


# Comparación del Desempeño Predictivo entre SVM y Técnicas Boosting para la Evaluación del Riesgo Crediticio en Cooperativa de Ahorro y Crédito Colombiana

**Trabajo de grado**

**Requisito para optar al grado de ciencia de datos**

**Tomás De Aza Márquez**

**Tutor: Ing. Julio Omar Palacio Niño, M.Sc**

**Pontificia Universidad Javeriana**

**Facultad de Ciencias**

**Bogotá, Colombia**

**2025**

## Agradecimientos

Para todas aquellas personas que han estado en el camino recorrido y han puesto un grano de arena que me permitió llegar a este punto de la carrera y de mi vida.

En primer lugar, infinitos agradecimientos a mi familia, Doris, Stillman y Alejandro; ellos han sido el pilar que me ha sostenido en cada paso, en cada duda y en cada decisión, no tengo más que palabras de agradecimiento, admiración y orgullo. Es ese amor incondicional brindado y tantas enseñanzas tanto personales como profesionales las que hacen posible el día de hoy entregar este trabajo de grado, de nuevo gracias.

En un segundo momento deseo agradecer al Ing. Julio Omar Palacio Niño que desde el primer día en la clase de bases de datos ha sido un ejemplo a seguir y un profesor excepcional que al enseñar con pasión y dedicación hace de cada estudiante que pasa por una silla de su salón un mejor profesional y una mejor persona, asimismo me gustaría dedicar mis palabras de agradecimiento al profesor John Jairo Corredor Franco, una persona excepcional que desde el primer momento nos orientó para ser los mejores profesionales del mañana, y son esas palabras las que recuerdo en estos momentos, gracias por su fe y motivación impuestas en mí.

Además agradecer a mis amigos de toda la carrera y a mi pareja sentimental, Samuel Peña, Mauricio Rodríguez, Sebastián Córdoba, Juan José Rodríguez, Daniel Becerra, Juan David López, Juan Carvajal e Isabella Díaz, varios de ellos son personas con las cuales tuve el honor de graduarme desde el colegio y otros los conocí en este camino universitario, a ti Isabella infinitas gracias por ese apoyo en estas etapas finales del trabajo, sin ti nada de esto sería posible; para mí todos han sido esa fuerza diaria que me ayuda a salir adelante, estas personas que no me dejan perder en la adversidad de una tarea o una situación, con ustedes mis más sinceros agradecimientos.

Por último, pero no menos importante, agradecer a cada persona que no ha sido nombrada acá, desde familiares, profesores, amigos y compañeros, cada persona que ha tenido un efecto en mi vida y que hacen realidad hoy este suceso espero que los logros por venir tanto suyos como míos sean compartidos y gozados en conjunto.

## Resumen

**Título**

Comparación del Desempeño Predictivo entre SVM y Técnicas Boosting para la Evaluación del Riesgo Crediticio en una Cooperativa de Ahorro y Crédito Colombiana.

**Objetivo**.

Comparar el desempeño de XGBoost y SVM-RBF para predecir impago en una cooperativa colombiana, bajo una metodología CRISP-DM con énfasis en validación y explicabilidad. Diseño: estudio aplicado con datos tabulares institucionales; pipeline de preparación, modelado y evaluación; análisis de explicabilidad con SHAP y Permutation Importance. Intervención/modelos: XGBoost y SVM-RBF con búsqueda de hiperparámetros y evaluación fuera de muestra. Mediciones principales: AUC-ROC, AUC-PR, F1, Balanced Accuracy y métricas orientadas al negocio (p. ej., Savings); se exige trazabilidad e interpretabilidad. Resultados clave: los desempeños inicialmente altos se explicaron por la influencia de una variable altamente asociada al evento de impago, identificada mediante explicabilidad (riesgo de fuga); al excluirla y rehacer los modelos, el rendimiento descendió y emergió un recall bajo en la clase minoritaria, elevando el costo por morosos no detectados. En consecuencia, se reentrenaron los modelos y se redefinió el criterio de decisión hacia umbralado orientado a costos con metas explícitas de recall, junto con monitoreo de deriva y revisiones de explicabilidad para estabilidad operativa. Conclusión: XGBoost y SVM-RBF son viables, pero su valor depende de datos sin fuga, umbrales ajustados a costos y la incorporación futura de variables longitudinales y macroeconómicas para sensibilidad al ciclo.


