Modelo de Scoring Crediticio con Machine Learning para el Sector de la Construcción: Una Aplicación con Random Forest

Yilver Jesus Gonzalez Grueso
Jose Alejandro Angulo Torres

Corporación Universitaria Autónoma del Cauca
Facultad de Ciencias Administrativas, Contables y Económicas
Programa de Administración de Empresas
Popayán – Cauca
2026

Diseño de un modelo de Scoring Crediticio con machine learning para el sector de la construcción

Yilver Jesús González Grueso

José Alejandro Angulo Torres

Trabajo de Grado modalidad Investigación para optar el título profesional de Administrador de Empresas

Director

Jorge Eduardo Orozco Álvarez

Corporación Universitaria Autónoma del Cauca

Facultad de Ciencias Administrativas, Contables y Económicas

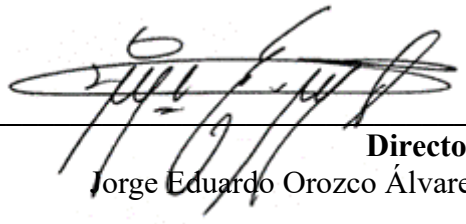
Programa de Administración de Empresas

Popayán – Cauca

2026

Nota de Aceptación

Una vez revisado y aprobado el documento final del trabajo de grado titulado ***Diseño de un modelo de Scoring con machine learning para el sector de la construcción***; realizado por los estudiantes Yilver Jesús González Grueso y José Alejandro Angulo, se autoriza para optar el título de profesionales en Administración de Empresas en la Corporación Universitaria Autónoma del Cauca.



Director.
Jorge Eduardo Orozco Álvarez

Jurado 1.
Herlan Alban Diaz

Jurado 2.
José Ignacio Vasquez Vargas

Popayán, 2026

Dedicatoria

Este trabajo de grado está dedicado con profundo agradecimiento a mis padres, Edinson González Grueso y Stella Grueso Aragón, quienes han sido mi apoyo incondicional, mi guía constante y el motor que me impulsó a perseverar en cada etapa de este proceso.

A ellos les debo no solo este logro académico, sino también los valores, la disciplina y la fortaleza que me permitieron culminar esta meta. Su esfuerzo y confianza han sido fundamentales en mi formación personal y profesional.

Yilver González

Dedico este trabajo de grado, en primer lugar, a mi madre, Ana Yolanda Torrez Bastidas, quien ha sido un pilar fundamental en mi vida. Gracias a su amor, su esfuerzo y su apoyo constante a lo largo de mi formación, hoy puedo alcanzar este logro. Su ejemplo de disciplina y compromiso ha sido, sin duda, una de mis mayores motivaciones.

También quiero dedicar este logro a la memoria de mi padre, José Nereo Angulo. Aunque ya no esté físicamente conmigo, su presencia sigue viva en mis recuerdos, en sus enseñanzas y en los valores que me dejó. Él ha sido una inspiración permanente para seguir adelante y no rendirme en la búsqueda de mis sueños.

José Angulo

Agradecimientos

Nosotros, Yilver González y José Angulo, queremos expresar nuestro más sincero agradecimiento a la Corporación Universitaria Autónoma del Cauca por habernos brindado la oportunidad de crecer tanto en lo académico como en lo profesional, dentro de un entorno caracterizado por la calidad, el compromiso y la excelencia. En esta institución no solo adquirimos conocimientos, sino que también fortalecimos nuestras habilidades y sentamos las bases de nuestro futuro profesional.

De manera especial, agradecemos al profesor Jorge Eduardo Orozco, director de este trabajo de grado, por su acompañamiento permanente, su orientación y sus valiosos aportes a lo largo de todo el proceso. Su experiencia, dedicación y rigurosidad fueron clave para el desarrollo y la culminación de este proyecto.

Finalmente, extendemos nuestro agradecimiento a todas aquellas personas que, de una u otra manera, hicieron posible la realización de este trabajo.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Resumen

Esta monografía incorpora el diseño de un modelo de scoring de crédito para una firma multinacional del sector de la construcción, haciendo técnicas de machine learning para llegar a la probabilidad de incumplimiento de sus clientes. Para ello, se cotejaron cuatro algoritmos: Árboles de Decisión, Random Forest, Gradient Boosting y Redes Neuronales, cada uno con sus propias características en términos de precisión, capacidad de generalización y manejo de grandes volúmenes de datos. La investigación incluyó la selección y evaluación de variables significativas a través del índice de Gini y técnicas de eliminación recursiva, con el objetivo de evitar el sobreajuste y fortalecer el modelo. Los resultados obtenidos permiten concluir que el algoritmo Random Forest presentó el mejor desempeño predictivo, alcanzando una precisión del 90% en la muestra de testeo, superando a los modelos de Árboles de Decisión (88%), Gradient Boosting y Redes Neuronales. Este hallazgo evidencia la capacidad superior del Random Forest para capturar relaciones no lineales y reducir el sobreajuste mediante la combinación de múltiples árboles de decisión. Asimismo, el modelo identificó 12 variables predictoras óptimas, destacándose el índice de endeudamiento como el factor más relevante en la explicación del fenómeno analizado. Esto sugiere que la estructura financiera y el nivel de apalancamiento constituyen elementos determinantes dentro del modelo predictivo. En consecuencia, se generó una herramienta de apoyo útil para fortalecer el proceso de evaluación del riesgo crediticio, contribuyendo a la optimización de las decisiones financieras de la compañía.

Palabras clave: Scoring de crédito, Machine learning, Impago, Modelos predictivos, Riesgo crediticio

Abstract

This monograph incorporates the design of a credit scoring model for a multinational firm in the construction sector, using machine learning techniques to predict the probability of default among its clients. To achieve this, four algorithms were compared: Decision Trees, Random Forest, Gradient Boosting, and Neural Networks, each with its own characteristics in terms of accuracy, generalization capability, and handling of large volumes of data. The research included the selection and evaluation of significant variables through the Gini index and recursive elimination techniques, with the objective of preventing overfitting and strengthening the model. The results obtained allow us to conclude that the Random Forest algorithm showed the best predictive performance, achieving 90% accuracy on the test sample, outperforming the Decision Tree model (88%), Gradient Boosting, and Neural Networks. This finding demonstrates the superior capacity of Random Forest to capture non-linear relationships and reduce overfitting by combining multiple decision trees. Additionally, the model identified 12 optimal predictor variables, with the debt ratio standing out as the most relevant factor in explaining the analyzed phenomenon. This suggests that financial structure and leverage level constitute determining elements within the predictive model. Consequently, a useful support tool was developed to strengthen the credit risk assessment process, contributing to the optimization of the company's financial decision-making.

Keywords: Credit scoring, Machine learning, Default, Predictive models, Credit risk

Introducción

La gestión financiera es muy importante para que una empresa pueda mantenerse y crecer en el tiempo. En este sentido, el análisis del riesgo financiero, especialmente el relacionado con la evaluación de créditos, juega un papel clave. Las decisiones que se toman al momento de otorgar un crédito pueden afectar directamente la estabilidad de la empresa, por lo que es necesario contar con herramientas que permitan analizar de forma adecuada la capacidad de pago de los solicitantes y así reducir el riesgo de incumplimiento.

Para hacer una buena evaluación del riesgo crediticio es fundamental contar con información confiable, usar correctamente los indicadores financieros y tener en cuenta variables del entorno. Sin embargo, muchas empresas todavía presentan dificultades porque utilizan modelos muy generales que no consideran las características específicas del sector al que pertenecen sus clientes. Esto puede llevar a tomar malas decisiones, aumentar la morosidad y afectar los resultados financieros.

Esta problemática se observa en la empresa objeto de estudio, una multinacional que financia productos a compañías del sector de la construcción. Actualmente, presenta un alto nivel de incumplimiento en los pagos, lo cual está relacionado con el uso de un modelo de evaluación crediticia que no se ajusta a las dinámicas del sector. Al no considerar estas particularidades, se han aprobado créditos a empresas con alta probabilidad de no pago, lo que pone en riesgo la estabilidad financiera de la organización.

Por esta razón, surge la necesidad de desarrollar un modelo de scoring crediticio que se adapte mejor a las características del sector construcción. La idea es incluir tanto variables financieras como factores del entorno y apoyarse en herramientas estadísticas y tecnológicas que permitan estimar de manera más precisa la probabilidad de incumplimiento. Con esto se busca mejorar la toma de decisiones, reducir los niveles de morosidad y fortalecer la gestión del riesgo dentro de la empresa.

Diversos estudios han resaltado la importancia de la gestión financiera en el éxito de las organizaciones, señalando que tomar decisiones basadas en análisis adecuados permite optimizar el uso de los recursos y disminuir riesgos. En el caso de los créditos, una mala evaluación puede generar pérdidas importantes, lo que hace evidente la necesidad de contar con modelos de scoring más ajustados a cada sector (Becerra Molina et al., 2023; Huacchillo Pardo et al., 2020).

El uso de modelos generales ha demostrado tener varias limitaciones, especialmente cuando no se consideran las condiciones propias de cada sector económico. En este caso, el modelo actual de la empresa ha permitido aprobar créditos a empresas con alta probabilidad de incumplimiento, lo cual se refleja en un nivel de morosidad cercano al 13% durante el año 2023. Esto evidencia la necesidad de diseñar un modelo más adecuado que permita mejorar las decisiones y reducir el riesgo.

Este estudio es de tipo aplicado y tiene un enfoque cuantitativo, ya que busca diseñar un modelo de scoring crediticio adaptado al sector construcción. Para ello, se analizarán variables financieras junto con factores del entorno que pueden influir en la capacidad de pago de las empresas. Sin embargo, existen algunas limitaciones, como la disponibilidad de información histórica y el acceso restringido a ciertos datos confidenciales.

El principal aporte de esta investigación es la propuesta de un modelo de scoring crediticio ajustado a las características del sector, lo que permitirá mejorar la calidad de las decisiones en la asignación de créditos. Además, este trabajo puede servir como referencia para otras empresas con problemas similares en la gestión del riesgo. Desde el punto de vista social, también es relevante, ya que puede contribuir a la estabilidad de las empresas del sector construcción, favoreciendo su crecimiento y generando efectos positivos en la economía y el empleo. En el contexto colombiano, este tipo de estudios es importante porque ayuda a mejorar las prácticas financieras y a crear un entorno económico más sólido.

Finalmente, el trabajo está organizado en cinco capítulos: el primero presenta el problema y los objetivos, el segundo desarrolla el marco teórico, el tercero explica la metodología, el cuarto muestra los resultados y su análisis, y el quinto expone las conclusiones y recomendaciones.

Capítulo 1: Problema de Investigación

1.1. Planteamiento del Problema

El problema identificado es un alto índice de incumplimiento de pagos (morosidad cercana al 13% en 2023), situación que enfrenta en la empresa Empresa Objeto De Estudio, la cual financia productos a compañías del sector construcción. Esta situación se acentúa por efecto de su modelo de evaluación crediticia actual es general y no se ajusta a las particularidades del mercado de la construcción, lo que ha llevado a otorgar créditos a empresas con alta probabilidad de impago, poniendo en riesgo la salud financiera de la compañía. En este trabajo abordaremos la respuesta a la siguiente pregunta de investigación:

¿Qué modelo de machine learning ofrece mayor precisión y capacidad de generalización para predecir la oportunidad de falla crediticio de las empresas clientes del sector construcción de Empresa Objeto De Estudio, y cuáles variables financieras y no financieras son relevantes dentro del modelo predictivo?

El uso de modelos generales para evaluar el riesgo crediticio ha demostrado no ser suficiente, ya que no tienen en cuenta las características propias de cada sector. En el caso del sector construcción, hay factores importantes como los ciclos financieros largos, la dependencia del flujo de caja, la sensibilidad a cambios económicos y los altos niveles de endeudamiento. Todo esto hace que los modelos tradicionales no sean tan precisos al momento de estimar el riesgo de incumplimiento. Por eso, diseñar un modelo de scoring ajustado a estas condiciones le permitiría a la empresa hacer evaluaciones más acordes a su realidad y tomar mejores decisiones.

Desde el punto de vista empresarial, implementar un modelo basado en Machine Learning es una buena oportunidad para mejorar la gestión del riesgo, optimizar la asignación de créditos y reducir la morosidad. Estas herramientas permiten identificar relaciones más complejas entre las variables, tanto financieras como no financieras, lo que ayuda a predecir mejor el comportamiento de pago de las empresas. Así, se logra usar mejor los recursos y disminuir posibles pérdidas.

En cuanto al impacto económico y social, mejorar la estabilidad financiera de las empresas del sector construcción puede traer beneficios importantes, ya que este sector es clave para la economía colombiana y la generación de empleo. Una mejor evaluación del crédito ayuda a que los proyectos continúen, fortalece a las empresas y contribuye al desarrollo de las regiones.

