Prototipo De Aplicación Web Basado En Machine Learning Para Optimizar La Gestión De Cartera De Clientes En Una Entidad Aseguradora Colombiana

Fundación Universitaria Agraria de Colombia

Facultad de Ingeniería

Programa de Ingeniería Industrial

Autor(a): Karen Dayanna Bermúdez Martínez, Lucerito Barbosa Ortiz

Asesor: Msc. Yeison Eduardo Conejo Sandoval

Trabajo De Grado Para Optar Al Título De Ingeniera Industrial

Bogotá D.C., Colombia

2026

Resumen

En la actualidad, el sector asegurador en Colombia enfrenta desafíos significativos asociados a la morosidad en el pago de pólizas, afectando la solvencia y liquidez de las entidades. Para abordar esta problemática, esta investigación desarrolló un prototipo de aplicación web basado en machine learning bajo la metodología CRISP-DM, orientado a estimar el nivel de riesgo crediticio (RiskScore en escala de 0 a 100) y a apoyar la transición de la gestión de cartera de un esquema reactivo a uno proactivo y segmentado.

A partir de una base de datos académica y sintética de solicitudes de crédito, con 20.000 registros y 33 variables predictoras, se evaluaron tres algoritmos de regresión supervisada (Regresión Lineal, Random Forest y XGBoost) utilizando una partición de entrenamiento y prueba de 80/20.

Los resultados evidenciaron que la Regresión Lineal presentó el mejor desempeño predictivo y de ajuste, alcanzando un Error Cuadrático Medio (MSE) de 0,150, un Error Absoluto Medio (MAE) de 0,140, una Raíz del Error Cuadrático Medio (RMSE) de 0,387 y un coeficiente de determinación (R^2) de 0,998, superando a XGBoost (R^2 = 0,996) y Random Forest (R^2 = 0,981). El modelo seleccionado fue empaquetado en un Pipeline de procesamiento e integrado en un prototipo de aplicación web interactivo desarrollado con el framework Streamlit, el cual permite a los analistas ingresar variables sociodemográficas y financieras para obtener la estimación del RiskScore en tiempo real. El elevado R2 se explica principalmente porque el RiskScore de la base es un puntaje sintético construido a partir de las mismas variables predictoras; por ello, los resultados demuestran la viabilidad técnica y metodológica del prototipo, pero no su desempeño sobre la cartera de una aseguradora. Los

beneficios operativos en la priorización de clientes, la asignación de recursos y la eficiencia de la cobranza se plantean como efectos esperados que deberán verificarse con datos reales.

Palabras clave: modelo predictivo, machine learning, Regresión Lineal, RiskScore, Streamlit, gestión de cartera, sector asegurador

Abstract

Currently, the insurance sector in Colombia faces significant challenges related to policy payment delinquency, impacting insurers' solvency and liquidity. To address this issue, this research developed a machine learning-based web application prototype following the CRISP-DM methodology, aimed at estimating credit risk levels (RiskScore on a 0 to 100 scale) and supporting the transition of portfolio management from a reactive approach to a proactive, segmented strategy.

Using a synthetic academic dataset of 20,000 credit application records and 33 predictive variables, three supervised regression algorithms (Linear Regression, Random Forest, and XGBoost) were evaluated under an 80/20 train-test split.

The results demonstrated that Linear Regression achieved the best overall predictive performance and fit, obtaining a Mean Squared Error (MSE) of 0.150, a Mean Absolute Error (MAE) of 0.140, a Root Mean Squared Error (RMSE) of 0.387, and a coefficient of determination (R2) of 0.998, outperforming XGBoost (R2 = 0.996) and Random Forest (R2 = 0.981). The selected model was exported within a data processing pipeline and integrated into an interactive web application developed with Streamlit, allowing analysts to input financial and sociodemographic variables to retrieve real-time RiskScore predictions. The

very high R2 is mainly explained by the fact that the dataset's RiskScore is a synthetic score built from the same predictors; therefore, the results demonstrate the technical and methodological feasibility of the prototype rather than its performance on an insurer's portfolio. Operational benefits in customer prioritization, resource allocation and collection efficiency are expected effects that must be verified with real data.

Keywords: predictive model, machine learning, Linear Regression, RiskScore, Streamlit, portfolio management, insurance sector.

Introducción

El sector asegurador cumple un papel estratégico en la economía colombiana, ya que permite a personas y empresas transferir y mitigar sus riesgos, al tiempo que canaliza recursos hacia la inversión. En la última década, las primas emitidas por la industria han mostrado una tendencia creciente, aunque con variaciones asociadas al comportamiento de la economía (Fasecolda, 2024). Este crecimiento ha incrementado el número de clientes y pólizas que deben administrarse, lo que convierte la gestión de cartera y de cobranza en uno de los procesos críticos para la sostenibilidad financiera de las compañías (Herrera Román, 2020).

El incumplimiento en el pago de las primas afecta directamente la liquidez, el flujo de caja y la solvencia de las aseguradoras, y las obliga a constituir provisiones sobre la cartera deteriorada de acuerdo con los lineamientos de la Superintendencia Financiera de Colombia (1995). Según Fasecolda (2024), cerca del 18 % de las pólizas presenta atrasos en los pagos, situación que evidencia la necesidad de pasar de esquemas de cobranza generalizados y reactivos a estrategias segmentadas según el nivel de riesgo de cada cliente.

Frente a este panorama, la analítica de datos y el aprendizaje automático (machine learning) se han consolidado como herramientas para anticipar el comportamiento de pago de los clientes. Estudios comparativos han demostrado que diversos algoritmos de aprendizaje supervisado ofrecen un desempeño competitivo frente a los métodos tradicionales en la estimación del riesgo crediticio (Lessmann et al., 2015), mientras que autores como Grize et al. (2020) y Eling et al. (2022), así como la OECD (2020), destacan que la inteligencia artificial está transformando la cadena de valor del sector asegurador, desde la suscripción hasta la gestión de la cartera.

En el contexto colombiano, la adopción de estas herramientas aún es incipiente. Trabajos como los de Hernández Solano y González Veloza (2022) y Libreros Medrano (2024) evidencian el potencial de los modelos predictivos para mejorar la segmentación de clientes y la asignación de recursos de cobranza. No obstante, su implementación exige atender la disponibilidad y calidad de los datos, la integración tecnológica con los sistemas existentes y el cumplimiento del régimen de protección de datos personales (Ley 1581, 2012).

Con base en lo anterior, este trabajo tiene como objetivo desarrollar un prototipo de aplicación web basado en machine learning, implementado en Python, que permita estimar el nivel de riesgo crediticio de los clientes mediante la predicción del RiskScore, siguiendo las fases de la metodología CRISP-DM (Chapman et al., 2000). Para ello se utilizó una base de datos de carácter académico obtenida de Kaggle (Zoppelletto, s. f.), con 20.000 registros de información financiera, crediticia y sociodemográfica, sobre la cual se entrenaron y compararon los algoritmos de Regresión Lineal, Random Forest (Breiman, 2001) y XGBoost (Chen & Guestrin, 2016). El modelo con mejor desempeño se desplegó en una aplicación web desarrollada con Streamlit (Streamlit Inc., 2024), concebida como un prototipo que,

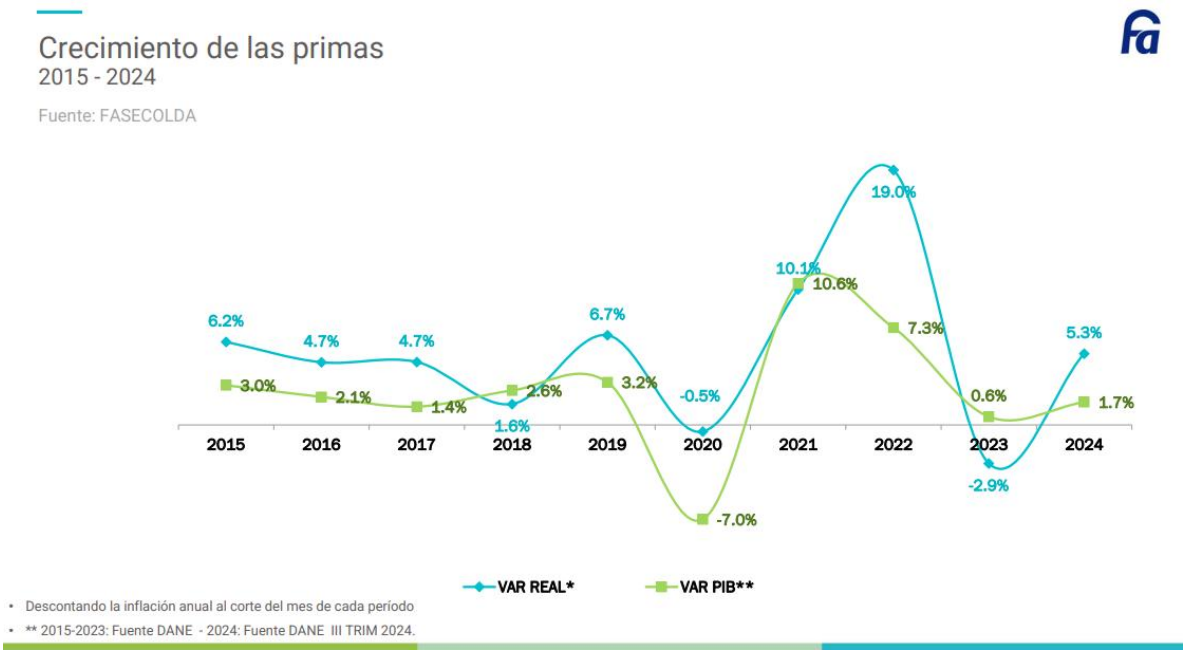
previa validación con datos reales, podría adaptarse a entidades del sector asegurador colombiano.

El documento se organiza de la siguiente manera: en primer lugar, se presentan el planteamiento del problema, la justificación, los objetivos y la metodología; posteriormente, se desarrolla el marco teórico, que incluye el marco legal y el estado del arte; luego, se exponen los resultados en tres capítulos que responden a cada uno de los objetivos específicos; y, finalmente, se presentan la discusión, las recomendaciones y las conclusiones del trabajo.

Planteamiento del problema

El sector asegurador colombiano atraviesa una situación compleja derivada del incremento en la morosidad de los pagos de pólizas, lo que impacta de manera directa la liquidez, la planeación financiera y la sostenibilidad de las compañías, la cual también es relacionada con el comportamiento y crecimiento de las primas frente a los diversos cambios que se presentan por situaciones socio- económicas en el país, tal como se evidencia en la Figura 1 (Fasecolda, 2024).

