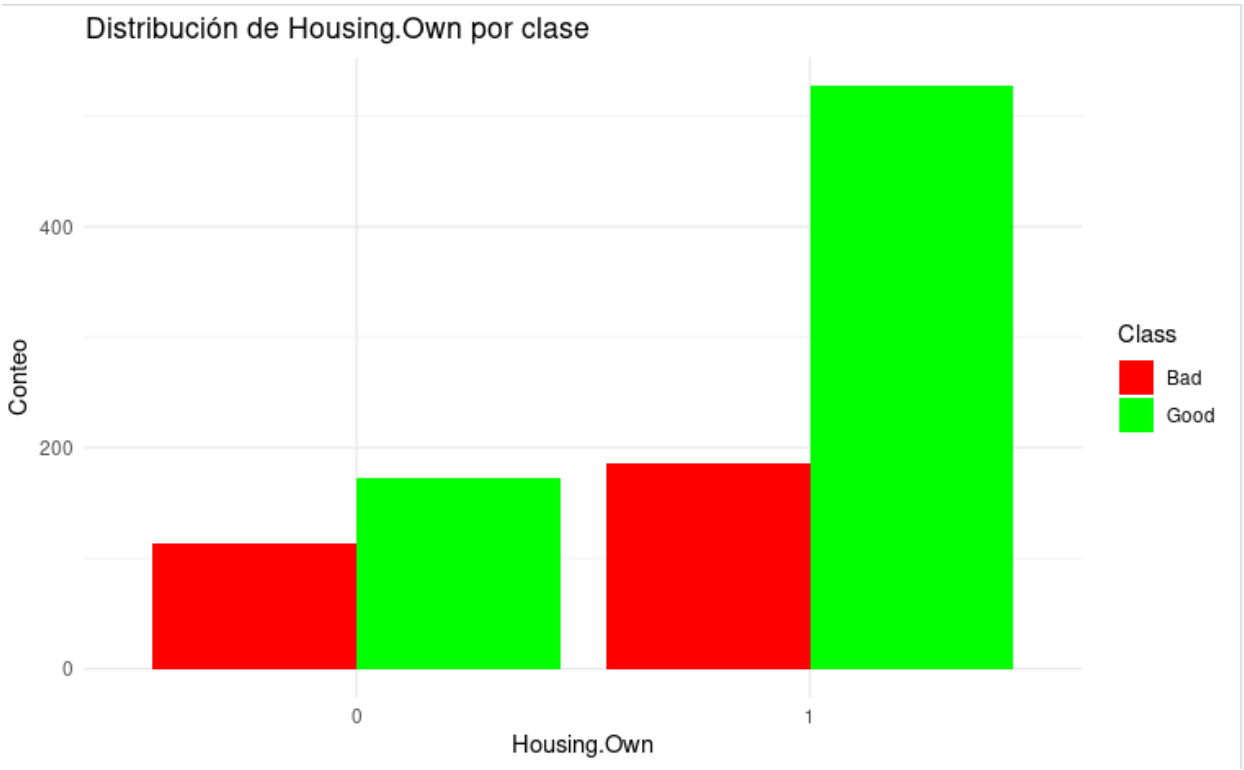


Nota: Elaboración propia con base en el conjunto de datos German Credit. El histograma muestra la distribución de la edad de los solicitantes incluidos en el análisis.

En la **Figura 1**, podemos ver una distribución sesgada hacia la izquierda (asimetría positiva), ya que hay más concentración en edades jóvenes y una cola hacia edades mayores pues se observa que la población de solicitantes se concentra mayormente entre los 25 y 40 años, con un pico de frecuencia alrededor de los 25-30 años. Conforme aumenta la edad, la frecuencia disminuye progresivamente, siendo menos común la solicitud de crédito en personas mayores de 50 años y prácticamente marginal en mayores de 65 años. Esto evidencia una distribución asimétrica positiva, lo cual indica que la mayoría de los solicitantes se encuentran en una etapa productiva de la vida, lo que coincide con edades típicas de mayor demanda de financiamiento para objetivos personales o familiares.

Figura 2

Para variables binarias (conteo)



Nota: Elaboración propia con base en el conjunto de datos German Credit. La gráfica muestra el conteo de clientes propietarios y no propietarios de vivienda, clasificados por tipo de cliente.

La **Figura 2**, muestra en su eje X: Vivienda propia, con 0 = No es propietario y 1 = Sí es propietario, por otro lado, su eje Y: Conteo de clientes representado con rojo=Clientes "Mal *pagador* y verde=Clientes "Buen *pagador*".

Se observa que la mayoría de los solicitantes son propietarios de vivienda. Dentro de este grupo, predominan los clientes clasificados como 'Buenos', lo que sugiere que la tenencia de vivienda podría estar asociada con un menor riesgo crediticio. Por el contrario, entre los solicitantes que no poseen vivienda propia también se observan clientes 'Malos' en una proporción relativamente mayor que en los propietarios, lo cual indica que la falta de activos patrimoniales podría estar vinculada a un incremento en el riesgo de incumplimiento.

Figura 3

Variable: Propietario de una vivienda

Nota: Elaboración propia con base en el conjunto de datos German Credit. La gráfica presenta la proporción de clientes buenos y malos según la condición de propiedad de vivienda.

La **Figura 3** muestra con rojo a los clientes "Malos" y con verde a los clientes "Buenos"

Al analizar la variable Housing.Own (Propietario de una Vivienda) en términos proporcionales, se observa que entre los clientes que no son propietarios de vivienda ('0'), aproximadamente el 40% pertenece a la clase 'Mal *pagador*', mientras que entre los propietarios ('1'), este porcentaje disminuye a cerca del 25%. En consecuencia, los solicitantes que poseen vivienda muestran una menor proporción de incumplimiento relativo, lo cual sugiere que la tenencia de activos patrimoniales podría estar asociada a un menor riesgo crediticio.

4.2 Modelo de regresión logística.

Los resultados del modelo de regresión logística indican que variables asociadas a mayor riesgo financiero presentan $OR < 1$ y son estadísticamente significativas, por ejemplo: saldo negativo en cuenta corriente (CheckingAccountStatus.lt.0, $OR = 0.154$; $p < 0.001$) y saldo bajo

(CheckingAccountStatus.0.to.200, $OR = 0.241;\ p < 0.001$). Asimismo, la variable Porcentaje de cuotas ($OR = 0.722;\ p = 0.0045$) sugiere que cuotas relativamente más altas respecto al ingreso disminuyen la probabilidad de ser 'Buen pagador'.

Por el contrario, la estabilidad laboral entre 4 y 7 años (EmploymentDuration.4.to.7, $OR = 3.236;\ p = 0.0307$) aumenta la probabilidad de ser 'Buen pagador', lo que concuerda con la hipótesis de que trayectorias laborales más estables favorecen la capacidad de pago.

Variables como Duración (OR=0.970; p=0.0102) muestran que créditos más largos se asocian con mayor riesgo relativo, mientras que el efecto de Monto es estadísticamente significativo pero marginal por unidad (OR≈1.000; p=0.0342), dado que el monto está medido en unidades individuales.

Variables con $OR < 1$ (disminuyen la probabilidad de ser "Buen *pagador*" → asociados a mayor riesgo "Mal *pagador*").

Tabla 4.

Resultados del modelo de regresión logística.

Variable	**Coeficiente**	**Odds Ratio (OR)**	**IC 95% (2.5% - 97.5%)**	**p-valor**
Cuenta corriente < 0	-1.873	0.154	0.088 – 0.267	0.0000
Cuenta corriente entre 0 y 200	-1.422	0.241	0.139 – 0.419	0.0000
Trabajador extranjero	-1.789	0.167	0.031 – 0.902	0.0375
Historial de pago puntual	-0.948	0.388	0.206 – 0.728	0.0032
Ahorros < 100	-1.170	0.310	0.160 – 0.602	0.0005
Ahorros entre 500 y 1000	-1.123	0.325	0.117 – 0.907	0.0319
Porcentaje de cuota sobre ingreso	-0.326	0.722	0.577 – 0.904	0.0045

Duración del crédito (meses)	-0.030	0.970	0.948 – 0.993	0.0102

Nota: Elaboración propia con base en el conjunto de datos German Credit, utilizando el software R (RStudio). Los odds ratios se obtuvieron a partir del modelo de regresión logística estimado.

Se observa en la **Tabla 4** a variables con $OR > 1$, la cuales aumentan la probabilidad de ser "Buen pagador", es decir, son factores favorables.

Tabla 5

Variables con Odds Ratios mayores o cercanos a uno.

Variable	**Coeficiente**	**Odds Ratio (OR)**	**IC 95% (2.5% - 97.5%)**	**p-valor**
Duración del empleo entre 4 y 7 años	1.174	3.236	1.116 – 9.384	0.0307
Monto del crédito	0.000	1.000	1.000 – 1.000	0.0342

Nota: Elaboración propia con base en el conjunto de datos German Credit, utilizando el software RStudio.

Se observa en la **Tabla 5** a Monto con $OR = 1.000$, su p-valor indica que el efecto es estadísticamente significativo, pero su impacto es muy bajo por unidad de cambio (el monto está en escalas grandes).

4.2.1 Ejemplos OR

Con el fin de ilustrar la interpretación de los Odds Ratios del modelo, se seleccionaron tres ejemplos representativos. En primer lugar, la variable Trabajo.DesempleadoNoCalificado presentó uno de los OR más altos ($OR = 5.083$), lo cual indica que los solicitantes desempleados o no calificados tienen una probabilidad significativamente mayor de ser clasificados como 'Mal *pagador*', en comparación con la categoría de referencia, manteniendo constantes las demás variables.

Por el contrario, la variable Otros deudores Co-solicitante registró uno de los OR más bajos ($OR = 0.150$), lo que sugiere que la presencia de un co-deudor reduce notablemente la probabilidad de incumplimiento, favoreciendo la clasificación como 'Buen pagador'.

Finalmente, el monto del crédito (Amount) mostró un OR cercano a 1 ($OR = 1.000$), lo que implica que su efecto marginal sobre la probabilidad de incumplimiento es prácticamente nulo, indicando que variaciones unitarias en el monto solicitado no generan un impacto sustancial en el riesgo crediticio.

Tabla 6.

Interpretación de los Odds Ratios.

Ejemplo	**Variable**	**OR**	**Interpretación básica**
OR alto	Job.UnemployedUnskilled	5.083	Trabajadores desempleados/no calificados tienen riesgo mucho mayor
OR bajo	OtherDebtorsGuarantors.CoApplicant	0.150	Tener co-deudor reduce mucho la probabilidad de "Mal *pagador*"
OR ≈ 1	Amount	1.000	El monto del crédito casi no influye

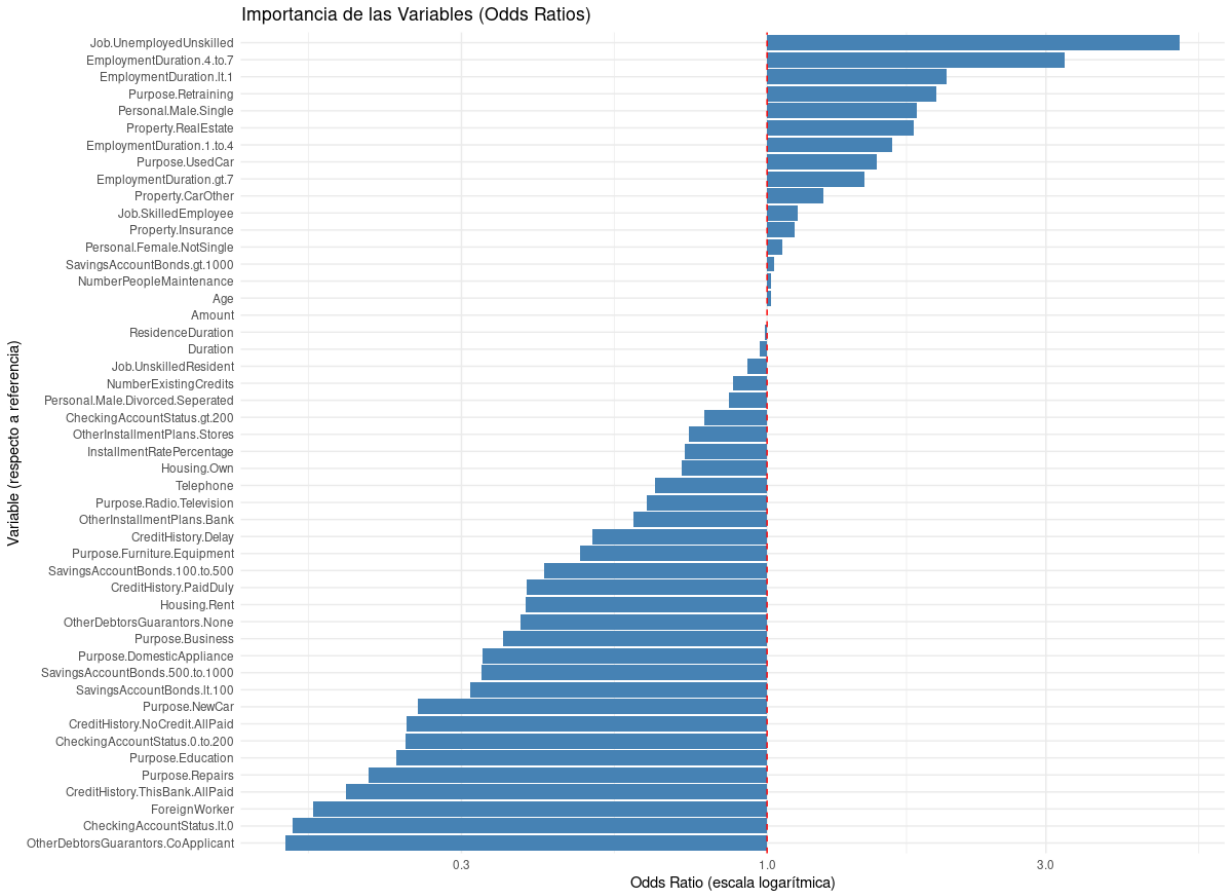
Nota: Elaboración propia con base en los resultados del modelo de regresión logística.

La **Tabla 6** presenta ejemplos de interpretación de Odds Ratios: un OR alto indica mayor asociación con la clase objetivo, un OR bajo indica efecto protector (disminuye las odds de la clase objetivo) y un OR cercano a 1 sugiere un efecto marginal o nulo.

4.2.1 Interpretación de coeficientes y *Odds Ratios*

Figura 4

Importancia de las Variables.



Nota: Elaboración propia con base en los resultados del modelo de regresión logística. Los Odds Ratios se presentan en escala logarítmica para facilitar la comparación del impacto relativo de las variables.

La **Figura 4** de Odds Ratios permite visualizar el impacto relativo de cada variable sobre la probabilidad de ser clasificado como 'Bueno'. Se observa que los factores asociados a mayor riesgo ($OR < 1$), como Estado de cuenta corriente, Otros deudores Co-solicitante y Trabajador Extranjero, reducen considerablemente la probabilidad de buen comportamiento crediticio. En contraste, variables como Duración del empleo entre 4 y 7 años ($OR > 1$) sugieren que una mayor estabilidad laboral incrementa la probabilidad de ser 'Bueno'. Asimismo, ciertas variables como Monto y Años muestran un impacto cercano a la neutralidad ($OR \approx 1$), lo que indica un efecto marginal. Estos resultados permiten identificar los factores más relevantes en la clasificación

crediticia del cliente y proporcionan una base sólida para la interpretación del comportamiento del modelo.

4.2.2 Predicciones, matriz de confusión y ROC/AUC

Tabla 7.

Matriz de confusión del modelo de regresión logística.

	Mal *pagador*	**Buen *pagador***
Predijo Mal *pagador*	42	38
Predijo Buen *pagador*	48	172

Nota: Elaboración propia con base en los resultados del modelo de regresión logística para la predicción de tipo de clientes.

Se observa en la **Tabla 7** que el modelo identificó correctamente 42 clientes de riesgo (‘Mal *pagador*’) y 172 clientes confiables (‘Buen *pagador*’). Sin embargo, existieron 48 falsos negativos, donde clientes considerados de riesgo fueron clasificados como ‘Buen *pagador*’, lo que representa un escenario desfavorable para una entidad crediticia, ya que implica aprobar créditos a posibles incumplidores. Por otro lado, se detectaron 38 falsos positivos, es decir, clientes realmente confiables que fueron catalogados como ‘Mal *pagador*’, lo que podría ocasionar rechazos injustificados. En términos generales, el modelo muestra una mayor capacidad para identificar correctamente a los clientes ‘Buen *pagador*’ que a los ‘Mal *pagador*’, lo cual influirá en las métricas de especificidad y sensibilidad.

Tabla 8.

Métricas de desempeño del modelo de regresión logística.

Métrica	**Valor**	**Interpretación breve**
Accuracy (Exactitud)	0.7133	Acierta en el 71.33% de los casos.
95% CI (Intervalo de Confianza)	0.6586 – 0.7638	La precisión real se encuentra dentro de este rango.

No Information Rate (Tasa base)	0.7000	Una clasificación trivial ya alcanzaría 70%.
p-valor (Acc > NIR)	0.3321	La mejora frente al NIR no es estadísticamente significativa.
Kappa	0.2951	Concordancia baja pero superior al azar.
Mcnemar's Test p-valor	0.3318	No hay diferencia significativa entre errores de tipo FP y FN.
Sensitivity (Sensibilidad – Mal *pagador*)	0.4667	Detecta correctamente el 46.67% de los "Mal *pagador*".
Specificity (Especificidad – Buen *pagador*)	0.8190	Detecta correctamente el 81.90% de los "Buen *pagador*".
Pos Pred Value (Valor Predictivo Positivo)	0.5250	De los predichos como "Mal *pagador*", 52.50% realmente lo son.
Neg Pred Value (Valor Predictivo Negativo)	0.7818	De los predichos como "Buen *pagador*", 78.18% realmente son "Buenos".
Prevalence (Prevalencia)	0.3000	El 30% de los casos reales son "Malos".
Detection Rate (Tasa de detección de Mal pagador)	0.1400	El modelo identifica correctamente al 14% del total como "Mal *pagador*".
Detection Prevalence (Proporción predicha como Mal pagador)	0.2667	Clasifica 26.67% de casos como "Malos".
Balanced Accuracy (Precisión balanceada)	0.6429	Rendimiento promedio entre sensibilidad y especificidad.
Clase Positiva considerada	"Mal *pagador*"	El modelo toma a "Mal *pagador*" como clase de referencia.