**Palabras clave:** Aprendizaje automático, Riesgo crediticio, XGBoost, SVM-RBF, Explicabilidad,

## Abstract

**Title**

Comparing Predictive Performance of SVM and Boosting Techniques for Credit Risk Assessment in a Colombian Savings and Credit Cooperative.

**Objective.**

To compare XGBoost and SVM-RBF for credit default prediction in a Colombian cooperative using the CRISP-DM framework with strict validation and explainability. Design: applied study on institutional tabular data; end-to-end preprocessing, modeling, and evaluation pipeline; model explainability via SHAP and Permutation Importance. Interventions/Models: XGBoost and SVM-RBF with hyperparameter search and held-out testing. Primary measures: AUC-ROC, AUC-PR, F1, Balanced Accuracy, and cost-oriented business metrics (e.g., Savings); transparency and interpretability are required. Key results: the initially high metrics were traced to a highly target-aligned variable, revealed by explainability (risk of leakage). After removing it and rebuilding models, performance dropped, exposing low recall on the minority (default) class and increasing the cost of undetected defaulters. Consequently, models were retrained and the decision policy shifted to cost-oriented thresholding with explicit recall targets, plus drift monitoring and periodic explainability checks to ensure operational stability. Conclusion: XGBoost and SVM-RBF are feasible baselines, but business value hinges on leakage-free data, cost-aligned thresholds, and future inclusion of longitudinal and macroeconomic predictors to achieve point-in-time, cycle-sensitive risk estimates.


**Keywords:** Machine learning, Credit risk, XGBoost, SVM-RBF, Explainability.

# Tabla de Contenidos

1. Abstract ..... 5

2. Introducción ..... 7

2.1. Planteamiento del Problema ..... 7

2.2. Justificación ..... 9

3. Objetivos ..... 11

3.1. Objetivo General ..... 11

3.2. Objetivos Específicos ..... 11

4. Preguntas de Investigación ..... 11

5. Marco Teórico ..... 12

5.1. Riesgo Crediticio a Nivel Global ..... 12

5.2. Riesgo Crediticio en Colombia ..... 13

5.3. Métodos Tradicionales de Pronóstico del Riesgo Crediticio ..... 14

5.4. Modelos Boosting (XGBoost) ..... 16

5.5. SVM con Kernels No Lineales ..... 18

5.6. Métricas de Evaluación del Desempeño Predictivo ..... 20

6. Metodología ..... 27

6.1. Comprensión del Negocio ..... 28

6.2. Comprensión de los Datos ..... 29

6.3. Preparación de los Datos ..... 38

6.4. Modelado ..... 42

6.5. Resultados ..... 43

7. Conclusiones ..... 53

8. Referencias ..... 56

# Introducción

## Planteamiento del Problema

El acceso al crédito es un factor determinante en la estabilidad económica y el desarrollo financiero de Colombia. Sin embargo, el proceso de aprobación crediticia presenta múltiples desafíos, principalmente debido a la dificultad de evaluar con precisión el riesgo de impago[1] de los solicitantes. Tradicionalmente, las entidades financieras han empleado modelos basados en reglas estáticas y puntajes de crédito que, aunque efectivos en algunos contextos, no logran captar con precisión la evolución del comportamiento financiero de los solicitantes ni la incidencia de las condiciones macroeconómicas en la estabilidad del crédito (Superintendencia Financiera de Colombia [SFC], 2024).

El sistema financiero colombiano se encuentra influenciado por factores económicos como la inflación, las tasas de interés y la variabilidad en el empleo formal e informal, los cuales pueden afectar la capacidad de pago de los solicitantes y, en consecuencia, incrementar la morosidad en los créditos. Según la Superintendencia Financiera de Colombia (2024), la cartera bruta de crédito alcanzó los $341,5 billones en julio de 2023, representando el 50,2% del total de la cartera financiera. Asimismo, indicadores macroeconómicos del Banco de la República (2024) han señalado que la inflación anual en septiembre de 2024 se situó en 5.81%, evidenciando una tendencia de desaceleración respecto a los períodos previos. Estas fluctuaciones económicas impactan directamente la calidad de la cartera crediticia y generan incertidumbre en las entidades financieras respecto a la estabilidad de sus modelos de evaluación crediticia.

[1] **Impago**, incumplimiento por parte de un deudor de las condiciones de pago acordadas en un crédito.