Figura 1. Crecimiento de las primas 2015–2024

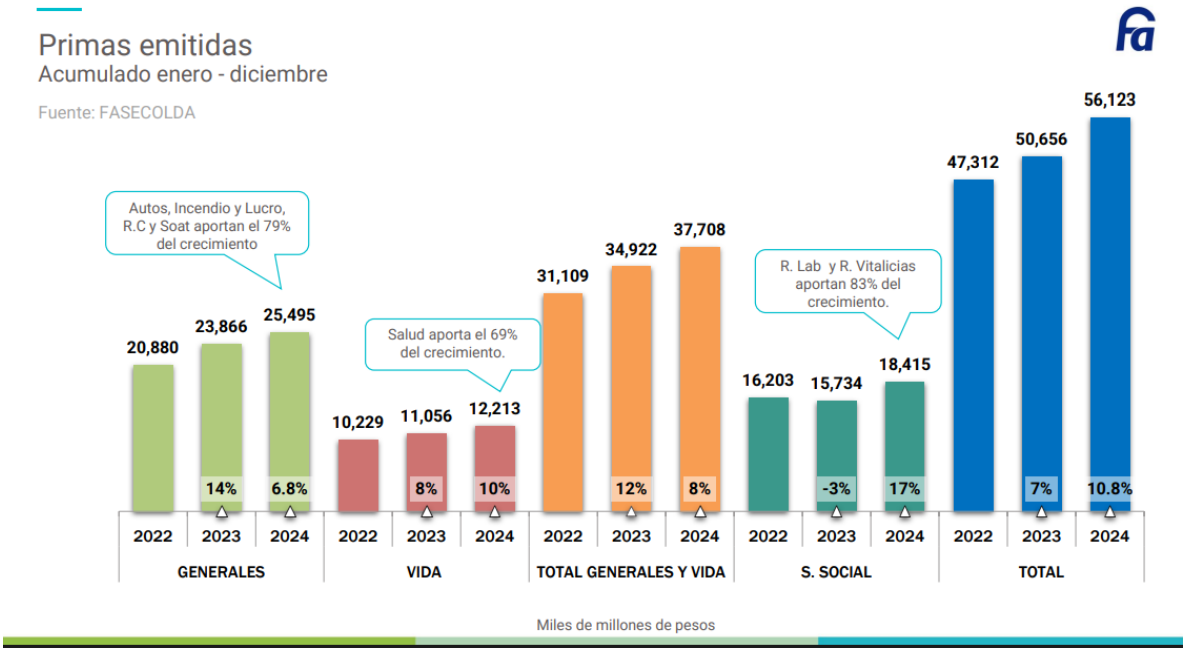


Fuente: Cifras de la Industria diciembre 2024, Fasecolda

Según la Federación de Aseguradores Colombianos (Fasecolda, 2024), cerca del 18% de las pólizas presentan atrasos en los pagos, generando pérdidas que superan los 350 mil millones de pesos anuales. De acuerdo con la Figura 1, se puede evidenciar la irregularidad en el crecimiento de las primas con bajas significativas hacia el año 2022 y 2023 lo cual refleja inestabilidad financiera. Este escenario obliga a las aseguradoras a provisionar recursos significativos para cubrir riesgos de cartera y eleva los costos de recuperación, reduciendo su competitividad en un mercado altamente dinámico.

Las primas emitidas varían según el ramo de seguro; no obstante, al contemplar la totalidad de los ramos entre los años 2022 y 2024, se observa la tendencia presentada a continuación:

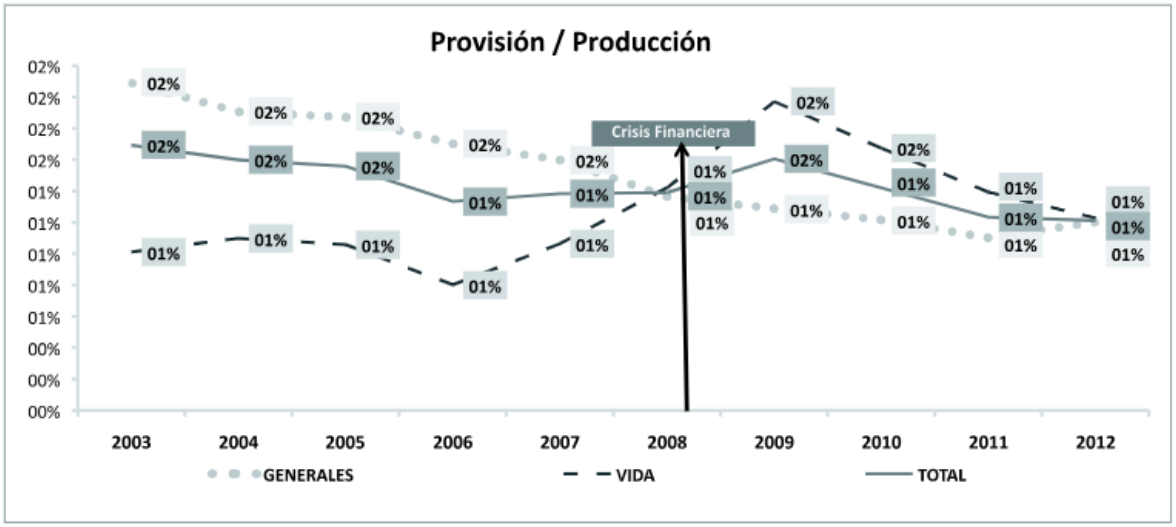
Figura 2. Primas emitidas 2022–2024



Fuente: FASECOLDA (2024). Cifras de la industria, diciembre de 2024.

Actualmente, los procesos de gestión de cartera se ejecutan en gran medida de forma generalizada, sin discriminar el nivel de riesgo de cada cliente (Herrera Román, 2020). Esto genera que los esfuerzos de cobranza sean costosos, poco efectivos y, en muchos casos, reactivos, debido a la ausencia de herramientas tecnológicas capaces de anticipar el comportamiento de pago. En consecuencia, la falta de segmentación y predicción limita la eficiencia operativa, incrementa las pérdidas financieras y afecta la estabilidad del sector asegurador. Por el contrario, la literatura muestra que el uso de datos y de algoritmos de aprendizaje automático permite mejorar la estimación del riesgo y la eficiencia de los procesos a lo largo de la cadena de valor del seguro (Eling et al., 2022; Grize et al., 2020).

Figura 3. Estados financieros transmitidos por las aseguradoras a la Superfinanciera



Fuente: Comportamiento de la cartera de la Industria aseguradora

Ante esta problemática surge la necesidad de implementar un modelo predictivo basado en machine learning que, utilizando datos históricos y características de los asegurados, permita clasificar a los clientes según su probabilidad de incumplimiento de pago, enfoque ampliamente validado en la literatura sobre calificación del riesgo crediticio (Baesens et al., 2016; Lessmann et al., 2015). De esta manera, la entidad aseguradora podrá asignar recursos de cobranza de forma más eficiente, diseñar estrategias diferenciadas (recordatorios, planes de pago, incentivos personalizados) y focalizar esfuerzos en los clientes de mayor riesgo.

La transformación esperada con la implementación de este modelo consiste en pasar de un esquema de cobranza reactivo y homogéneo a un sistema proactivo, segmentado y basado en datos, que contribuya a optimizar la gestión de cartera, reducir los índices de morosidad, fortalecer el flujo de caja y mejorar la sostenibilidad financiera de las aseguradoras, en línea con los principios de adecuada administración del riesgo exigidos por la Superintendencia Financiera de Colombia (1995). Además, permitirá mejorar la

planeación estratégica y la toma de decisiones, apoyando una cultura organizacional orientada hacia la innovación y la eficiencia.

¿Cómo desarrollar un prototipo de aplicación web basado en machine learning para la gestión financiera que facilite la estimación y reducción del impacto de la morosidad en el pago de pólizas sobre la liquidez y viabilidad de las aseguradoras en Colombia?

Justificación

La gestión eficiente de la cartera de clientes es un factor crítico para garantizar la sostenibilidad, competitividad y rentabilidad de las compañías aseguradoras. En el contexto colombiano, el incremento sostenido en la morosidad de pagos ha generado impactos negativos en la liquidez, el margen operativo y la estabilidad financiera del sector, obligando a las empresas a destinar recursos significativos a procesos de cobranza que, en su mayoría, se ejecutan de manera reactiva y poco segmentada.

La implementación de un modelo predictivo basado en machine learning representa una transformación sustancial en la forma de administrar la cartera. Este cambio implica pasar de un esquema tradicional, homogéneo y correctivo, a un sistema proactivo, segmentado y guiado por datos, capaz de:

- Anticipar el riesgo de incumplimiento mediante el análisis de patrones históricos y variables asociadas al comportamiento de pago.

- Clasificar a los clientes según su probabilidad de morosidad, priorizando esfuerzos y recursos en los casos de mayor riesgo.

- Optimizar las estrategias de cobranza, diseñando acciones diferenciadas como recordatorios personalizados, planes de pago flexibles o incentivos específicos.

- Contribuir a reducir los índices de morosidad, fortaleciendo el flujo de caja y la estabilidad financiera de la aseguradora.

Desde la perspectiva académica, este trabajo de grado de pregrado en Ingeniería Industrial constituye una oportunidad para aplicar de manera integrada conocimientos en gestión de procesos, análisis de datos, estadística, programación y optimización de recursos. La propuesta fortalece competencias clave en gestión de riesgos, eficiencia operativa, transformación digital y toma de decisiones basadas en datos, alineándose con las tendencias globales y con las demandas actuales del mercado laboral. Además, potencia el perfil del ingeniero industrial como profesional capaz de liderar proyectos de innovación y mejora continua en sectores estratégicos como el financiero y asegurador.

En el ámbito profesional, la investigación ofrece un procedimiento metodológico replicable que, una vez validado con datos reales, podría adaptarse a diferentes compañías del sector asegurador, contribuyendo a elevar los estándares de gestión financiera y a fomentar una cultura organizacional orientada a la innovación, la eficiencia y la sostenibilidad.

Desde el punto de vista institucional, el proyecto se encuentra alineado con los Objetivos de Desarrollo Sostenible (ODS), especialmente con los relacionados con la industria, innovación e infraestructura (ODS 9) y el trabajo decente y crecimiento económico (ODS 8), ya que promueve el uso de herramientas tecnológicas orientadas a la eficiencia y la

sostenibilidad. Además, impulsa la transformación digital dentro del sector, contribuyendo al desarrollo de prácticas empresariales más competitivas y responsables.

Finalmente, el impacto social potencial de esta propuesta se reflejaría en la estabilidad del sistema asegurador, lo que favorecería el respaldo a los asegurados ante eventualidades. En la medida en que se logren optimizar los costos operativos, las aseguradoras podrían ofrecer primas más competitivas y ampliar el acceso a seguros, favoreciendo la inclusión financiera y la protección patrimonial de familias y empresas. De esta manera, el proyecto no solo constituye un aporte académico y profesional, sino también un mecanismo de impacto económico y social positivo para el país.

Objetivos

Objetivo general

Desarrollar un prototipo de aplicación web basado en técnicas de aprendizaje automático, implementado en Python, que permita estimar el nivel de riesgo crediticio de los clientes a partir de sus características financieras, crediticias y sociodemográficas, mediante la predicción del RiskScore y siguiendo las fases de la metodología CRISP-DM**.**

Objetivos específicos

1. Preparar una base de datos de carácter académico con información financiera, crediticia y sociodemográfica mediante procesos de revisión, limpieza, transformación y estructuración para el modelado del RiskScore.

2. Evaluar el desempeño de modelos de aprendizaje automático en la predicción del RiskScore mediante métricas de error y ajuste, identificando el algoritmo con mejores resultados.

3. Desplegar el modelo de Regresión Lineal seleccionado en una aplicación web desarrollada con Streamlit, permitiendo el ingreso de datos del solicitante y la generación de una predicción del RiskScore.

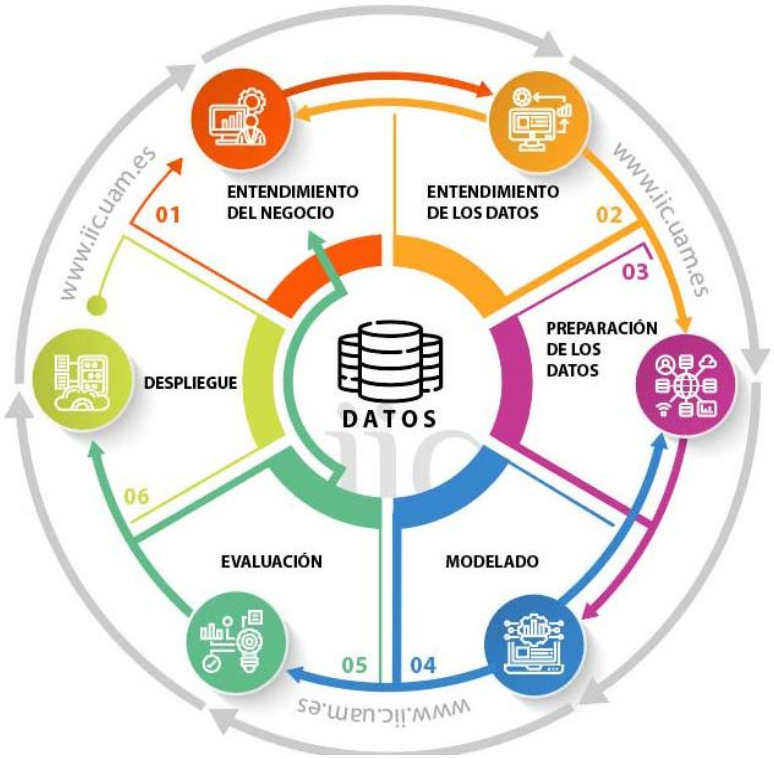
Metodología

El desarrollo de este trabajo de grado tiene un enfoque mixto y corresponde a una investigación no experimental. El componente cuantitativo se evidencia en el tratamiento estadístico de los datos, el entrenamiento de los modelos de aprendizaje automático y su evaluación mediante métricas de error y ajuste (MAE, MSE, RMSE y R^2); el componente cualitativo se refleja en la revisión documental del contexto del sector asegurador, del marco normativo y del estado del arte, así como en la interpretación de los resultados para la toma de decisiones en la gestión de cartera (Creswell & Plano Clark, 2018; Hernández-Sampieri & Mendoza, 2018). Se llevará a cabo siguiendo la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), un estándar ampliamente utilizado en proyectos de analítica avanzada y minería de datos. Esta metodología proporciona un marco estructurado, iterativo y flexible que garantiza la calidad, trazabilidad y replicabilidad del proceso de construcción del modelo predictivo.