Finalmente, este estudio aborda un problema real del contexto colombiano utilizando herramientas de análisis más avanzadas. Su aporte no solo es útil para la empresa estudiada, sino que también puede servir como referencia para otras organizaciones del sector. De esta manera, se combina lo académico con lo práctico, aportando al desarrollo económico.

1.2. Justificación

El uso de modelos generales para evaluar el riesgo crediticio ha demostrado no ser suficiente, ya que no tienen en cuenta las características propias de cada sector. En el caso del sector construcción, hay factores importantes como los ciclos financieros largos, la dependencia del flujo de caja, la sensibilidad a cambios económicos y los altos niveles de endeudamiento. Todo esto hace que los modelos tradicionales no sean tan precisos al momento de estimar el riesgo de incumplimiento. Por eso, diseñar un modelo de scoring ajustado a estas condiciones le permitiría a la empresa hacer evaluaciones más acordes a su realidad y tomar mejores decisiones.

Desde el punto de vista empresarial, implementar un modelo basado en Machine Learning es una buena oportunidad para mejorar la gestión del riesgo, optimizar la asignación de créditos y reducir la morosidad. Estas herramientas permiten identificar relaciones más complejas entre las variables, tanto financieras como no financieras, lo que ayuda a predecir mejor el comportamiento de pago de las empresas. Así, se logra usar mejor los recursos y disminuir posibles pérdidas.

En cuanto al impacto económico y social, mejorar la estabilidad financiera de las empresas del sector construcción puede traer beneficios importantes, ya que este sector es clave para la economía colombiana y la generación de empleo. Una mejor evaluación del crédito ayuda a que los proyectos continúen, fortalece a las empresas y contribuye al desarrollo de las regiones.

Finalmente, este estudio aborda un problema real del contexto colombiano utilizando herramientas de análisis más avanzadas. Su aporte no solo es útil para la empresa estudiada, sino que también puede servir como referencia para otras organizaciones del sector. De esta manera, se combina lo académico con lo práctico, aportando al desarrollo económico.

1.3. Objetivos

1.3.1. Objetivo General

Diseñar y validar un modelo de scoring crediticio basado en técnicas de machine learning que permita estimar la probabilidad de incumplimiento de los clientes de la empresa objeto de estudio en el sector construcción, mediante la comparación de algoritmos de aprendizaje supervisado y la selección estadística de variables financieras con mayor capacidad predictiva.

1.3.2. Objetivos Específicos

- Identificar la importancia relativa de variables financieras (como endeudamiento y liquidez) y no financieras (antigüedad y comportamiento historico) que influyen en el comportamiento crediticio de las empresas del sector.
- Comparar el rendimiento de diferentes algoritmos de aprendizaje supervisado (Árboles de Decisión, Random Forest, Gradient Boosting y Redes Neuronales) para determinar cuál ofrece mayor precisión y capacidad de generalización.
- Validar el modelo seleccionado mediante herramientas estadísticas y el uso de datos históricos para fortalecer el proceso de toma de decisiones financieras de la organización mediante reglas operativas de decisión.

Capítulo 2: Estado del Arte

La evolución de la gestión financiera y la necesidad de mitigar el riesgo de impago han impulsado el desarrollo de herramientas analíticas cada vez más sofisticadas. Históricamente, la evaluación crediticia se ha basado en modelos generales; sin embargo, la literatura reciente subraya que la falta de personalización frente a las particularidades de sectores específicos (como el de la construcción) suele derivar en decisiones financieras erróneas y en un incremento de la morosidad. En este contexto, cada vez más investigadores han empezado a utilizar herramientas como el machine learning y modelos estadísticos más avanzados para mejorar la toma de decisiones. A lo largo de distintos estudios, se han propuesto soluciones que van desde modelos predictivos aplicados al factoring y a cooperativas de ahorro en América Latina, hasta el uso de redes neuronales combinadas con algoritmos genéticos en sistemas bancarios de Asia. Todo esto muestra que existe un interés creciente por desarrollar modelos más precisos y adaptados a diferentes realidades económicas.

En esta línea, se presenta una revisión de antecedentes que sirven como base para este estudio. Se analizan diferentes enfoques utilizados por varios autores para abordar el problema del scoring crediticio, especialmente usando algoritmos como árboles de decisión, Random Forest, Gradient Boosting y redes neuronales. Esto permite tener una mejor idea de cómo está el tema actualmente y, al mismo tiempo, ayuda a construir una base sólida para proponer un modelo más ajustado a las necesidades del sector construcción.

2.1. Antecedentes

2.1.1. Antecedentes Internacionales

Los antecedentes internacionales presentados en esta monografía forman parte de la revisión de literatura y abordan el uso de modelos de scoring y técnicas de machine learning en diversos contextos globales.

En el Ecuador se propuso un modelo de scoring de originación para una Cooperativa de Ahorro y Crédito (COAC) del segmento 3. El objetivo era identificar clientes potenciales de manera efectiva y excluir a aquellos con alto riesgo, adaptando las herramientas a las normativas

locales de la Superintendencia de Economía Popular y Solidaria (SEPS) (Tulcanaza Aguilar, 2021).

En Chile desarrollaron un modelo de evaluación crediticia para la empresa Fantasía S.A. con el fin de reducir el peligro de impagos y mejorar la liquidez. El modelo demostró una eficiencia del 81,82% en la correcta evaluación de los créditos otorgados (Leal Fica et al., 2018).

En Turquía valoraron el comportamiento del sector bancario turco utilizando un enfoque de balanced scorecard y el método Analytic Network Process (ANP). El estudio analizó 33 bancos y determinó que el factor financiero es el más relevante (65.7%), seguido por la visión del cliente (22.1%) (Dinçer et al., 2020).

En el continente asiático se implementaron un modelo de scoring de crédito utilizando una red neuronal probabilística (PNN) optimizada con algoritmos genéticos. Este estudio comparó ocho técnicas de machine learning y concluyó que el modelo optimizado ofrecía una mayor precisión y velocidad de convergencia para predecir incumplimientos en sistemas bancarios débiles (Rastegar y Faraji, 2023).

2.1.2. Antecedentes Nacionales

Los antecedentes nacionales presentados en el documento se enfocan en la evolución modelos de scoring adaptados a sectores específicos dentro del contexto financiero colombiano.

Los estudios destacados son los del grupo Factoring de Occidente S.A.S., donde la meta del análisis fue implementar un modelo de scoring crediticio para optimizar la evaluación del riesgo de los clientes de esta empresa. Utilizaron una metodología de recolección y análisis de datos históricos junto con técnicas estadísticas avanzadas para crear un modelo predictivo basado en variables clave del negocio del factoring. Los resultados indicaron que el modelo fue altamente efectivo para ver clientes con alto impacto de impago, mejorando la precisión en la evaluación y reduciendo pérdidas financieras (Santiago Sandoval & Urán Vélez, 2021).

En el sector de servicios públicos - Gas el estudio analizó el comportamiento crediticio en una empresa dedicada a la instalación de tuberías de gas y revisión de pagos de medidores. El autor buscaba mejorar los índices de morosidad y sanear deudas antiguas. El análisis determinó que las variables más importantes para predecir el incumplimiento eran el número de refinanciaciones de la deuda y la duración de los periodos en mora. Curiosamente, el estudio encontró que variables

exógenas como la inflación y el desempleo no tenían un impacto estadísticamente significativo en este modelo específico (Garzon Escobar, 2020).

En la empresa que se estudió se menciona como un antecedente directo de la problemática nacional en el sector de la construcción. El documento señala que la empresa enfrentaba una morosidad del 13% en 2023 debido al uso de modelos de evaluación generales que no consideraban las particularidades de sus clientes constructores (Orozco Alvarez, 2025).

2.1.3. Antecedentes

Shmueli y Koppius (2011) fueron de los primeros en destacar la importancia de la analítica predictiva como una herramienta útil para anticipar el comportamiento del mercado y apoyar la toma de decisiones en las organizaciones. Su principal aporte fue plantear que estos modelos no deben entenderse únicamente como técnicas estadísticas, sino como herramientas estratégicas que contribuyen a la generación de valor.

Por su parte, Davenport y Harris (2017) reforzaron esta idea al evidenciar el impacto de la analítica avanzada en los procesos estratégicos empresariales. Según estos autores, el uso de datos permite mejorar la toma de decisiones y fortalecer la competitividad en distintos sectores, especialmente en el financiero y asegurador.

En el contexto del riesgo crediticio, Dastile et al. (2020) realizaron una comparación entre modelos estadísticos tradicionales y técnicas de *machine learning*. Sus resultados mostraron que los métodos de ensamble tienden a ser más precisos que los modelos individuales. Además, señalaron que el *deep learning* tiene un gran potencial en el *credit scoring*, aunque su aplicación aún es limitada.

Posteriormente, Markov et al. (2022) desarrollaron una revisión sistemática de la literatura reciente, con el objetivo de actualizar el panorama del *credit scoring*. Su estudio permitió identificar la evolución de los enfoques metodológicos, así como la importancia de adaptar los modelos a los cambios del mercado y a la incorporación de nuevas fuentes de datos.

Finalmente, Barbaglia et al. (2023) demostraron que los algoritmos de *machine learning* pueden superar a los modelos tradicionales en la predicción del incumplimiento de préstamos hipotecarios. Asimismo, destacaron la relevancia de integrar variables relacionadas con el prestatario, las características del préstamo y las condiciones económicas locales. En este sentido,

identificaron factores clave como la tasa de interés y el entorno económico regional, lo que evidencia la necesidad de desarrollar modelos ajustados a contextos específicos.

2.2. Marco Conceptual o Teórico

El presente estudio se sustenta en la integración de la gestión del riesgo financiero y las ciencias de la computación, particularmente en el campo del aprendizaje automático. A continuación, se presentan los principales ejes conceptuales que permiten comprender la estructura del modelo propuesto y su fundamento teórico.

El estudio de González López (2018) se centró en desarrollar un modelo estadístico para predecir el no pago de créditos en el sector del factoring, con el objetivo de mejorar los procesos de aprobación. Para ello, utilizó una metodología que incluyó el entendimiento del negocio, la organización de los datos, el modelamiento y el análisis de resultados, lo que permitió adaptar técnicas de credit scoring del sector bancario a este contexto.

Por su parte, Tulcazana Aguilar (2021) propuso lineamientos para diseñar un modelo de scoring en una cooperativa de ahorro y crédito en Ecuador, buscando identificar clientes con buen perfil y filtrar aquellos con mayor riesgo de incumplimiento. Este trabajo tuvo en cuenta el crecimiento de estas entidades y aspectos regulatorios, apoyándose en la recopilación, análisis de datos y validación del modelo.

De igual forma Leal fica et al. (2018) desarrollaron un modelo de evaluación crediticia para una empresa chilena que presentaba problemas de liquidez, logrando mejorar la gestión del crédito y reducir el riesgo de impago, con resultados positivos en la aprobación de créditos.

Además, se desarrolló una métrica que combina diferentes indicadores para obtener resultados más confiables en el scoring crediticio, destacándose el buen desempeño de modelos optimizados con algoritmos genéticos y PSO frente a otros métodos.

En el contexto colombiano, Santiago Sandoval y Urán Valez (2021) diseñaron un modelo de scoring crediticio para una empresa de factoring, logrando mejorar la identificación de clientes con mayor riesgo de incumplimiento y reduciendo las pérdidas financieras asociadas.

Por otro lado, Garzón Escobar (2020) analizó el comportamiento de pago en una empresa de distribución de gas, encontrando que factores como las refinanciaciones y los retrasos en los

pagos influyen significativamente en el incumplimiento, mientras que variables como la inflación no resultaron relevantes dentro del modelo.

El riesgo crediticio se entiende como la posibilidad de que un cliente no cumpla con sus pagos, lo que puede afectar la estabilidad financiera de una empresa, por lo que es importante contar con herramientas que permitan identificar estos riesgos desde etapas tempranas.

El credit scoring es una metodología que asigna una puntuación a los clientes con el fin de estimar su probabilidad de pago, facilitando la toma de decisiones y reduciendo la subjetividad en la evaluación.

Finalmente, el machine learning permite analizar grandes volúmenes de datos e identificar patrones más complejos que los métodos tradicionales, siendo una herramienta clave para mejorar la predicción del riesgo crediticio.

- Árboles de Decisión: Modelos que cortan los datos en subconjuntos apoyados en reglas de decisión lógica. Son altamente interpretables (Chang et al., 2016).

* Random Forest: Una técnica de "ensamble" que combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste (overfitting) (Behr y Weinblat, 2017).

- Gradient Boosting: Un método que construye modelos de forma secuencial, donde cada nuevo modelo intenta corregir los errores de los anteriores, siendo muy eficaz para problemas de clasificación complejos (Bai et al., 2022).
- Redes Neuronales: Modelos inspirados en la estructura biológica del cerebro, capaces de aprender relaciones extremadamente complejas entre variables, aunque con una menor interpretabilidad ("caja negra") (Sariev y Germano, 2020).