En este contexto, el presente estudio se propone explorar y aplicar tanto métodos estadísticos avanzados y modelos de Machine Learning para poder predecir ese impago de las personas a través de variables financieras, socioeconómicas y demográficas.

Uno de los principales problemas que enfrentan los modelos tradicionales es la rigidez en la asignación de pesos a ciertas variables, como el historial crediticio o el nivel de ingresos, sin considerar que su relevancia puede cambiar según las condiciones económicas. Al permitir la actualización periódica de estos pesos con base en evidencia empírica, se logrará un sistema más equitativo y preciso para la aprobación crediticia.

Además, este estudio cobra relevancia en el ámbito de la inclusión financiera. En Colombia, una parte significativa de la población no cuenta con un historial crediticio sólido o tiene ingresos informales, lo que limita su acceso al financiamiento. Mediante los modelos a desarrollar se logrará tener en cuenta otros factores que no sean el Score crediticio[2] únicamente, esto es importante debido a que en la base de datos de la empresa en la que se constatan todas las solicitudes de crédito de un periodo en específico solamente el 25% de las personas tienen un Score de más de 800 lo que es considerado por la Cooperativa de Ahorro y Crédito trabajada como un valor bueno.

A pesar del papel fundamental que desempeña el acceso al crédito en el desarrollo económico y la inclusión financiera en Colombia, las metodologías tradicionales empleadas por las entidades financieras para evaluar el riesgo crediticio —basadas en reglas estáticas y puntajes fijos como el score crediticio— presentan limitaciones importantes.

Estas herramientas, al no considerar adecuadamente las condiciones macroeconómicas cambiantes ni el comportamiento evolutivo de los solicitantes, resultan poco eficaces para anticipar

[2] **Score Crediticio**, es un puntaje que es otorgado por una entidad que recopila factores de una persona para calificar su historial de pagos.

el impago, especialmente en poblaciones con ingresos informales o historial crediticio insuficiente. Esta rigidez metodológica no solo compromete la capacidad predictiva de los modelos actuales, sino que perpetúa barreras de exclusión para sectores vulnerables de la población. En este contexto, se requiere la implementación de enfoques alternativos que, mediante técnicas avanzadas de análisis de datos y modelos de aprendizaje automático, permitan construir sistemas de evaluación crediticia más dinámicos, sensibles a la variación de las variables relevantes y capaces de mejorar simultáneamente la precisión predictiva, la estabilidad financiera de las entidades y la equidad en el acceso al crédito.

**Justificación**

El presente estudio se fundamenta en la necesidad de optimizar los procesos de aprobación crediticia en Colombia, asegurando que las decisiones tomadas por las entidades financieras sean más precisas, flexibles y adaptables a las condiciones económicas cambiantes. La relevancia de este trabajo radica en su impacto directo en la estabilidad del sistema financiero, la reducción de la morosidad y la promoción de la inclusión financiera (Banco de la República, 2024).

El sector crediticio en Colombia enfrenta desafíos estructurales relacionados con la asignación eficiente del crédito. La falta de modelos que permitan ajustar dinámicamente la ponderación de las variables más influyentes en la toma de decisiones crediticias genera ineficiencias en la aprobación de préstamos. Esto puede llevar tanto a la concesión de créditos a individuos con alto riesgo de incumplimiento como a la exclusión de personas que, bajo un análisis más flexible, podrían ser sujetos de crédito viables.

De acuerdo con la Superintendencia Financiera de Colombia (2024), la cartera vencida representa actualmente el 10,8% del total, con una morosidad particularmente alta en los créditos

de consumo (12,3%). Esta situación evidencia que persisten desafíos significativos en la predicción del impago, lo que refuerza la pertinencia de este estudio para desarrollar modelos más adaptativos y robustos ante los cambios económicos.

Desde una perspectiva metodológica, este estudio es innovador porque busca usar modelos LDA y Regresiones Logísticas usados en el mundo, usar modelos de Machine Learning que permitan entender los datos de diferentes formas tanto separando los grupos de manera lineal como buscar planos no lineales que hagan esa separación con SVM. Esto se espera que permita generar modelos de decisión más robustos y transparentes. Al identificar las variables más determinantes en la aprobación del crédito se logrará una toma de decisiones más fundamentada en datos reales y tendencias económicas actuales.