El proyecto se desarrollará en seis fases:

Figura 4. Esquema del ciclo CRISP-DM estándar

Fuente: Instituto de Ingeniería del conocimiento



1. Comprensión del negocio: En esta fase se realizará un análisis del contexto del sector asegurador colombiano, con énfasis en la problemática de la morosidad en el pago de pólizas. Se identificarán los objetivos del proyecto y los indicadores clave de desempeño (KPIs) relacionados con la gestión de cartera. El propósito es alinear el modelo predictivo con la problemática estudiada y definir claramente los requerimientos funcionales de la solución.

2. Comprensión de los datos: Se recopilará y explorará la información histórica de clientes, pagos, características de las pólizas y variables relevantes. Se evaluará la calidad de los datos, identificando valores faltantes, inconsistencias o posibles sesgos.

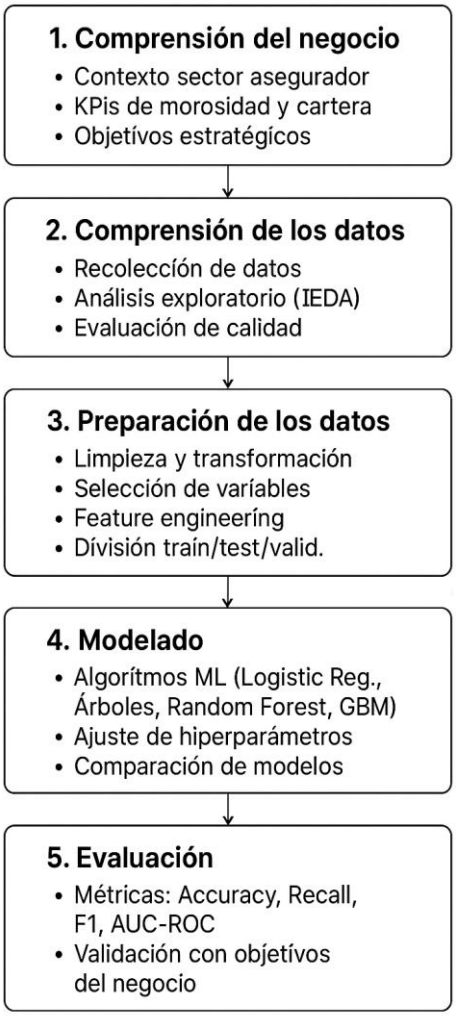
3. Preparación de los datos: En esta etapa se llevará a cabo la limpieza, transformación y normalización de los datos, así como la selección de variables relevantes para el entrenamiento del modelo.

4. Modelado: Se implementarán y entrenarán algoritmos de aprendizaje automático en Python. Se ajustarán los parámetros buscando maximizar el rendimiento del modelo y su capacidad predictiva.

5. Evaluación: Se validará el desempeño del modelo utilizando métricas estadísticas apropiadas. Además, se analizará su capacidad para clasificar adecuadamente el riesgo de morosidad y su coherencia con los objetivos planteados.

6. Despliegue: En esta fase se desarrollará una aplicación web interactiva utilizando el framework Streamlit, que permitirá visualizar los resultados del modelo predictivo. La aplicación servirá como prototipo funcional para la consulta de predicciones, generación de reportes y visualización de segmentaciones de riesgo, facilitando la interpretación de los resultados por parte del usuario final.

Figura 5. Fases de la metodología CRISP-DM aplicadas al proyecto



Fuente: Autores

Marco teórico

El marco teórico de este trabajo se estructura en cinco ejes fundamentales: gestión de cartera en el sector asegurador, modelos predictivos aplicados a la gestión de cartera, machine learning en el sector asegurador, marco legal y estado del arte.. Estos ejes se interrelacionan para sustentar la propuesta de un modelo predictivo que optimice la administración de la cartera de clientes en una compañía de seguros en Colombia.

1. Gestión de cartera en el sector asegurador

La gestión de cartera se define como el conjunto de procesos, estrategias y herramientas orientadas a administrar las obligaciones financieras de los clientes, garantizando el recaudo oportuno y minimizando el riesgo de incobrabilidad (Herrera Román, 2020). En el sector asegurador, esta gestión se centra en el seguimiento y control de las primas de seguros emitidas, cuyo incumplimiento afecta directamente la liquidez y la capacidad de respuesta de la compañía.

En Colombia, la morosidad en pólizas de seguros ha mostrado una tendencia creciente. Según la Federación de Aseguradores Colombianos – Fasecolda (2024), el promedio nacional de morosidad alcanzó el 18 %, lo que representa pérdidas anuales superiores a 350 mil millones de pesos (ver Tabla 1). Este escenario obliga a las aseguradoras a provisionar recursos significativos para cubrir riesgos de cartera, incrementando los costos de recuperación y reduciendo su competitividad.

Tabla 1. Indicadores de morosidad en pólizas de seguros en Colombia (2021–2024)

Año	Número total de pólizas (millones)	Pólizas en mora (miles)	Porcentaje de morosidad (%)
2021	3,2	420	13,1
2022	3,5	520	14,8
2023	3,8	630	16,5
2024	4,1	738	18,0

Fuente: Fasecolda (2024).

Una gestión de cartera eficiente no solo impacta la estabilidad financiera, sino que también fortalece la relación con los clientes, mejora la planeación estratégica y permite diseñar políticas de cobranza más efectivas y segmentadas.

2. Modelos predictivos aplicados a gestión de cartera

Los modelos predictivos son herramientas analíticas que, a partir de datos históricos, permiten estimar la probabilidad de que un evento ocurra en el futuro (Hernández Solano & González Veloza, 2022). En el contexto de la gestión de cartera, estos modelos permiten:

- Clasificar clientes según su probabilidad de incumplimiento.

- Priorizar acciones de cobranza en función del riesgo.

- Optimizar recursos humanos y financieros.

- Reducir pérdidas y mejorar la recuperación de cartera.

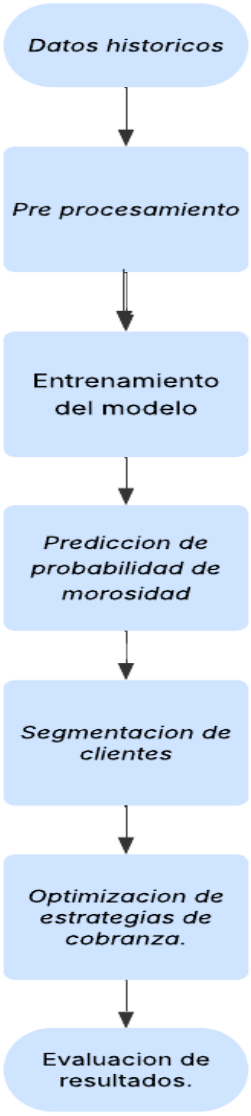
Entre los algoritmos más utilizados para la predicción de morosidad se encuentran los presentados en la Tabla 2.

Tabla 2. Algoritmos más utilizados en la predicción de morosidad

Algoritmo	Características	Ventajas
Regresión logística	Modelo estadístico lineal para variables binarias	Interpretabilidad y rapidez
Random Forest	Conjunto de árboles de decisión	Alta precisión y manejo de datos no lineales
XGBoost	Algoritmo de gradient boosting optimizado	Excelente rendimiento en datasets desbalanceados
Redes neuronales	Modelos inspirados en el cerebro humano	Capacidad de modelar relaciones complejas

Fuente: Elaboración propia con base en Breiman (2001), Chen y Guestrin (2016) y Hastie et al. (2017).

Figura 6. Flujo conceptual del modelo predictivo propuesto



Fuente: Elaboración propia.

3. Machine learning en el sector asegurador

El uso de **machine learning** es una rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y realizar predicciones sin ser programados explícitamente para cada caso (James et al., 2021). Su aplicación en el sector asegurador ha demostrado beneficios como:

- Automatización de tareas repetitivas en la gestión de cartera.

- Mayor precisión en la detección de patrones de comportamiento de pago.

- Personalización de estrategias de cobranza y planes de pago.

- Segmentación dinámica de clientes según riesgo.

Un ejemplo relevante es el uso de XGBoost, que ha mostrado gran capacidad para manejar conjuntos de datos desbalanceados, comunes en escenarios de morosidad (Libreros Medrano, 2024).

4. Marco legal

Este trabajo se encuentra dentro del conjunto de normativas vigentes en Colombia que regulan el sector asegurador, la gestión financiera, los datos personales y el uso de recursos tecnológicos para poder tomar decisiones. Dichas normas construyen el soporte jurídico que respalda el desarrollo del modelo predictivo propuesto por este trabajo de grado que permitirá en esta ocasión optimizar la gestión de cartera de clientes en una entidad aseguradora.

De acuerdo con la Constitución Política de Colombia, el artículo 333 establece la libertad económica y la iniciativa privada dentro de los límites del bien común. Asimismo, el artículo 335 señala que las actividades financiera, bursátil, aseguradora y cualquier otra relacionada con el manejo de recursos del público son de interés público y deben ser reguladas y vigiladas por el Estado.

El Decreto 663 de 1993 (Estatuto Orgánico del Sistema Financiero) regula las actividades de las entidades aseguradoras en Colombia, estableciendo los principios de solvencia, liquidez y adecuada administración del riesgo, aspectos que se encuentran directamente relacionados con la gestión de cartera y el control de la morosidad. El desarrollo de modelos predictivos orientados a anticipar el incumplimiento de pagos se ajusta con los objetivos de estabilidad financiera y la adecuada gestión del riesgo que se contemplan en este estatuto.

Ley 1328 de 2009-Regimen de protección al consumidor financiero establece los derechos y deberes de los consumidores financieros y las obligaciones de las entidades vigiladas por la superintendencia Financiera de Colombia. Esta ley está promoviendo practicas transparentes y responsables en la gestión de las relaciones con los clientes, incluyendo los procesos de cobranza. El modelo propuesto en este trabajo de grado permite el diseño de estrategias de gestión de carteras más segmentadas y personalizadas, contribuyendo a un trato equitativo y adecuado de los asegurados.

Ley 1266 de 2008- Habeas data financiero la cual regula el manejo de la información financiera, crediticia, comercial y de servicios. Esta norma es fundamental para el desarrollo del modelo predictivo propuesto, ya que establece los lineamientos para la recolección,

tratamiento y uso de datos relacionados con el comportamiento de pago de los clientes. El uso de información histórica para la construcción del modelo se realiza bajo los principios de veracidad, seguridad y confidencialidad establecidos por esta ley.

Ley 1581 de 2012 y decreto 1377 de 2013- Protección de datos personales, esta ley establece el régimen general de protección de datos personales en Colombia, estas normas son especialmente relevantes para la investigación, dado que el modelo predictivo usa datos de clientes para la clasificación del riesgo de morosidad. El cumplimiento de esta norma garantiza que el tratamiento de la información se realice con confidencialidad y fines legítimos, asegurando la protección de los derechos de los titulares de los datos.