Evaluación y Selección de Variables, para que un modelo sea robusto, no debe incluir variables irrelevantes que generen ruido.

Índice de Gini: Utilizado para medir la capacidad de discriminación del modelo; cuanto más alto sea el índice, mejor diferencia el modelo entre un cliente moroso y uno cumplidor.

Eliminación Recursiva de Características (RFE): Técnica que selecciona las variables más significativas eliminando repetidamente las de menor importancia, asegurando que el modelo final sea eficiente y generalizable.

Probabilidad de Incumplimiento (PD): Es la medida específica que el modelo intenta calcular. Se expresa como un valor entre 0 y 1, donde valores cercanos a 1 indican una alta certeza de que el cliente entrará en mora. Esta métrica es el insumo principal para decidir los límites de crédito y las condiciones comerciales para los clientes de la empresa.

2.3. Marco Legal

Para construir el Marco Legal, es necesario identificar las normativas que regulan el riesgo crediticio, la protección de datos y la actividad financiera, tanto en el contexto general como en el colombiano. En este sentido, en Colombia se destacan la *Ley 1266 de 2008*, relacionada con el habeas data financiero (Congreso de Colombia, 2008), y la *Ley 1581 de 2012*, sobre la protección de datos personales, las cuales establecen lineamientos para el tratamiento de la información crediticia (Congreso de Colombia, 2012). Asimismo, el Estatuto Orgánico del Sistema Financiero y las disposiciones de la Superintendencia Financiera regulan la gestión del riesgo en las entidades financieras.

Aunque este documento se centra en la parte técnica, un modelo de scoring de crédito debe alinearse con un esquema normativo.

El diseño e implementación de un modelo de scoring crediticio para la empresa EMPRESA OBJETO DE ESTUDIO está sujeto a un conjunto de disposiciones legales que garantizan la transparencia, el manejo ético de la información y la gestión del riesgo.

Dado que el modelo utiliza datos históricos de clientes para predecir comportamientos futuros, se rige principalmente por:

- Ley 1266 de 2008 (Ley de Habeas Data): Regula el manejo de la información contenida en bases de datos personales, especialmente la financiera, crediticia y comercial. El modelo debe asegurar que el uso de la información para el cálculo del score cuente con la autorización del titular (Congreso de Colombia, 2008).
- Ley 1581 de 2012: Marco general de protección de datos personales en Colombia, que establece los principios de finalidad, libertad, veracidad y seguridad en el tratamiento de la información (Congreso de Colombia, 2012).

En este documento se hace referencia a la mitigación del riesgo de impago, lo cual se alinea con las directrices de los entes reguladores:

- Circular Externa 100 de 1995: Establece el Sistema de Administración de Riesgo Crediticio (SARC). Aunque EMPRESA OBJETO DE ESTUDIO sea una firma comercial y no necesariamente una entidad financiera regulada, el SARC proporciona los estándares técnicos para el diseño de modelos de otorgamiento de crédito (Superintendencia Financiera de Colombia, 1995).
- Normas Internacionales de Información Financiera (NIIF 9): Esta normativa internacional exige a las empresas reconocer las "pérdidas crediticias esperadas". El modelo de scoring propuesto sirve como herramienta técnica para dar cumplimiento a esta exigencia contable, al predecir la probabilidad de incumplimiento de manera anticipada (IASB, 2014).

El uso de algoritmos de Machine Learning (como Redes Neuronales o Random Forest) mencionados en el artículo debe evitar sesgos algorítmicos:

- Constitución Política de Colombia (Artículo 13): Garantiza la igualdad y prohíbe la discriminación. En términos de scoring, esto implica que las variables seleccionadas (como las depuradas mediante el índice de Gini en el estudio) no deben basarse en criterios discriminatorios (raza, religión, etc.), sino exclusivamente en factores de capacidad y comportamiento de pago.

Y está también, la Reglamentación del Sector Comercio:

- Código de Comercio de Colombia: Regula las obligaciones mercantiles y los contratos de suministro o venta a crédito, que son la base de la relación entre EMPRESA OBJETO DE ESTUDIO y sus clientes del sector de la construcción.

Capítulo 3: Metodología

La investigación siguió un diseño cuantitativo con un enfoque comparativo y aplicado. La presente investigación se clasifica como aplicada, dado que tiene como propósito el desarrollo de una solución práctica orientada a mejorar el proceso de análisis del peligro crediticio en una organización del sector construcción. Asimismo, adopta un enfoque cuantitativo, al fundamentarse en el análisis estadístico y computacional de datos históricos para la construcción y validación de modelos predictivos (Hernández Sampieri et al, 2014).

Se utilizaron datos históricos de operaciones comerciales de una organización privada del sector construcción con más de trescientos clientes. Se recopilaron indicadores financieros (balance general, estados de resultados) y datos no financieros (antigüedad, sector, tamaño).

Para la selección de variables se aplicó el Índice de Gini para medir la importancia de las variables y una técnica de eliminación recursiva para identificar el número óptimo de indicadores (resultando en 12 variables finales) y evitar el sobreajuste (overfitting).

Se empleó el software RStudio para implementar cuatro modelos como son los árboles de decisión, Randon Forest, gradient Boosting y Redes Neuronales. Los Árboles de Decisión se utilizaron para segmentar datos bajo reglas jerárquicas simples, Random Forest para reducir la varianza combinando múltiples árboles, Gradient Boosting con un enfoque secuencial que corrige errores de modelos anteriores y las Redes Neuronales Para capturar relaciones complejas y no lineales mediante capas ocultas.

Se dividió la muestra en un 90% para entrenamiento y un 10% para testeo. Se realizó validación cruzada (2 particiones y 15 repeticiones) para ajustar los hiperparámetros de cada modelo y asegurar su confiabilidad.

3.1. Tipo y Enfoque de Investigación

La presente investigación es de tipo aplicada, dado que su propósito central es generar un modelo de scoring crediticio, que dé solución a un problema real de gestión financiera en una organización del sector construcción, orientado a la toma de decisiones operativas.

El enfoque adoptado es cuantitativo, pues el fenómeno de estudio —la probabilidad de incumplimiento de pago— se aborda mediante la recolección, procesamiento estadístico y análisis

numérico de datos financieros históricos. Este enfoque resulta pertinente dado que los datos disponibles son de naturaleza cuantitativa (indicadores financieros de razón e intervalo) y el fenómeno admite medición precisa a través de probabilidades y métricas de clasificación estadística.

El alcance del estudio es explicativo-predictivo: explicativo porque identifica y jerarquiza las variables financieras y no financieras que determinan el riesgo de incumplimiento; predictivo porque el modelo resultante estima la probabilidad de impago de nuevos clientes a partir de sus características financieras observadas.

3.2. Diseño de Investigación

El diseño de esta investigación es no experimental, puesto que no se manipulan variables ni se interviene en el entorno de la organización. Los datos utilizados corresponden a operaciones comerciales ya realizadas, sobre las cuales no es posible ni pertinente ejercer control experimental.

Dentro de los diseños no experimentales, la investigación adopta un corte retrospectivo de panel longitudinal: se trabaja con registros financieros de múltiples empresas clientes a lo largo de un período de tiempo determinado, lo que permite observar la evolución del comportamiento crediticio y capturar patrones dinámicos de pago e incumplimiento.

La unidad de análisis es la empresa cliente de EMPRESA OBJETO DE ESTUDIO que haya tenido al menos una operación de crédito activa durante el período de estudio, caracterizada a través de sus indicadores financieros y su historial de cumplimiento.

3.3. Población y Muestra

La población objeto de estudio está conformada por la totalidad de empresas clientes del sector de la construcción que mantuvieron o mantienen relaciones comerciales de crédito con Empresa Objeto De Estudio durante los últimos tres años. Estas empresas son personas jurídicas, principalmente del sector construcción colombiano, que financian la adquisición de materiales e insumos a través de crédito directo otorgado por la firma.

Los datos utilizados para el desarrollo del modelo comprenden 350 registros de empresas **clientes**, identificados a partir de la base de datos histórica de Empresa Objeto De Estudio. Esta cifra incluye tanto empresas que han cumplido con sus obligaciones de pago (clientes en buen

estado) como empresas que han incurrido en incumplimiento (clientes en mora), lo cual es un requisito fundamental para entrenar un modelo de clasificación supervisada. La distribución de la variable objetivo refleja el contexto real de la cartera, donde una minoría de clientes —consistente con la tasa de morosidad del 13% reportada para 2023— corresponde a la clase de incumplimiento.

Período de análisis: Los datos corresponden al período comprendido entre los años 2022 a 2025 en los que Empresa Objeto De Estudio dispone de información financiera consolidada de sus clientes, priorizando la vigencia de los estados financieros y la disponibilidad de registros de comportamiento de pago. La información fue suministrada directamente por la organización a través de su sistema de gestión comercial y financiera.

3.4. Fuentes e Instrumentos de Recolección de Datos

Fuente primaria: La principal fuente de información fue la base de datos interna de operaciones crediticias de Empresa Objeto De Estudio, que contiene los registros históricos de crédito de sus clientes del sector construcción. Esta base integra dos tipos de información:

- Información financiera: Estados financieros de los clientes (balance general y estado de resultados), de los cuales se extrajeron los indicadores financieros utilizados como variables predictoras del modelo. Los indicadores calculados incluyen razones de liquidez (razón corriente, prueba ácida), indicadores de endeudamiento, rentabilidad (ROA, ROE, margen neto, margen operacional) y métricas de recuperación de cartera, entre otros.
- Información de comportamiento crediticio: Historial de pagos, registros de mora, número de refinanciaciones y reportes en centrales de riesgo, los cuales permiten construir la variable dependiente del modelo (impago: sí/no).

Instrumento: Dado el carácter cuantitativo y de fuente secundaria de la investigación, no se aplicaron encuestas ni entrevistas. El instrumento de recolección fue una matriz de extracción de datos financieros, estructurada con las variables identificadas en la revisión de literatura como predictoras relevantes del incumplimiento crediticio. Esta matriz fue validada por el director del proyecto, quien tiene conocimiento directo del proceso de evaluación crediticia de Empresa Objeto De Estudio.

3.5. Procesamiento y Análisis de Datos

La investigación siguió un diseño cuantitativo con enfoque comparativo y aplicado. Se utilizaron datos históricos de operaciones comerciales de una organización privada del sector construcción. Se recopilaron indicadores financieros (balance general, estados de resultados) y datos no financieros (antigüedad, sector, tamaño).

Para la selección de variables se aplicó el Índice de Gini para medir la importancia de las variables y una técnica de eliminación recursiva para identificar el número óptimo de indicadores —resultando en 12 variables finales— y evitar el sobreajuste (*overfitting*).

Se empleó el software RStudio para implementar cuatro modelos: Árboles de Decisión, Random Forest, Gradient Boosting y Redes Neuronales. Los Árboles de Decisión se utilizaron para segmentar datos bajo reglas jerárquicas simples; Random Forest para reducir la varianza combinando múltiples árboles; Gradient Boosting con un enfoque secuencial que corrige errores de modelos anteriores; y las Redes Neuronales para capturar relaciones complejas y no lineales mediante capas ocultas.

Se dividió la muestra en un 90% para entrenamiento y un 10% para testeo. Se realizó validación cruzada con 2 particiones y 15 repeticiones para ajustar los hiperparámetros de cada modelo y asegurar su confiabilidad estadística. El desempeño de cada algoritmo se evaluó a través de la métrica de precisión global, comparando su comportamiento así como la muestra de entrenamiento como en la de testeo, con el fin de identificar el modelo con mayor capacidad de generalización.

3.6. Consideraciones Éticas

El desarrollo de esta investigación implicó el manejo de información financiera y comercial de carácter sensible, perteneciente tanto a Empresa Objeto De Estudio como a sus clientes del sector construcción. Por esta razón, se adoptaron las siguientes medidas para garantizar el uso ético y responsable de los datos:

Confidencialidad y anonimización: En cumplimiento de la *Ley 1266 de 2008* (Ley de Habeas Data) y la *Ley 1581 de 2012* (Protección de Datos Personales), el nombre de la empresa objeto de estudio se ha omitido deliberadamente a solicitud de sus propietarios, utilizándose en su

lugar el seudónimo "EMPRESA OBJETO DE ESTUDIO" a lo largo de todo el documento. Los registros individuales de los clientes fueron tratados de forma agregada y anonimizada, sin que sea posible identificar empresas específicas a partir de los resultados presentados.

Autorización institucional: La recolección y uso de los datos fue autorizada formalmente por los directivos de Empresa Objeto De Estudio, quienes facilitaron el acceso a la base de datos histórica bajo el entendido de que la información sería utilizada exclusivamente con fines académicos y de investigación, y que los resultados serían presentados de manera que no comprometan la confidencialidad de la organización ni de sus clientes.

Uso exclusivamente académico: Los datos obtenidos no fueron compartidos con terceros, no se utilizaron con fines comerciales y su procesamiento se realizó en entornos locales seguros (RStudio instalado en equipos de los investigadores), sin almacenamiento en plataformas de nube de acceso público.