En Colombia, más del 60% de las entidades financieras han comenzado a adoptar modelos de machine learning para evaluación de riesgo crediticio (MinTIC, 2024). No obstante, la falta de explicabilidad de algunos modelos avanzados ha generado barreras regulatorias, lo que subraya la necesidad de metodologías transparentes y justificables como las que este estudio propone.

En términos sociales, el impacto de este estudio es significativo. Un acceso al crédito más equitativo y eficiente contribuye al desarrollo económico, permitiendo que un mayor número de personas acceda a financiamiento para emprendimientos, educación o consumo. Además, un mejor análisis del riesgo crediticio beneficia a las entidades financieras, reduciendo la morosidad y mejorando la sostenibilidad del sistema financiero (Banca de las Oportunidades, 2024).

# Objetivos

## Objetivo General

Evaluar el desempeño predictivo de distintos modelos de aprendizaje de máquina para la predicción del impago en una cooperativa de ahorro y crédito bogotana.

## Objetivos Específicos

Destacar los aspectos financieros, socioeconómicos, y demográficos de clientes de una cooperativa de ahorro y crédito bogotana.

Aplicar modelos de machine learning como lo son SVM no Lineal y XGBoost para la predicción del impago.

Comparar el desempeño predictivo de los modelos mediante métricas como el AUC-ROC, F1-Score, Savings e Índice de Gini.

# Preguntas de Investigación

¿Cuáles son las características financieras, socioeconómicas y demográficas más relevantes entre los clientes que presentan impago y aquellos que no lo hacen en una cooperativa de ahorro y crédito en Bogotá?

¿Qué nivel de precisión y capacidad predictiva presentan los modelos SVM no lineal y XGBoost al ser aplicados en la predicción del impago crediticio?

¿Existen diferencias significativas en el desempeño de los modelos seleccionados cuando se evalúan mediante métricas como AUC-ROC, F1-Score, Savings e Índice de Gini?

¿Qué variables tienen mayor influencia en la predicción del impago según los modelos aplicados, y cómo puede interpretarse esta importancia en términos de decisiones crediticias?

¿Qué ventajas y limitaciones presentan los modelos de machine learning propuestos frente a los métodos estadísticos tradicionales en el contexto de riesgo crediticio?

## Marco Teórico

### Riesgo Crediticio a Nivel Global

El riesgo crediticio se refiere a la posibilidad de pérdida que enfrenta una institución financiera ante el incumplimiento de un deudor. Su gestión deficiente puede comprometer la estabilidad de bancos y economías completas, como quedó evidenciado en la crisis financiera de 2008-2009 (Bhatt et al., 2023). Por ello, su predicción ha sido objeto de constante evolución metodológica.

Tradicionalmente, este riesgo se evaluaba mediante herramientas estadísticas como la regresión logística o el análisis discriminante, complementadas con la experiencia de analistas humanos (Shi et al., 2022). No obstante, en la última década, el auge del aprendizaje automático ha impulsado un cambio significativo: una revisión sistemática de 76 estudios reveló que los modelos de aprendizaje profundo y los métodos en ensamble superan con frecuencia a los métodos clásicos en precisión predictiva (Shi et al., 2022).

Esta transformación ha sido posible gracias a la disponibilidad de grandes bases de datos, mayor capacidad computacional y una presión regulatoria creciente por mejorar las prácticas de gestión del riesgo (Barboza et al., 2017). En consecuencia, algoritmos como árboles de decisión y máquinas de soporte vectorial (SVM). Roy y Vasa (2025) destacan que esta tendencia es mundial, con aportes metodológicos desde distintas regiones, lo que ha enriquecido el ecosistema de modelos y favorecido su comparación bajo múltiples métricas.

Así, el análisis y pronóstico del riesgo crediticio ya no se limita a modelos lineales estáticos, sino que explora herramientas más complejas y flexibles, capaces de adaptarse a entornos económicos dinámicos y estructuras de datos heterogéneas.

**Riesgo Crediticio en Colombia**

En Colombia, la evaluación del riesgo crediticio ha seguido enfoques similares a los globales, empleando por décadas modelos estadísticos como la regresión logística y los scorecards3 tradicionales, dentro del marco regulatorio local (Salazar Vergara, 2022). Sin embargo, en los últimos años han surgido iniciativas académicas que exploran el uso de técnicas de aprendizaje automático como alternativa o complemento a estos métodos.