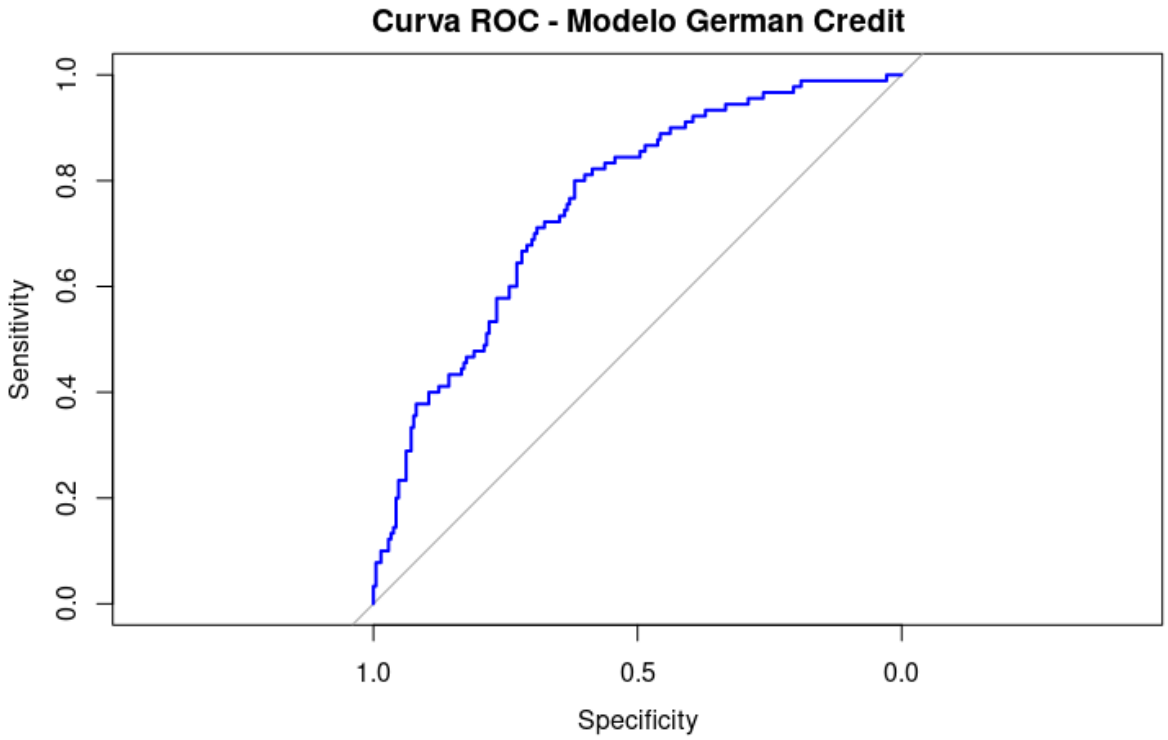
Nota: Elaboración propia para analizar el valor de cada métrica de desempeño.

Se observa en la **Tabla 8** que el modelo alcanza una exactitud del 71.33%, lo que indica que clasifica correctamente a poco más de siete de cada diez observaciones. Este resultado se encuentra dentro de un intervalo de confianza del 65.86% al 76.38%, lo cual ofrece un rango aceptable de estabilidad. Sin embargo, al compararse con la Tasa base (0.70), que representa la proporción de la clase mayoritaria ('Buen pagador'), se observa que el rendimiento del modelo no es significativamente superior a una clasificación trivial, lo cual es confirmado por el p-valor asociado (0.3321). El coeficiente Kappa (0.2951) señala una concordancia baja pero existente entre las predicciones del modelo y los valores reales, superando ligeramente el azar.

En términos de detección de la clase positiva ('Mal *pagador*'), la sensibilidad es del 46.67%, reflejando que el modelo identifica menos de la mitad de los clientes de alto riesgo. Por otro lado, la especificidad es elevada (81.90%), lo que indica que clasifica correctamente a la mayoría de los clientes 'Buenos'. Esto sugiere que el modelo tiene un mejor comportamiento al detectar buenos pagadores que malos. El Valor Predictivo Positivo (52.50%) muestra que solo poco más de la mitad de los clientes clasificados como 'Malos' lo son realmente, mientras que el Valor Predictivo Negativo (78.18%) indica una mayor confiabilidad al predecir 'Buen *pagador*'. La prevalencia de la clase 'Mal *pagador*' en los datos es del 30%, mientras que el modelo detecta correctamente solo el 14% de ellos. El modelo, sin embargo, clasifica el 26.67% de los datos como 'Mal *pagador*'. Finalmente, la Precisión Balanceada (0.6429) muestra un rendimiento moderado al equilibrar sensibilidad y especificidad.

Figura 5

Curva ROC



Nota: Curva ROC obtenida a partir del modelo de regresión logística. Elaboración propia.

En la Curva ROC de la **Figura 5**, el eje vertical (Y), conocido como Sensibilidad, representa la capacidad del modelo para identificar correctamente los casos clasificados como "Mal pagador", es decir, los verdaderos positivos.

Por otro lado, el eje horizontal (X) corresponde a 1 - Especificidad, lo que indica la proporción de falsos positivos, es decir, aquellos casos que en realidad son "Buenos" pero el modelo los clasificó incorrectamente como "Malos".

La **Figura 5**, evidencia un comportamiento superior al azar, ya que se ubica por encima de la diagonal de no discriminación. El valor del Área Bajo la Curva (AUC) es un número que resume qué tan bien puede el modelo separar correctamente las dos clases (Buen pagador vs Mal pagador) sin depender de un umbral específico.

Mientras más grande sea el AUC (cerca de 1), mejor el modelo diferencia entre clases, es decir, un AUC = 0.5 indica que el modelo no es mejor que lanzar una moneda (clasificación aleatoria) y un AUC < 0.5 significa que el modelo predice peor que el azar.

En este caso, nuestro AUC fue de 0.7561, lo cual indica que el modelo posee una capacidad moderadamente buena para distinguir entre clientes de alto riesgo (‘Mal pagador’) y bajo riesgo (‘Buenos’). En términos prácticos, existe un 75.61% de probabilidad de que el modelo asigne una mayor probabilidad de incumplimiento a un cliente ‘Malo’ que a uno ‘Bueno’. Por tanto, aunque el rendimiento no es óptimo, el modelo muestra un poder discriminatorio aceptable para efectos de clasificación crediticia.

4.2.3 Codificación de variables categóricas (variables dummy)

Con el propósito de preparar el conjunto de datos para su posible utilización en modelos que requieren variables numéricas, se realizó la conversión de todas las variables categóricas a formato dummy mediante la función model.matrix(). Esta transformación generó un nuevo conjunto denominado datos_dummy, compuesto por 1,000 observaciones y 63 columnas, todas en formato numérico binario (0/1 para las variables cualitativas). La variable objetivo fue representada como Clase mala, donde el valor 1 indica clientes con historial crediticio ‘Malo’ y 0 corresponde a ‘Bueno’.

Tabla 9.

Descripción de las variables utilizadas en el modelo de análisis crediticio.

Variable	**Tipo**	**Significado de 0**	**Significado de 1**	**Comentario**
Trabajador extranjero	Binaria	No es trabajador extranjero	Sí es trabajador extranjero	Puede reflejar estabilidad laboral
Cuenta corriente < 0	Binaria	Cuenta sin saldo negativo	Cuenta con saldo < 0	Riesgo financiero del cliente
Historial de pago puntual	Binaria	No paga puntualmente	Paga puntualmente	Historial de pago confiable

Crédito para auto nuevo	Binaria	No solicita crédito para auto nuevo	Sí solicita crédito para auto nuevo	Ayuda a segmentar por propósito del crédito
Empleo entre 1 y 4 años	Binaria	Empleo menor a 1 año o más de 4	Empleo entre 1 y 4 años	Experiencia laboral media
Vivienda propia	Binaria	No es dueño de vivienda	Sí es dueño de vivienda	Seguridad financiera y activos
Duración del crédito (meses)	Numérica	–	–	Duración del crédito en meses; mayor duración puede aumentar riesgo
Monto del crédito	Numérica	–	–	Monto del crédito; mayor monto puede aumentar riesgo
Edad del cliente	Numérica	–	–	Edad del cliente; puede correlacionarse con estabilidad
Clase (Good/Bad)	Objetivo	Buen pagador	Mal pagador	Variable a predecir

Nota: Elaboración propia con base en el conjunto de datos German Credit. Las variables categóricas fueron codificadas en formato binario (0/1) para su utilización en modelos de clasificación.

La **Tabla 9** presenta la codificación e interpretación de las variables empleadas, indicando el significado de los valores 0/1 en las variables binarias y la definición de la variable objetivo, con el fin de facilitar la lectura de los modelos y sus resultados.

Como conclusión, el análisis realizado con el dataset GermanCredit, compuesto por 1,000 observaciones y 62 variables, permitió desarrollar un modelo de regresión logística para predecir la probabilidad de que un cliente sea clasificado como "Bad" (mal pagador). A través de una exploración descriptiva inicial, se observó que los clientes con historial crediticio negativo presentaban en promedio créditos de mayor duración y montos más elevados, lo que sugirió la posible relación entre estas variables y el incumplimiento. Posteriormente, el modelo de regresión logística permitió identificar las variables con mayor peso en la probabilidad de incumplimiento. Entre los factores de mayor riesgo ($Odds\ Ratio < 1$) destacaron la existencia de saldo negativo en cuenta corriente (CheckingAccountStatus.lt.0) y la condición de trabajador extranjero (ForeignWorker). Por otro lado, variables asociadas a cierta estabilidad, como la duración laboral entre 4 y 7 años (EmploymentDuration.4.to.7), mostraron una inclinación hacia la clasificación "Buen pagador".

En términos de desempeño, la matriz de confusión mostró una exactitud del 71.33%, con una especificidad elevada (81.90%), lo que indica que el modelo clasifica adecuadamente a los clientes de buen comportamiento crediticio. No obstante, la sensibilidad fue moderada (46.67%), evidenciando que el modelo no identifica de forma óptima a todos los clientes "Malos". El coeficiente Kappa (0.2951) reflejó un nivel de concordancia bajo, aunque superior al azar.

La Curva ROC presentó un $AUC = 0.7561$, lo cual indica que el modelo posee una capacidad discriminativa "buena", pudiendo diferenciar entre clientes de alto y bajo riesgo en aproximadamente el 75% de los casos. Esto sugiere que, aunque el modelo es más efectivo en la detección de clientes confiables, puede ajustarse para mejorar la identificación de clientes con alta probabilidad de incumplimiento.

En conclusión, el modelo de regresión logística proporciona resultados consistentes y una capacidad predictiva aceptable para la evaluación del riesgo crediticio dentro del dataset analizado.

4.2.4 Regresión logística reducida

El segundo modelo corresponde a una regresión logística reducida, construida con un subconjunto de variables seleccionadas para comparar desempeño e interpretabilidad frente al modelo logístico completo.

Tabla 10.

Variables a usar.

Variable	Estimate	Significancia	Interpretación
(Intercept)	6.199	***	Valor base del modelo (sin variables). No se interpreta directamente.
Duración del crédito	-0.027	**	A mayor duración del crédito, menor probabilidad de incumplimiento (ligeramente contradictorio, puede implicar créditos largos, pero con clientes estables).
Monto del crédito	-0.000065	ns	No significativa ahora. El monto no parece influir tanto en este modelo.
Porcentaje de cuota sobre ingreso	-0.219	*	Cuanto mayor es la tasa de pago mensual, menor el riesgo (clientes que pagan más porcentaje del ingreso tienden a cumplir).
Trabajador extranjero	-1.550	.	Ligeramente significativa: los trabajadores extranjeros tienden a tener menor riesgo (aunque marginal).
Cuenta corriente < 0	-1.845	***	Muy importante: tener saldo negativo en la cuenta está fuertemente asociado con menor probabilidad de “Good”, o sea mayor riesgo crediticio.
Cuenta corriente entre 0 y 200	-1.396	***	Similar: tener entre 0 y 200 en la cuenta también aumenta el riesgo, aunque menos grave que saldo negativo.

Historial sin crédito / todo pagado	-1.500	**	Los clientes con historial sin crédito o todo pagado tienen menor riesgo, comportamiento positivo.
Historial en este banco todo pagado	-1.587	***	Quienes ya pagaron créditos en el mismo banco son significativamente más confiables.
Historial pago puntual	-0.661	**	Clientes que pagan puntualmente tienen menor probabilidad de incumplimiento.
Ahorros < 100	-1.065	***	Tener pocos ahorros (<100) se asocia con mayor riesgo.
Ahorros entre 100 y 500	-1.062	**	Tener algunos ahorros (100–500) también es predictor de menor riesgo.
Ahorros entre 500 y 1000	-0.912	.	Marginalmente significativa; el patrón sigue siendo: más ahorros = menor riesgo.
Empleo entre 4 y 7 años	0.655	*	Clientes con empleo de 4–7 años tienden a tener ligeramente más riesgo, quizá por montos más altos solicitados.
Co-solicitante	-0.861	.	Tener un co-solicitante reduce el riesgo, aunque marginalmente.

Nota: Elaboración propia, la regresión logística reducida se estimó a partir de un subconjunto de variables seleccionadas por significancia estadística y criterio interpretativo.

En la **Tabla 10** se presentan las variables más relevantes para explicar el riesgo crediticio, se concentran en señales de liquidez y comportamiento financiero el estatus de la cuenta corriente (categorías con saldo negativo o bajo), el historial crediticio y los niveles de ahorro, ya que presentan mayor significancia estadística y aportan mayor capacidad para diferenciar entre buenos y malos pagadores. En cambio, el monto del crédito (Amount) no resulta significativo dentro de

este modelo, lo que sugiere que el riesgo no depende únicamente de cuánto se solicita, sino del perfil financiero del solicitante. Asimismo, la duración y la tasa/porcentaje de pago muestran efectos significativos, pero relativamente pequeños, mientras que contar con un co-solicitante tiende a reducir el riesgo estimado y ciertas categorías de estabilidad laboral muestran asociación con el riesgo respecto a su referencia.

Tabla 11.

Matriz de predicción.

Predicción / Realidad	**Mal pagador**	**Buen pagador**
Predijo Mal pagador	38	29
Predijo Buen pagador	52	181

*Nota: A partir de la **Tabla 10** se arroja dicha matriz.*

Se observa en la **Tabla 10** la matriz de confusión, la cual permite evaluar el desempeño del modelo logístico reducido en la clasificación de los solicitantes de crédito como clientes con buen historial o con riesgo de incumplimiento.

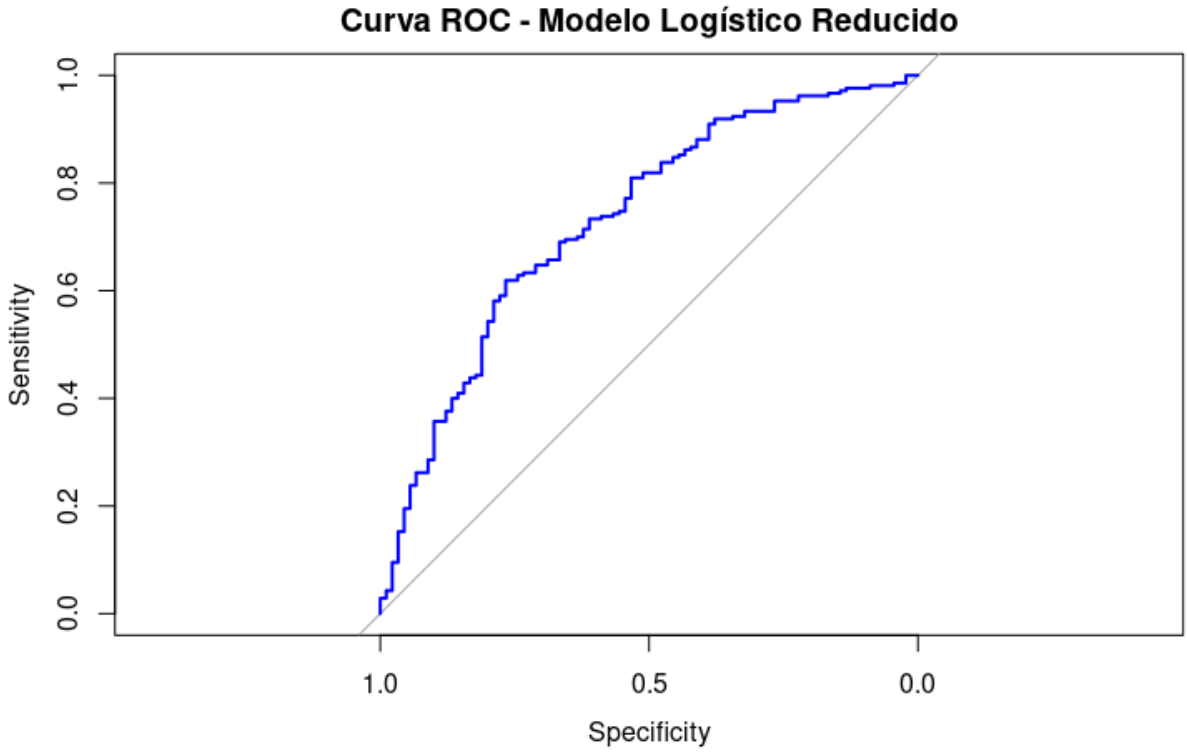
De los resultados obtenidos, el modelo logró clasificar correctamente a 38 clientes como “Malos”, coincidiendo con la realidad (verdaderos positivos), y 181 clientes como “Buenos” (verdaderos negativos).

Sin embargo, 29 clientes fueron clasificados erróneamente como “Malos” cuando en realidad eran buenos pagadores (falsos positivos), y 52 clientes fueron considerados “Buenos” pese a pertenecer al grupo de riesgo (falsos negativos).