Circular básica contable y financiera- Superintendencia financiera de Colombia define los lineamientos para el registro, control y provisión de cartera en las entidades aseguradoras. Esta norma establece los criterios para la medición del riesgo crediticio y la constitución de provisiones, lo cual está directamente relacionado con la necesidad de contar con herramientas analíticas que permitan anticipar el deterioro de la cartera y mejorar notablemente la toma de decisiones financieras.

Normativa sobre gobierno y gestión de riesgos de la superintendencia financiera de Colombia promueve el uso de metodologías y herramientas que fortalezcan la gestión integral de riesgos en las entidades aseguradoras. El modelo predictivo basado en machine learning propuesto en este trabajo se constituye como un mecanismo de apoyo a la gestión del riesgo de cartera, alineado con las buenas prácticas de control interno y gestión financiera.

Normas técnicas y lineamientos internacionales como la ISO31000 sobre la gestión del riesgo y las buenas prácticas en analítica de datos y modelos predictivos sirven como

referencia técnica para el diseño del modelo propuesto. Estas normas respaldan el uso de metodologías estructuradas y basadas en datos para la identificación, análisis y mitigación de riesgos financieros.

En conclusión, este conjunto de normas mencionadas anteriormente proporciona el respaldo jurídico para el desarrollo del modelo predictivo propuesto, garantizando que la gestión de cartera se realice dentro de los principios de legalidad, transparencia, protección de datos y estabilidad financiera. De esta manera, el trabajo de grado no solo aporta una solución técnica e innovadora al problema de la morosidad, sino que también se encuentra alineado con el marco normativo que regula el sector asegurador colombiano.

5. Estado del arte

Se realizó una revisión de literatura académica en bases de datos como Scopus, IEEE Xplore y ScienceDirect, complementada con publicaciones sectoriales de Fasecolda.

5.1. Antecedentes internacionales

A nivel global, aseguradoras en mercados como Estados Unidos, España y Reino Unido han reportado la aplicación de modelos de machine learning para anticipar incumplimientos y apoyar la gestión de cobranza; en la Tabla 3 se resumen los estudios revisados y los hallazgos informados por sus autores.

Tabla 3. Modelos predictivos aplicados a la gestión de cartera en el contexto internacional

Autor / Año	Algoritmo	Sector	Hallazgos clave
Lee & Kim (2021)	Random Forest	Seguros de vida	Redujo morosidad en un 25 % mediante segmentación de clientes.
Hernández et al. (2022)	XGBoost	Seguros generales	Mejoró la precisión de predicción al 91 %.

Autor / Año	Algoritmo	Sector	Hallazgos clave
Wang et al. (2023)	Redes neuronales	Seguros médicos	Identificó patrones de incumplimiento con datasets desbalanceados.
Patel (2024)	LightGBM	Banca y seguros	Logró optimizar la recuperación de cartera en un 30 %.
Lopes, Ladeira & Carvalho (2017)	Árboles de decisión y SVM	Microfinanzas y seguros	Predijo con 85 % de exactitud la probabilidad de recuperación de créditos morosos mediante machine learning supervisado.
Appel et al. (2020)	Random Forest	Seguros corporativos	Desarrolló un modelo para predecir tiempos de pago y reducir cuentas vencidas en 20 %.
Kureljusic & Metz (2023)	Regresión logística y Boosting	Seguros comerciales	Determinó que la integración de ML en accounts receivable mejora la eficiencia de cobranza en 28 %.
Maisog et al. (2019)	Ensemble Learning	Seguros de salud	Implementó ML sobre 48 M de registros y logró identificar asegurados de alto riesgo con AUC > 0,87.
Quan et al. (2024)	Redes neuronales profundas	Seguros empresariales	Propuso un sistema InsurTech que redujo pérdidas por impago en un 18 % mediante predicción temprana.

Fuente: Elaboración propia con base en los estudios citados en la tabla.

5.2. Antecedentes nacionales

En Colombia, el uso de modelos predictivos en el sector asegurador está en una fase de adopción temprana. Investigaciones como la de Libreros Medrano (2024) han demostrado que la implementación de modelos de machine learning puede mejorar significativamente la segmentación de clientes y la asignación de recursos de cobranza. Hernández Solano y González Veloza (2022) identificaron que el modelo Linear Discriminant Analysis obtuvo el mejor desempeño en su estudio para predecir incumplimientos en seguros de vida. En la Tabla 4 se presentan algunas implementaciones reportadas en aseguradoras del país.

Tabla 4. Implementaciones de modelos predictivos en aseguradoras colombianas

Entidad	Algoritmo	Objetivo principal	Impacto reportado
Seguros Bolívar	Regresión logística	Predecir morosidad temprana	Disminución de morosidad en 12 %
SURA	XGBoost	Segmentar clientes por riesgo	Incremento del 15 % en recuperación de cartera
Mapfre Colombia	Random Forest	Optimizar procesos de cobranza	Reducción del tiempo de gestión en 20 %
Liberty Seguros	Árboles de decisión	Identificar clientes propensos a cancelar pólizas	Reducción del 10 % en tasa de cancelación
Axa Colpatria	Redes neuronales	Predecir incumplimiento en primas de seguros de vida	Mejora del 18 % en exactitud de predicción
Allianz Colombia	LightGBM	Estimar probabilidad de atraso en pagos de pólizas empresariales	Incremento del 22 % en eficiencia de cobranza
HDI Seguros	SVM (Máquinas de soporte vectorial)	Clasificar clientes según riesgo crediticio	Reducción del 14 % en impagos recurrentes
Equidad Seguros	K-Means + Regresión logística	Analizar segmentación por comportamiento de pago	Mejora del 16 % en estrategias de retención
Fasecolda (2024, estudio sectorial)	Ensemble (combinado)	Evaluar impacto del ML en la gestión de riesgo de cartera	Evidenció que las aseguradoras con ML mejoran su rentabilidad en 11 % frente a las que usan métodos tradicionales

Fuente: Fasecolda (2024).

5.3. Brechas identificadas

Tras revisar la literatura y ejemplos del país, se encuentran tres brechas clave:

1. Falta de modelos ajustados:

La mayoría de los modelos se crean sobre fundamentos generales, sin tener en cuenta las particularidades de cada compañía de seguros.

2. Integración tecnológica limitada:

Muchas organizaciones todavía manejan su cartera utilizando métodos manuales o con poca automatización.

3. Poca simulación de escenarios:

Existen pocos análisis que utilicen conjuntos de datos simulados representativos para comprobar el modelo antes de que se ponga en práctica.

6. Resultados

Capítulo 1. Obtención y preparación de los datos para la aplicación

El desarrollo de los objetivos específicos se realizó a partir de la metodología CRISP-DM, considerando las etapas de comprensión de los datos, preparación de la información, modelado, evaluación y despliegue. El propósito fue construir un prototipo de aprendizaje automático capaz de apoyar la estimación del riesgo crediticio mediante la predicción del RiskScore. Aunque el contexto del proyecto se relaciona con la gestión de cartera de una aseguradora colombiana, la base utilizada es de carácter académico, de naturaleza sintética, y fue descargada de Kaggle. Por tanto, sus registros corresponden a solicitudes de crédito y no a pólizas ni primas de una aseguradora, lo que define el alcance exploratorio de los resultados que se presentan en este capítulo.

Descripción general del proceso de obtención de datos

Este objetivo se desarrolló mediante la preparación de una base de datos con información financiera, crediticia y sociodemográfica de solicitantes de crédito. El proceso incluyó la identificación y descarga de la fuente, la revisión de la estructura de los datos, la limpieza de registros, la transformación de variables y la organización de la información requerida para el modelado del RiskScore. La información fue utilizada como aproximación académica a los datos que podrían emplearse en procesos de análisis y gestión de cartera.

Limitaciones de la correspondencia entre los datos y el problema de investigación

La base utilizada no proviene de una entidad aseguradora. De acuerdo con su autor, corresponde a un conjunto de datos sintético de 20.000 registros, generado mediante código para facilitar el desarrollo de modelos de evaluación del riesgo y de aprobación de créditos (Zoppelletto, s. f.). En consecuencia, existe una brecha entre la unidad de análisis de los datos (el solicitante de un crédito) y la unidad de análisis del problema (el tomador de una póliza con primas pendientes de pago), que limita el alcance de los resultados en los siguientes aspectos:

Ausencia de variables propias del negocio asegurador: la base no incluye el ramo, el valor y la periodicidad de la prima, la forma de recaudo, la antigüedad de la póliza, el canal de venta ni el historial de pago de primas, que son determinantes directos de la morosidad en seguros.

Variable objetivo no observada: el RiskScore no registra un incumplimiento real en el pago de primas, sino un puntaje de riesgo crediticio generado de forma sintética; por ello, el modelo estima un indicador aproximado (proxy) del riesgo y no la probabilidad de mora de una póliza.

Contexto geográfico y temporal: los datos no representan a la población asegurada colombiana ni sus condiciones socioeconómicas, de modo que no es posible extrapolar los valores obtenidos al mercado nacional.

Ausencia de datos de gestión: la base no contiene información sobre acciones de cobranza, costos, tiempos de recaudo ni resultados de recuperación, por lo que no permite medir el efecto del prototipo sobre la eficiencia del proceso de cobranza.

Por estas razones, los resultados de este trabajo deben interpretarse como una demostración de la viabilidad técnica y metodológica del flujo CRISP-DM y del prototipo, y no como evidencia de su desempeño sobre la cartera de una aseguradora. Su transferencia al sector asegurador exige reentrenar y validar el modelo con datos reales, como se plantea en las recomendaciones.

La fuente original consultada fue Kaggle. La base utilizada en el proyecto corresponde a una versión descargada y posteriormente adaptada para el análisis, denominada BASE_CREDITOS_CORREGIDA_RISKSCORE_0_100.xlsx. Las características generales de la fuente se presentan en la Tabla 5.

Fuente de los datos

Tabla 5. Fuente y características generales de los datos

Elemento	Descripción
Fuente	Kaggle
Nombre de la base original	Financial Risk for Loan Approval
Autor	Lorenzo Zoppelletto
Enlace	https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval/data
Archivo utilizado	BASE_CREDITOS_CORREGIDA_RISKSCORE_0_100.xlsx

Elemento	Descripción
Transformaciones realizadas	Corrección y adaptación de la base original; limpieza, tratamiento de valores faltantes y estandarización de RiskScore en una escala de 0 a 100.

Fuente: Elaboración propia con base en Zoppelletto (s. f.).

Durante la etapa de preparación de los datos se creó una copia de la base original para realizar las transformaciones sin modificar el archivo fuente. Posteriormente, se estandarizó la variable ApplicationDate al formato de fecha y se convirtió TotalDebtToIncomeRatio a formato numérico. También se realizó una limpieza específica de la variable RiskScore, debido a la presencia de registros con formatos inconsistentes, como valores que contenían más de un punto decimal. Finalmente, se aplicaron las transformaciones a los registros y se verificaron los tipos de datos mediante la función info(), con el propósito de comprobar que las variables estuvieran correctamente estructuradas para las siguientes etapas del modelamiento. En la Tabla 6 se resumen las transformaciones realizadas.

Tabla 6. Transformaciones realizadas a las variables de la base de datos

Variable	Problema	Solución
ApplicationDate	Fechas almacenadas en un formato que debía estandarizarse	Conversión a formato fecha
TotalDebtToIncomeRatio	Valores que podían estar almacenados como texto	Conversión a numérico
RiskScore	Valores con formato inconsistente, incluyendo múltiples puntos	Limpieza y conversión a número
Toda la base	Riesgo de modificar los datos originales	Se creó una copia
Base después de limpiar	Necesidad de comprobar las transformaciones	df_modelo.info()

Fuente: Elaboración propia.