Conflicto de intereses: Los autores declaran que no existe conflicto de intereses en el desarrollo de esta investigación. La participación del director del proyecto como fuente de información sobre el contexto de la empresa fue debidamente documentada y acotada al rol de asesoría técnica.

Capítulo 4: Resultados y Discusión

En este capítulo se presentan los resultados obtenidos a partir de la aplicación y comparación de diferentes modelos de Machine Learning para predecir el riesgo crediticio en la empresa objeto de estudio, utilizando datos históricos del sector de la construcción y evaluando cuatro modelos principales para analizar su capacidad de clasificación y precisión.

A partir del uso del índice de Gini, se identificó cuáles variables financieras tenían mayor importancia dentro del modelo, destacándose el nivel de endeudamiento como uno de los factores más influyentes. Además, se aplicaron técnicas como la eliminación recursiva de variables para seleccionar las más relevantes, reducir el ruido en los datos y evitar el sobreajuste, llegando a un total de 12 variables con buen nivel explicativo.

Cada modelo fue evaluado mediante validación cruzada y ajuste de hiperparámetros con el fin de mejorar su desempeño. Por ejemplo, el modelo basado en árboles de decisión alcanzó una precisión cercana al 85%, mostrando buenos resultados en la clasificación de los niveles de riesgo.

Estos resultados permitieron no solo identificar el modelo más confiable desde el punto de vista estadístico, sino también construir una herramienta que puede apoyar la toma de decisiones financieras y ayudar a disminuir los niveles de morosidad en la empresa.

En cuanto a las variables, cada una aporta información importante sobre la situación financiera de las empresas. El nivel de endeudamiento permite analizar las obligaciones frente a los recursos propios, el ROA muestra qué tan bien los activos generan ganancias, y la razón corriente indica la capacidad de cumplir con obligaciones a corto plazo, mientras que la prueba ácida ofrece una visión más estricta de la liquidez al no considerar inventarios.

También, el ROE permite ver la rentabilidad desde el punto de vista de los accionistas, y tanto el margen como el margen neto reflejan la eficiencia en la generación de utilidades. El indicador de recuperación se enfoca en la gestión de cartera, y la concentración de pasivos a corto plazo ayuda a identificar posibles riesgos en la estructura financiera.

Además, la rentabilidad operacional muestra los resultados del negocio antes de impuestos e intereses, y el margen operacional permite evaluar la eficiencia de las operaciones frente a las ventas. Por último, la variable de riesgo aporta información clave sobre el comportamiento crediticio del solicitante.

En general, todos estos modelos de scoring crediticio permiten tener una base sólida para analizar el riesgo, combinando información financiera con herramientas analíticas que ayudan a tomar decisiones más claras y mejor fundamentadas.

4.1. Modelos de scoring implementados para el análisis de riesgo de crédito

4.1.1. Árboles de decisión

Los árboles de decisión son una técnica de aprendizaje supervisado muy utilizada en problemas de clasificación y regresión, ya que permiten organizar la información en grupos más homogéneos mediante reglas de decisión, formando una estructura similar a un árbol que facilita la predicción del riesgo de impago.

Su funcionamiento consiste en seleccionar en cada paso las variables que mejor separan los datos según la variable objetivo, en este caso el incumplimiento. Para lograrlo, se utilizan medidas como la impureza de Gini o la entropía en clasificación, y el error cuadrático medio en regresión, lo que ayuda a encontrar las mejores divisiones y formar grupos más uniformes.

En este modelo, cada nodo representa una condición basada en una variable, y a partir de allí se generan ramas que dividen los datos en subconjuntos. Este proceso se repite varias veces, creando una estructura cada vez más detallada hasta llegar a los nodos finales o hojas, donde se agrupan observaciones con características similares, especialmente en cuanto al riesgo de impago.

Una de las principales ventajas de los árboles de decisión es que son fáciles de interpretar, ya que permiten entender claramente cómo influyen las variables en la predicción. Sin embargo, también tienen una desventaja importante, que es el sobreajuste, es decir, que el modelo se ajusta demasiado a los datos de entrenamiento y puede perder precisión con nuevos datos. Para evitar esto, se suelen aplicar técnicas como la poda, que consiste en eliminar ramas que no aportan información relevante al modelo.

4.1.2. Random Forest

El Random Forest es una técnica de aprendizaje supervisado que combina varios árboles de decisión con el objetivo de mejorar la precisión y lograr resultados más confiables. A diferencia de un solo árbol, que puede presentar problemas de sobreajuste, este modelo utiliza un conjunto

de árboles y luego combina sus resultados, generalmente por medio de votaciones o promedios, lo que permite obtener predicciones más estables.

Su funcionamiento inicia creando varios árboles de decisión a partir de diferentes subconjuntos de datos. Esto se hace mediante una técnica llamada bootstrap sampling, donde cada árbol se entrena con una muestra distinta tomada del conjunto original, permitiendo que algunos datos se repitan y otros no sean seleccionados. Esta variación ayuda a reducir la varianza del modelo y mejora su capacidad para trabajar con datos nuevos, además de evitar que se ajuste demasiado a un solo conjunto de datos.

Adicionalmente, en cada nodo de los árboles se selecciona de forma aleatoria un grupo de variables para realizar las divisiones. Esto hace que los árboles sean diferentes entre sí, reduciendo la relación entre ellos y fortaleciendo el modelo en general, lo que mejora su capacidad de predicción.

Una vez que todos los árboles han sido entrenados, el modelo combina sus resultados para obtener una predicción final (Behr y Weinblat, 2017). En tareas de clasificación, como la predicción de impago, se utiliza generalmente el voto mayoritario: la clase que más se repite entre los árboles es la que se toma como resultado final. En cambio, en problemas de regresión, se calcula el promedio de las predicciones de todos los árboles.

Una de las grandes ventajas de Random Forest es su resistencia al sobreajuste. Al basarse en múltiples árboles y promediar sus resultados, el modelo se vuelve menos sensible a las particularidades del conjunto de entrenamiento, logrando así una mejor generalización frente a nuevos datos (Behr y Weinblat, 2017). También destaca por su capacidad para trabajar con muchas variables y detectar relaciones complejas entre ellas.

Otro aspecto útil es que permite identificar la importancia de cada variable dentro del modelo, lo cual facilita entender qué factores influyen más en la predicción del impago. Esto se logra evaluando cuánto mejora la precisión cuando una variable se utiliza en las divisiones de los árboles.

Sin embargo, una de sus limitaciones es la interpretabilidad. A diferencia de un árbol de decisión simple, resulta más difícil explicar cómo se obtiene una predicción específica, ya que esta es el resultado de la combinación de múltiples árboles.

4.1.3. Gradient Boosting

El Gradient Boosting es otro método de ensamble que, al igual que el Random Forest, combina varios modelos para mejorar las predicciones, pero se diferencia en que los entrena de forma secuencial, es decir, cada nuevo modelo busca corregir los errores del anterior, lo que lo hace muy útil para problemas como la predicción del riesgo de impago.

La idea principal de este método es convertir modelos simples, conocidos como modelos débiles (generalmente árboles pequeños), en un modelo más fuerte al combinarlos. Este proceso se realiza paso a paso, donde en cada iteración se entrena un nuevo árbol enfocado en corregir los errores del modelo anterior.

El proceso comienza con un modelo inicial sencillo, como un promedio en regresión o una predicción básica en clasificación. Luego, se calculan los errores entre lo que el modelo predijo y los valores reales, y se entrena un nuevo árbol para reducir esos errores. Este procedimiento se repite varias veces hasta alcanzar un buen nivel de precisión o el número de iteraciones definido.

Una de sus principales ventajas es que permite encontrar relaciones complejas entre las variables, incluso cuando no son lineales, ya que va corrigiendo los errores poco a poco y detectando patrones que otros modelos más simples no logran identificar.

Además, es un modelo flexible porque se puede adaptar a diferentes tipos de problemas, como clasificación y regresión, dependiendo de la función de pérdida que se utilice.

Sin embargo, también tiene algunas desventajas, ya que puede presentar sobreajuste si no se ajustan bien sus parámetros, como la tasa de aprendizaje o el número de iteraciones, lo que puede hacer que funcione muy bien con los datos de entrenamiento pero no tan bien con datos nuevos.

4.1.4. Redes Neuronales

Las redes neuronales son modelos que se inspiran en el funcionamiento del cerebro humano y se utilizan para reconocer patrones y hacer predicciones. Son especialmente útiles cuando se trabaja con grandes cantidades de datos y cuando existen relaciones complejas entre las variables, como en el caso de la predicción del riesgo de impago.

Capa de entrada: Recibe los datos iniciales, como las variables financieras, donde cada nodo representa una característica del conjunto de datos.

Capas ocultas: Procesan la información, realizando cálculos y aplicando funciones de activación como ReLU o la sigmoide, lo que permite al modelo identificar relaciones más complejas entre los datos.

Capa de salida: Genera la predicción final, y en problemas de clasificación se suele usar una función sigmoide para obtener probabilidades entre 0 y 1.

El entrenamiento de una red neuronal se basa en dos procesos principales: la propagación hacia adelante y la retropropagación. Primero, los datos pasan por toda la red y se obtiene una predicción; luego, esta se compara con el valor real mediante una función de pérdida. A partir del error obtenido, se ajustan los pesos de las conexiones usando métodos como el gradiente descendente, repitiendo este proceso varias veces hasta mejorar el rendimiento del modelo.

Las redes neuronales se destacan porque pueden modelar relaciones complejas y no lineales entre las variables, lo que las hace muy útiles en problemas donde los datos no siguen patrones simples. Además, tienen la capacidad de trabajar con grandes volúmenes de información y muchas variables al mismo tiempo, lo que mejora su utilidad en el análisis del riesgo crediticio.

No obstante, presentan algunas desventajas. Una de las principales es su baja interpretabilidad: entender cómo toman decisiones puede ser difícil debido a su estructura interna. Además, su entrenamiento suele requerir bastante capacidad computacional y tiempo, especialmente en modelos grandes. También es necesario ajustar cuidadosamente varios hiperparámetros, como el número de capas, neuronas y la tasa de aprendizaje, lo que puede resultar un proceso complejo.

4.2. Importancia relativa de variables financieras (como endeudamiento y liquidez) y no financieras que influyen en el comportamiento crediticio de las empresas del sector.

Para dar cumplimiento al primer objetivo, se da un proceso analítico riguroso que permite filtrar, entre una gran cantidad de datos, aquellas variables que realmente tienen un peso significativo en la predicción del impago.

El desarrollo de este objetivo se realizó mediante la selección de inicial de variables Aplicación del Índice de Gini los siguientes pasos técnicos:

1. **Selección Inicial de Variables (Criterio de Experto y Literatura)**

El proceso se inició identificando variables tanto financieras como no financieras basadas en la literatura previa y en la base de datos de la empresa:

- Financieras: Liquidez, niveles de endeudamiento, rentabilidad y flujo de caja.
- No financieras: Comportamiento histórico de pagos, antigüedad de la relación comercial, sector específico de la construcción y ubicación geográfica.

2. **Aplicación del Índice de Gini**

Se usa el Índice de Gini como herramienta de medición de pureza. Este se utilizó para evaluar qué tan bien cada variable individual (por ejemplo, el índice de liquidez) podía "discriminar" o separar a los clientes que cumplen con sus pagos de aquellos que caen en mora. Variables con un Índice de Gini más alto fueron consideradas como de mayor importancia relativa.

3. **Eliminación Recursiva de Características (RFE)**

Para refinar la importancia y evitar el ruido de variables redundantes, se aplicó la técnica Recursive Feature Elimination (RFE). Este proceso consiste en:

- Entrenar el modelo con todas las variables.
- Asignar un peso o importancia a cada una.
- Eliminar las menos importantes de forma recursiva hasta quedarse con el conjunto óptimo que maximiza la precisión del modelo sin sobreajustarlo.

4. **Evaluación de la Capacidad Predictiva**

Se determinó la importancia relativa observando cómo la exclusión de ciertas variables afectaba el rendimiento de los algoritmos de Machine Learning (específicamente en Random Forest y Gradient Boosting).

El estudio permitió identificar que el comportamiento histórico de mora y ciertos ratios de endeudamiento tenían una correlación más fuerte con el incumplimiento que otras variables económicas externas.