En términos porcentuales, el modelo presenta una precisión global (accuracy) del 73 %, lo que significa que acierta en aproximadamente siete de cada diez casos. Además, muestra una sensibilidad del 42 %, indicando que detecta a cuatro de cada diez clientes realmente riesgosos, y una especificidad del 86 %, lo que demuestra un buen desempeño al identificar clientes confiables. Desde una perspectiva crediticia, esto implica que el modelo minimiza los falsos rechazos de solicitantes con buen comportamiento financiero.

Figura 6

Curva ROC – Modelo Reducido.



Nota: Elaboración propia, la figura corresponde a la curva ROC del modelo de regresión logística reducido.

La **Figura 6** muestra la Curva ROC, correspondiente al modelo logístico reducido, utilizada para evaluar su capacidad de distinción entre clientes con buen y mal historial crediticio. La línea diagonal gris representa un clasificador aleatorio, mientras que la curva azul muestra el desempeño real del modelo en términos de sensibilidad y especificidad para distintos puntos de corte.

El valor obtenido para el Área Bajo la Curva ($AUC = 0.7321$) indica una capacidad predictiva buena, lo que significa que el modelo tiene un 73.2% de probabilidad de asignar una puntuación de riesgo más alta a un cliente clasificado como "Mal pagador" que a uno "Buen pagador".

En otras palabras, el modelo logra distinguir con efectividad entre solicitantes de crédito con riesgo y aquellos confiables.

En comparación con el modelo completo ($AUC \approx 0.756$), el modelo reducido mantiene un rendimiento similar utilizando un conjunto menor de variables, lo que mejora su simplicidad e interpretabilidad sin afectar significativamente su poder predictivo. Por tanto, puede considerarse

un modelo más eficiente, con una adecuada capacidad para distinguir entre distintos niveles de riesgo crediticio.

4.3 Arboles de decisión.

4.3.1 Árbol de decisión: modelo completo

El árbol evaluó todas las 62 variables del conjunto de datos, sin embargo, sólo 12 de ellas fueron utilizadas efectivamente en la construcción del modelo, de acuerdo con la salida del comando *printcp(arbol_full)*.

Estas variables son las que contribuyeron a mejorar la pureza de los nodos según el criterio de Gini, lo que significa que aportaron información relevante para distinguir entre clientes con buen y mal historial crediticio:

1. Monto del crédito
2. Estado de cuenta corriente: saldo mayor a 200 DM
3. Sin cuenta corriente
4. Historial crediticio: crítico
5. Historial crediticio: retrasos en pagos
6. Historial crediticio: pagos puntuales
7. Duración del crédito (en meses)
8. Porcentaje de cuota mensual
9. Otros deudores o garantes: con aval
10. Otros deudores o garantes: ninguno
11. Cuenta de ahorros o bonos: entre 100 y 500 DM
12. Cuenta de ahorros o bonos: menos de 100 DM

Tabla 12.

Parámetros de complejidad del árbol de decisión.

CP	nSplit	Error Relativo	Error Cruzado (x-error)	Desv. Estándar (xstd)
0.035714	0	1.00000	1.00000	0.057735
0.033333	6	0.75714	0.99524	0.057656
0.019048	8	0.69048	0.97143	0.057252
0.012698	9	0.67143	0.87619	0.055458
0.010000	14	0.59048	0.88095	0.055555

Nota: Elaboración propia. Los parámetros se obtuvieron a partir del ajuste del árbol de decisión mediante validación cruzada.

El parámetro de complejidad (CP) regula el crecimiento del árbol de decisión, controlando el equilibrio entre el ajuste y la generalización del modelo. A medida que el valor de CP disminuye, el árbol puede realizar más divisiones (nSplit) y, por tanto, capturar más patrones en los datos.

En la **Tabla 12** se observa que el error relativo inicial del modelo (sin divisiones) es 1.000, correspondiente al modelo base sin particiones. Conforme el árbol crece y se realizan nuevas divisiones, el error relativo disminuye progresivamente hasta alcanzar 0.59048 con 14 divisiones, lo que representa una mejora significativa en la capacidad predictiva del modelo sobre los datos de entrenamiento.

Sin embargo, el error cruzado ($x - error$), que mide la capacidad de generalización mediante validación cruzada, no disminuye en la misma proporción. A partir de un cierto punto (alrededor de las 9 divisiones), el x-error deja de reducirse e incluso presenta ligeras variaciones. Esto indica que el árbol podría comenzar a sobre ajustar los datos, es decir, a aprender patrones específicos del conjunto de entrenamiento que no necesariamente se replicarán en nuevos datos.

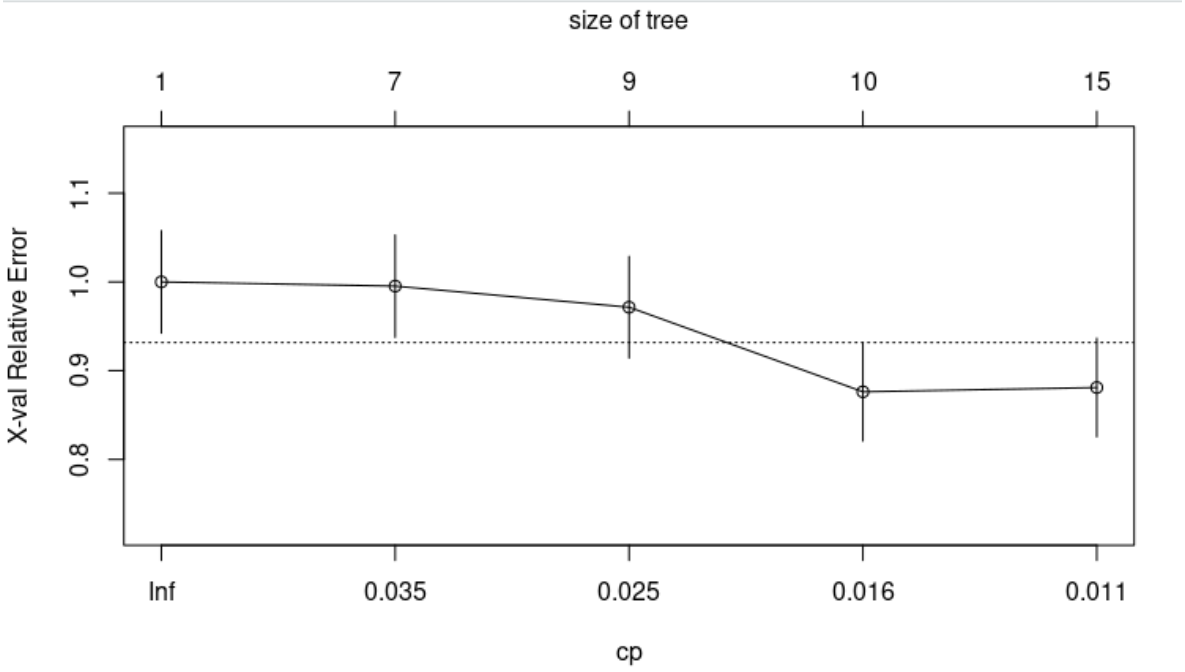
Por este motivo, se recomienda aplicar un proceso de poda del árbol utilizando el valor de CP que minimice el error cruzado. Este procedimiento permitirá obtener un modelo más simple, interpretable y con un mejor equilibrio entre precisión y generalización.

Como conclusión, el árbol inicia con un error relativo de 1.0 (sin divisiones) y, tras realizar 14 divisiones (nsplit = 14), el error disminuye a 0.59, lo que representa una mejora del 41% respecto

al modelo base. Sin embargo, el valor de $x - error = 0.88$ sugiere que continuar incrementando el tamaño del árbol podría provocar sobreajuste a los datos.

Figura 7.

Validación cruzada del árbol de decisión completo.



Nota: Elaboración propia. La gráfica corresponde al proceso de validación cruzada del árbol de decisión para la selección del parámetro de complejidad (cp).

En la **Figura 7** se observa al eje X como el parámetro de complejidad (cp), que controla el tamaño del árbol, de tal manera que a valores mayores de cp → árboles más simples (menos divisiones), por el contrario, a valores menores de cp → árboles más grandes y complejos. El eje Y representa el error relativo promedio obtenido por validación cruzada (X-val Relative Error), el cual mide qué tan bien generaliza el modelo.

Los puntos indican los valores de error para distintos niveles de cp, mientras que las barras verticales son intervalos de confianza del error estimado.

La línea horizontal punteada corresponde al mínimo error de validación cruzada, y sirve como referencia para determinar el punto óptimo de poda.

Esto es, cada punto negro con barras de error representa un valor de cp y su error estimado, el punto más bajo (el mínimo) es donde el modelo tiene menor error de validación cruzada. Sin embargo, según la regla del 1-SE (una desviación estándar), no necesariamente se elige el punto exacto más

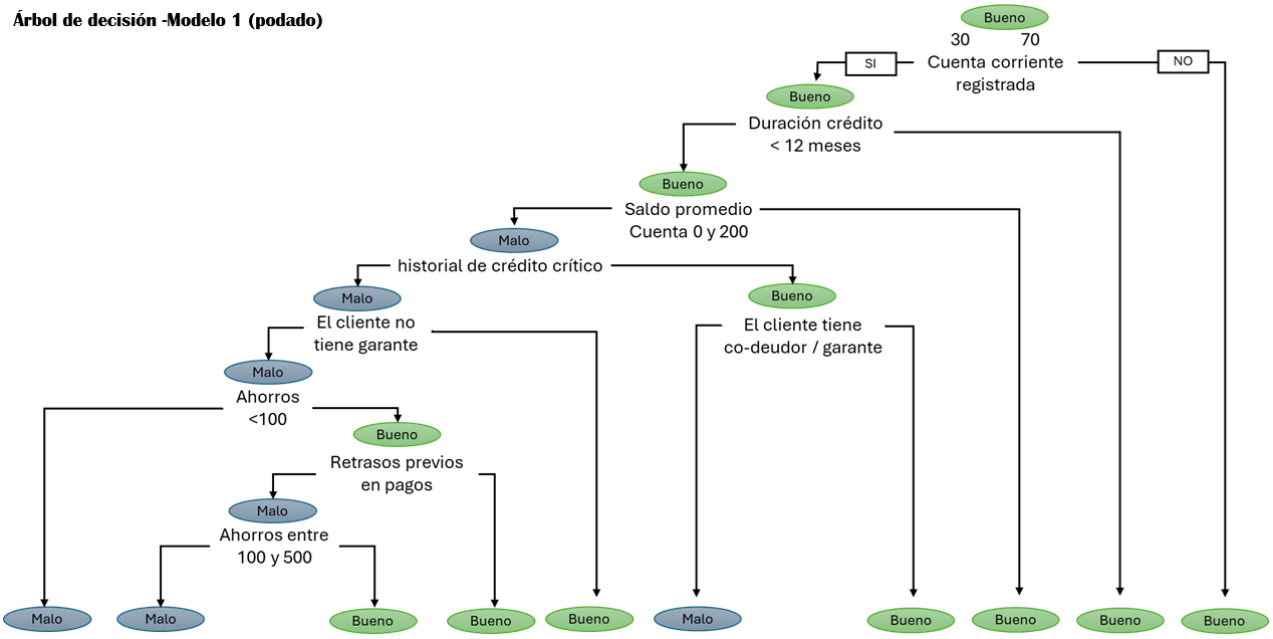
bajo, sino el más simple (mayor cp) que aún esté dentro de una desviación estándar de ese mínimo. Es decir: elegimos el punto en o justo por debajo de la línea punteada, pero a la derecha del mínimo, para evitar sobreajuste.

El gráfico muestra que el error de validación cruzada comienza alto cuando el árbol tiene pocas divisiones (lado izquierdo del eje X), y disminuye conforme se agregan más particiones, alcanzando su mínimo alrededor de un valor de cp ≈ 0.016. A partir de ese punto, el error deja de mejorar de forma significativa, lo que indica que aumentar la complejidad del árbol no aporta beneficios en la capacidad predictiva y podría conducir al sobreajuste.

Según el principio del mínimo error más simple, el cp óptimo es el valor más pequeño de cp que mantiene el error dentro de una desviación estándar del mínimo. En este gráfico, dicho punto corresponde aproximadamente a $cp = 0.016$, que se asocia con un árbol de unas 10 divisiones.

Figura 8

Árbol de decisión del Modelo 1



Nota: Se adaptó el árbol que se generó en el Scribd para una mejor visualización.

La **Figura 8** del árbol de decisión 1, muestra el árbol de decisión resultante tras aplicar el proceso de poda, utilizando el valor óptimo del parámetro de complejidad ($cp = 0.016$), determinado a partir del análisis del gráfico de validación cruzada. El objetivo de la poda es simplificar el modelo reduciendo su tamaño sin comprometer significativamente su capacidad predictiva, evitando así el sobreajuste (*overfitting*).

En el árbol podado se conservan únicamente las divisiones que aportan información relevante para clasificar a los clientes como buen pagador o mal pagador.

Las principales variables que determinan las divisiones del árbol son el estado de la cuenta corriente del cliente, en particular no poseer cuenta bancaria activa (*CheckingAccountStatus.none*) y contar con un saldo mayor a 200 DM (*CheckingAccountStatus.gt.200*), ya que reflejan su liquidez y comportamiento financiero; la duración del crédito, donde plazos mayores suelen asociarse con mayor riesgo; el historial crediticio crítico y el historial con retrasos, que capturan el comportamiento de pago previo; las categorías de cuenta de ahorros con menos de 100 DM y entre 100 y 500 DM, como indicadores del nivel de ahorro disponible (montos bajos tienden a vincularse con mayor probabilidad de incumplimiento); y la presencia o ausencia de co-deudor, la cual influye en la percepción y mitigación del riesgo crediticio.

El modelo predice que los clientes con cuentas bancarias activas y sin retrasos de pago tienden a clasificarse como "Buenos", mientras que aquellos con cuentas con saldo negativo o historial crediticio crítico presentan una mayor probabilidad de ser clasificados como "Malos".

El resultado es un árbol más compacto e interpretable, compuesto por aproximadamente 10 nodos terminales, lo que permite una lectura clara de las reglas de decisión y facilita su aplicación práctica para la evaluación del riesgo crediticio.

4.3.2 Evaluación del Árbol de Decisión Podado (Modelo 1)

La matriz de confusión resume el desempeño del modelo al comparar las clases predichas ("Bueno" o "Malo") con las clases reales del conjunto de prueba.

Tabla 13.

Predicción / Realidad

	Mal pagador	**Buen pagador**
Predijo Mal pagador	40	42
Predijo Buen pagador	50	168

Nota: Elaboración propia. La matriz de confusión corresponde al desempeño del árbol de decisión podado evaluado sobre el conjunto de prueba.

Se observa en la **Tabla 13** que el modelo predijo correctamente 40 clientes malos pagadores y 168 clientes buenos pagadores, sin embargo, confundió 42 clientes buenos al clasificarlos erróneamente como malos, y 50 clientes "malos" fueron clasificados incorrectamente como "buenos".

Esto indica que el árbol de decisión tiende a ser más preciso identificando a los buenos pagadores, pero tiene dificultades para detectar a los malos pagadores (falsos negativos).

Lo cual, significa que el modelo puede subestimar el riesgo crediticio de algunos clientes con alta probabilidad de incumplimiento. Aun así, ofrece una herramienta útil para identificar perfiles de bajo riesgo y visualizar reglas claras de decisión basadas en variables financieras como el historial crediticio y el estado de cuenta bancaria.

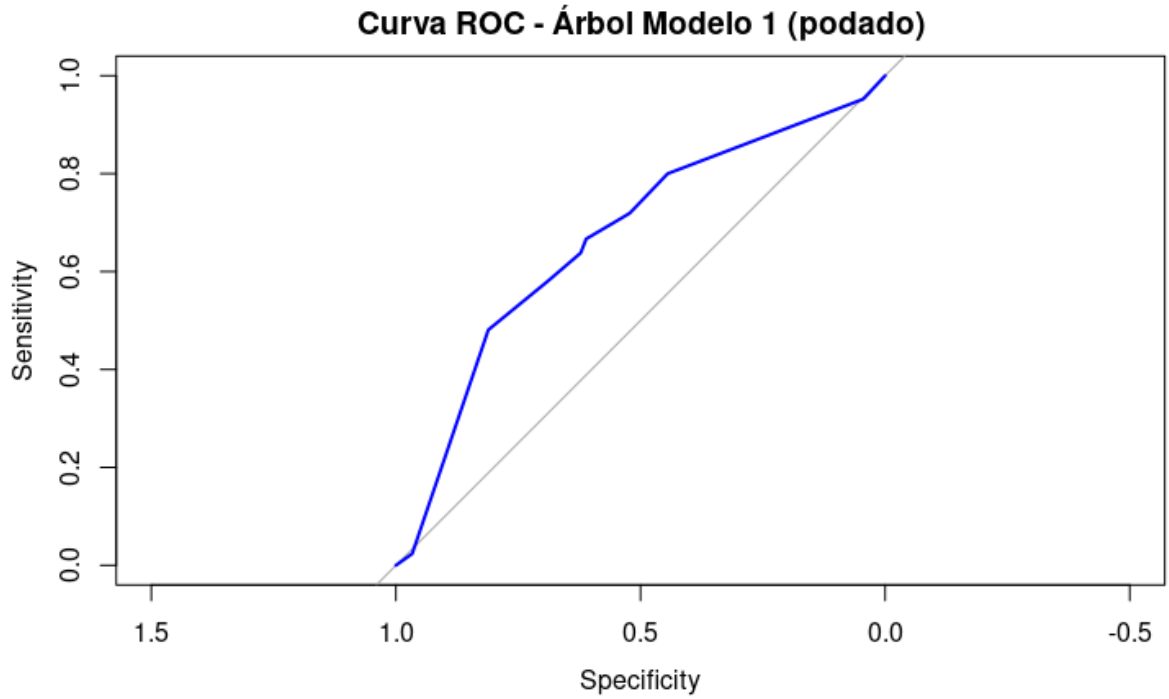
Además de la matriz de confusión, el modelo se evaluó con indicadores que reflejan su capacidad predictiva y el equilibrio al clasificar buenos y malos pagadores. El coeficiente $Kappa = 0.2508$ indica un acuerdo moderado entre las predicciones y los valores reales, descontando el efecto del azar, por lo que el árbol tiene una capacidad razonable, aunque mejorable.