Tratamiento y limpieza de los datos

Después de descargar la información, se realizó una revisión inicial de la estructura de la base para identificar el número de registros, las variables disponibles, sus tipos de datos y la presencia de valores faltantes. En particular, se verificó que no existieran celdas vacías en las variables utilizadas durante el análisis, debido a que estos valores podían generar errores en la ejecución del código y afectar el entrenamiento de los modelos.

Para apoyar esta actividad se utilizó inteligencia artificial como herramienta de asistencia en la generación y revisión de un código en Python. El código permitió identificar automáticamente los valores nulos mediante funciones de análisis de datos y aplicar un procedimiento de imputación de acuerdo con la naturaleza de cada variable. En las variables numéricas se utilizó el método definido en el código del proyecto —por ejemplo, media o mediana— y en las variables categóricas se aplicó el método correspondiente, como la moda o una categoría de identificación de datos faltantes.

Posteriormente, se ejecutó una segunda verificación para confirmar que no permanecieran celdas vacías en las variables seleccionadas. También se revisaron los tipos de datos, la consistencia de las categorías, los rangos de las variables numéricas y la codificación de las variables binarias. Las variables categóricas fueron transformadas a un formato compatible con los algoritmos de aprendizaje automático.

El tratamiento realizado permitió obtener una base estructurada y preparada para las etapas de análisis exploratorio, división entre datos de entrenamiento y prueba y construcción

de los modelos predictivos. La inteligencia artificial se utilizó únicamente como apoyo técnico; las decisiones sobre la limpieza y la imputación fueron revisadas por el equipo investigador. La Tabla 7 describe la metadata de las variables incluidas en la base.

Metadata de la base de datos

Tabla 7. Metadata de la base de datos

Variable	Tipo de dato	Tipo de variable	Descripción	Unidad o valores esperados
ApplicationDate	Fecha	Temporal	Fecha de solicitud de crédito.	Fecha calendario
Age	Entero	Numérica	Edad del solicitante.	Años
AnnualIncome	Decimal	Numérica	Ingreso anual estimado.	Moneda
CreditScore	Entero/decimal	Numérica	Puntaje crediticio.	Escala de la fuente
Experience	Entero	Numérica	Experiencia laboral.	Años
LoanAmount	Decimal	Numérica	Valor solicitado.	Moneda
LoanDuration	Entero	Numérica	Duración del crédito.	Meses o unidad de la fuente
NumberOfDependents	Entero	Numérica discreta	Número de dependientes.	Conteo
MonthlyDebtPayments	Decimal	Numérica	Pagos mensuales de deuda.	Moneda
EmploymentStatus	Texto	Categórica nominal	Situación laboral.	Categorías de la base
EducationLevel	Texto	Categórica ordinal	Nivel educativo.	Categorías ordenables
MaritalStatus	Texto	Categórica nominal	Estado civil.	Categorías de la base
HomeOwnershipStatus	Texto	Categórica nominal	Tenencia de vivienda.	Categorías de la base
LoanPurpose	Texto	Categórica nominal	Propósito del crédito.	Categorías de la base
BankruptcyHistory	Binaria	Categórica binaria	Antecedentes de bancarrota.	0 = no; 1 = sí
PreviousLoanDefaults	Binaria	Categórica binaria	Incumplimientos anteriores.	0 = no; 1 = sí

Variable	Tipo de dato	Tipo de variable	Descripción	Unidad o valores esperados
LoanApproved	Binaria	Categórica binaria	Aprobación del crédito.	0 = no; 1 = sí
DebtToIncomeRatio	Decimal	Numérica	Relación deuda-ingreso.	Proporción o porcentaje
CreditCardUtilizationRate	Decimal	Numérica	Utilización de tarjetas.	Proporción o porcentaje
TotalAssets	Decimal	Numérica	Total de activos.	Moneda
TotalLiabilities	Decimal	Numérica	Total de pasivos.	Moneda
NetWorth	Decimal	Numérica	Activos menos pasivos.	Moneda
RiskScore	Decimal	Numérica continua	Puntaje de riesgo crediticio.	0 a 100

Fuente: Elaboración propia con base en Zoppelletto (s. f.).

Capítulo 2. Aplicación y evaluación de modelos de machine learning para la predicción del RiskScore

Estrategia de entrenamiento y evaluación de los modelos

Este objetivo se desarrolló mediante la implementación y evaluación de tres algoritmos de regresión supervisada: Regresión Lineal, Random Forest Regressor y XGBoost Regressor. Los tres modelos utilizaron la misma variable objetivo continua, RiskScore, y fueron evaluados bajo una estrategia común de entrenamiento y prueba.

Para el entrenamiento, la base se dividió en un 80 % para entrenamiento y un 20 % para prueba, utilizando random_state=42. Las variables numéricas fueron imputadas mediante la mediana y estandarizadas con StandardScaler; las variables categóricas fueron imputadas con la moda y codificadas mediante OneHotEncoder. El desempeño se valoró con MAE, MSE, RMSE y R^2, lo que permitió comparar el error de predicción y la capacidad de ajuste de cada modelo.

Construcción y alcance de la variable RiskScore

La variable objetivo RiskScore no es un registro observado de incumplimiento, sino un puntaje generado. Según la documentación de la fuente, la base es sintética y fue creada para dos tareas: la regresión del RiskScore, entendido como un puntaje asociado a la probabilidad de incumplimiento o de inestabilidad financiera del solicitante, y la clasificación binaria de la aprobación del crédito (Zoppelletto, s. f.). El código de generación publicado con la base calcula el puntaje a partir de componentes de riesgo asignados a variables como la tasa de utilización de tarjetas de crédito, la historia crediticia y otras características financieras del solicitante, es decir, a partir de las mismas variables que se emplean como predictoras en este trabajo.

Sobre la variable original, se realizó dos intervenciones: la corrección de los registros con formato inconsistente descrita en la Tabla 6 y su expresión en una escala de 0 a 100. El RiskScore utilizado en el análisis ya venía precalculado en el archivo original BASE_CREDITOS_CORREGIDA_RISKSCORE_0_100.xlsx. Por tanto, en el cuaderno de Python no se implementó una transformación matemática adicional para escalar esta variable al rango de 0 a 100.

En Python se realizaron operaciones de limpieza y preparación de la variable: conversión a formato numérico, tratamiento de registros con múltiples puntos decimales e imputación de valores faltantes mediante la mediana. Estas operaciones preparan la variable para el análisis, pero no constituyen el procedimiento original de escalamiento.

Posteriormente, el RiskScore se clasificó en tres categorías mediante umbrales explícitos: riesgo bajo para puntuaciones menores o iguales a 40; riesgo medio para puntuaciones mayores que 40 y menores o iguales a 70; y riesgo alto para puntuaciones superiores a 70.

Esta forma de construcción tiene una implicación metodológica central: dado que el RiskScore se obtiene a partir de variables incluidas en el conjunto de predictores, los modelos no aprenden un fenómeno de mora observado, sino que reconstruyen la regla con la que se generó el puntaje. Por ello, las métricas de error y de ajuste que se reportan en este capítulo miden la capacidad de los algoritmos para reproducir dicha regla y no su capacidad para anticipar el incumplimiento real en el pago de primas.

Para resumir el flujo lógico del entrenamiento, evaluación y selección del modelo con mejor desempeño, la Figura 7 sintetiza el algoritmo desarrollado en Python. Este proceso abarca desde la carga y preparación de las variables predictoras hasta la persistencia del modelo definitivo mediante la librería Joblib.

Figura 7. Pseudocódigo – Objetivo Específico 2

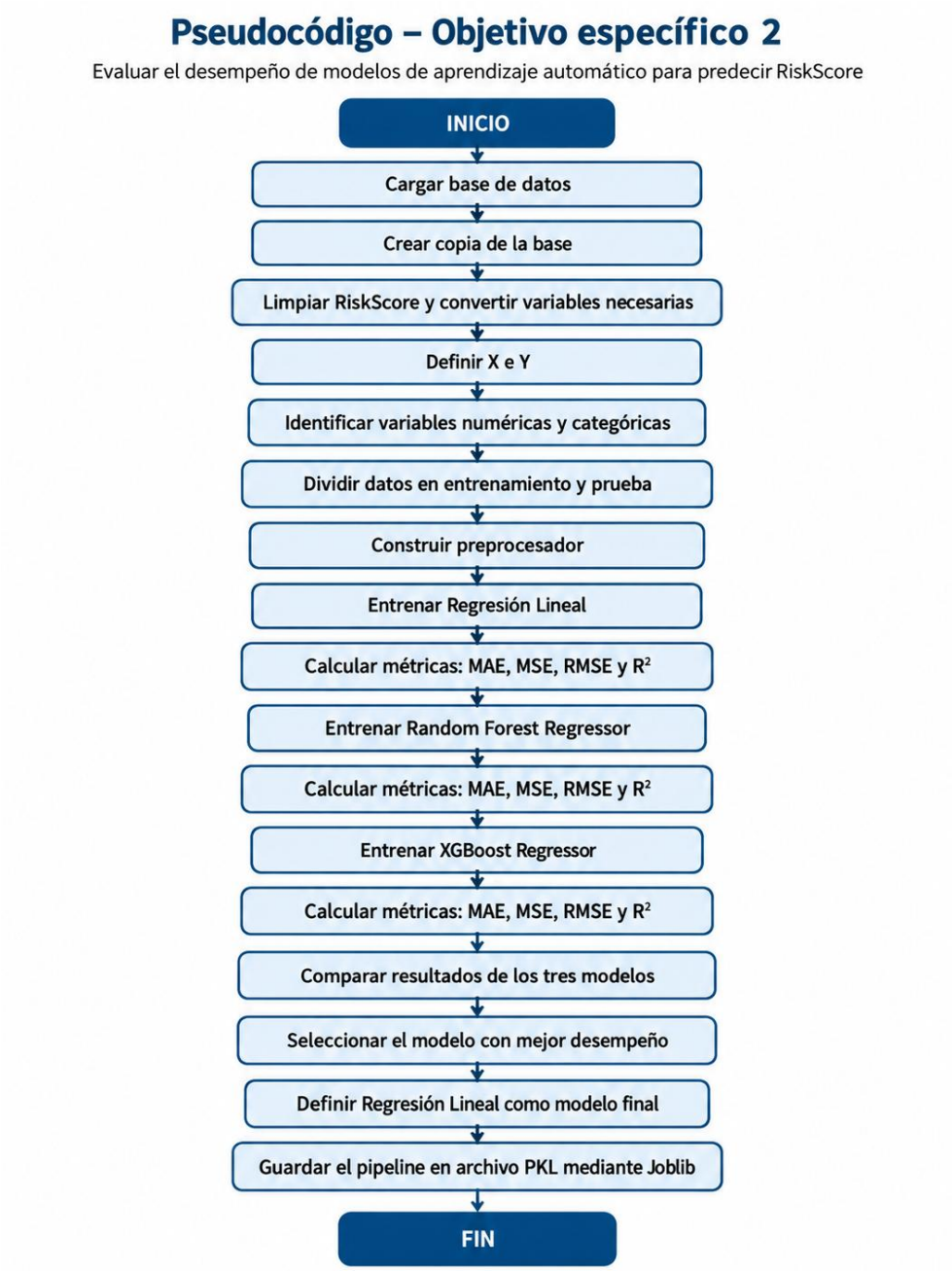


Figura. Pseudocódigo resumido del objetivo específico 2.

Fuente: Elaboración propia

Modelo 1: Regresión Lineal

La Regresión Lineal es un algoritmo supervisado utilizado para estimar una variable continua a partir de una combinación de variables predictoras. En este proyecto la variable objetivo es RiskScore, por lo que el modelo busca aproximar su valor a partir de las características financieras, crediticias y sociodemográficas del solicitante.

La implementación se integró en un Pipeline que incluye el preprocesamiento de variables numéricas y categóricas. De esta forma, las transformaciones aplicadas durante el entrenamiento quedan asociadas al modelo y pueden reutilizarse posteriormente durante la predicción en la aplicación web.