5. Comparación entre Algoritmos

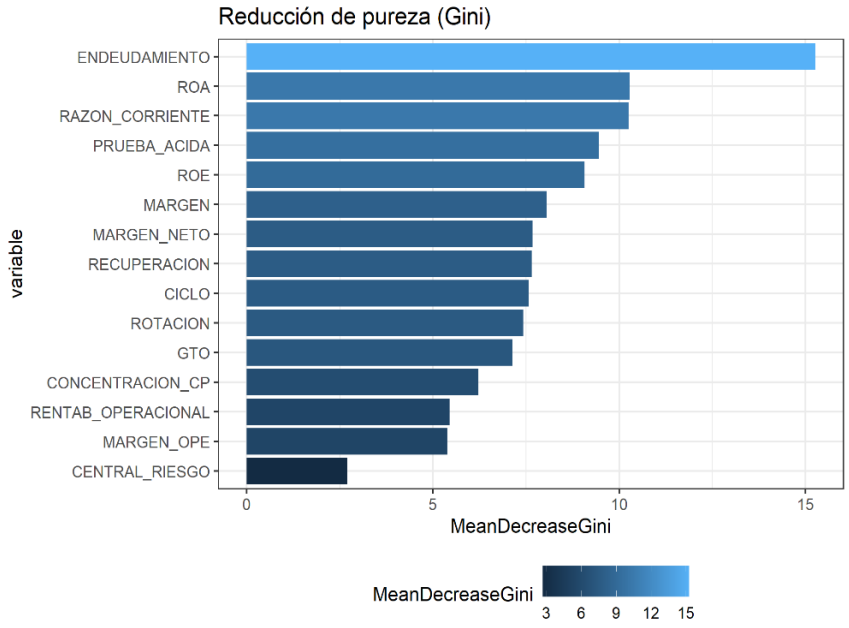
Finalmente, la importancia de las variables se validó al comparar cómo los cuatro algoritmos (Árboles de Decisión, Random Forest, Gradient Boosting y Redes Neuronales) reaccionaban ante ellas. Esto permitió confirmar que el modelo no solo dependía de la liquidez inmediata, sino de una combinación de factores que incluyen la estabilidad financiera a largo plazo y la disciplina de pago previa.

La importancia relativa no se asignó de forma subjetiva, sino que fue el resultado de un filtrado estadístico y algorítmico (Gini + RFE) que permitió destacar las variables financieras y no financieras con mayor poder de predicción.

Luego de hacer el análisis de la importancia relativa de los indicadores financieros de las empresas se evidenciaron los resultados que se ilustran en la siguiente figura:

Figura 1

Importancia de cada variable explicativa



Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por *Orozco* Álvarez, (2025).

La figura 1 muestra la importancia de las variables dentro de un modelo de Random Forest estimado con toda la muestra, utilizando la métrica conocida como *mean decrease Gini*. El Índice de Gini es una medida que se usa para evaluar qué tan homogéneos son los datos dentro de un nodo de un árbol de decisión. En términos sencillos, un nodo es más "puro" cuando la mayoría de los datos que contiene pertenecen a una misma clase. Si todos los elementos son de la misma clase, el nodo tiene una pureza total y el valor del Gini es 0.

Por el contrario, valores más altos del Índice de Gini indican mayor mezcla o impureza en los datos. A partir de esto, el *Mean Decrease Gini* mide cuánto contribuye cada variable a reducir esa impureza en promedio dentro del modelo. Es decir, evalúa en qué medida una variable ayuda a que las divisiones de los árboles sean más precisas al clasificar los datos.

En este sentido, las variables que presentan un mayor *Mean Decrease Gini* son las más relevantes, ya que tienen un papel importante en mejorar la calidad de las divisiones y, por ende, el desempeño del modelo. En cambio, aquellas con valores bajos o cercanos a cero aportan poco a la clasificación.

De acuerdo con los resultados observados en la figura 1, se puede destacar lo siguiente:

- Endeudamiento se posiciona como la variable más influyente según el Índice de Gini, lo que indica que su uso mejora significativamente la capacidad del modelo para clasificar correctamente.
- ROA y RAZÓN_CORRIENTE también tienen un peso importante, ya que contribuyen de manera notable a mejorar la pureza de los nodos.
- PRUEBA_ACIDA, ROE, MARGEN, MARGEN_NETO, y RECUPERACION presentan una importancia intermedia: ayudan al modelo, aunque en menor medida que las anteriores.
- Por último, CONCENTRACION_CP, RENTAB_OPERACIONAL, MARGEN_OPE y CENTRAL_RIESGO muestran un impacto reducido, lo que sugiere que su aporte al desempeño del modelo es limitado.

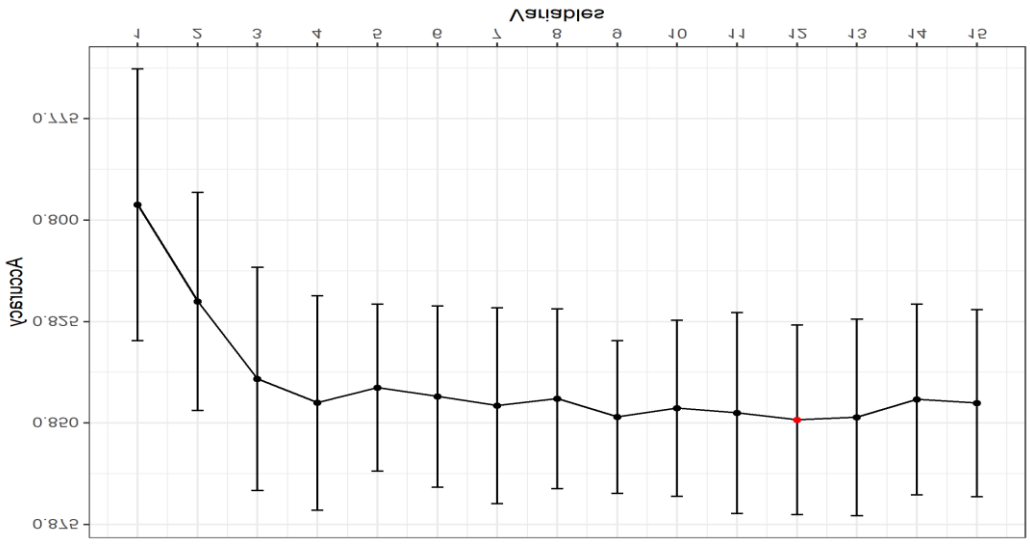
Es importante aclarar que este análisis no implica que las variables menos importantes deban eliminarse automáticamente. Más bien, lo que proporciona es una visión comparativa sobre el peso de cada variable dentro del modelo. Para determinar si conviene excluir algunas, se

emplean técnicas específicas de selección de características, como la eliminación recursiva. Este tipo de métodos busca identificar el conjunto óptimo de variables, eliminando aquellas que no aportan o incluso pueden afectar negativamente el rendimiento del modelo, ayudando así a evitar el sobreajuste. El procedimiento de eliminación recursiva se desarrolla generalmente de la siguiente manera:

1. Entrenamiento del modelo: Se entrena el modelo de Random Forest utilizando todas las variables disponibles.
2. Evaluación de importancia: Se calcula la relevancia de cada variable, comúnmente mediante métricas como la disminución del Gini o el impacto en la precisión al alterar sus valores.
3. Eliminación de la variable menos importante: Se retira la variable con menor importancia.
4. Repetición del proceso: El modelo se vuelve a entrenar con el nuevo conjunto de variables y se recalculan sus importancias. Este proceso se repite de forma iterativa.
5. Selección del mejor subconjunto de variables: Se elige el conjunto de variables que ofrece el mejor desempeño predictivo.

Figura 2

Cantidad óptima de variables



Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

De acuerdo con los resultados, la cantidad óptima de variables es 12, es decir, se deben eliminar 3 variables. Estas corresponden a RENTAB_OPERACIONAL, CONCENTRACION_CP y CENTRAL_RIESGO. Estas variables son precisamente 3 de las menos importantes de acuerdo con la disminución promedio de la pureza (figura 1).

En esta parte se estiman los modelos, ajustando los hiperparámetros para encontrar la mejor configuración posible. Para hacerlo, se utiliza el 90% de los datos para entrenar el modelo y el 10% restante se deja para probar qué tan bien funciona con datos nuevos. Además, con los datos de entrenamiento se aplica validación cruzada. Esto consiste en dividir los datos en dos partes y repetir el proceso 15 veces, de manera que el modelo se entrena y evalúa varias veces con diferentes combinaciones. En total, se realizan 30 evaluaciones, lo que ayuda a obtener resultados más confiables y no depender de una sola división de los datos.

4.3. Comparación de los rendimientos de diferentes algoritmos de aprendizaje supervisado (Árboles de Decisión, Random Forest, Gradient Boosting y Redes Neuronales) para determinar cuál ofrece mayor precisión y capacidad de generalización.

El objetivo de comparar el rendimiento de los algoritmos de aprendizaje supervisado se desarrolló a través de un proceso estructurado que incluyó la configuración, el ajuste de parámetros y la evaluación estadística de cada modelo.

Para la Implementación y Configuración de los Algoritmos se utilizaron cuatro algoritmos distintos, cada uno con un enfoque técnico específico para manejar los datos del sector de la construcción, arboles de decisión, Random Forest, Gradient Boosting y Redes Neuronales.

Para los Árboles de Decisión se empleó un modelo de aprendizaje supervisado simple para segmentar datos en grupos homogéneos mediante reglas de decisión jerárquicas.

El modelo Random Forest se implementó como una técnica de ensamble que combina múltiples árboles para reducir la varianza y mejorar la robustez frente al sobreajuste.

El Gradient Boosting se utilizó un enfoque secuencial donde cada nuevo modelo busca corregir los errores de los anteriores, ideal para capturar patrones complejos no lineales y con las Redes Neuronales se diseñó un modelo de aprendizaje profundo con capas de entrada, ocultas y

de salida, inspirado en el cerebro humano, para modelar relaciones altamente complicadas entre las variables financieras.

Para que la comparación fuera justa y precisa, se ajustaron los "hiperparámetros" (configuraciones internas) de los modelos para encontrar su versión óptima.

Los modelos se estimaron utilizando el 90% de la muestra total, reservando el 10% restante para las pruebas finales de testeo y se realizó una validación cruzada con 2 particiones y 15 repeticiones, lo que significa que cada modelo fue ajustado y evaluado 30 veces mediante remuestreos para asegurar la estabilidad de los resultados. Por ejemplo, en el modelo de Boosting, se ajustaron parámetros como el número de árboles (n.trees), la tasa de aprendizaje (shrinkage) y la profundidad de las divisiones (interaction.depth).

El rendimiento se determinó mediante métricas estadísticas procesadas en el software RStudio:

Para lograr la Precisión de los Árboles se estableció que el modelo de árboles base alcanzó una precisión del 85% en el total de los remuestreos. El proceso permitió validar cuál algoritmo presentaba la mayor confiabilidad estadística para predecir la probabilidad de impago, considerando tanto su exactitud en los datos de entrenamiento como su capacidad para generalizar con datos nuevos (el 10% reservado para testeo).

Para mejorar el rendimiento de todos los algoritmos comparados, se aplicó una técnica de eliminación recursiva que determinó que la cantidad óptima de variables para el modelo era de 12. Al eliminar las variables que no aportaban valor (como la rentabilidad operacional o la central de riesgo), se evitó el sobreajuste, permitiendo que la comparación se basara en los factores financieros y no financieros más potententes.

Modelos de árboles

El modelo de árboles no contiene hiperparámetros, por lo cual, su estimación es simple. En el total de remuestreos, el modelo de árboles tiene una precisión de 85%.

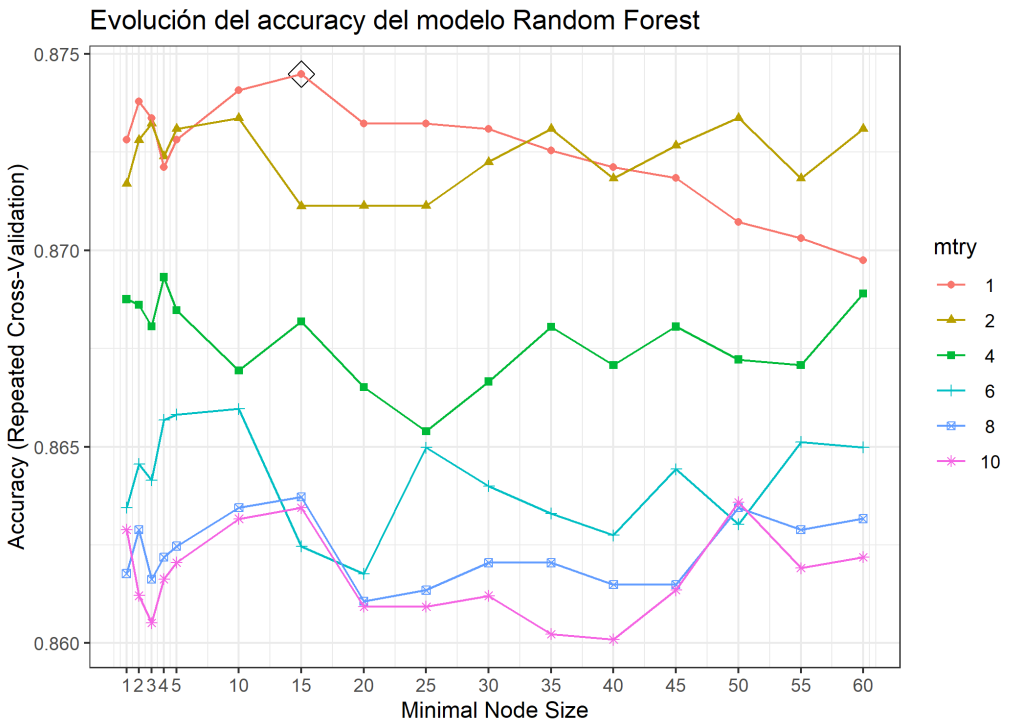
Random forest

Para este modelo se ajustan los hiperparámetros “mínimo tamaño de los nodos” y “mtry”. Este último puede tomar los valores {1, 2, 4, 6, 8, 10} y el tamaño mínimo de los nodos {1, 2, 3,

4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 50, 55, 60}. La combinación de estos valores produce un total de 96 posibles modelos.