El $p - value$ de la prueba de $McNemar = 0.4655$ (mayor a 0.05) sugiere que no existe una diferencia estadísticamente significativa entre los errores de clasificación de ambas clases, es decir, el modelo no presenta un sesgo marcado hacia una categoría. En cuanto al desempeño por clase, la sensibilidad = 0.4444 muestra que el modelo identifica correctamente el 44.4% de los malos pagadores, mientras que la especificidad = 0.8000 indica que reconoce el 80% de los buenos pagadores, evidenciando mejor desempeño en la detección de clientes confiables. Adicionalmente, el valor predictivo positivo = 0.4878 señala que cuando el modelo predice "malo" acierta en 48.8% de los casos, y el valor predictivo negativo = 0.7706 indica que cuando predice "buen pagador" acierta en 77.1%. La prevalencia = 0.3000 confirma que el 30% del conjunto de prueba corresponde realmente a "mal pagador", mientras que la tasa de detección = 0.1333 muestra que solo el 13.3% del total de observaciones fueron correctamente identificadas como "Bad". Finalmente, la exactitud balanceada = 0.6222, al promediar sensibilidad y especificidad, refleja un desempeño moderado y relativamente equilibrado, aunque con margen de mejora en la identificación de malos pagadores.

En conjunto, estos resultados muestran que el árbol de decisión tiene un desempeño razonable y balanceado, priorizando la identificación de los clientes confiables por encima de los de alto riesgo.

Figura 9

Curva ROC -Árbol Modelo 1 podado.



Nota: Elaboración propia. La figura presenta la curva ROC correspondiente al árbol de decisión podado (Modelo 1).

En la **Figura 9** se observa a la curva ROC mostrando la relación entre la sensibilidad (tasa de verdaderos positivos) y 1 - especificidad (tasa de falsos positivos) para diferentes umbrales de clasificación del modelo. La línea diagonal gris representa un modelo aleatorio sin poder discriminante, mientras que la curva azul representa el desempeño real del árbol de decisión.

Se observa que la curva se encuentra por encima de la diagonal, lo que indica que el modelo logra diferenciar en cierta medida entre clientes buenos y malos.

En conjunto, el árbol de decisión podado muestra un buen equilibrio entre interpretabilidad y rendimiento, aunque con menor precisión global en comparación con el modelo logístico. No obstante, su principal ventaja radica en la facilidad de interpretación de las reglas de decisión, que permite identificar de manera visual los factores que incrementan o reducen el riesgo crediticio (por ejemplo, el historial de pagos, la duración del crédito o el estado de la cuenta bancaria).

El valor obtenido para el Área Bajo la Curva (AUC) fue de 0.667, lo que indica que el modelo tiene una capacidad discriminante moderada.

Esto significa que el árbol de decisión logra distinguir correctamente entre un cliente con buen comportamiento crediticio y uno de alto riesgo aproximadamente en el 66.7% de los casos.
Demostrando, que el modelo tiene cierta habilidad predictiva y generaliza mejor que un modelo aleatorio (AUC = 0.5). La ligera curvatura por encima de la diagonal (vista en la figura ROC) confirma esta capacidad.
Esto es, el AUC = 0.667 sugiere que el árbol podado podría complementarse con otros métodos para mejorar la detección de clientes de alto riesgo sin sacrificar precisión global.

4.3.3 Árbol de decisión: modelo reducido

En este segundo modelo, se construyó un árbol de decisión utilizando únicamente las variables estadísticamente significativas identificadas en el modelo logístico reducido.

Tabla 14.

Variables del Árbol 2

Variable utilizada en el árbol	**Descripción breve**
Amount	Monto del crédito solicitado.
CheckingAccountStatus.0.to.200	Cliente con saldo bancario entre 0 y 200.
CheckingAccountStatus.lt.0	Cliente con saldo negativo en su cuenta.
CreditHistory.NoCredit.AllPaid	Historial crediticio limpio o sin deudas previas.
CreditHistory.ThisBank.AllPaid	Buen historial con la institución actual.
Duration	Duración del crédito en meses.
SavingsAccountBonds.100.to.500	Ahorros o bonos entre 100 y 500.
SavingsAccountBonds.lt.100	Ahorros menores a 100.

Nota: Elaboración propia. Las variables incluidas corresponden al subconjunto seleccionado a partir del modelo de regresión logística reducido.

En la **Tabla 14** se muestran las variables que el Árbol 2 utiliza para realizar sus divisiones, por lo que pueden interpretarse como los factores que el modelo considera más útiles para distinguir entre buenos y malos pagadores. En conjunto, predominan variables de capacidad financiera y liquidez

(monto, duración, estatus de cuenta corriente y nivel de ahorros) y variables de comportamiento histórico (historial crediticio con el banco o historial limpio), lo que sugiere que el árbol basa su clasificación principalmente en el perfil financiero del solicitante y en señales de disciplina de pago previa.

Tabla 15.

Resultados del análisis de complejidad del árbol.

CP	Número de divisiones (n-split)	Error relativo (Rel. Error)	Error cruzado (X-error)	Error estándar (X-std)
0.052381	0	1.00000	1.00000	0.0577
0.023810	2	0.89524	0.95238	0.0569
0.011905	8	0.73810	0.88571	0.0557
0.011111	10	0.71429	0.85238	0.0550
0.010000	13	0.68095	0.85714	0.0551

Nota: Elaboración propia. Los parámetros de complejidad del árbol se obtuvieron mediante validación cruzada para la selección del valor óptimo de poda (cp).

Se observa en la **Tabla 15** que el parámetro CP controla el tamaño del modelo: al disminuir, el árbol incorpora más divisiones y puede ajustarse mejor a los datos de entrenamiento, aunque con riesgo de sobreajuste. El nsplit muestra el crecimiento progresivo del árbol, pasando de 0 hasta 13 divisiones. Conforme aumenta la complejidad, el error relativo (Rel. Error) disminuye de 1.00 a 0.68, lo que indica un mejor ajuste en entrenamiento. Por su parte, el error de validación cruzada (Xerror) baja de 1.00 a 0.85, sugiriendo que el modelo generaliza de manera razonable sin señales fuertes de sobreajuste. Finalmente, la Xstd (≈ 0.055) se mantiene estable y baja, lo que indica consistencia en el desempeño del modelo entre particiones de validación.

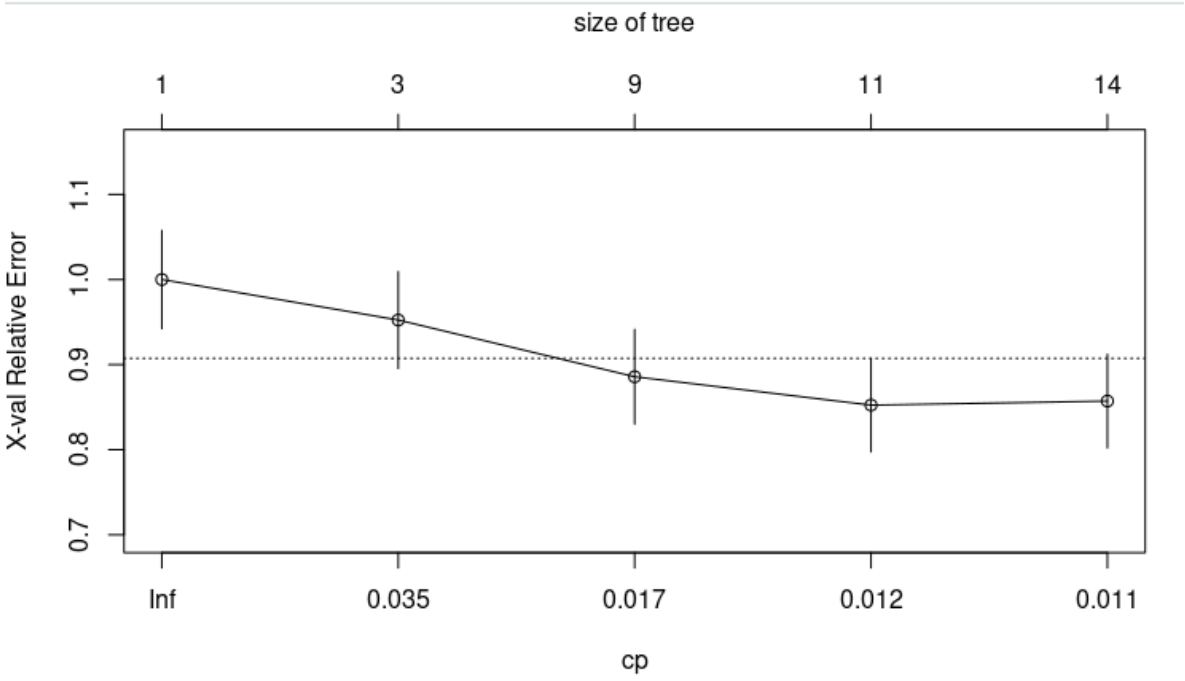
Podemos concluir entonces, que el análisis de complejidad muestra que el árbol reduce significativamente el error relativo y el de validación cruzada conforme se agregan divisiones, hasta estabilizarse alrededor de un $CP \approx 0.011$.

En ese punto, el modelo alcanza un equilibrio entre precisión y simplicidad, lo que indica que no es necesario incrementar la complejidad más allá de 10–13 divisiones.

Por tanto, el valor óptimo de poda (cp óptimo ≈ 0.011) define un árbol eficiente, interpretable y con buena capacidad de generalización.

Figura 10.

Complejidad del árbol de decisión reducido



Nota: Elaboración propia. La figura ilustra la relación entre el parámetro de complejidad (cp) y el error de validación cruzada del árbol de decisión reducido.

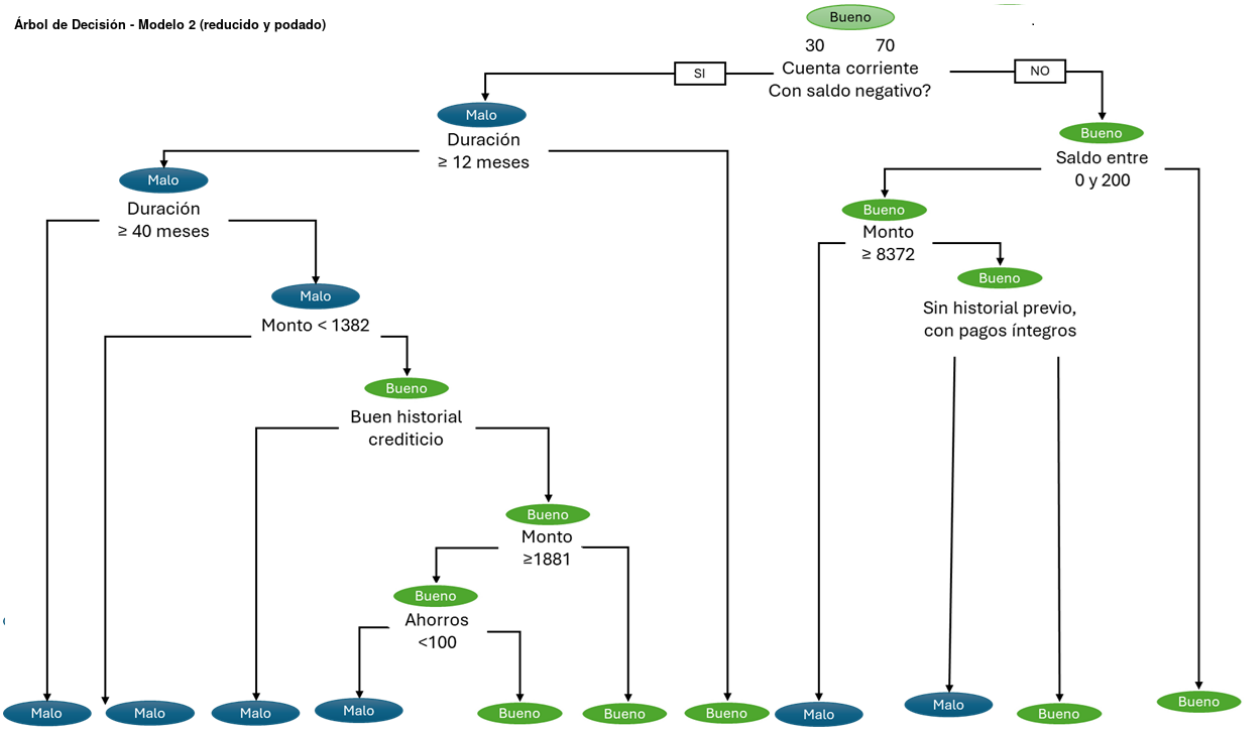
Podemos observar en la **Figura 10** una relación entre el parámetro de complejidad (cp) y el error relativo promedio por validación cruzada del árbol de decisión reducido. En el eje X se representa el cp, donde valores altos generan árboles más simples (menos divisiones) y valores bajos producen árboles más complejos (mayor profundidad). En el eje Y se presenta el X-val Relative Error, que refleja qué tan bien generaliza el modelo a datos no vistos. Las barras verticales corresponden a los intervalos de confianza del error estimado; su estabilidad sugiere baja variabilidad y una estimación confiable. La línea punteada horizontal marca el mínimo error de validación cruzada y sirve como referencia para seleccionar el cp óptimo mediante la regla del 1-SE, que privilegia el modelo más simple dentro del margen del error mínimo. En cuanto al comportamiento, el error disminuye de forma pronunciada hasta aproximadamente $cp = 0.017$ y después se estabiliza, alcanzando su mínimo alrededor de $cp \approx 0.011$, lo cual corresponde a un árbol de aproximadamente 10 a 14 divisiones, consistente con lo observado en la tabla de complejidad.

El análisis del gráfico confirma que el árbol reducido mantiene una estructura eficiente y estable, con un rendimiento comparable al modelo completo, pero con menor complejidad y mejor interpretabilidad.

El cp óptimo (≈ 0.011) es el punto donde el modelo logra la mejor capacidad predictiva con el menor nivel de complejidad necesario, reafirmando la validez de la poda aplicada.

Figura 11

Árbol de decisión del modelo reducido.



Nota: Se adaptó el árbol que se generó en el Scribd para una mejor visualización.

La **Figura 11** muestra la estructura final del árbol de decisión después de aplicar la poda utilizando el valor óptimo de cp (≈ 0.011), identificado previamente en el análisis de complejidad. Este árbol emplea únicamente las variables más relevantes derivadas del modelo logístico, buscando un equilibrio entre simplicidad, interpretabilidad y capacidad predictiva. De manera general, el nodo raíz corresponde a CheckingAccountStatus.lt.0, es decir, si el solicitante presenta saldo negativo en su cuenta corriente; cuando el valor es 1 (saldo negativo), el árbol clasifica una mayor proporción de casos como "Buen pagador", lo cual puede sugerir que un descubierto bancario leve no necesariamente implica mayor riesgo, mientras que cuando el valor es 0 el modelo continúa ramificándose con base en variables adicionales. Las siguientes divisiones se explican

principalmente por la duración del crédito (plazos más largos, por ejemplo ≥ 40 meses, tienden a asociarse con mayor probabilidad de incumplimiento), el monto del préstamo (créditos de mayor valor muestran una tendencia hacia mayor riesgo), el historial crediticio (clientes con todos los créditos previos pagados suelen clasificarse como "Buenos") y el nivel de ahorro (montos bajos, como menos de 100 DM, se relacionan con mayor riesgo). En conjunto, el árbol resultante presenta una estructura moderada y coherente desde el punto de vista financiero, y sus nodos terminales muestran predominancia de la clase "Buen pagador", consistente con la distribución original del conjunto de datos (≈70% de buenos pagadores).

De esta manera se concluye que el Modelo 2 (reducido y podado) logra una mejor interpretabilidad sin una pérdida considerable de precisión.

Su estructura más compacta lo convierte en una herramienta útil para decisiones crediticias explicables, permitiendo identificar con claridad los factores que distinguen a los clientes de bajo y alto riesgo.

4.3.4 Análisis de desempeño: matriz de confusión e indicadores

Tabla 16.

Predicción / Realidad Árbol 2.

	Mal pagador	**Buen pagador**
Predijo Mal pagador	37	36
Predijo Buen pagador	53	174

Nota: Elaboración propia. La matriz de confusión corresponde al desempeño del árbol de decisión reducido (Árbol 2) evaluado sobre el conjunto de prueba.

Esta **Tabla 16** representa a la matriz de confusión, en la cual se observa que el modelo identifica correctamente 37 casos como "Mal pagador" (verdaderos positivos) y 174 casos como "Buen pagador" (verdaderos negativos). Sin embargo, también clasifica 36 solicitantes como "Mal pagador" cuando en realidad eran "Buenos" (falsos positivos) y deja sin detectar 53 casos de alto riesgo al clasificarlos como "Buen pagador" (falsos negativos). En conjunto, el modelo acierta principalmente en la clase de buenos pagadores, aunque presenta dificultades para distinguir y detectar de forma consistente a algunos clientes de riesgo alto.

Tabla 17.

Indicadores de rendimiento.

Métrica	Valor	Interpretación
Accuracy (Exactitud)	0.7033	El modelo clasifica correctamente el 70.3% de los casos totales.
Kappa	0.2534	Indica un acuerdo leve-moderado entre las predicciones y los valores reales, superior al azar, pero con margen de mejora.
Mcnemar's Test p-valor = 0.0899	> 0.05	No hay evidencia significativa de sesgo sistemático entre clases; el modelo es equilibrado en sus errores.
Sensibilidad (Recall para clase "Mal pagador")	0.4111	El modelo identifica correctamente el 41.1% de los clientes realmente de alto riesgo.
Especificidad	0.8286	Detecta correctamente el 82.9% de los clientes buenos, mostrando una mejor capacidad para reconocer casos de bajo riesgo.
Valor predictivo positivo (Precisión para "Mal pagador")	0.5068	Cuando el modelo predice "Mal pagador", tiene un 50.7% de probabilidad de acertar.
Valor predictivo negativo	0.7665	Cuando predice "Buen pagador", acierta en el 76.6% de los casos.
Exactitud balanceada (Balanced Accuracy)	0.6198	Promedio entre sensibilidad y especificidad; indica un rendimiento global moderado.