Modelo 2: Random Forest Regressor

Random Forest Regressor es un método de aprendizaje supervisado basado en un conjunto de árboles de decisión. Cada árbol genera una estimación y el modelo combina sus resultados mediante el promedio, reduciendo la sensibilidad asociada a un único árbol de decisión.

En el proyecto se configuró un bosque de 200 árboles con random_state=42. Su desempeño se comparó con la Regresión Lineal mediante las mismas particiones de datos y las mismas métricas de evaluación.

Modelo 3: XGBoost Regressor

XGBoost Regressor es un algoritmo de ensamble basado en Gradient Boosting. Los árboles se construyen de manera secuencial y cada nuevo árbol busca reducir los errores residuales de las predicciones anteriores.

Para el proyecto se configuraron 200 estimadores, una profundidad máxima de 6 y una tasa de aprendizaje de 0,1. El modelo fue evaluado sobre el mismo conjunto de prueba utilizado por los demás algoritmos para mantener una comparación consistente.

Conclusión de los resultados de los modelos

A partir de los resultados obtenidos, se compararon tres modelos de regresión para predecir la variable RiskScore: Regresión Lineal, Random Forest Regressor y XGBoost Regressor. Los resultados de las métricas de evaluación se presentan en la Tabla 8.

Tabla 8. Métricas de desempeño de los modelos de regresión en el conjunto de prueba

Modelo	MAE	MSE	RMSE	R^2
Regresión Lineal	0,1398	0,1497	0,3867	0,9979
Random Forest Regressor	0,8425	1,4869	1,2194	0,9806
XGBoost Regressor	0,3774	0,2951	0,5432	0,9960

Fuente: Elaboración propia a partir de los resultados obtenidos en Python.

El modelo de Regresión Lineal presentó el mejor desempeño general. Obtuvo el menor valor de MAE, con 0,1398, lo que significa que, en promedio, las predicciones se alejaron aproximadamente 0,14 puntos del valor real del RiskScore. También alcanzó el menor MSE, con 0,1497, y el menor RMSE, con 0,3867. Estas métricas indican que el modelo generó errores de predicción reducidos.

Además, la Regresión Lineal obtuvo un coeficiente de determinación R^2 de 0,9979, lo que indica que el modelo explica aproximadamente el 99,79 % de la variabilidad observada en el RiskScore. Este resultado evidencia una relación muy alta entre los valores reales y los valores estimados por el modelo.

El modelo XGBoost Regressor presentó el segundo mejor desempeño, con un MAE de 0,3774, un RMSE de 0,5432 y un R^2 de 0,9960. Aunque sus resultados fueron satisfactorios y su capacidad predictiva fue elevada, sus errores fueron superiores a los obtenidos por la Regresión Lineal.

Por su parte, Random Forest Regressor obtuvo los resultados menos favorables entre los tres modelos evaluados. Aunque alcanzó un R^2 de 0,9806, este valor fue inferior al de la Regresión Lineal y XGBoost. Asimismo, presentó los mayores valores de MAE, MSE y RMSE, lo que indica un mayor error promedio en la estimación del puntaje de riesgo.

En consecuencia, se seleccionó la Regresión Lineal como el modelo con mejor desempeño para la predicción del RiskScore. La elección se fundamenta en que obtuvo los valores más bajos en las métricas de error y el valor más alto de R^2. Por lo tanto, este modelo será utilizado como modelo principal para integrarlo en el prototipo de aplicación web.

Verificación del coeficiente de determinación obtenido

El R^2 de 0,9979 obtenido por la Regresión Lineal es inusualmente alto para un problema de riesgo crediticio, en el que los modelos aplicados a datos reales suelen alcanzar niveles de ajuste considerablemente menores. Por esta razón, antes de interpretar el resultado se definió el siguiente protocolo de verificación:

Revisión de fuga de información (data leakage): se identifican las variables predictoras que intervienen en la construcción del RiskScore o que se derivan de él. En particular, se verifica si LoanApproved hace parte de las 33 variables predictoras; de ser así, debe excluirse, puesto que en la fuente la aprobación del crédito se relaciona directamente con el puntaje de riesgo. Del mismo modo, se reentrena el modelo sin las variables derivadas (TotalDebtToIncomeRatio, MonthlyLoanPayment, InterestRate, BaseInterestRate y NetWorth) para cuantificar su aporte al ajuste.

Validación cruzada: se estiman el promedio y la desviación estándar de MAE, RMSE y R^2 mediante validación cruzada de cinco particiones ($k = 5$), con el fin de comprobar que el resultado no depende de una única partición 80/20.

Modelo de referencia: se compara el desempeño de los modelos con un predictor ingenuo que asigna a todos los registros la media del RiskScore del conjunto de entrenamiento, para dimensionar la mejora atribuible al aprendizaje.

Análisis de residuos por categoría: se revisa la distribución de los errores en cada categoría de riesgo, con énfasis en los registros de riesgo medio y alto, que son los de mayor interés para la gestión de cobranza.

Los resultados de este protocolo son los siguientes:

Figura 8. Tabla consolidada de análisis de verificación del RISKSCORE

	Dimensión del Análisis	Métrica / Segmento	Valor / R^2	Error Promedio (MAE)	Error Máximo o StdDev (RMSE)
0	Desempeño General	Nuestro Modelo (Regresión Lineal)	0.9980	0.1398	0.3869 (RMSE)
1	Desempeño General	Modelo Baseline (Predecir Media)	-0.0001	6.8185	8.6367 (RMSE)
2	Validación Cruzada (5 Folds)	R^2 Promedio (Estabilidad)	0.997898	-	0.000240 (Std Dev R^2)
3	Validación Cruzada (5 Folds)	RMSE Promedio (Estabilidad)	0.393284	-	0.023715 (Std Dev RMSE)
4	Análisis por Categoría	Riesgo Bajo (Score <= 40)	N = 3,363 casos	0.1114	6.7918 (Max Error)
5	Análisis por Categoría	Riesgo Medio (40 < Score <= 70)	N = 637 casos	0.2898	6.7698 (Max Error)

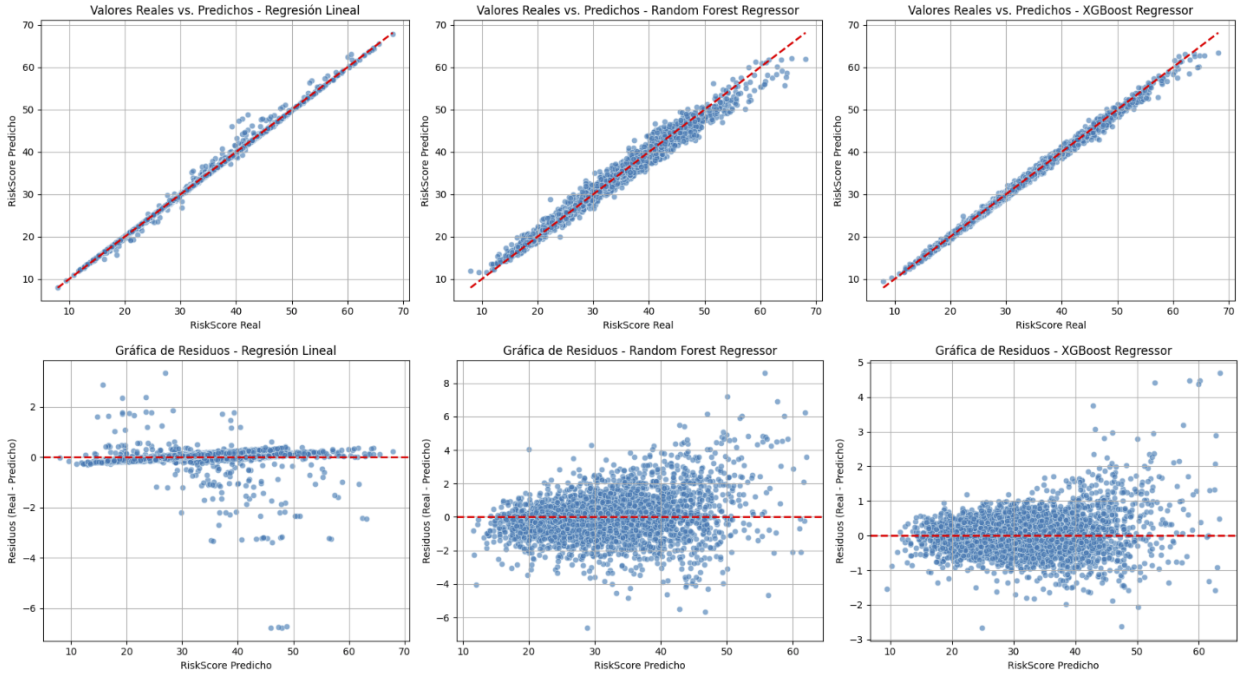
Fuente: Elaboración propia a partir de los resultados obtenidos en Python.

Con el fin de confirmar la validez del resultado obtenido se verifica que LoanApproved forma parte de las 33 variables predictoras que pueden incorporar información de RiskScore, por tal motivo se tiene:

- Variable objetivo (y): el modelo intenta predecir RiskScore.
- Variables predictoras (X): todas las columnas de df_modelo_risk, excepto LoanApproved, RiskScore y ApplicationDate.
- Exclusión de LoanApproved: evita utilizar directamente esa variable para predecir el RiskScore.

En cualquier caso, el alto valor de R^2 confirma que el procedimiento de modelado reproduce de forma consistente la estructura del RiskScore sintético, pero no constituye evidencia del desempeño que tendría el modelo sobre la cartera de una entidad aseguradora.

Figura 9. Valores reales frente a valores predichos por los modelos de regresión

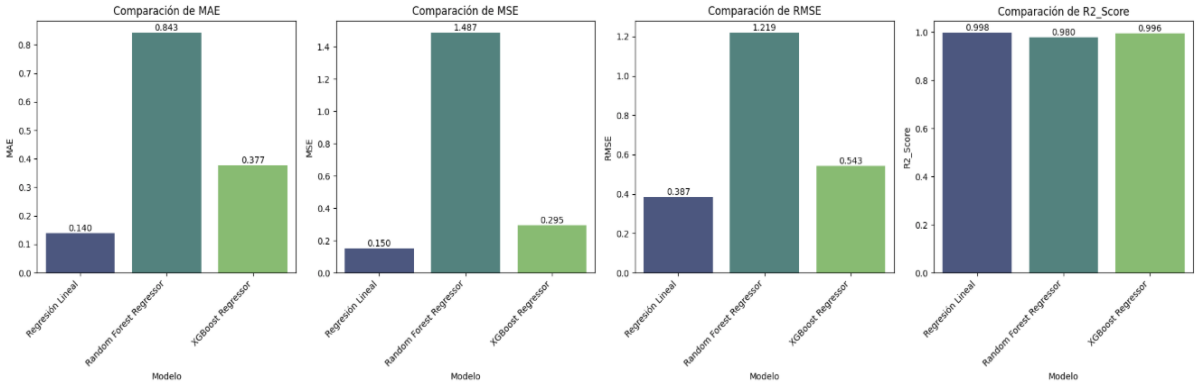


Fuente: Elaboración propia a partir de los resultados obtenidos en Python.

Las gráficas de la Figura 9 muestran que los tres modelos lograron aprender la relación entre las variables predictoras y el RiskScore. Sin embargo, la regresión lineal presentó la mejor correspondencia entre los valores reales y predichos, debido a que sus puntos se ubicaron más cerca de la línea de predicción perfecta y sus residuos se concentraron principalmente alrededor de cero.

XGBoost también presentó un comportamiento favorable, aunque con una dispersión ligeramente mayor. Random Forest mostró los errores más amplios y una menor precisión en los valores altos del puntaje de riesgo.

Figura 10. Comparación de las métricas de evaluación de los modelos de regresión



Fuente: Elaboración propia a partir de los resultados obtenidos en Python.

La Figura 10 presenta la comparación de tres modelos de regresión utilizados para predecir la variable RiskScore.