Figura 3

Hiperparámetros óptimos del modelo Random Forest



Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

La combinación óptima de parámetros corresponde a {mtry=1, min.node.size= 15}.

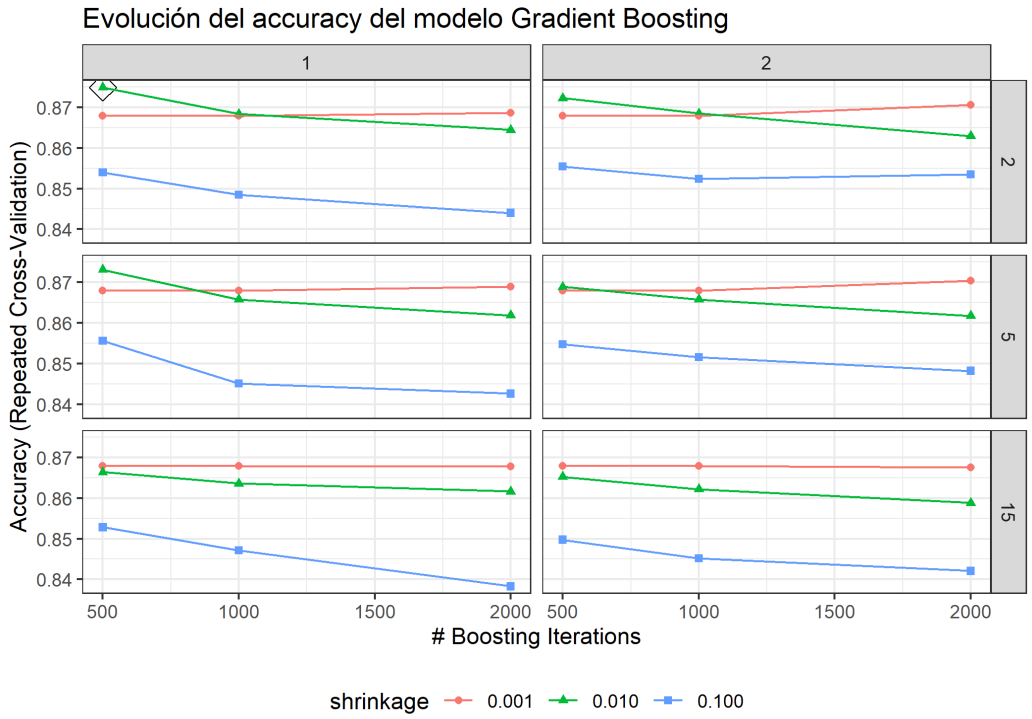
Boosting

Para este modelo se ajustan varios hiperparámetros con el objetivo de encontrar la mejor combinación posible. En este caso, se trabajan *interaction.depth* (que indica el número de divisiones del árbol), *n.trees* (la cantidad de árboles que conforman el modelo), *shrinkage* (la tasa de aprendizaje) y *n.minobsinnode* (el número mínimo de observaciones que debe tener un nodo para poder dividirse). Estos parámetros pueden tomar distintos valores: *interaction.depth* = {1, 2}, *n.trees* = {500, 1000, 2000}, *shrinkage* = {0.001, 0.01, 0.1} y *n.minobsinnode* = {2, 5, 15}, lo que da un total de 54 combinaciones posibles. Luego de evaluar todos estos modelos, cuyos resultados

se muestran en la siguiente gráfica, se encontró que la mejor configuración corresponde a *interaction.depth* = 1, *n.trees* = 1000, *n.minobsinnode* = 2 y *shrinkage* = 0.01.

Figura 4

Hiperparámetros óptimos del modelo Boosting



Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

4.4. Validación del modelo seleccionado mediante herramientas estadísticas y el uso de datos históricos para fortalecer el proceso de toma de decisiones financieras de la organización.

El cumplimiento de este objetivo se desarrolló mediante un proceso técnico estructurado de validación cruzada y testeo, utilizando el software estadístico RStudio para procesar los datos históricos de la organización.

Para fortalecer la toma de decisiones, se utilizaron registros históricos de operaciones comerciales de la organización. Estos datos se dividieron estratégicamente:

Muestra de Entrenamiento (90%): Utilizada para la construcción y ajuste inicial de los modelos.

Muestra de Testeo (10%): Reservada exclusivamente para validar la capacidad de generalización del modelo con datos que no "conoció" durante el entrenamiento.

Con el fin de asegurar la estabilidad estadística y que los resultados no fueran producto del azar, se aplicó una técnica de validación cruzada sobre la muestra de entrenamiento:

Se realizaron 30 remuestreos (2 particiones con 15 repeticiones) por cada algoritmo evaluado.

Cada modelo fue ajustado y evaluado 30 veces para obtener una métrica de precisión promedio confiable.

La validación incluyó un proceso de "tuneo" o ajuste fino de los parámetros internos de los algoritmos para encontrar su configuración más eficiente:

Random Forest: Se probaron 96 combinaciones posibles de parámetros (como el tamaño mínimo de nodos y número de variables), identificando la configuración óptima en {mtry=1, min.node.size= 15}.

Boosting: Se evaluaron 54 modelos posibles ajustando la tasa de aprendizaje y el número de árboles, logrando la mayor efectividad con 1,000 árboles y una tasa de 0.01.

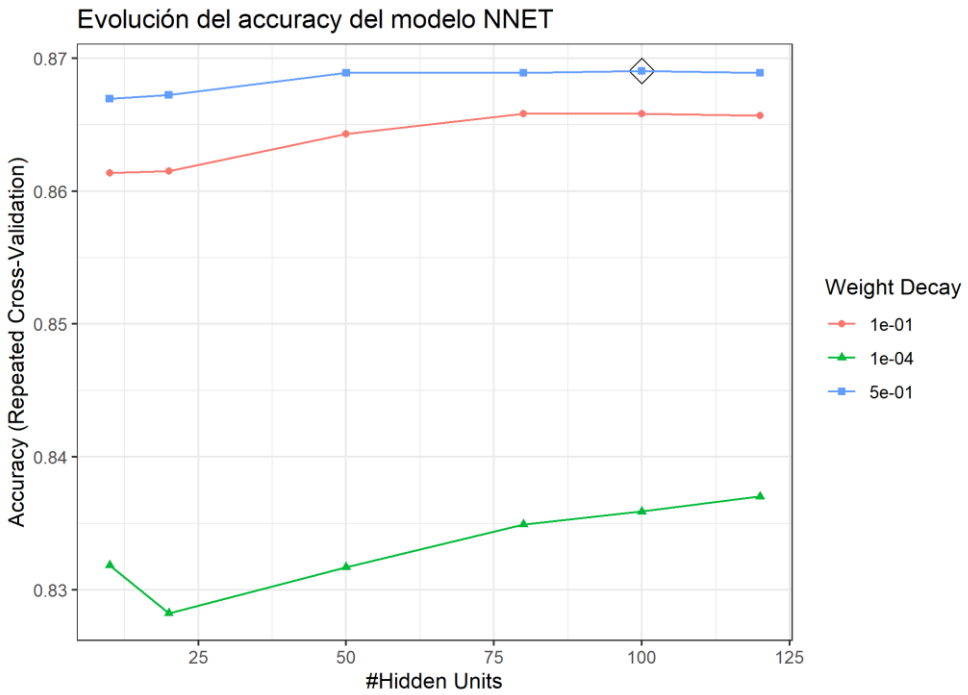
Redes Neuronales: Se compararon 54 arquitecturas diferentes variando el número de neuronas y la regularización (decay).

Redes neuronales

Para este modelo se ajustan los hiperparámetros *size* (que representa el número de neuronas en la capa oculta) y *decay* (que controla la regularización durante el entrenamiento de la red). Estos parámetros pueden tomar distintos valores: *size* = {10, 20, 50, 80, 100, 120} y *decay* = {0.0001, 0.1, 0.5}, lo que da un total de 18 combinaciones posibles. Después de evaluar todos estos modelos, cuyos resultados se presentan en la siguiente gráfica, se encontró que la mejor configuración corresponde a *size* = 100 y *decay* = 0.5.

Figura 5

Hiperparámetros óptimos del modelo de Redes Neuronales

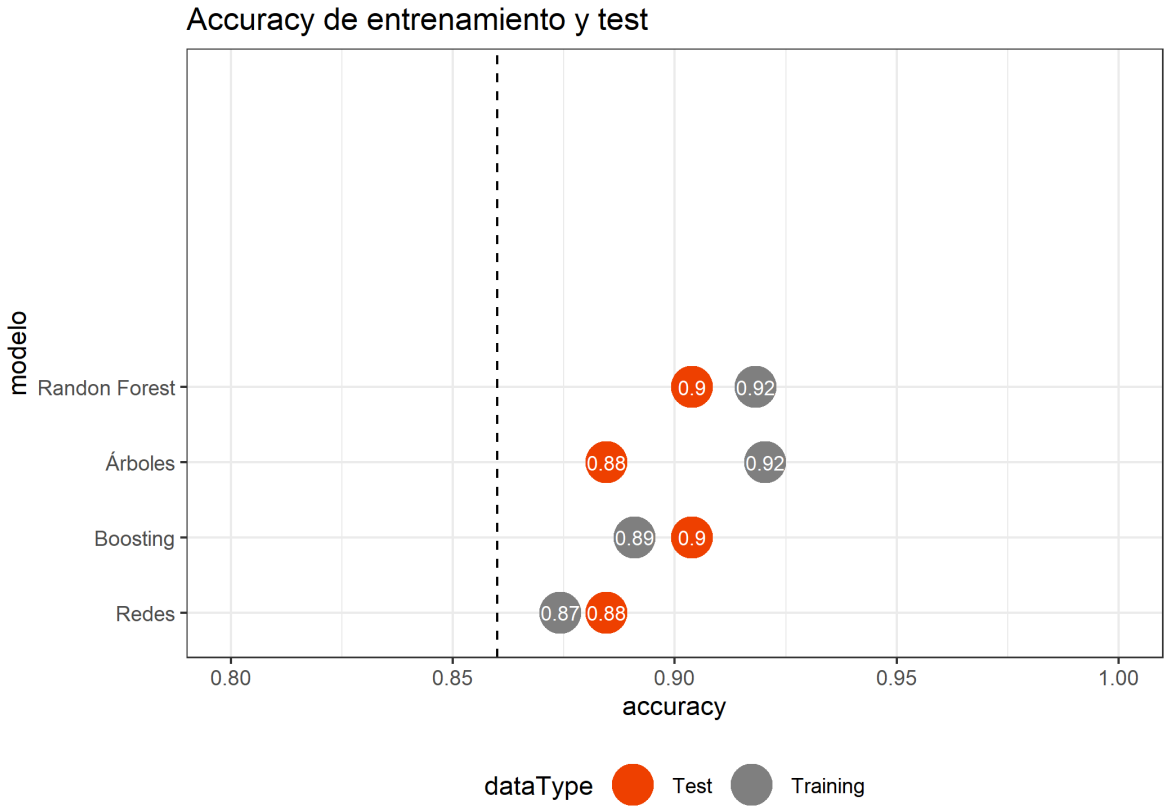


Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

Una vez que se ajustan todos los modelos, se comparan sus niveles de precisión tanto en los datos de entrenamiento como en los de prueba. Según los resultados, en la muestra de entrenamiento los modelos que mejor desempeño tienen son el árbol de decisión y el random forest, ambos con una precisión del 92%. Sin embargo, al evaluarlos con la muestra de testeo, el modelo de árbol baja su precisión al 88%, mientras que el random forest solo baja al 90%. Por esta razón, se decide escoger el modelo de random forest para estimar la PI, ya que mantiene un mejor desempeño con datos nuevos.

Figura 6

Comparación de modelos en la muestra de entrenamiento y testeo.