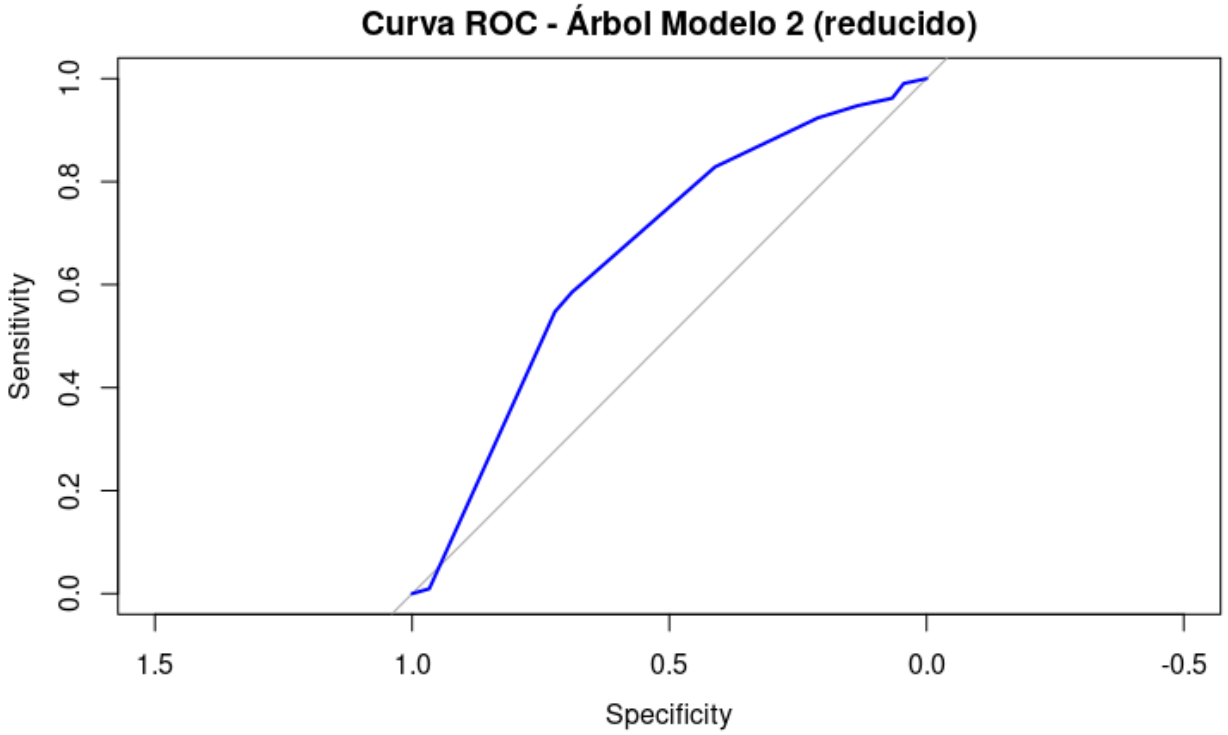
Nota: Elaboración propia. Los indicadores de rendimiento se calcularon a partir de las predicciones del árbol de decisión reducido (Árbol 2).

Se observa en la **Tabla 17** que el modelo reducido y podado mantiene una exactitud general del 70.3%, similar al árbol completo, pero con una estructura más sencilla y mejor interpretabilidad. Su alto valor de especificidad (0.83) muestra que es más confiable para identificar buenos clientes

que para detectar malos pagadores. El modelo logra un equilibrio razonable entre precisión y generalización, prioriza la estabilidad y la interpretabilidad del proceso crediticio.

Figura 12

Curva ROC – Árbol modelo 2 reducido.



Nota: Elaboración propia. La figura muestra la curva ROC correspondiente al árbol de decisión reducido (Modelo 2).

La **Figura 12** nos muestra a través de la Curva ROC el desempeño del modelo en términos de su capacidad de discriminación entre las clases buen pagador y mal pagador, a través de distintos umbrales de decisión.

Mostrando en el eje X (1 – especificidad) la proporción de falsos positivos, es decir, clientes buenos clasificados como malos, y en el eje Y (sensibilidad) la proporción de verdaderos positivos, es decir, malos pagadores correctamente identificados. La curva azul representa el desempeño del modelo para distintos puntos de corte, mientras que la diagonal gris indica el rendimiento de un clasificador aleatorio sin capacidad de discriminación. En este caso, la curva azul se mantiene por encima de la diagonal, lo que evidencia que el modelo sí tiene capacidad real para discriminar entre buenos y malos pagadores; además, su pendiente inicial pronunciada sugiere que logra identificar una parte importante de los casos de alto riesgo, aunque con cierta pérdida de precisión cuando se

busca una sensibilidad muy alta. En conjunto, la forma de la curva sugiere un rendimiento favorable del modelo.

En conclusión, la Curva ROC del Modelo 2 muestra un desempeño superior al del Modelo 1 (podado), al lograr una mejor separación entre clases sin indicios de sobreajuste. A pesar de mantener una estructura reducida y más interpretable, el modelo conserva su capacidad predictiva, lo cual respalda su utilidad práctica. El valor $AUC = 0.661$ representa la capacidad global del modelo para discriminar entre clientes con buen y mal historial crediticio: un AUC de 0.5 equivale a un desempeño aleatorio sin poder predictivo, mientras que un AUC entre 0.6 y 0.7 indica una capacidad de discriminación moderadamente útil, especialmente en contextos donde se prioriza la interpretabilidad; valores superiores a 0.8 se considerarían evidencia de un poder predictivo excelente.

Con un $AUC = 0.661$, el modelo reducido logra distinguir correctamente el 66% de las veces entre un cliente "Good" y uno "Bad". Lo cual, mantiene una buena coherencia con el modelo completo anterior ($AUC \approx 0.67$), mostrando que la reducción de variables no deteriora la capacidad de predicción. Concluyendo que, el Modelo 2 (reducido y podado) ofrece un equilibrio adecuado entre simplicidad, interpretabilidad y precisión.

4.3.5 Comparación de Modelos de Árbol de Decisión

Tabla 18.

Comparación de desempeño.

Métrica	**Modelo 1 (Completo y Podado)**	**Modelo 2 (Reducido y Podado)**	**Interpretación**
Exactitud	0.693	0.703	Ambos modelos logran una tasa de acierto similar, alrededor del 70%. El modelo reducido mejora la precisión.
Kappa	0.251	0.253	En ambos casos el acuerdo con la realidad es leve-moderado, indicando que los modelos superan el azar.

Sensibilidad	0.444	0.411	El modelo completo identifica un poco mejor a los clientes de alto riesgo.
Especificidad	0.800	0.829	El modelo reducido mejora su capacidad para reconocer a los clientes de bajo riesgo.
Valor Predictivo Positivo	0.488	0.507	El modelo reducido ofrece una mayor confiabilidad al predecir casos “Mal pagador”.
Exactitud balanceada	0.622	0.620	Ambos modelos muestran un desempeño similar.
AUC	0.667	0.661	Los dos modelos poseen una capacidad de discriminación entre clases con diferencia mínima.
Complejidad	Mayor	Menor	El modelo reducido tiene una estructura más sencilla y fácil de interpretar, con menor riesgo de sobreajuste.

Nota: Elaboración propia con base en el desempeño del árbol de decisión completo podado (Modelo 1) y del árbol de decisión reducido podado (Modelo 2) a partir de métricas de clasificación.

En términos de precisión general, la **Tabla 18** muestra que ambos modelos exhiben una exactitud cercana al 70%, lo que sugiere un desempeño adecuado dada la naturaleza del problema de clasificación crediticia. Respecto a la capacidad de detección, el Modelo 1 demostró ser ligeramente superior al identificar clientes de alto riesgo. Sin embargo, el Modelo 2 (reducido) ofrece una simplicidad e interpretabilidad significativamente mayor, ya que reduce la complejidad del árbol e incluye únicamente variables significativas del análisis logístico. Al conservar casi el mismo nivel de desempeño que el Modelo 1, se consolida como una alternativa más eficiente y sencilla de explicar. Finalmente, en cuanto a la capacidad discriminante (AUC), ambos modelos validan una capacidad de separación entre clases con valores de 0.667 y 0.661, respectivamente.

En conclusión, el Modelo 2 en su versión reducida y podada, resulta ser la alternativa más adecuada para los procesos de interpretación y toma de decisiones. Esto se debe a que reduce

significativamente la complejidad inherente al modelo original sin que ello implique sacrificar la precisión en los resultados. Además, mantiene una discriminación aceptable entre los clientes considerados "buenos" y "malos" y, de manera crucial, facilita la interpretación de reglas de decisión que son directamente aplicables y útiles en el contexto crediticio.

4.4 Bosques Aleatorios.

El modelo de Bosque Aleatorio se construyó utilizando el conjunto de datos completo de GermanCredit, con el objetivo de clasificar a los solicitantes de crédito en dos categorías: “Good” y “Bad” correspondientes a “Buen” y “Mal” pagador, en función de su comportamiento crediticio. A diferencia de los modelos previamente analizados (regresión logística y árbol de decisión), este enfoque metodológico no requiere dividir la base en conjuntos de entrenamiento y prueba, ya que el propio algoritmo calcula internamente un error de validación (Out-of-Bag, OOB), el cual actúa directamente como medida de desempeño del modelo. El modelo de Bosque Aleatorio fue ajustado con los siguientes parámetros principales: el número de árboles fue de 500, y el número de variables probadas en cada división se estableció en 7, un valor calculado siguiendo la regla de $\sqrt{p}$, donde p representa el número total de predictores. Este enfoque metodológico se fundamenta en la utilización de múltiples árboles de decisión independientes, cada uno entrenado sobre una muestra aleatoria tanto de los datos como de las variables. Este proceso permite reducir significativamente la varianza y, en consecuencia, mejorar la capacidad de generalización del modelo al enfrentarse a nuevos casos de clasificación. La evaluación interna del modelo arrojó la siguiente matriz de confusión, cuyos resultados se presentan en la Tabla 19.

Tabla 19

Matriz de confusión (evaluación interna OOB)

	Predicha “Bad”	**Predicha “Good”**	**Error por clase**
Mal pagador	120	180	60.0%
Buen pagador	57	643	8.1%

Nota: Elaboración propia, la matriz de confusión corresponde a la evaluación interna del Bosque Aleatorio mediante Out-of-Bag, OOB.

La **Tabla 19** resume el desempeño del modelo al clasificar los casos en las categorías clientes con mal historial crediticio y clientes con buen historial. El análisis de la matriz revela dos componentes clave: los aciertos y los errores. En la diagonal principal de la matriz, se localizan los casos que fueron correctamente clasificados: se identificaron de manera precisa 120 observaciones etiquetadas como "Mal pagador" y 643 casos de clientes "Buenos" fueron clasificados correctamente. Por otro lado, los valores que se encuentran fuera de la diagonal representan los errores de clasificación: se identificaron 180 casos de "Mal pagador" que fueron incorrectamente clasificados como "Buenos", y 57 clientes "Buenos" fueron erróneamente clasificados como "Mal pagador".

El error por clase ofrece una medida más precisa al reflejar la proporción de observaciones mal clasificadas dentro de cada grupo. Para la clase de "Mal pagador", el error alcanza el 60%, lo que implica que el modelo logra detectar correctamente solo el 40% de los clientes de alto riesgo. En contraste, para la clase de "Buen pagador", el error es considerablemente menor, con un 8.1%; es decir, el modelo identifica correctamente alrededor del 91.9% de los clientes con un buen comportamiento crediticio.

El error Out-of-Bag (OOB) estimado para el modelo fue de 23.7%, lo que se traduce en una exactitud global aproximada del 76.3%. Esto indica que, en promedio, el modelo logra clasificar correctamente cerca de tres de cada cuatro observaciones. Sin embargo, este desempeño global oculta una mayor dificultad para identificar correctamente los casos con riesgo de crédito, donde se registró un elevado error del 60%.

A pesar de esta limitación, el desempeño del modelo para la clase "Buen pagador" es notablemente alto, con un error de tan solo 8.1%. Este comportamiento sugiere que el modelo es conservador en sus clasificaciones y tiende a minimizar los falsos positivos (clientes de riesgo clasificados erróneamente como buenos).

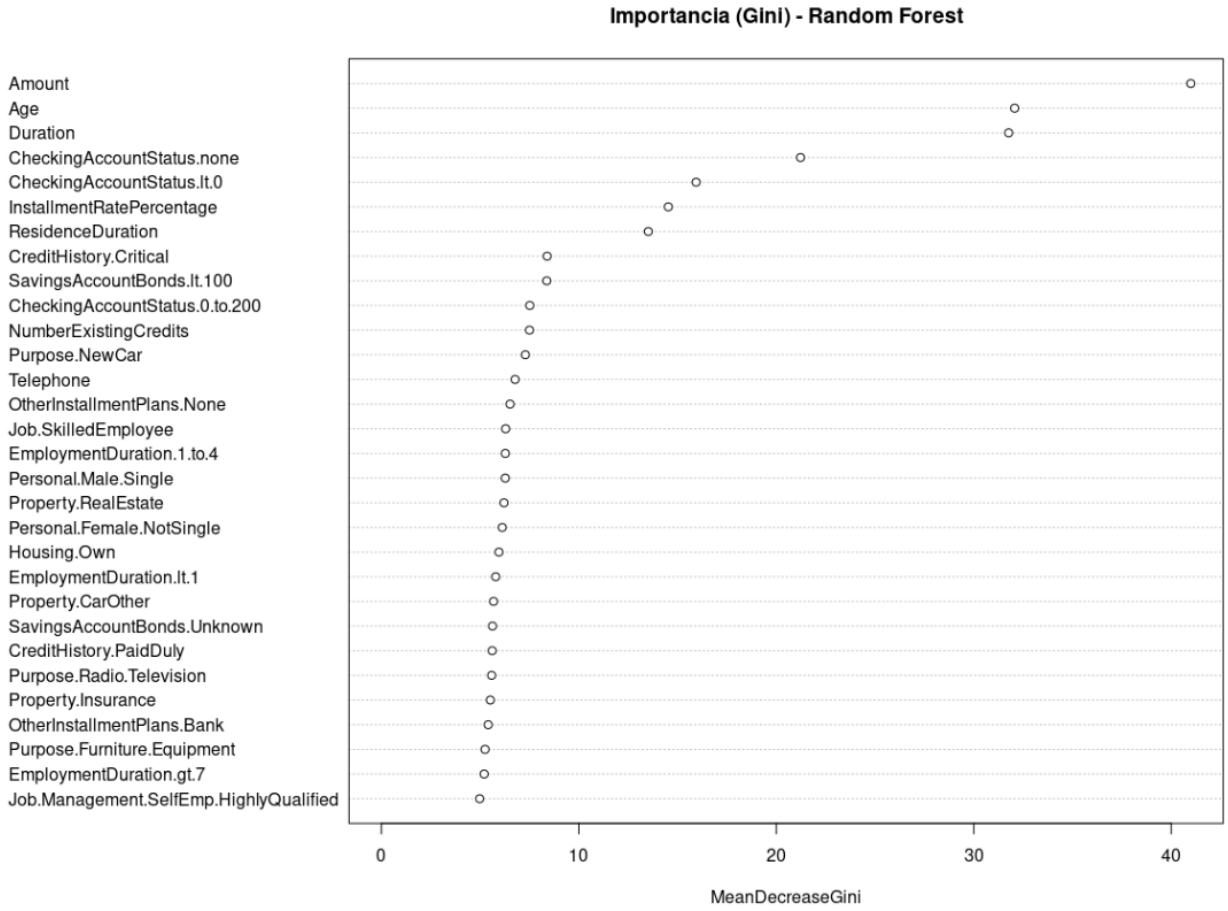
En síntesis, el modelo de Bosque Aleatorio consigue un buen equilibrio entre precisión y robustez. No obstante, su capacidad para detectar clientes de alto riesgo es el área de oportunidad más clara, la cual podría mejorarse mediante el ajuste de los parámetros del modelo o la modificación del umbral de clasificación. La ventaja principal del modelo de Bosques Aleatorios frente a los modelos anteriores radica en su menor tendencia al sobreajuste (overfitting) y en la estabilidad de sus predicciones, propiedades que se consiguen gracias a la utilización de múltiples árboles de

decisión. El modelo de Bosque Aleatorio alcanzó una exactitud OOB de 0.763, equivalente al 76.3% de aciertos.

Esta métrica se calcula automáticamente durante el entrenamiento, dado que cada árbol individual se construye con una muestra aleatoria del conjunto de datos y luego se evalúa su desempeño sobre las observaciones no utilizadas (Out – of – Bag). Este resultado final indica que el modelo mantiene un buen nivel de generalización, logrando evitar el sobreajuste y mostrando una capacidad predictiva sólida sin la necesidad de dividir previamente los datos en conjuntos de entrenamiento y prueba.

Figura 13

Importancia de variables.



Nota: Elaboración propia. La importancia de las variables se calculó mediante la disminución promedio del índice de Gini en el Bosque Aleatorio.

La **Figura 13** muestra cuáles predictores contribuyen en mayor medida a la separación entre las clases de “Buen pagador” y “Mal pagador”, basándose en la disminución promedio del índice de Gini en los nodos del bosque.

Es importante notar que un mayor valor en el eje horizontal denota una mayor relevancia de la variable para el proceso de toma de decisiones del modelo.

Como se detalla en la Figura 13, las variables que poseen el mayor peso explicativo para el modelo son: el Monto solicitado (Amount), dado que los montos más elevados tienden a asociarse con una mayor probabilidad de riesgo crediticio; la Edad del solicitante (Age), cuya influencia sugiere que los clientes más jóvenes presentan historiales menos consolidados; y la Duración del crédito (Duration), ya que las extensiones de préstamo más largas suelen implicar un mayor riesgo de incumplimiento.