La comparación gráfica de las métricas de evaluación permitió identificar que la Regresión Lineal presentó el mejor desempeño en la predicción del RiskScore. Este modelo obtuvo un MAE de 0,140, un MSE de 0,150, un RMSE de 0,387 y un coeficiente de determinación (R^2) de 0,998. Estos resultados indican que las predicciones se encuentran muy próximas a los valores reales y que el modelo explica aproximadamente el 99,8 % de la variabilidad del puntaje de riesgo. Aunque XGBoost también presentó resultados favorables, sus métricas fueron ligeramente inferiores. Random Forest mostró el mayor nivel de error y el menor coeficiente de determinación. En consecuencia, se seleccionó la Regresión Lineal como el modelo final para la estimación del riesgo crediticio.

Tabla 9. Distribución de los registros por categoría de riesgo (RiskCategory)

Categoría de riesgo	Cantidad de registros	Proporción (%)
Riesgo bajo	16.885	84,425
Riesgo medio	3.111	15,555
Riesgo alto	4	0,020
Total	20.000	100,000

Fuente: Elaboración propia a partir de los resultados obtenidos en Python.

La Tabla 9 muestra la cantidad y el porcentaje de clientes pertenecientes a cada categoría de riesgo, creada a partir de la variable RiskScore.

La distribución de la variable Risk Category evidenció que la mayor parte de los registros pertenece a la categoría de riesgo bajo, con 16.885 observaciones, equivalentes al 84,425 % del total. La categoría de riesgo medio está conformada por 3.111 registros, correspondientes al 15,555 %, mientras que la categoría de riesgo alto cuenta únicamente con 4 observaciones, equivalentes al 0,020 %. En total, la base está compuesta por 20.000 registros.

Capítulo 3. Despliegue del modelo de Regresión Lineal en una aplicación web

Persistencia e integración del modelo en la aplicación

Después de seleccionar la Regresión Lineal como el modelo con mejor desempeño, se generó un nuevo archivo de persistencia denominado mejor_modelo_riskscore_regresion_lineal.pkl. El archivo almacena el Pipeline completo, incluyendo las transformaciones aplicadas a las variables numéricas y categóricas y el modelo de Regresión Lineal, lo que permite reproducir durante la predicción el mismo procesamiento utilizado en el entrenamiento.

El prototipo web se desarrolló en Python mediante Streamlit, sin utilizar Docker. La aplicación carga el modelo con Joblib y presenta un formulario para registrar las principales características del solicitante. Con el fin de simplificar la interacción, las variables de uso frecuente se muestran directamente en la interfaz y las variables complementarias se agrupan en una sección de parámetros avanzados. Algunas variables derivadas, como MonthlyIncome, NetWorth, BaseInterestRate, InterestRate, MonthlyLoanPayment y TotalDebtToIncomeRatio, se calculan automáticamente antes de realizar la predicción.

Al ejecutar la opción de cálculo, la aplicación construye un DataFrame con las 33 variables predictoras en el mismo orden utilizado durante el entrenamiento y envía el registro al método predict() del Pipeline. El resultado se presenta como un valor estimado de RiskScore y, adicionalmente, puede visualizarse el conjunto de datos utilizado en la predicción para facilitar la verificación del registro ingresado. Los componentes del prototipo se resumen en la Tabla 10.

Tabla 10. Componentes del prototipo de despliegue

Componente	Implementación	Resultado
Framework web	Streamlit	Interfaz web sencilla para el ingreso de datos y la consulta de la predicción.
Modelo desplegado	Regresión Lineal almacenada en un Pipeline	Uso del modelo seleccionado a partir de las métricas de evaluación.
Persistencia	Joblib / archivo .pkl	Carga directa del modelo entrenado dentro de la aplicación.
Preprocesamiento	SimpleImputer, StandardScaler y OneHotEncoder	Transformación consistente de variables numéricas y categóricas.
Entrada	Formulario con variables principales y parámetros avanzados	Captura organizada de la información del solicitante.
Salida	Predicción de RiskScore	Visualización del puntaje estimado de riesgo crediticio.
Ejecución	streamlit run app.py	Despliegue funcional del prototipo en entorno local.

Fuente: Elaboración propia.

La arquitectura de ejecución del prototipo web interactivo desarrollado en Streamlit se estructuró mediante un flujo lógico que captura los datos del usuario, realiza el cálculo automático de variables derivadas y ejecuta el Pipeline previamente exportado. La Figura 10 detalla la secuencia algorítmica implementada en el archivo app.py para el funcionamiento del prototipo web.

Figura 11. Pseudocódigo – Objetivo Específico 3

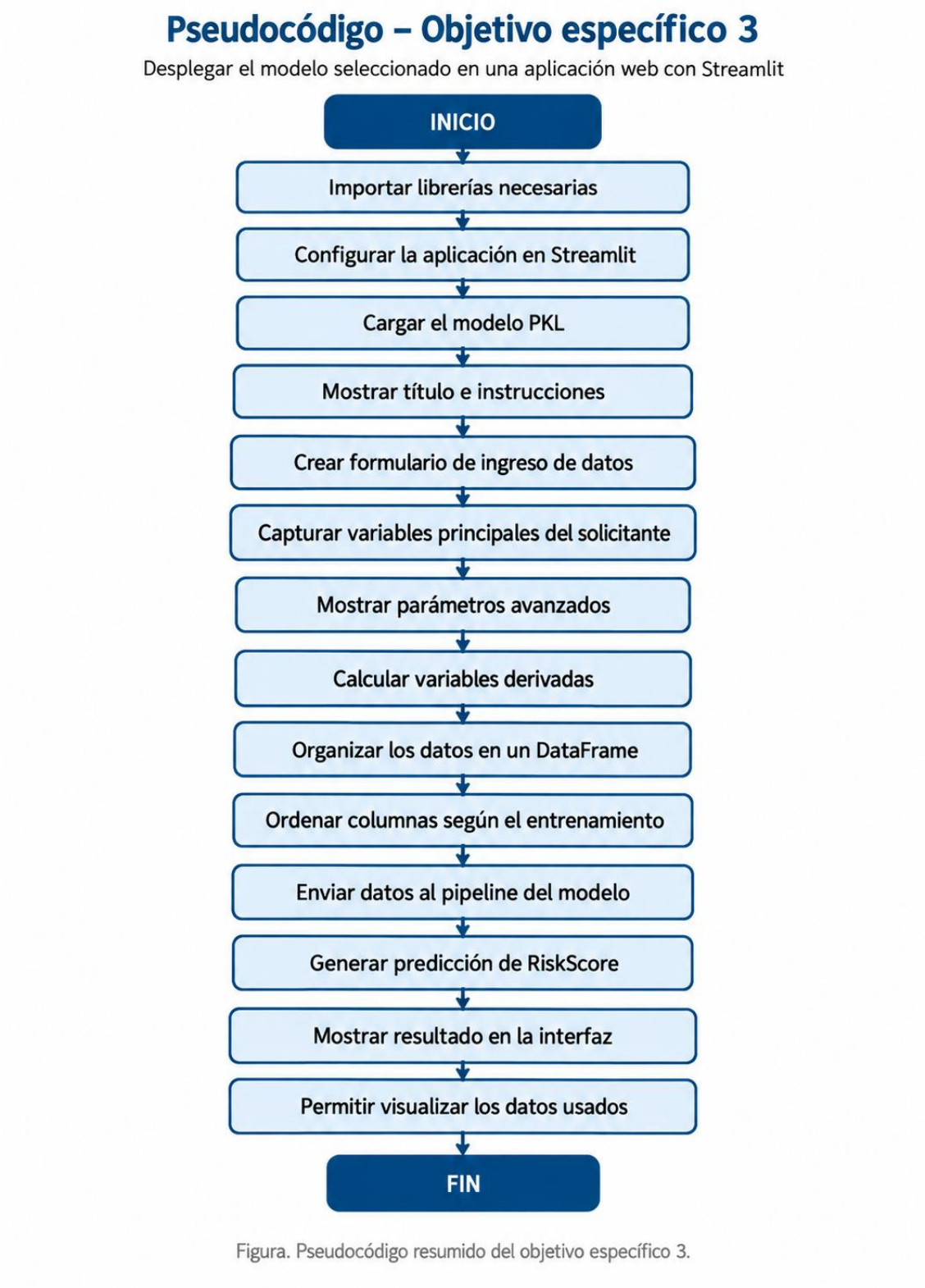


Figura. Pseudocódigo resumido del objetivo específico 3.

Fuente: Elaboración propia

Funcionamiento del prototipo web

Como resultado del tercer objetivo se obtuvo un prototipo web funcional que integra el modelo de Regresión Lineal seleccionado con una interfaz desarrollada en Streamlit. El prototipo permite ingresar la información del solicitante, procesar las variables de entrada mediante el Pipeline entrenado y obtener la estimación del RiskScore. En esta etapa el despliegue se encuentra validado en entorno local; por tanto, una publicación posterior en un servicio web puede considerarse como una actividad adicional si el proyecto requiere acceso externo.

Las Figuras 12 a 14 documentan visualmente el funcionamiento de la aplicación a partir de capturas obtenidas durante la ejecución del prototipo, evidenciando la interfaz y el resultado de la predicción.

Figura 12. Vista general de la aplicación web en Streamlit



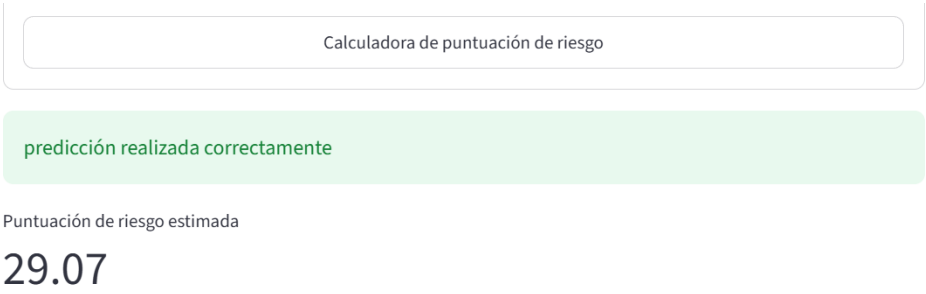
Fuente: Elaboración propia a partir de la ejecución del prototipo.

Figura 13. Formulario de ingreso de información del solicitante



Fuente: Elaboración propia a partir de la ejecución del prototipo.

Figura 14. Resultado de la predicción del RiskScore



Fuente: Elaboración propia a partir de la ejecución del prototipo.

Alineación con la Fase de Despliegue (CRISP-DM) e Impacto Operativo: El desarrollo del prototipo en Streamlit representa la materialización de la fase de despliegue de CRISP-DM; la herramienta fue concebida para apoyar la transición de la gestión de cartera de una entidad aseguradora desde un esquema tradicional de cobranza reactivo y homogéneo a un sistema proactivo, segmentado y guiado por analítica de datos.

Utilidad para la Toma de Decisiones Financieras: Al desplegar el modelo en la interfaz web, los analistas de cartera y tomadores de decisiones pueden interpretar de forma inmediata el nivel de riesgo del solicitante (categorizado en riesgo bajo, medio o alto). Esto puede servir de base para diseñar y priorizar estrategias de recuperación diferenciadas —tales como recordatorios automáticos, planes de pago flexibles o acuerdos personalizados—, con el propósito de mejorar la asignación de recursos operativos y contribuir a la liquidez de la compañía, efectos que deberán verificarse en un entorno real.

Consideraciones de Replicabilidad y Marco Normativo: La implementación local mediante Streamlit reduce la exposición de la información financiera y sociodemográfica a terceros; sin embargo, el cumplimiento del marco legal colombiano en materia de Habeas Data y protección de datos personales requiere, además, la autorización de los titulares, el control de accesos y una política de tratamiento de la información. Asimismo, la arquitectura modular del prototipo en Python facilita que esta solución pueda adaptarse a diversas entidades del sector asegurador, previa validación con sus propios datos, promoviendo la innovación tecnológica e industrial en línea con los objetivos de Desarrollo Sostenible.