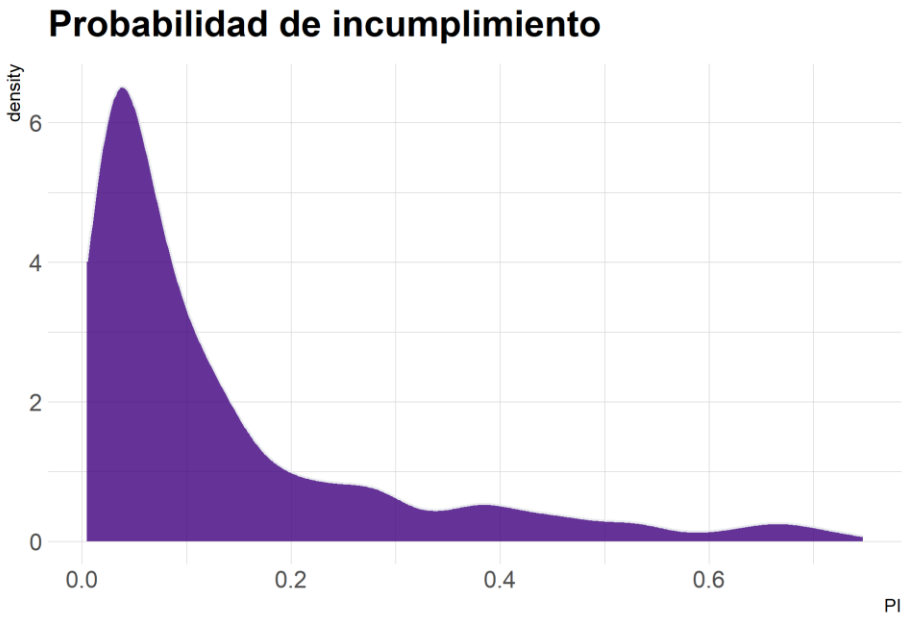
Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

Estimación del porcentaje de impago y cálculo del scoring

Una vez identificado el modelo más confiable, se procede a calcular la probabilidad de incumplimiento (PI) y el scoring. A continuación, se muestra la distribución de estas probabilidades de incumplimiento:

Figura 7

Distribución de la Probabilidad de Incumplimiento



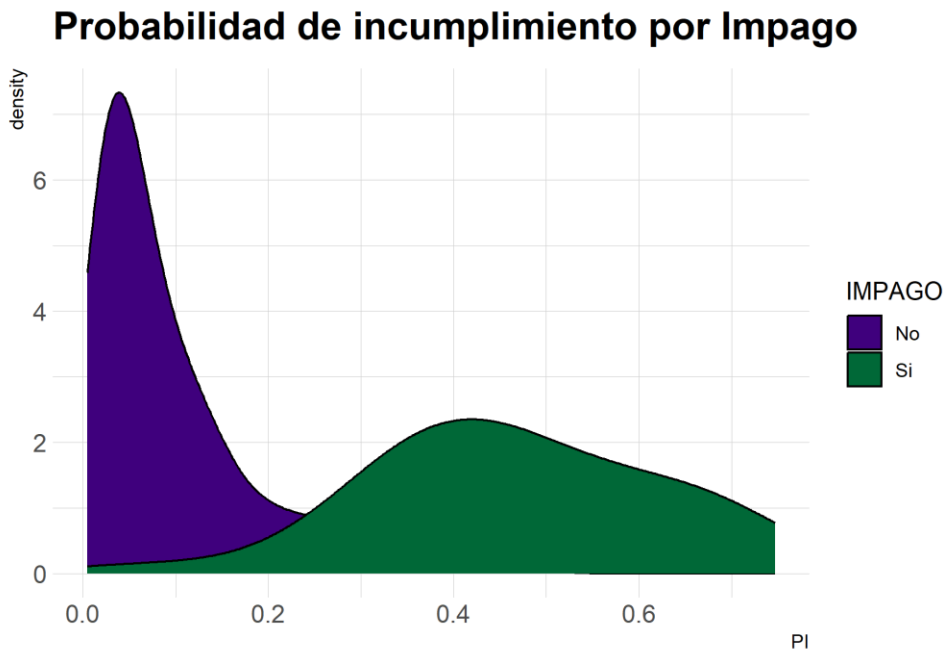
Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por *Orózco* Álvarez, (2025).

En la figura 7 se presentan la repartición de las probabilidades de mora por empresa, previstas con el modelo de random forest.

La mayor parte de las empresas se concentra en probabilidades de incumplimiento cercanas a 0, lo que indica que la mayoría tiene un bajo riesgo de no pagar sus préstamos. A medida que la probabilidad de incumplimiento aumenta, la curva baja rápidamente, lo que muestra que son pocas las empresas que superan el 20% de probabilidad. Sin embargo, aunque sean pocas, es importante prestar atención a las empresas que están en la parte derecha de la distribución, ya que presentan un mayor riesgo y pueden afectar el portafolio de créditos.

Figura 8

Distribución de la PI por Impago



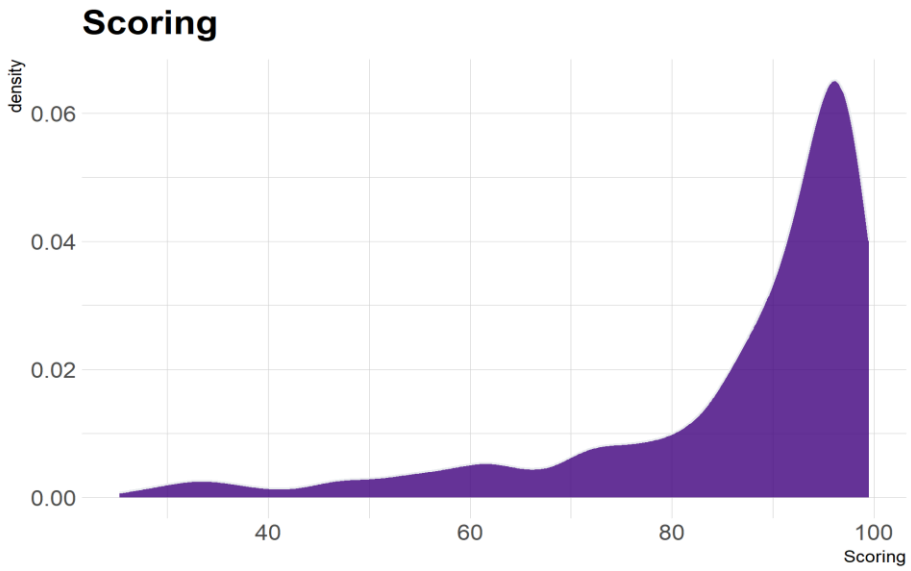
Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

Después, se presenta la probabilidad de incumplimiento (PI) diferenciando entre las empresas que han incumplido y las que no. La curva azul representa a las empresas que no han dejado de pagar, donde la mayoría de los valores se concentran en probabilidades cercanas a 0, lo que indica bajo riesgo. En cambio, la curva verde corresponde a las empresas que sí han incumplido, y en este caso los valores se concentran en probabilidades más altas, aproximadamente entre 0.4 y 0.6. Como es lógico, el modelo asigna mayores probabilidades de incumplimiento a las empresas que efectivamente han fallado en sus pagos. Finalmente, con base en la PI se construye el índice de scoring, que de forma sencilla se define como su complemento, es decir, la probabilidad de que la empresa sí cumpla con sus obligaciones.

$$Scoring = (1 - PI) * 100$$

Figura 9

Distribución del Scoring



Nota. Tomado de Implementación de un modelo de scoring de crédito para empresa objeto de estudio, por Orózco Álvarez, (2025).

La figura 9 muestra la distribución del Scoring, que corresponde al complemento de la probabilidad de incumplimiento (PI) llevado a una escala de 0 a 100. En el eje X se ubican los valores del Scoring, calculados como (1 - PI) * 100, mientras que en el eje Y se observa la densidad, es decir, cómo se concentran las empresas en esos valores. La figura 9 indica que la mayoría de los datos se agrupa en valores altos, especialmente entre 90 y 100, lo que significa que la mayor parte de las empresas tiene una alta probabilidad de cumplir con sus obligaciones. La curva aumenta rápidamente hacia estos valores y alcanza su punto más alto cerca de 100, lo que confirma que predominan empresas con bajo riesgo de incumplimiento y, por lo tanto, con altos niveles de Scoring.

Los resultados de esta investigación confirman y amplían lo que ya han mostrado otros estudios, al evidenciar que el algoritmo Random Forest tiene un mejor desempeño frente a otras técnicas de machine learning. Además, resaltan la importancia de seleccionar adecuadamente las variables y de aplicar buenos procesos de validación al momento de construir modelos predictivos

para el scoring de crédito. En esta misma línea, Liaw y Wiener (2002), en uno de los trabajos más representativos sobre Random Forest, demostraron que este algoritmo es capaz de trabajar con grandes volúmenes de datos y de identificar las variables más relevantes mediante el uso del índice de Gini.

Este estudio va en la misma línea al mostrar que variables como el endeudamiento, el ROA y la razón corriente son fundamentales para mejorar la precisión del modelo. Por otro lado, Chen y Guestrin (2016), en su trabajo sobre Gradient Boosting, destacan que este algoritmo logra mejorar el desempeño del modelo al ir corrigiendo los errores de forma iterativa. Aunque en este caso Random Forest mostró un mejor balance entre precisión y capacidad de generalizar a nuevos datos, también es importante tener en cuenta que Gradient Boosting puede ser una muy buena opción cuando existen relaciones complejas y no lineales entre las variables.

Finalmente, Kuhn y Johnson (2013), en su libro sobre machine learning, resaltan la importancia de usar validación cruzada y de elegir bien los modelos para evitar el sobreajuste. Siguiendo estas recomendaciones, en este estudio se aplicó un proceso cuidadoso de ajuste de hiperparámetros y validación. Gracias a esto, se encontró que Random Forest no solo tuvo buenos resultados en los datos de entrenamiento, sino también en los de prueba, lo que demuestra la importancia de realizar estos procesos de forma adecuada para lograr modelos más confiables y que funcionen bien con datos nuevos.

Regla de decisión

Tabla 1

Reglas de Decisión

Regla de decisión	**# Empresas**
Scoring <65%, En seguimiento	63
Scoring 65%-80%, A comité	56
Scoring >80%, Aprobado	410

Nota. Datos tomados de Orozco Álvarez (2025). Implementación de un modelo de scoring de crédito para empresa objeto de estudio

Esta regla de decisión clasifica a los solicitantes en diferentes categorías de crédito, basándose en la variable Scoring, que refleja su puntuación crediticia.

- Scoring < 65: El solicitante se encuentra en seguimiento, ya que presenta un puntaje bajo que indica un mayor riesgo de crédito.
- Scoring entre 65 y 80: El caso debe ser revisado por un comité para realizar una evaluación más detallada antes de tomar una decisión final.
- Scoring > 80: El crédito es aprobado, ya que el puntaje es alto y esto indica un bajo riesgo.

Finalmente, la validación permitió identificar que el modelo de Árboles de Decisión alcanzó una precisión del 85% en los remuestreos. Al validar estadísticamente qué variables financieras (como el endeudamiento y el ROA) son los mejores predictores de impago, la organización pudo generar una herramienta de apoyo que reduce la subjetividad y optimiza la evaluación del riesgo crediticio, mitigando así el índice de morosidad que presentaba la compañía.

Los anteriores resultados del estudio tienen implicaciones estratégicas, operativas y financieras relevantes para la empresa objeto de estudio. En primer lugar, la identificación del modelo Random Forest como el algoritmo con mejor equilibrio entre precisión (90% en testeo) y capacidad de generalización evidencia la viabilidad de incorporar herramientas de machine learning en el proceso formal de evaluación crediticia. Esto permite reemplazar criterios subjetivos o exclusivamente tradicionales por un sistema cuantitativo robusto, sustentado en validación estadística y ajuste óptimo de hiperparámetros.

Desde una perspectiva de gestión del riesgo, la aplicación de una técnica de eliminación recursiva que redujo el modelo a 12 variables clave implica una mayor eficiencia analítica y una disminución del riesgo de sobreajuste. La exclusión de variables sin poder predictivo significativo —como algunas relacionadas con rentabilidad operacional o reportes externos— demuestra que no necesariamente un mayor número de indicadores mejora la precisión, sino que la calidad y relevancia de los mismos es determinante. Esto optimiza los procesos internos de recolección de información y reduce costos operativos asociados al análisis crediticio.

En términos de toma de decisiones comerciales, la construcción del índice de Scoring como complemento de la Probabilidad de Incumplimiento (PI) permite traducir resultados estadísticos

complejos en una herramienta práctica y fácilmente interpretable para la gerencia. La regla de decisión propuesta (seguimiento, comité o aprobación automática) facilita la segmentación objetiva del portafolio, permitiendo:

Automatizar la aprobación de clientes con bajo riesgo (Scoring > 80), reduciendo tiempos de respuesta y mejorando la competitividad.

Enfocar recursos analíticos y humanos en los casos intermedios (65–80), donde la evaluación cualitativa agrega valor.

Implementar estrategias de monitoreo preventivo para clientes con mayor riesgo (Scoring < 65), mitigando la morosidad futura.