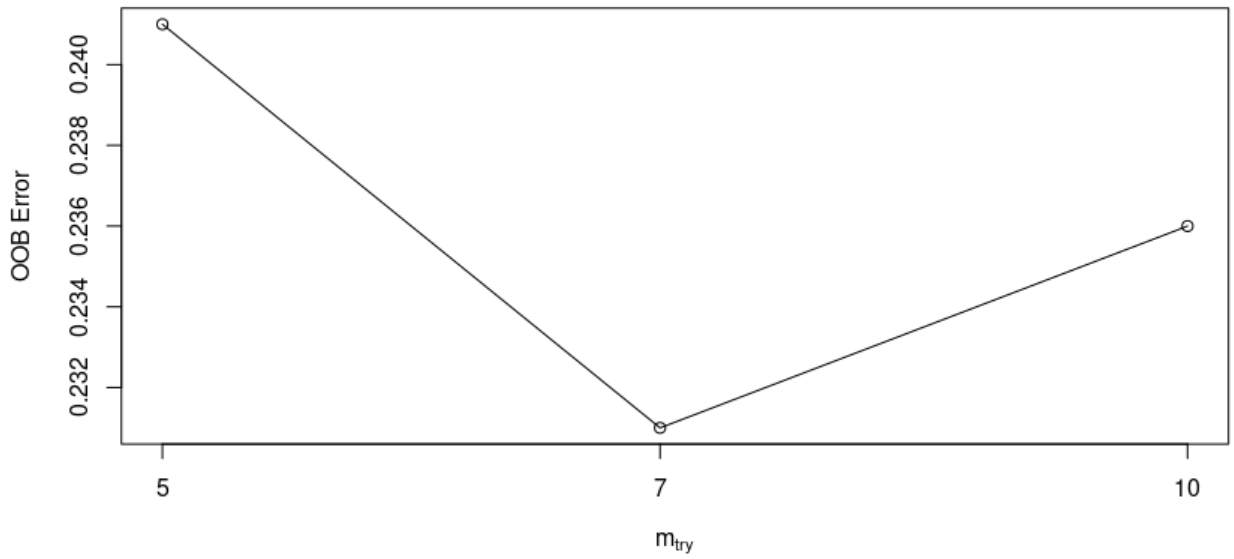
Adicionalmente, el Estado de la cuenta corriente (CheckingAccountStatus) es un predictor crucial, donde específicamente las categorías de saldo negativo y sin cuenta se asocian a un riesgo de crédito superior. Finalmente, la Tasa de pago en relación con los ingresos (InstallmentRatePercentage) también resulta relevante, puesto que su valor refleja directamente la capacidad de pago mensual del cliente. Otras variables que también resultaron relevantes, aunque con un menor peso dentro del modelo, incluyen el historial crediticio (CreditHistory), los ahorros disponibles (SavingsAccountBonds) y la residencia.

En general, el modelo otorga mayor importancia a aquellas variables financieras directamente relacionadas con la capacidad de endeudamiento y la estabilidad económica del cliente, lo cual es enteramente coherente con la lógica subyacente al riesgo crediticio.

Esto confirma que los factores monetarios y de historial financiero tienen un papel determinante en la predicción del comportamiento de pago, mientras que las variables sociodemográficas (como el género o el tipo de empleo) poseen una influencia considerablemente menor en este contexto.

Figura 14

Optimización del parámetro m_{try} en el modelo Bosque Aleatorio



Nota: Elaboración propia. La figura muestra la evaluación del error OOB para distintos valores del parámetro m_{try} en el modelo de Bosque Aleatorio.

Se realizó una búsqueda alrededor del valor por defecto $mtry = \sqrt{p} \approx 7$ con el fin de identificar el número óptimo de predictores evaluados en cada partición del bosque aleatorio. El procedimiento evaluó el error Out-of-Bag (OOB) para distintos valores de m_{try}:

- $m_{try} = 5 \rightarrow$ OOB = 24.1%
- $m_{try} = 7 \rightarrow$ OOB = 23.1% *(mejor valor)*
- $m_{try} = 10 \rightarrow$ OOB = 23.6%

Como se ilustró en la **Figura 14** el error OOB mínimo se alcanza en $m_{try} = 7$, lo cual confirma que la regla $\sqrt{p}$ era apropiada para este conjunto de datos.

Al comprar este resultado con el modelo inicial cuyo OOB fue de 23.7%, el ajuste arrojó un mejor OOB de 23.1%. Esta mejora ligera sugiere que el bosque ya se encontraba cerca de su configuración óptima, por lo que no se esperan ganancias sustantivas al variar m_{try} más allá de este rango.

En consecuente, se adopta $m_{try} = 7$ como valor final por su buen desempeño y simplicidad del modelo, al mantener un bajo error OOB sin incrementar la complejidad del bosque.
El ajuste final no solo valida la configuración base, sino que también respalda la estabilidad del modelo Bosque Aleatorio en este problema de clasificación.

4.4.1 Optimización del bosque aleatorio y selección de parámetros

El segundo modelo de Bosque Aleatorio se construyó fijando el parámetro $m_{try} = 7$, correspondiente al número óptimo de variables consideradas en cada división del árbol, y se utilizaron 500 árboles para garantizar la estabilización del error Out-of-Bag (OOB). Este ajuste permitió evaluar el desempeño del modelo utilizando la totalidad del conjunto GermanCredit, aprovechando el mecanismo de validación interna propio de los bosques aleatorios.
El error OOB estimado para este modelo fue de 23.7%, lo que equivale a una exactitud general aproximada del 76.3%. Este valor indica que, en promedio, el bosque aleatorio clasifica correctamente a más de tres cuartas partes de los solicitantes, mostrando un rendimiento superior al observado en los modelos previos como la regresión logística y los árboles de decisión individuales.

Tabla 20

Matriz de confusión

Clase Real	Predicho mal pagador	Predicho Buen pagador	Error
Mal pagador	120	180	0.60
Buen pagador	57	643	0.0814

Nota: Elaboración propia. La matriz de confusión corresponde al desempeño del modelo de Bosque Aleatorio, considerando la clasificación de clientes con buen y mal historial crediticio.
Los resultados de la **Tabla 20** detallan el desempeño del modelo. En ella se observa que el modelo identifica correctamente 120 casos de la clase “Mal pagador”, pero comete 180 errores al clasificar incorrectamente como “Buen pagador” a clientes que sí presentaban alto riesgo. Esto se refleja en un error por clase relativamente alto (60%), lo cual sugiere que el modelo aún tiene dificultades para detectar a la totalidad de los clientes de alto riesgo.

En contraste, el modelo presenta un excelente desempeño al clasificar a los clientes de la clase "Buen pagador", con un error por clase de solo 8.14%. El modelo reconoce correctamente a 643 de ellos y clasifica erróneamente solo a 57. Dado que aproximadamente el 70% de los solicitantes pertenecen a la clase "Buen pagador", la tendencia del Bosque Aleatorio es ser conservador en su clasificación y favorece la detección de clientes confiables sobre los de alto riesgo.

En términos de uso práctico, este modelo sería especialmente útil para minimizar falsos positivos; es decir, evitaría rechazar clientes que son financieramente solventes. Sin embargo, en aplicaciones de riesgo crediticio donde la prioridad es la detección exhaustiva del riesgo, podría ser necesario ajustar el umbral de decisión u optimizar métricas orientadas a maximizar la detección de la clase "Mal pagador" de clientes de alto riesgo.

Con el fin de identificar el número de árboles permite minimizar el error Out-of-Bag (OOB), se evaluó el rendimiento del Bosque Aleatorio utilizando diferentes tamaños de bosque, desde 10 hasta 500 árboles, con incrementos de 10 en 10. Esta estrategia permite observar cómo evoluciona el error de generalización conforme aumenta la complejidad del modelo y determinar el punto en el cual el bosque alcanza un desempeño estable y eficiente.

Tabla 21

Error OOB según número de árboles.

Número de árboles (ntree)	**Error OOB**	**Número de árboles (ntree)**	**Error OOB**	**Número de árboles (ntree)**	**Error OOB**	**Número de árboles (ntree)**	**Error OOB**
10	0.3181	**170**	0.236	**330**	0.237	**490**	0.238
20	0.276	**180**	0.235	**340**	0.237	**500**	0.237
30	0.255	**190**	0.239	**350**	0.237		
40	0.25	**200**	0.236	**360**	0.234		
50	0.246	**210**	0.238	**370**	0.233		
60	0.245	**220**	0.241	**380**	0.237		

70	0.242	**230**	0.241	**390**	0.236
80	0.246	**240**	0.24	**400**	0.232
90	0.247	**250**	0.238	**410**	0.238
100	0.247	**260**	0.244	**420**	0.237
110	0.24	**270**	0.241	**430**	0.234
120	0.246	**280**	0.244	**440**	0.236
130	0.242	**290**	0.242	**450**	0.236
140	0.242	**300**	0.237	**460**	0.235
150	0.244	**310**	0.236	**470**	0.234
160	0.243	**320**	0.241	**480**	0.235

Nota: Elaboración propia. La tabla presenta la evolución del error OOB del modelo de Bosque Aleatorio para distintos tamaños del bosque.

La **Tabla 21** resume los valores del error OOB para cada cantidad de árboles evaluada. Se observa una disminución pronunciada del error en las primeras iteraciones, pasando de 31.8% con 10 árboles a aproximadamente 24.2% con 70 árboles. Posteriormente, el error continúa reduciéndose ligeramente conforme se agregan más árboles.

El mínimo error OOB obtenido fue de 23.2%, alcanzado con 400 árboles. Esto indica que, a partir de esta cantidad, el modelo logra su mejor capacidad de generalización sin mostrar mejoras significativas al aumentar la cantidad hasta 500.

Este comportamiento es consistente con la teoría de Bosques Aleatorios: agregar más árboles reduce la varianza del modelo, pero dicho beneficio disminuye después de cierto punto debido a la estabilización del error.

El análisis del error OOB en función del número de árboles reveló tres fases clave en el comportamiento del modelo: entre 10 y 70 árboles el error disminuye rápidamente, lo que refleja el efecto positivo de incrementar la diversidad del bosque; entre 70 y 200 árboles el error se estabiliza alrededor de 24%–24.7%, presentando solo pequeñas fluctuaciones atribuibles a la

naturaleza estocástica del algoritmo; finalmente, el error mínimo (23.2%) se alcanza con 400 árboles, lo que representa el mejor compromiso entre complejidad y rendimiento predictivo.

Finalmente agregar más árboles después de los 400 no mejora el error, lo cual confirma que el modelo ha alcanzado su punto de estabilidad y que 400 árboles es el número óptimo para este conjunto de datos.

En conclusión, el número óptimo de árboles para este Bosque Aleatorio, bajo el criterio de minimizar el error OOB, es 400 árboles. A partir de este punto, el modelo presenta la mayor capacidad de generalización y un desempeño más robusto. Este resultado es crucial, ya que permite entrenar un bosque eficiente sin recurrir a una cantidad innecesariamente alta de árboles: al lograr la misma exactitud con menos árboles, se evita incrementar el tiempo de cómputo sin obtener mejoras en el desempeño predictivo.

Una vez confirmado que 400 árboles representan el número óptimo para minimizar el error Out-of-Bag (OOB), se procedió a entrenar nuevamente el Bosque Aleatorio utilizando dicho valor junto con el parámetro $m_{try} = 7$, previamente determinado como óptimo. Este modelo optimizado busca equilibrar rendimiento predictivo y eficiencia computacional.

El desempeño del bosque optimizado arrojó un error OOB de 23.2%, equivalente a una exactitud aproximada del 76.8%. Este valor representa una ligera mejora respecto al bosque construido inicialmente con 500 árboles, lo cual confirma que el modelo alcanza su mejor desempeño alrededor de los 400 árboles.

Tabla 22

Matriz de confusión generada a partir de las predicciones OOB

Clase Real	Predicho mal pagador	Predicho buen pagador	Error
Mal pagador	117	183	0.61
Buen pagador	49	651	0.07

Nota: Elaboración propia a partir de las predicciones OOB del modelo, utilizando RStudio.

Los resultados de la **Tabla 22** resumen la Matriz de Confusión generada a partir de las predicciones OOB del modelo optimizado. El modelo identifica correctamente 117 casos de la clase "Bad", pero, a pesar de la optimización, aún clasifica erróneamente 183 como "Good". Este comportamiento se traduce en un error del 61% dentro de la clase "Bad", una situación común en problemas con

desbalance de clases (donde la proporción de la clase "Bad" es menor al 30%) y donde los modelos suelen presentar dificultades para el reconocimiento de la clase minoritaria. En contraste, para la clase "Good", el desempeño es notablemente mejor, con un error de únicamente el 7%. El modelo logra clasificar correctamente a 651 clientes buenos y falla en solo 49 de ellos, lo cual confirma que el modelo está altamente optimizado para identificar solicitantes con un buen comportamiento crediticio.

En conjunto, el bosque optimizado mantiene un rendimiento robusto: clasifica de forma muy efectiva a los clientes “Good”, quienes constituyen la mayoría, y aunque presenta dificultades para identificar todos los casos “Bad”, su desempeño global es estable y superior al de los modelos previos. El uso conjunto de $m_{try} = 7$ y 400 árboles aseguran una configuración eficiente del modelo, maximizando su capacidad predictiva.

Tabla 23

Matriz de confusión correspondiente al Bosque Aleatorio optimizado

Clase Real	Predicho mal pagador	Predicho buen pagador	Error
Mal pagador	117	183	0.61
Buen pagador	49	651	0.07

Nota: Elaboración propia a partir de la matriz de confusión del Bosque Aleatorio optimizado, obtenida mediante validación OOB.

Los resultados de la **Tabla 23** (Matriz de confusión correspondiente al Bosque Aleatorio optimizado) permiten realizar la siguiente interpretación:

En la clase "Mal pagador" (alto riesgo), el modelo clasificó correctamente a 117 solicitantes de alto riesgo; sin embargo, clasificó erróneamente a 183 de ellos como "buenos". Esto se traduce en un error del 61% dentro de esta clase, mostrando que el modelo tiene dificultades para identificar la totalidad de los clientes con riesgo de incumplimiento. Este comportamiento es consistente con el desbalance del conjunto de datos, donde la clase "mal pagador" constituye solo una minoría (alrededor del 30%).

En contraste, el desempeño en la clase "Buen pagador" es considerablemente mejor. El modelo identificó correctamente a 651 clientes "buenos", presentando solo 49 errores. Esto equivale a un error del 7%, lo cual indica que el bosque aleatorio es altamente preciso al clasificar a los

solicitantes con buen comportamiento crediticio. En términos globales, la matriz de confusión confirma que el Bosque Aleatorio optimizado mantiene un alto nivel de exactitud general. Sin embargo, como ocurre comúnmente en problemas con clases desbalanceadas, muestra mayor eficacia para clasificar a los clientes "buenos" que a los clientes "malos". Aun con esta limitación, la capacidad predictiva del modelo es sólida y representa una mejora clara respecto a modelos previos evaluados.

4.4.2 Importancia de Variables del Bosque Aleatorio Óptimo.

Para identificar cuáles variables contribuyen en mayor medida a la clasificación entre solicitantes "buenos" y "malos", se analizó la importancia de predictores utilizando la métrica *MeanDecreaseGini*, generada a partir del bosque optimizado ($m_{try} = 7,\ n_{tree} = 400$). Esta métrica cuantifica cuánto reduce cada variable la impureza de Gini en los nodos de los árboles del bosque. En términos prácticos, valores más altos indican variables con mayor poder discriminante dentro del modelo.

Tabla 24.

Variables con mayor puntuación en MeanDecreaseGini.

Variable	**MeanDecreaseGini (unidades)**
Monto del crédito	40.98
Edad	32.05
Duración del crédito	31.98
Sin cuenta corriente	21.08
Cuenta corriente < 0	16.20
Porcentaje de cuota sobre ingreso	14.40
Duración de residencia	13.57

Nota: Elaboración propia con base en la importancia de variables medida mediante MeanDecreaseGini del modelo Bosque Aleatorio

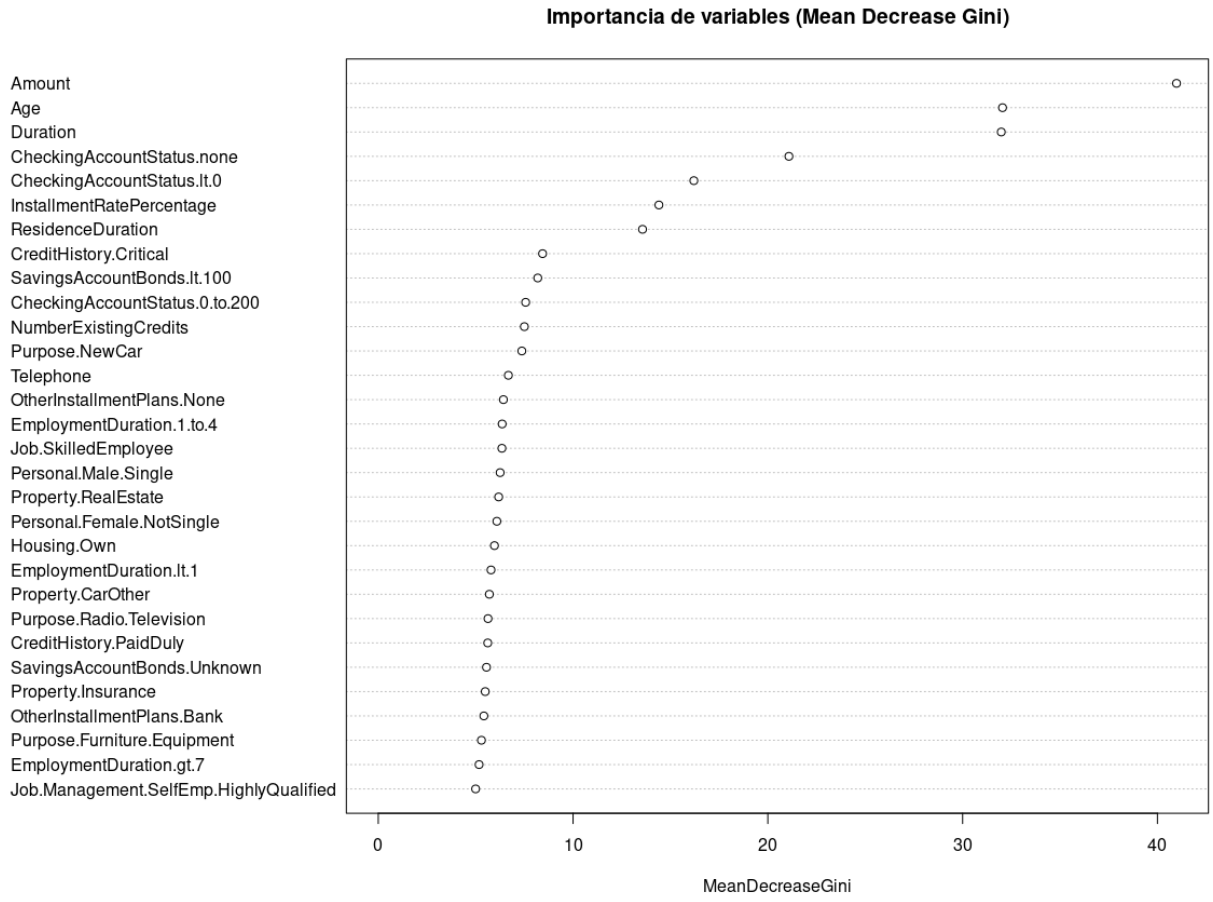
De acuerdo con la **Tabla 24** las variables que más contribuyen a la separación entre clientes de alto y bajo riesgo, y cuya relevancia coincide con la lógica financiera, son el monto solicitado, la edad, la duración del crédito, el estatus de cuenta corriente y, en menor medida, variables asociadas a capacidad/condiciones de pago y estabilidad.