Propuesta de uso operativo del prototipo en el proceso de cobranza

Desde la ingeniería industrial, el valor del prototipo no se limita a la precisión de la predicción, sino a su capacidad para convertirse en un insumo de decisión dentro del proceso de gestión de cartera. Para ello se propone integrar el RiskScore en tres decisiones operativas: la priorización de clientes, la asignación de recursos y el diseño del flujo de cobranza.

Priorización de clientes: la cartera vencida se ordena de forma descendente según un índice de prioridad que combina el riesgo estimado y la exposición económica, por ejemplo,

IP = (RiskScore / 100) × saldo en mora. De esta manera, los primeros clientes de la lista son aquellos con mayor riesgo relativo y mayor valor expuesto, lo que reemplaza la gestión por orden de vencimiento o por antigüedad de la deuda.

Asignación de recursos: cada categoría de riesgo se asocia con un canal y una intensidad de gestión diferenciados. Los clientes de riesgo bajo se atienden mediante canales automatizados de bajo costo (mensajes de texto, correo electrónico y recordatorios previos al vencimiento); los de riesgo medio, mediante contacto preventivo de un analista y ofrecimiento de acuerdos de pago; y los de riesgo alto, mediante gestión personalizada, negociación y, de ser necesario, escalamiento a cobro prejurídico. Aplicado a la distribución de la Tabla 9, este esquema concentraría la gestión con analistas en 3.115 de los 20.000 registros (15,575 %), mientras que el 84,425 % restante se atendería por canales automatizados. Este cálculo es ilustrativo y depende de los umbrales de categoría definidos, por lo que debe recalcularse con la cartera real de la aseguradora.

Rediseño del flujo de cobranza: el proceso pasa de una secuencia homogénea — vencimiento, contacto masivo y escalamiento por antigüedad de la deuda— a una secuencia segmentada en la que la carga de datos, el cálculo del RiskScore, la asignación de la categoría y la generación de la lista priorizada preceden a la gestión, y los resultados de cada contacto retroalimentan el modelo.

Indicadores para medir el beneficio operativo: para verificar el efecto del prototipo se proponen indicadores de proceso y de resultado, como el índice de morosidad, la tasa de recuperación de cartera, los días promedio de cartera, el costo de cobranza por peso recuperado, el número de gestiones por analista y el tiempo de ciclo de la gestión. Estos

indicadores deben medirse antes y después de la implementación, o comparando un grupo de control con un grupo gestionado con apoyo del prototipo.

Los beneficios descritos en esta sección corresponden a una propuesta de uso y a efectos esperados; no fueron verificados experimentalmente en este trabajo, puesto que la base utilizada no contiene información sobre acciones, costos ni resultados de cobranza.

7. Discusión

Los resultados muestran que la Regresión Lineal obtuvo el mejor desempeño en la predicción del RiskScore (MAE = 0,1398; MSE = 0,1497; RMSE = 0,3867; R^2 = 0,9979), por encima de XGBoost (R^2 = 0,9960) y de Random Forest (R^2 = 0,9806), como se presenta en la Tabla 8. Este hallazgo contrasta con buena parte de la literatura revisada, en la que los modelos de ensamble suelen superar a los modelos lineales en problemas de riesgo crediticio (Chen & Guestrin, 2016; Lessmann et al., 2015), y con los estudios del sector asegurador que reportan mejores resultados con algoritmos no lineales (Tablas 3 y 4).

La explicación más probable de este resultado es que el RiskScore de la base utilizada es una variable sintética construida a partir de una combinación, en esencia lineal, de las variables financieras y crediticias incluidas en el conjunto de datos, por lo que un modelo lineal puede reproducir su estructura con un error muy bajo. En contraste, los modelos basados en árboles aproximan las relaciones mediante particiones escalonadas y tienden a perder precisión en los extremos de la distribución (Hastie et al., 2017; James et al., 2021), lo cual es coherente con la mayor dispersión que presentó Random Forest en los valores altos del RiskScore (Figura 9).

Asimismo, un coeficiente de determinación cercano a 1 debe interpretarse con cautela. Un ajuste tan alto puede indicar que algunas variables predictoras, en especial las derivadas, como TotalDebtToIncomeRatio o MonthlyLoanPayment, contienen información muy cercana a la utilizada para construir la variable objetivo, lo que podría sobreestimar el desempeño esperado con datos reales (fuga de información o data leakage). Por ello, los resultados deben entenderse como una validación del procedimiento metodológico y del prototipo, más que como una estimación del desempeño que tendría el modelo sobre la cartera de una aseguradora (James et al., 2021).

Es necesario, además, diferenciar dos niveles de resultado. El primero es el desempeño predictivo, medido mediante MAE, MSE, RMSE y R^2 sobre el conjunto de prueba, que sí fue verificado experimentalmente con las limitaciones ya descritas. El segundo corresponde a los beneficios operativos —reducción de la morosidad, aumento de la recuperación de cartera, disminución del costo de cobranza y mejor uso de la capacidad de los analistas—, que no fueron medidos, debido a que la base no contiene información sobre acciones ni resultados de cobranza. Estos beneficios deben entenderse como efectos esperados, cuya verificación requiere una prueba piloto en una entidad aseguradora con un grupo de control gestionado bajo el esquema actual y un grupo gestionado con el apoyo del prototipo.

Desde la perspectiva de la ingeniería industrial, la selección de la Regresión Lineal ofrece ventajas prácticas: su interpretabilidad facilita explicar a los analistas de cartera y a los entes de control el efecto de cada variable sobre el riesgo, su costo computacional es bajo y su integración en un Pipeline permite un despliegue sencillo y reproducible. Esta transparencia resulta relevante frente a los principios de trato justo al consumidor financiero

establecidos en la Ley 1328 de 2009 y frente a los lineamientos de gestión del riesgo de la norma ISO 31000.

Por último, la distribución de las categorías de riesgo (Tabla 9) evidencia un marcado desbalance: el 84,425 % de los registros corresponde a riesgo bajo, el 15,555 % a riesgo medio y solo el 0,020 % a riesgo alto (4 registros). Esta situación limita la capacidad del modelo para caracterizar a los clientes de mayor riesgo, que son precisamente los de mayor interés para la gestión de cobranza, y coincide con la problemática de los datos desbalanceados señalada en el estado del arte (Libreros Medrano, 2024).

8. Recomendaciones

Validar el modelo con datos reales de una entidad aseguradora colombiana, previa a validación de la información y en cumplimiento de la Ley 1266 de 2008 y la Ley 1581 de 2012, incorporando variables propias del negocio asegurador, como el ramo, el valor de la prima, la periodicidad de pago, la antigüedad de la póliza y el historial de recaudo.

Fortalecer la evaluación del modelo mediante validación cruzada (k-fold), análisis de importancia de variables y revisión de posibles fugas de información, retirando o recalculando las variables derivadas que estén directamente relacionadas con la construcción del RiskScore, con el fin de obtener una estimación más robusta del desempeño.

Optimizar los hiperparámetros de Random Forest y XGBoost mediante búsqueda en malla o aleatoria, y evaluar modelos lineales regularizados (Ridge y Lasso), para confirmar si la ventaja de la Regresión Lineal se mantiene bajo condiciones de comparación más exigentes.

Tratar el desbalance de la categoría de riesgo alto mediante técnicas de remuestreo o de ponderación de clases, y complementar la estimación del RiskScore con un modelo de clasificación que estime directamente la probabilidad de incumplimiento en el pago de las primas.

Publicar el prototipo en un servicio en la nube con autenticación de usuarios, cifrado de la información y registros de auditoría, e integrarlo con los sistemas de información de cartera de la aseguradora para automatizar la carga de datos y la generación de alertas.

Definir un esquema de monitoreo y reentrenamiento periódico del modelo, junto con indicadores de gestión como el índice de morosidad, la tasa de recuperación de cartera, los días promedio de cartera y el costo de cobranza por cliente, que permitan medir el impacto operativo del prototipo en el proceso de cobranza, mediante una prueba piloto que compare un grupo de clientes gestionado con el prototipo frente a un grupo de control gestionado con el esquema actual.

9. Conclusiones

Se preparó una base de datos de carácter académico y sintético con 20.000 registros y 33 variables predictoras de tipo financiero, crediticio y sociodemográfico, mediante procesos de revisión, limpieza, imputación de valores faltantes, transformación de variables y estandarización del RiskScore en una escala de 0 a 100. Este proceso permitió obtener una base consistente y sin valores faltantes, y confirmó la importancia de la fase de preparación de los datos dentro de la metodología CRISP-DM.

Se evaluaron tres algoritmos de regresión supervisada bajo una misma partición 80/20 y las mismas métricas de error y ajuste. La Regresión Lineal presentó el mejor desempeño,

con un MAE de 0,1398, un MSE de 0,1497, un RMSE de 0,3867 y un R^2 de 0,9979, seguida de XGBoost y Random Forest, por lo que fue seleccionada como modelo final tanto por su precisión como por su interpretabilidad. No obstante, el R^2 cercano a 1 refleja principalmente que el RiskScore sintético se construyó a partir de las mismas variables predictoras, por lo que este desempeño valida el procedimiento de modelado y no debe extrapolarse a datos reales de cartera.

Se desplegó el modelo seleccionado en un prototipo de aplicación web desarrollado con Streamlit, que carga el Pipeline entrenado y permite ingresar la información del solicitante para obtener en tiempo real la estimación del RiskScore. El prototipo fue validado en entorno local y constituye una base funcional para su posterior integración en los procesos de gestión de cartera.

En conjunto, el trabajo cumple el objetivo general y demuestra la viabilidad técnica de apoyar la transición de la gestión de cartera de un esquema reactivo y homogéneo a uno proactivo, segmentado y basado en datos. Sin embargo, dado que los resultados se obtuvieron con una base académica, su aplicación operativa en una entidad aseguradora requiere una validación previa con datos reales y el cumplimiento del marco normativo colombiano en materia de protección de datos.

Desde la ingeniería industrial, el aporte del proyecto consiste en integrar el análisis de datos con el diseño del proceso de cobranza, mediante una propuesta de priorización de clientes basada en un índice de riesgo y exposición, una asignación diferenciada de canales y analistas por categoría de riesgo y un conjunto de indicadores para medir la eficiencia del proceso. Cabe precisar que el desempeño predictivo del modelo fue verificado mediante

métricas de error y ajuste, mientras que los beneficios operativos —reducción de la morosidad, aumento de la recuperación y disminución del costo de cobranza— constituyen efectos esperados que no fueron medidos en este trabajo y que deberán evaluarse mediante una prueba piloto con datos reales.

Bibliografía

Baesens, B., Rösch, D., & Scheule, H. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS. Wiley.

Breiman, L. (2001). Random forests. Machine Learning.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM

Congreso de la República de Colombia. (2008). Ley 1266 de 2008. Por la cual se dictan las disposiciones generales del hábeas data. Diario Oficial.

Congreso de la República de Colombia. (2009). Ley 1328 de 2009. Por la cual se dictan normas en materia financiera, de seguros, del mercado de valores y otras disposiciones. Diario Oficial.

Congreso de la República de Colombia. (2012). Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE.

Eling, M., Nuessle, D., & Staubli, J. (2022). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. The Geneva Papers on Risk and Insurance – Issues and Practice, 47(2), 205–241.

Federación de Aseguradores Colombianos. (2024). *Informe del sector asegurador colombiano*. FASECOLDA.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grize, Y.-L., Fischer, W., & Lützelschwab, C. (2020). Machine learning applications in nonlife insurance. Applied Stochastic Models in Business and Industry

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hernández-Sampieri, R., & Mendoza, C. P. (2018). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta. McGraw-Hill Interamericana.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research.

OECD. (2020). The impact of big data and artificial intelligence (AI) in the insurance sector. OECD Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Streamlit Inc. (2024). *Streamlit documentation*.

Superintendencia Financiera de Colombia. (1995). Circular Básica Contable y Financiera (Circular Externa 100 de 1995).

Zoppelletto, L. (s. f.). Financial risk for loan approval [Conjunto de datos]. Kaggle.