Adicionalmente, la distribución observada de la Probabilidad de Incumplimiento —donde la mayoría de las empresas presenta baja PI— indica que el portafolio tiene un perfil predominantemente saludable. Sin embargo, la existencia de una "cola derecha" con probabilidades superiores al 20% resalta la necesidad de políticas diferenciadas de seguimiento, provisiones y límites de crédito para evitar concentraciones de riesgo.

Capítulo 5: Conclusiones y Recomendaciones

5.1. Conclusiones

El endeudamiento se consolidó como la variable financiera con mayor importancia relativa para predecir el incumplimiento de pago, según la métrica de mean decrease gini. Otros indicadores clave que mejoran significativamente la capacidad de clasificación del modelo son el ROA (Retorno sobre Activos) y la Razón Corriente. Variables como la rentabilidad operacional y el reporte en centrales de riesgo mostraron un impacto relativamente bajo en la pureza de los nodos del modelo para este sector específico.

Se determinó que la configuración óptima del modelo, en términos de maximización de la precisión y estabilidad predictiva, se alcanza con un conjunto de 12 variables. La aplicación de técnicas de eliminación recursiva permitió depurar el modelo mediante la exclusión de tres variables —Rentabilidad Operacional, Concentración a Corto Plazo y Central de Riesgo— que no evidenciaron un aporte estadísticamente significativo al poder predictivo. Este proceso de selección no solo mejoró la parsimonia del modelo, sino que también contribuyó a mitigar el riesgo de sobreajuste (*overfitting*), fortaleciendo su capacidad de generalización y su aplicabilidad en escenarios reales de evaluación crediticia.

El modelo base de Árboles de Decisión alcanzó una precisión promedio del 85% durante los procesos de validación cruzada y remuestreo, evidenciando un desempeño adecuado como aproximación inicial en la estimación del riesgo crediticio; sin embargo, el ajuste sistemático de hiperparámetros mediante RStudio permitió optimizar algoritmos de mayor, logrando configuraciones más robustas, mayor capacidad de generalización y estabilidad estadística en las predicciones. La comparación empírica confirmó que los modelos ensamblados y no lineales superan al modelo base en precisión y reducción del sobreajuste, fortaleciendo la confiabilidad del sistema de scoring. Desde una perspectiva estratégica, la implementación de este modelo personalizado posibilita la transición desde esquemas de evaluación generalistas hacia una herramienta analítica adaptada a las particularidades financieras y operativas del sector construcción, aspecto fundamental en un contexto donde el índice de morosidad alcanzó aproximadamente el 13% en 2023, contribuyendo así a mejorar la gestión integral del riesgo,

optimizar la toma de decisiones crediticias y favorecer la sostenibilidad financiera de la organización.

Esta monografía concluye que confirma que el análisis de los sistemas de credit scoring ha modificado significativamente la manera en las cuales las entidades financieras gestionan el riesgo de impago. Al realizar un análisis comparativo entre modelos de machine learning se demuestra que la capacidad predictiva y la robustez varía ampliamente entre métodos, siendo Random Forest el que muestra un mejor desempeño en términos de precisión y confiabilidad para el contexto específico de análisis empresarial.

Se encuentra que, aunque los Árboles de Decisión son útiles por su facilidad de interpretación, presentan limitaciones relacionadas con el sobreajuste. En contraste, el modelo Random Forest supera estas debilidades al reducir la varianza del modelo mediante el ensamblaje de múltiples árboles, lo que resulta en resultados más estables. Por otro lado, el Gradient Boosting se destaca en la corrección iterativa de errores, siendo especialmente útil en situaciones donde hay relaciones no lineales complejas. Las Redes Neuronales, por su parte, son capaces de captar estructuras complejas en los datos, aunque a menudo a costa de la interpretabilidad.

El modelo elegido, Random Forest, demuestra su capacidad para distinguir adecuadamente entre empresas con alto y bajo riesgo crediticio. Se observa que la mayoría de las empresas analizadas tienen una baja probabilidad de incumplimiento, lo que sugiere que la cartera empresarial es relativamente sólida. Sin embargo, se identifican algunos casos específicos de alto riesgo que requieren atención especial para mitigar su impacto en el portafolio general. Esta clara diferenciación entre niveles de riesgo respalda la efectividad del modelo en la asignación de créditos y la gestión del riesgo financiero.

Además, se ha comprobado que crear un índice de scoring basado en la probabilidad de incumplimiento ofrece una medida directa, comprensible y práctica para la toma de decisiones. Este índice permite a organizaciones como Empresa Objeto de Estudio fortalecer sus políticas de crédito mediante una segmentación más precisa de los solicitantes, optimizando así la asignación de recursos y reduciendo la exposición al riesgo.

El estudio tuvo como objetivo complementario desarrollar un modelo predictivo que se pueda aplicar en el sector empresarial colombiano, utilizando técnicas modernas de análisis de datos. Sin embargo, es importante señalar que su implementación enfrenta ciertas limitaciones,

como el tamaño y la calidad de la base de datos, la representatividad de la muestra y la falta de integración de datos no estructurados, lo que podría influir en la capacidad de generalizar los resultados. Además, interpretar modelos complejos puede requerir conocimientos técnicos específicos, lo que podría ser un obstáculo para que los usuarios no técnicos los adopten de inmediato.

5.2. Recomendaciones

Basado en el contenido y las conclusiones del documento "Diseño de un modelo de Scoring con machine learning", se pueden establecer las siguientes recomendaciones para la empresa y para futuras investigaciones:

Se recomienda a la Empresa Objeto De Estudio sustituir o complementar sus métodos de evaluación tradicionales por el modelo de scoring desarrollado. Dado que el modelo fue entrenado específicamente con datos del sector de la construcción, permitirá reducir el índice de morosidad (que se situaba en el 13%) al identificar con mayor precisión a los clientes de alto riesgo.

Al momento de solicitar documentación a nuevos clientes o renovar cupos de crédito, el área financiera debe poner especial énfasis en los indicadores de endeudamiento, el ROA y la Razón Corriente. El estudio demostró que estas son las variables financieras con mayor peso predictivo para evitar el impago.

Los modelos de Machine Learning pueden perder precisión con el tiempo debido a cambios en la economía o en el comportamiento del mercado. Se recomienda:

Realizar una validación de la precisión del modelo al menos una vez al año.

Reentrenar el algoritmo incorporando nuevos datos históricos para que el sistema aprenda de las tendencias de pago más recientes.

Integrar el modelo (específicamente el algoritmo que haya resultado óptimo tras la validación) en el sistema ERP de la compañía. Esto permitiría obtener una calificación de riesgo instantánea, estandarizando los criterios de aprobación y eliminando la subjetividad del analista humano.

Para mejorar aún más la capacidad de generalización del modelo, se sugiere:

Explorar la inclusión de variables macroeconómicas (como el PIB del sector construcción o tasas de interés) para observar si mejoran la predicción en momentos de crisis económica.

Probar técnicas de ensamble más avanzadas o arquitecturas de redes neuronales más profundas si el volumen de datos de la empresa sigue creciendo.

Es fundamental capacitar al equipo de crédito y cartera en la interpretación de los resultados del modelo. Aunque el algoritmo da una probabilidad de impago, el factor humano sigue siendo clave para entender contextos excepcionales que el dato frío podría no capturar.

Si bien el modelo desarrollado demostró un desempeño predictivo satisfactorio mediante la utilización de variables financieras y no financieras a nivel microeconómico, una de las principales limitaciones del estudio radica en la no incorporación sistemática de variables macroeconómicas específicas del sector construcción colombiano. En este sentido, se recomienda que investigaciones futuras integren indicadores sectoriales y macroeconómicos que puedan capturar efectos cíclicos y condiciones estructurales del entorno económico. Entre estos podrían considerarse el Índice de Construcción, las licencias o permisos de obra aprobados, el comportamiento del PIB del sector construcción, así como variables relacionadas con tasas de interés, inflación sectorial y dinámica del mercado inmobiliario en Colombia.

Referencias

Bai, M., Zheng, Y., & Shen, Y. (2022). Gradient boosting survival tree with applications in credit scoring. *Journal of the Operational Research Society, 73*(1), 39-55. https://doi.org/10.1080/01605682.2021.1919035

Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting loan default in europe with machine learning. *Journal of Financial Econometrics*, *21*(2), 569-596. https://doi.org/10.1093/jjfinec/nbab010

Becerra Molina, E., Ojeda Orellana, R., Lituma Yascaribay, M., & Carrasco Ruano, T. (2023). Análisis de estados financieros, como herramienta útil para la gestión económica tras la pandemia COVID 19. *Revista Universidad y Sociedad, 15*(3), 263-272. https://rus.ucf.edu.cu/index.php/rus/article/view/3745

Behr, A., & Weinblat, J. (2017). Default patterns in seven eu countries: A random forest approach. *International Journal of the Economics of Business*, *24*(2), 181-222. https://doi.org/10.1080/13571516.2016.1252532

Congreso de Colombia. (2008, diciembre 31). Ley 1266 de 2008. *Por la cual se dictan las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial, de servicios y la proveniente de terceros países y se dictan otras disposiciones.*

Congreso de Colombia. (2012, octubre 17). ley 1581 de 2012. *Por la cual se dictan disposiciones generales para la protección de datos personales.*

Chang, Y.-C., Chang, K.-H., Chu, H.-H., & Tong, L.-I. (2016). Establishing decision tree-based short-term default credit risk assessment models. *Communications in Statistics - Theory and Methods*, *45*(23), 6803-6815. https://doi.org/10.1080/03610926.2014.968730

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Davenport, T. H., & Harris, J. (2017). *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning.* Harvard Business Review Press.

Dinçer , H., Yüksel, S, S., & Martínez, L. (2020). Analysis of Balanced Scorecard-based efficiencies of Turkish banking sector with the Analytic Network Process. International Journal of Decision Sciences & Applications, 1(1), 1-12. https://doi.org/10.20525/ijdsa.v1i1.1415

Garzón Escobar, H. F. (2020). *Modelo scoring de comportamiento de cartera de crédito para una empresa de servicios públicos domiciliarios* [tesis de maestría, Universidad de Santander]. Repositorio Institucional UDES. https://repositorio.udes.edu.co/entities/publication/c21409a5-8d8d-4b01-b70b-60743fa4943d

González López, P. P. (2018). *Desarrollo de modelo credit scoring para admisión de facturas en factoring* [Trabajo de grado pregrado, Universidad de Chile]. Repositorio Académico Universidad de Chile. https://repositorio.uchile.cl/handle/2250/149160

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill.

Huacchillo Pardo, L. A., Ramos Farroñan, E. V., & Pulache Lozada, J. L. (2020). La gestión financiera y su incidencia en la toma de decisiones financieras. *Revista Universidad y Sociedad, 12*(2), 356-362. https://rus.ucf.edu.cu/index.php/rus/article/view/1528

IASB. (2014). *NIIF 9 Instrumentos Financieros.*

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer. https://link.springer.com/book/10.1007/978-1-4614-6849-3

Leal Fica, A. L., Aranguiz Casanova, M. A., & Gallegos Mardones, J. (2018). Análisis de riesgo crediticio, propuesta del modelo credit scoring. *Revista Facultad de Ciencias Económicas*, *26*(1), 181–207. https://doi.org/10.18359/rfce.2666

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science*, *8*, 180-201. https://doi.org/10.1016/j.jfds.2022.07.002

Orozco Alvarez, J. E. (2025). *Implemetacion de un modelo de scoring de credito para Feretweros Internacional*. México: UBJ.

Rastegar, M., & Faraji, M. (2023). Development of a Hybrid Credit Scoring Model for the Banking System. *Scientia Iranica*. https://doi.org/10.24200/sci.2023.60399.6778

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC. Disponible en: https://rstudio.com.

Santiago Sandoval, C., & Urán Vélez, A. (2021). *Modelo de credit scoring para la empresa Grupo Factoring de Occidente S.A.S*. [Tesis de maestría, Universidad EAFIT]. Repositorio Institucional Universidad Eafit. https://repository.eafit.edu.co/entities/publication/aac0e017-aabf-4d22-b4bd-164a9052eeb1

Sariev, E., & Germano, G. (2020). Bayesian regularized artificial neural networks for the estimation of the probability of default. *Quantitative Finance*, *20*(2), 311-328. https://doi.org/10.1080/14697688.2019.1633014

Superintendencia Financiera de Colombia. (1995). *Circular Externa 100 de 1995*.

Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *Revista electronica SSRN, 35*(3), 553-572. https://doi.org/10.2139/ssrn.1606674

Thomas, L. C. (2009). *Consumer Credit Models: Development, Evaluation, and Validation*. Oxford University Press.

Tulcanaza Aguilar, M. A. (2021). *Propuesta de un modelo de score de originación para la cartera de consumo de una cooperativa de ahorro y crédito del segmento 3 en el Ecuador* [Tesis de maestria, Universidad Andina Simón Bolívar]. Repositorio Institucional UASB. https://repositorio.uasb.edu.ec/handle/10644/8115