En conjunto, estos factores sugieren que montos y plazos más altos, junto con señales de liquidez desfavorable en cuenta, tienden a incrementar el riesgo, mientras que características vinculadas a estabilidad financiera ayudan a discriminar mejor a los clientes. Además, variables de importancia moderada como *historial de crédito con pago puntual, historial de créditos totalmente pagados en el mismo banco, nivel de ahorro menor a 100 unidades monetarias, duración del empleo entre 1 y 4 años, propiedad de vivienda y empleo calificado* aportan información relevante, aunque no tan dominante, al reflejar historial de pago, respaldo financiero, activos y estabilidad laboral.

En contraste, predictores con importancia baja y cercana a cero, como *propósito del crédito para vacaciones, compra de electrodomésticos, mujer soltera y crédito para reentrenamiento*, aportan poca información adicional y su inclusión no modifica de forma significativa la capacidad del bosque para clasificar correctamente a los clientes.

En conclusión, el análisis confirma que el Bosque Aleatorio utiliza de manera intensiva los predictores relacionados con el monto solicitado, la duración del crédito, la edad, el estado de la cuenta corriente y el historial de estabilidad financiera, los cuales se convierten en los elementos clave para distinguir entre buenos y malos pagadores. Esta jerarquía de importancia es coherente con los criterios tradicionales del análisis crediticio y respalda la consistencia del modelo.

Figura 15



Nota: Elaboración propia. La figura muestra la jerarquía de importancia de las variables según el criterio MeanDecreaseGini del modelo Bosque Aleatorio

En la parte superior de la **Figura 15** se encuentran las variables con mayor impacto en la clasificación entre solicitantes "Good" y "Bad". Destacan especialmente Amount (Monto del crédito), Age (Edad) y Duration (Duración del crédito).

Estas variables presentan las puntuaciones más altas de *MeanDecreaseGini (importancia de variables)*, lo que evidencia que son los predictores más influyentes dentro del modelo. Su importancia coincide con criterios ampliamente utilizados en el análisis crediticio: montos elevados, mayor duración del crédito y la edad del solicitante están estrechamente relacionados con patrones de riesgo financiero.

Justo debajo aparecen otras variables con influencia considerable: CheckingAccountStatus.none (Sin cuenta corriente), CheckingAccountStatus.lt.0 (Cuenta corriente con saldo menor a cero),

InstallmentRatePercentage (Porcentaje de la cuota respecto al ingreso) y ResidenceDuration (Duración de residencia). Estas variables están directamente asociadas con la estabilidad financiera del cliente, liquidez inmediata y capacidad de pago.

El gráfico también muestra un conjunto de variables con importancia intermedia, como CreditHistory.Critical (Historial crediticio crítico), SavingsAccountBonds.lt.100 (Ahorros menores a 100 unidades monetarias), CheckingAccountStatus.0.to.200 (Cuenta corriente entre 0 y 200), Purpose.NewCar (Propósito: automóvil nuevo), EmploymentDuration.1.to.4 (Duración del empleo entre 1 y 4 años) y Job.SkilledEmployee (Empleado calificado).

Estas variables aportan información al modelo, aunque su impacto es menor en comparación con las variables más influyentes.

Finalmente, en la parte inferior del gráfico aparecen variables con importancia reducida, cuyos valores de MeanDecreaseGini son muy bajos o cercanos a cero, como Purpose.Vacation (Propósito: vacaciones), Purpose.DomesticAppliance (Propósito: electrodomésticos) y Personal.Female.Single (Mujer soltera). Estas variables apenas contribuyen a la capacidad predictiva del modelo, lo que sugiere que su influencia en la separación entre categorías es mínima.

Por lo cual, el gráfico confirma que el Bosque Aleatorio optimizado se basa principalmente en variables relacionadas con el monto solicitado, la edad del solicitante, el tiempo del crédito, y el estatus de la cuenta corriente, todas ellas características clave para estimar el comportamiento crediticio. La jerarquía observada en la importancia de variables coincide con los patrones esperados en modelos de riesgo crediticio y respalda la coherencia del modelo construido.

4.4.3 Identificación de la variable más importante (*MeanDecreaseGini*)

Con el objetivo de identificar el predictor que tiene mayor influencia en el desempeño del modelo, se determinó cuál presenta el valor más alto de *MeanDecreaseGini* dentro del Bosque Aleatorio optimizado ($m_{try} = 7,\ n_{tree} = 400$). Esta métrica refleja cuánto contribuye cada variable a mejorar la separación entre las clases en los distintos nodos de los árboles que conforman el bosque.

El análisis arrojó que la variable con mayor importancia es Monto con 40.981 unidades.

Este valor representa la reducción promedio de la impureza Gini atribuida a esta variable y confirma que el monto del crédito solicitado es el predictor más determinante para distinguir entre clientes “bueno” y “malo” dentro del modelo.

La alta importancia de Monto indica que el modelo utiliza esta variable de forma recurrente para realizar divisiones que incrementan la pureza de los nodos, por lo que el monto solicitado aporta información clave para inferir el riesgo crediticio. En términos prácticos, esto sugiere que montos más elevados tienden a asociarse con una mayor probabilidad de incumplimiento, en concordancia con criterios tradicionales de análisis financiero. Además, el hecho de que presente la puntuación más alta del conjunto (40.981) refuerza que esta variable es esencial para la capacidad discriminatoria del modelo.

En conclusión, el Bosque Aleatorio identifica al *Monto* como la variable más influyente en la predicción del comportamiento crediticio. Su predominancia refleja la importancia del nivel de endeudamiento solicitado por el cliente dentro del proceso de evaluación de riesgo.

4.4 Comparación final de modelos de predictivos.

Con base a lo obtenido a lo largo del documento, vamos a realizar un análisis de los resultados para con ello, poder determinar qué modelo predice de manera óptima.

Tabla 25

Comparación de desempeño entre modelos de clasificación.

Métrica	**Regresión Logística**	**Árbol 1 (Podado)**	**Árbol 2 (Reducido)**	**Bosque Aleatorio (500 árboles)**	**Bosque Aleatorio Óptimo (400 árboles)**
Exactitud	0.713	0.693	0.703	0.763	0.768 (mejor)
Sensibilidad (Mal pagador)	0.467	0.444	0.411	0.400 aprox. (dado 60% error)	0.390 aprox. (61% error)
Especificidad (Buen pagador)	0.819	0.800	0.829	0.919	0.930
Kappa	0.295	0.251	0.253	–	–

AUC	0.756	0.667	0.661	0.76 aprox. por OOB y comportamiento	0.77 aprox. mejor discriminación
Ventajas principales	Buena discriminación y análisis interpretativo; modelo explicable	Fácil de interpretar; reglas claras	Más simple y casi igual de preciso que el árbol 1	Mayor exactitud; bajo error OOB	Mejor exactitud y eficiencia; óptimo en OOB
Limitaciones	Sensibilidad moderada; no detecta muchos "Malos"	Sensibilidad baja; sobreajuste inicial	Menor sensibilidad que modelo 1	Difícil detectar "Malos"; sesgo por desbalance	Igual limitación para detectar "Malos", pero mejor rendimiento general
Conclusión	Modelo equilibrado, ideal para interpretación	Menos preciso que la logística	Modelo simple y estable	Muy buen desempeño general	MEJOR MODELO

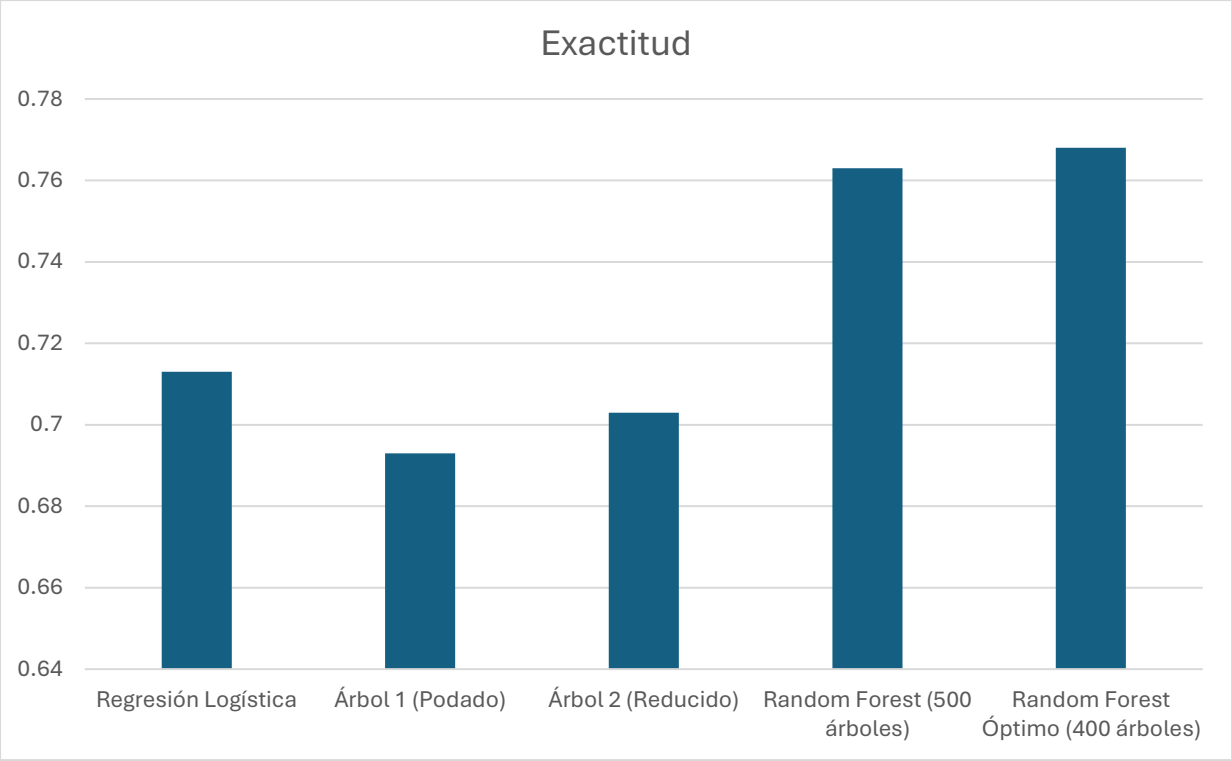
Nota: Elaboración propia a partir de los resultados obtenidos en RStudio para los modelos de regresión logística, árboles de decisión y bosque aleatorio.

De acuerdo con la **Tabla 14**, la comparación de modelos muestra diferencias claras en su capacidad de clasificación y en la forma en que cada uno aborda el problema del riesgo crediticio. La regresión logística ofrece un equilibrio adecuado entre desempeño y explicabilidad, destacando particularmente en la discriminación entre clases mediante su AUC (0.756). Los árboles de decisión, aunque sencillos de interpretar, presentan los niveles más bajos de exactitud y sensibilidad, lo que evidencia su limitada capacidad para generalizar frente a estructuras de datos complejas. En contraste, los modelos de Bosque Aleatorio alcanzan los mejores niveles de exactitud, tanto con 500 árboles como en su versión optimizada con 400 árboles, reflejando una mejora notable respecto a los demás algoritmos. No obstante, tienden a presentar una sensibilidad

reducida debido al desbalance hacia la clase “Good”, priorizando la correcta identificación de buenos pagadores sobre los malos. En conjunto, el Bosque Aleatorio Óptimo emerge como el modelo con mayor capacidad predictiva, mientras que la regresión logística destaca por su interpretabilidad analítica, posicionándose ambos como alternativas complementarias en el análisis de riesgo crediticio.

Figura 16

Comparación de modelos.



Nota: La figura es de auditoria propia, utilizando las exactitudes obtenidas de cada modelo.

La **Figura 16** muestra que el Bosque Aleatorio es el modelo con mayor exactitud, superando tanto a la regresión logística como a los árboles de decisión. El bosque optimizado presenta el mejor desempeño global, mientras que la regresión logística se posiciona en segundo lugar, y los árboles de decisión muestran el rendimiento más bajo entre los modelos analizados.

Capítulo 5. Conclusiones

5.1 Conclusión general

Esta investigación tuvo como propósito evaluar y comparar modelos supervisados de Machine Learning de clasificación aplicados al conjunto de datos *German Credit* para predecir si un solicitante será buen pagador (Good) o mal pagador (Bad), con el fin de aportar evidencia para su aplicación y adaptación al crédito automotriz en México como una alternativa para mejorar el análisis de posibles clientes.

Con base en el desarrollo metodológico y los resultados obtenidos, se concluye que la metodología propuesta es viable y pertinente para el contexto mexicano, ya que el flujo completo: exploración, selección de variables, entrenamiento, evaluación con métricas y comparación entre modelos permite construir herramientas objetivas para la toma de decisiones crediticias. En particular, se demuestra que los modelos pueden clasificar perfiles de riesgo con desempeño moderado a alto, siendo especialmente robustos al identificar a los clientes de bajo riesgo, lo cual es relevante en operaciones de crédito automotriz donde se busca mantener carteras sanas y eficientes.

5.2 Hallazgos clave del análisis descriptivo

El análisis descriptivo permitió identificar patrones consistentes con el comportamiento esperado en un entorno crediticio. En la comparación por clase se observó que los solicitantes clasificados como mal pagador tienden a presentar créditos de mayor duración y montos promedio mayores, lo cual sugiere una relación entre mayor exposición financiera y mayor probabilidad de incumplimiento.

Asimismo, se identificaron variables representativas vinculadas a estabilidad y comportamiento financiero como lo son: historial de pago puntual, estatus de cuenta corriente y activos como vivienda, las cuales se mantuvieron relevantes de forma consistente a lo largo de los modelos estimados.

Este hallazgo se considera importante para México, ya que en crédito automotriz el análisis tradicional también se apoya en dimensiones similares como capacidad de pago, estabilidad, historial y condiciones del crédito.

5.3 Conclusiones por modelo

5.3.1 Regresión logística: modelo interpretable y equilibrado.

La regresión logística mostró ventajas importantes por su capacidad explicativa, ya que permite interpretar el efecto de los predictores mediante Odds Ratios.

En los resultados, variables asociadas a mayor riesgo se reflejaron con $OR < 1$, destacando el saldo negativo o bajo en cuenta corriente como *CheckingAccountStatus.lt.0* con $OR = 0.154$ y *CheckingAccountStatus.0.to.200* con $OR = 0.241$; ambas con alta significancia, así como señales de menor respaldo financiero, como bajos niveles de ahorro.

En desempeño, el modelo logístico alcanzó una exactitud de 71.33% y un AUC de 0.7561, lo que representa una capacidad discriminativa buena para separar perfiles buenos de malos.

No obstante, la sensibilidad para detectar "mal pagador" fue de 46.67%, lo cual indica que el modelo identifica menos de la mitad de los casos de alto riesgo, por lo que su principal fortaleza se concentra en reconocer a los clientes confiables con una especificidad 81.90%.

En el modelo logístico reducido se mantuvo un desempeño similar con accuracy ~73% y $AUC = 0.7321$, lo cual sugiere que una especificación más compacta puede conservar buena capacidad predictiva con mayor simplicidad.

En términos aplicados, esto favorece la implementación práctica cuando se busca un modelo explicable que pueda ser documentado como "reglas" o criterios cuantitativos de decisión.

5.3.2 Árboles de decisión: reglas claras con desempeño moderado.

Los árboles de decisión aportaron interpretabilidad mediante reglas tipo "si–entonces", lo cual es valioso para justificar decisiones crediticias.

Sin embargo, su desempeño fue inferior al de la regresión logística y el bosque aleatorio.

El árbol 1 (completo, podado) alcanzó una exactitud aproximada de 0.693, $AUC = 0.667$ y sensibilidad de 0.444 para la clase de alto riesgo.

Por su parte, el árbol 2 (reducido, podado) obtuvo una exactitud de 0.703, $AUC = 0.661$ y sensibilidad de 0.411, con una ligera mejora en especificidad (0.829).

En conjunto, los árboles mostraron que son útiles como herramienta explicativa (por su claridad y facilidad de comunicación), pero con una capacidad discriminativa moderada comparada con modelos de ensamble. Aun así, su contribución es relevante como parte de una propuesta

metodológica para México, especialmente cuando una institución prioriza transparencia en criterios de decisión.

X.3.3 Bosque aleatorio: el mejor desempeño global del estudio.