Un caso destacado es el de Borrero-Tigreros y Bedoya-Leiva (2020), quienes aplicaron algoritmos como redes neuronales, árboles de decisión y máquinas de soporte vectorial a datos reales de una entidad financiera colombiana. Sus resultados muestran que el árbol de decisión obtuvo el mejor desempeño con un AUC de 0.8829, superando ligeramente a las redes neuronales y las SVM. Este hallazgo evidencia que incluso modelos relativamente simples de ML pueden superar a enfoques tradicionales cuando se configuran adecuadamente.

Por su parte, Salazar Vergara (2022) diseñó un scorecard logístico para predecir incumplimientos en fondos de empleados, logrando segmentar con éxito a solicitantes de alto riesgo pese al tamaño reducido de la muestra (n = 384). Este estudio confirma que los métodos clásicos conservan relevancia, especialmente en contextos donde la interpretabilidad y la disponibilidad limitada de datos son claves.

[3] **Scorecard,** herramientas de evaluación de riesgo que, a través de un sistema de puntuación predice la probabilidad con la cual un prestatario puede incumplir sus obligaciones.

Ambos casos revelan un punto central: en el contexto colombiano coexisten enfoques tradicionales y modernos, y comparar sus desempeños es necesario para fundamentar la adopción de nuevas metodologías. Además, muestran que los métodos de aprendizaje automático pueden ser aplicables localmente sin requerir bases masivas, siempre que se ajusten adecuadamente a las características del problema y los datos disponibles.

**Métodos Tradicionales de Pronóstico del Riesgo Crediticio**

Uno de los enfoques más antiguos en la clasificación de riesgo crediticio es el Análisis Discriminante Lineal (LDA), introducido por Fisher (1936). Este modelo proyecta las observaciones a un espacio en el que las clases se separan linealmente, bajo los supuestos de normalidad multivariada y covarianza homogénea entre clases. La función discriminante se expresa como:

$$\delta(x) = x^{\top}\Sigma^{(-1)}(\mu_1 - \mu_0) + ln\left(\frac{\pi_1}{\pi_0}\right) \tag{1}$$

Donde $\mu_k$ son las medias por clase, $\Sigma$ la matriz de covarianza común y $\pi_k$ las probabilidades a priori. La decisión es lineal y óptima bajo estos supuestos (McLachlan, 1992).

Históricamente, LDA fue clave en el desarrollo de herramientas como el Z-score de Altman (1968) para predecir quiebras corporativas y fue ampliamente utilizado en scorecards bancarios durante las décadas de 1970–1980. Aunque su uso ha disminuido con la llegada de modelos más flexibles, estudios recientes aún reportan desempeños competitivos: Reis et al. (2023) encontraron un AUC = 0.863 con sensibilidad del 88.9% en una cartera real.

A pesar de sus limitaciones —como la sensibilidad a violaciones de normalidad o covarianza desigual— LDA sigue siendo un referente histórico por su interpretabilidad y eficiencia

computacional. Su importancia metodológica radica en haber sentado las bases del análisis discriminante y la clasificación supervisada.

La regresión logística binaria es un modelo probabilístico ampliamente utilizado para la clasificación de eventos dicotómicos, como el incumplimiento crediticio. Su objetivo es modelar la probabilidad condicional de la clase positiva (default) como una función logística de los predictores:

$$logit\big(P(Y=1|X=x)\big) = log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

La probabilidad de incumplimiento se recupera mediante la función inversa del logit:

$$P(\,Y=1 \mid X=x\,) = \frac{1}{1+exp\left[-\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right]} \tag{3}$$

Los parámetros $\beta_j$ se estiman mediante máxima verosimilitud, optimizando la función:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P(y_i|x_i;\beta) = \prod_{i=1}^{n}\left(\frac{1}{1+e^{-x_i^T\beta}}\right)^{y_i}\left(1-\frac{1}{1+e^{-x_i^T\beta}}\right)^{1-y_i} \tag{4}$$

Este modelo no requiere supuestos de normalidad, homocedasticidad ni independencia entre predictores. Sin embargo, asume: Linealidad del logit en los predictores, independencia condicional de las observaciones dado $X$, tamaño muestral suficiente para la validez de inferencias asintóticas y ausencia de multicolinealidad severa para la estabilidad de $\hat{\beta}$.