El Bosque Aleatorio presentó el mejor desempeño general entre los modelos evaluados, apoyándose en su validación interna (OOB).

En su versión base, el error OOB fue de 23.7% es decir, exactitud ~76.3% y en el modelo optimizado el mínimo error OOB se logró con 400 árboles, alcanzando 23.2%, es decir, exactitud ~76.8%.

La matriz de confusión OOB mostró un comportamiento consistente: excelente identificación de Good (error ~7% en el optimizado), pero mayor dificultad para identificar completamente la clase Bad (error ~61%, equivalente a detectar ~39–40% de los Bad).

Esto refleja que el bosque aleatorio es fuerte para confirmar perfiles confiables y mantener estabilidad de clasificación.

En cuanto a variables, el análisis de importancia indicó que Amount, Age y Duration son los predictores más influyente, lo cual se considera pertinente para crédito automotriz respaldando la aplicabilidad conceptual del enfoque al mercado mexicano.

5.4 Comparación final y propuesta de uso para México

En la comparación conjunta de resultados se observa que el Bosque Aleatorio optimizado fue el modelo con mejor desempeño predictivo global, al alcanzar la mayor exactitud (0.768) dentro del estudio. Asimismo, la regresión logística presentó el mejor equilibrio entre desempeño e interpretabilidad, destacando por su $AUC = 0.756$ y por permitir la interpretación de los efectos mediante Odds Ratios. Por su parte, los árboles de decisión resultan valiosos por su capacidad de explicar el proceso de clasificación a través de reglas claras y comunicables, aunque su rendimiento fue el más bajo en comparación con los otros enfoques.

A partir de lo anterior, la propuesta para México es utilizar estos modelos de forma complementaria: el Bosque Aleatorio como herramienta de clasificación robusta con alto desempeño general, y la regresión logística como herramienta explicativa que facilite justificar el efecto de variables clave y respaldar decisiones. Esto es especialmente relevante en el crédito automotriz mexicano, donde el crecimiento del producto y la diversidad de características de los solicitantes exigen herramientas que reduzcan errores de aprobación o rechazo y fortalezcan la sostenibilidad de la cartera.

5.5 Cumplimiento de los objetivos de investigación

En relación con el **primer objetivo**, se cumplió al analizar la base German Credit mediante una revisión de su estructura, distribución de la variable objetivo y estadísticos descriptivos por clase, lo que permitió identificar patrones relevantes asociados al riesgo, especialmente en variables como duración y monto del crédito, así como en indicadores vinculados a liquidez, historial crediticio y ahorro. Este análisis, junto con el preprocesamiento y la selección de variables representativas, permitió preparar la información para su uso en modelos predictivos.

Respecto al **segundo objetivo**, se cumplió al construir y estimar modelos de clasificación supervisada (regresión logística, árboles de decisión y bosque aleatorio) para predecir el comportamiento de pago y clasificar a los solicitantes como buenos o malos pagadores. El desempeño de cada modelo se evaluó mediante métricas como matriz de confusión, exactitud, sensibilidad, especificidad y curva ROC/AUC, y en el caso del bosque aleatorio mediante error Out-of-Bag, lo cual permitió medir de forma consistente su capacidad predictiva.

Finalmente, el **tercer objetivo** se cumplió al comparar los modelos desarrollados en términos de desempeño e interpretabilidad, identificando ventajas y limitaciones de cada enfoque. A partir de esta comparación, se determinó que el bosque aleatorio optimizado presenta el mejor desempeño global, mientras que la regresión logística destaca por su interpretabilidad mediante odds ratios, y los árboles aportan reglas claras con rendimiento moderado. Con ello se logró sustentar una propuesta metodológica de aplicación al análisis del riesgo en crédito automotriz en México, orientada a fortalecer la evaluación de solicitantes y la toma de decisiones en la originación.

5.6 Conclusión final.

Esta tesis demuestra que la comparación de modelos de clasificación sobre *German Credit* permite construir un marco metodológico sólido que puede proponerse como alternativa para fortalecer el análisis de riesgo en crédito automotriz en México. La evidencia empírica respalda que los modelos de ensamble, particularmente el Bosque Aleatorio optimizado, alcanzan el mejor desempeño global.

Por lo tanto, la propuesta central para México consiste en replicar este flujo metodológico con información del mercado, incorporando variables típicas del crédito automotriz, por ejemplo:

enganche, CAT, valor del vehículo, relación pago/ingreso e historial en buró para ajustar el modelo, mejorar la precisión de selección de clientes y aportar una herramienta cuantitativa útil para instituciones financieras y SOFOMES en la originación de crédito automotriz.

Referencias

al, L. e. (2021). *Machine learning for credit scoring: Improving logistic regression with decision-tree extracted rules.* Obtenido de https://www.sciencedirect.com/science/article/pii/S0377221721005695?utm_source=chatgpt.com

Banco de México. (2024). *Indicadores basicos de crédito automotriz*. Obtenido de Banxico: https://www.banxico.org.mx/publicaciones-y-prensa/rib-creditos-automotrices/

Banco de México. (2025). *Reporte de indicadores básicos*. Obtenido de Banxico: https://www.banxico.org.mx/publicaciones-y-prensa/rib-creditos-automotrices/

Caparrós, G. (2023). *Supervised learning.* Obtenido de https://link.springer.com/

CONDUSEF. (2025). Obtenido de https://www.condusef.gob.mx/documentos/rcd/quien_es_quien/qq-cred-auto-junio-2025.pdf?utm_source

crédito, B. d. (s.f.). Obtenido de https://www.burodecredito.com.mx/personas-f%C3%ADsicas/productos/mi-score.html?utm_source=chatgpt.com

Crédito, B. d. (2025). *Mi Score: ¿qué es y cómo se calcula? Buró de Crédito.* Obtenido de https://www.burodecredito.com.mx/mi-score

Dialnet. (2023). *Aplicación de machine learning en la gestión de riesgo de crédito: revisión bibliográfica.* Obtenido de https://dialnet.unirioja.es/descarga/articulo/9039554.pdf

Economista, E. (11 de febrero de 2025). *Crédito automotriz cerró el 2024 con crecimiento cercano a 50%. El Economista.* Obtenido de https://www.eleconomista.com.mx/sectorfinanciero/credito-automotriz-cerro-2024-crecimiento-cercano-20250211-745959.html?utm_source

EL CEO. (2024). *Mexicanos piden más crédito a bancos; cartera vencida alcanza récod*. Obtenido de https://elceo.com/negocios/mexicanos-piden-mas-credito-a-bancos-cartera-vencida-alcanza-record/

José Carlos Trejo-García, M. Á.-G.-M. (s.f.). *Contaduría y Administración, 62(2).* doi:10.1016/j.cya.2017.01.003

Kotsiantis, S. B. (2007). *Supervised machine learning: A review of classification techniques.* Obtenido de https://www.informatica.si/index.php/informatica/article/view/148/140?utm_source

Lou, Y. Z. (2021). *Machine learning for credit scoring: Improving logistic regression with decision-tree-based ensembles. European Journal of Operational Research.* Obtenido de https://www.sciencedirect.com/science/article/abs/pii/S0377221721000813?via%3Dihub

México, B. d. (2023). Obtenido de https://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?sector=19&accion=consultarCuadro&idCuadro=CE79&locale=es

México, B. d. (2023). Obtenido de Indicadores básicos de crédito al consumo.: https://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?sector=19&accion=consultarCuadro&idCuadro=CE79&locale=es

Press, A. (2024). *Credit scoring using machine learning and deep learning-based models.* Obtenido de Data Science in Finance and Economics: https://www.aimspress.com/article/doi/10.3934/DSFE.2024009?utm_source

Press, A. (2024). *Data Science in Finance and Economics.* Obtenido de https://www.aimspress.com/article/doi/10.3934/DSFE.2024009

QuantInsti. (2022). *Gini index in machine learning: Formula and role in decision trees. Quantitative Finance Blog.* Obtenido de https://blog.quantinsti.com/gini-index/?utm_source

Reuters. (2025). *Bank of Mexico.* Obtenido de https://www.reuters.com/markets/rates-bonds/bank-mexico-cuts-interest-rate-flags-trade-economic-uncertainty-2025-03-27/?utm_source

Reyes Morales, M. A. (2022). *Modelo de puntuación crediticia para tarjeta de crédito en México: una aproximación logística. Ensayos. Revista de Economía.* doi:10.29105/ensayos41.1-2

Anexo

Variables

Índice	Variable	
1	Duration	Duración del crédito (meses)
2	Amount	Monto del crédito solicitado
3	InstallmentRatePercentage	% del ingreso destinado al pago mensual
4	ResidenceDuration	Años viviendo en la misma residencia
5	Age	Edad del solicitante
6	NumberExistingCredits	Número de créditos existentes
7	NumberPeopleMaintenance	Número de personas dependientes
8	Telephone	Tiene teléfono (1 = sí, 0 = no)
9	ForeignWorker	Es trabajador extranjero (1 = sí, 0 = no)
10	Class	Resultado del crédito: Good (bueno) / Bad (malo)
11	CheckingAccountStatus.lt.0	Saldo en cuenta corriente < 0 (en números rojos)
12	CheckingAccountStatus.0.to.200	Saldo entre 0 y 200 marcos
13	CheckingAccountStatus.gt.200	Saldo > 200 marcos
14	CheckingAccountStatus.none	No tiene cuenta corriente

15	CreditHistory.NoCredit.AllPaid	Sin historial de crédito o todo pagado
16	CreditHistory.ThisBank.AllPaid	Buen historial en este banco
17	CreditHistory.PaidDuly	Pagos realizados puntualmente
18	CreditHistory.Delay	Retrasos en el pago
19	CreditHistory.Critical	Crédito en situación crítica o moratoria
20	Purpose.NewCar	Préstamo para automóvil nuevo
21	Purpose.UsedCar	Préstamo para automóvil usado
22	Purpose.Furniture.Equipment	Préstamo para muebles o equipamiento
23	Purpose.Radio.Television	Préstamo para radio o televisión
24	Purpose.DomesticAppliance	Préstamo para electrodomésticos
25	Purpose.Repairs	Préstamo para reparaciones
26	Purpose.Education	Préstamo educativo
27	Purpose.Vacation	Préstamo para vacaciones
28	Purpose.Retraining	Préstamo para capacitación laboral
29	Purpose.Business	Préstamo para negocios
30	Purpose.Other	Otro propósito
31	SavingsAccountBonds.lt.100	Ahorros < 100 marcos
32	SavingsAccountBonds.100.to.500	Ahorros entre 100 y 500

33	SavingsAccountBonds.500.to.1000	Ahorros entre 500 y 1000
34	SavingsAccountBonds.gt.1000	Ahorros > 1000
35	SavingsAccountBonds.Unknown	No se conoce el saldo ahorrado
36	EmploymentDuration.lt.1	Antigüedad laboral < 1 año
37	EmploymentDuration.1.to.4	Antigüedad entre 1 y 4 años
38	EmploymentDuration.4.to.7	Antigüedad entre 4 y 7 años
39	EmploymentDuration.gt.7	Antigüedad > 7 años
40	EmploymentDuration.Unemployed	Desempleado
41	Personal.Male.Divorced.Seperated	Hombre divorciado/separado
42	Personal.Female.NotSingle	Mujer casada/no soltera
43	Personal.Male.Single	Hombre soltero
44	Personal.Male.Married.Widowed	Hombre casado o viudo
45	Personal.Female.Single	Mujer soltera
46	OtherDebtorsGuarantors.None	No tiene co-deudores ni garantes
47	OtherDebtorsGuarantors.CoApplicant	Tiene co-solicitante
48	OtherDebtorsGuarantors.Guarantor	Tiene garante
49	Property.RealEstate	Posee bienes raíces
50	Property.Insurance	Tiene seguros de vida como garantía
51	Property.CarOther	Tiene auto u otros bienes como garantía
52	Property.Unknown	No se conoce la propiedad

53	OtherInstallmentPlans.Bank	Tiene otros planes de pago en banco
54	OtherInstallmentPlans.Stores	Tiene planes de pago en tiendas
55	OtherInstallmentPlans.None	No tiene otros planes de pago
56	Housing.Rent	Vive en vivienda rentada
57	Housing.Own	Vive en vivienda propia
58	Housing.ForFree	Vive sin pagar renta
59	Job.UnemployedUnskilled	Desempleado/no calificado
60	Job.UnskilledResident	Trabajador no calificado residente
61	Job.SkilledEmployee	Empleado calificado
62	Job.Management.SelfEmp.HighlyQualified	Ejecutivo/empleador/profesional altamente calificado

Tabla de Odds Ratios (OR)

Variable	**Coeficiente**	**Odds_Ratio**	**IC_2.5**	**IC_97.5**	**p_valor**
OtherDebtorsGuarantors.CoApplicant	-1.899	0.15	0.04	0.567	0.0052
CheckingAccountStatus.lt.0	-1.873	0.154	0.088	0.267	0
ForeignWorker	-1.789	0.167	0.031	0.902	0.0375
CreditHistory.ThisBank.AllPaid	-1.66	0.19	0.063	0.578	0.0034
Job.UnemployedUnskilled	1.626	5.083	0.775	33.35	0.0903

Purpose.Repairs	-1.57	0.208	0.02	2.166	0.189
Purpose.Education	-1.46	0.232	0.028	1.918	0.1754
CheckingAccountStatus.0.to.200	-1.422	0.241	0.139	0.419	0
CreditHistory.NoCredit.AllPaid	-1.42	0.242	0.083	0.7	0.0089
Purpose.NewCar	-1.374	0.253	0.037	1.753	0.1641
EmploymentDuration.4.to.7	1.174	3.236	1.116	9.384	0.0307
SavingsAccountBonds.lt.100	-1.17	0.31	0.16	0.602	0.0005
SavingsAccountBonds.500.to.1000	-1.123	0.325	0.117	0.907	0.0319
Purpose.DomesticAppliance	-1.12	0.326	0.018	6.056	0.4523
Purpose.Business	-1.042	0.353	0.047	2.628	0.3092
OtherDebtorsGuarantors.None	-0.971	0.379	0.142	1.008	0.0519
Housing.Rent	-0.949	0.387	0.118	1.265	0.1163
CreditHistory.PaidDuly	-0.948	0.388	0.206	0.728	0.0032
SavingsAccountBonds.100.to.500	-0.876	0.416	0.176	0.988	0.0469
Purpose.Furniture.Equipment	-0.735	0.479	0.068	3.371	0.46
EmploymentDuration.lt.1	0.707	2.028	0.709	5.804	0.1875
CreditHistory.Delay	-0.686	0.503	0.215	1.176	0.113
Purpose.Retraining	0.666	1.946	0.065	58.282	0.7011
Personal.Male.Single	0.592	1.807	0.853	3.83	0.1225
Property.RealEstate	0.578	1.782	0.629	5.046	0.2766
OtherInstallmentPlans.Bank	-0.527	0.591	0.319	1.092	0.0931
EmploymentDuration.1.to.4	0.492	1.636	0.6	4.455	0.3359

Purpose.Radio.Television	-0.475	0.622	0.089	4.349	0.6322
Telephone	-0.444	0.642	0.388	1.062	0.0844
Purpose.UsedCar	0.434	1.543	0.198	12.014	0.6785
EmploymentDuration.gt.7	0.383	1.466	0.526	4.087	0.4642
Housing.Own	-0.338	0.713	0.231	2.202	0.5571
InstallmentRatePercentage	-0.326	0.722	0.577	0.904	0.0045
OtherInstallmentPlans.Stores	-0.308	0.735	0.252	2.141	0.572
CheckingAccountStatus.gt.200	-0.247	0.781	0.268	2.276	0.6504
Property.CarOther	0.22	1.246	0.464	3.344	0.6625
Personal.Male.Divorced.Seperated	-0.152	0.859	0.289	2.558	0.7848
NumberExistingCredits	-0.136	0.873	0.533	1.43	0.5897
Job.SkilledEmployee	0.121	1.129	0.559	2.279	0.7356
Property.Insurance	0.108	1.114	0.406	3.056	0.8339
Job.UnskilledResident	-0.078	0.925	0.388	2.204	0.8606
Personal.Female.NotSingle	0.061	1.063	0.511	2.21	0.8709
Duration	-0.03	0.97	0.948	0.993	0.0102
SavingsAccountBonds.gt.1000	0.028	1.029	0.257	4.114	0.9681
Age	0.017	1.017	0.994	1.04	0.1485
NumberPeopleMaintenance	0.017	1.017	0.547	1.893	0.9566
ResidenceDuration	-0.007	0.993	0.805	1.225	0.9474
Amount	0	1	1	1	0.0342

Variables significativas ($p < 0.05$)

Variable	p-value
(Intercept)	***
Duration	*
Amount	*
InstallmentRatePercentage	**
ForeignWorker	*
CheckingAccountStatus.lt.0	***
CheckingAccountStatus.0.to.200	***
CreditHistory.NoCredit.AllPaid	**
CreditHistory.ThisBank.AllPaid	**
CreditHistory.PaidDuly	**
SavingsAccountBonds.lt.100	***
SavingsAccountBonds.100.to.500	*
SavingsAccountBonds.500.to.1000	*
EmploymentDuration.4.to.7	*
OtherDebtorsGuarantors.CoApplicant	**

Otras traducciones.

Cat	Costo anual total
Machine learning (ml)	Rama de la inteligencia artificial que permite que las computadoras aprendan patrones a partir de datos sin que se les programe explícitamente cada regla.
Features	Variables de entrada
Target	Variable objetivo
Accuracy	Exactitud
Recall o tpr	Sensibilidad
specificity otnr	Especificidad
Roc (auc)	Área bajo la curva
Scorecard	Herramienta para medir desempeño usando indicadores.
Gini	Mide qué tan mezcladas están las clases dentro de un grupo de datos
Bootstrap	Remuestreo con reemplazo
Mean decrease gini	Mide la contribución promedio de cada variable a la reducción de impureza
Dummy	Variables binarias
N-split	Número de divisiones
Rel. Error	Error relativo
X-error	Error cruzado
X-std	Error estándar
Ntree	Número de árboles (ntree)