Desde los años 1980, la regresión logística se consolidó como el método estándar para credit scoring en entidades financieras, en parte gracias al modelo O-score propuesto por Ohlson (1980), que mostró alta precisión en la predicción de quiebras empresariales. Su popularidad

también se debe a su interpretabilidad: cada $\beta_j$ puede interpretarse como el log-odds incremental asociado al predictor $x_j$ , facilitando su aplicación en entornos regulados.

En contextos colombianos, su vigencia se refleja en trabajos como el de Salazar Vergara (2022), quien diseñó un scorecard logístico para otorgamiento de créditos, y en el estudio de Borrero-Tigreros y Bedoya-Leiva (2020), donde la regresión logística mostró desempeño competitivo frente a modelos de aprendizaje automático.

A pesar de su robustez y trazabilidad, la forma funcional lineal del logit puede limitar su capacidad de capturar relaciones no lineales o interacciones complejas. Esto ha motivado el uso de modelos más flexibles —como árboles de decisión o ensambles— que, aunque menos transparentes, pueden lograr mejor desempeño predictivo en datos complejos.

**Modelos Boosting (XGBoost)**

Los modelos de boosting se basan en la idea de construir un modelo predictivo de forma secuencial con base en modelos simples, generalmente árboles de decisión, donde cada árbol corrige los errores cometidos por los anteriores. Este enfoque permite alcanzar un alto desempeño, ya que combina múltiples predictores débiles para formar un modelo fuerte capaz de reducir el sesgo y la varianza simultáneamente.

XGBoost (Chen & Guestrin, 2016) es una implementación eficiente del algoritmo de Gradient Boosting de árboles de decisión, ampliamente adoptado en credit scoring por su alto desempeño. El modelo consiste en construir un ensamble de árboles aditivos, donde cada nuevo árbol corrige los errores residuales del conjunto acumulado, optimizando una función de pérdida diferenciable.

Para un conjunto de datos $(x_i, y_i)_{i=1}^{N}$, el objetivo general que XGBoost optimiza en la iteración t es:

$$\mathcal{L}^{(t)} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{l}\left(\mathrm{y_i}, \widehat{\mathrm{y}_\mathrm{i}}^{(\mathrm{t-1})} + \mathrm{f_t}(\mathrm{x_i})\right) + \Omega(\mathrm{f_t})$$

(5)

Donde $\iota$ es la función de pérdida (usualmente log-loss) y $\Omega(f_t)$ es el término de regularización:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

(6)

Con $T$ número de hojas y $\omega_j$ los valores de predicción de cada hoja. Mediante una expansión de segundo orden de Taylor, la pérdida se aproxima por:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N} \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

(7)

Donde:

$$g_i = \frac{\partial \iota(y_i, \widehat{y_\iota}^{(t-1)})}{\partial \widehat{y_\iota}^{(t-1)}}$$

(8)

$$h_i = \frac{\partial^2 \iota\left(y_i, \widehat{y_\iota}^{(t-1)}\right)}{\partial \left(\widehat{y_\iota}^{(t-1)}\right)^2}$$

(9)

Estas derivadas guían la construcción del árbol $f_t$ que minimiza la pérdida regularizada.

XGBoost incluye además técnicas como shrinkage[4], submuestreo de observaciones y columnas, y paralelización del crecimiento de árboles, lo cual mejora su eficiencia y capacidad de generalización.

El modelo es no paramétrico y no impone supuestos sobre la distribución de los datos, pero requiere un adecuado ajuste de hiperparámetros para evitar sobreajuste. Típicamente se optimizan la profundidad máxima de los árboles, la tasa de aprendizaje, y el número de iteraciones usando búsqueda en malla (grid search).

Numerosos estudios respaldan su uso en riesgo crediticio. Por ejemplo, El-Qadi et al. (2022) implementaron XGBoost con ajuste de hiperparámetros para series de tiempo crediticias, y Borrero-Tigreros y Bedoya-Leiva (2020) mostraron que incluso un solo árbol de decisión ya superaba a redes neuronales y SVM en un caso colombiano (AUC ≈ 0.8829), lo que sugiere que un ensamble como XGBoost podría lograr aún mejores resultados. Además, Chen et al. (2025) reportan que XGBoost mejora el F1, G-Mean y AUC frente a modelos tradicionales, incluso en contextos con clases desbalanceadas.

Por último, XGBoost puede ser interpretado mediante herramientas como SHAP (Lundberg & Lee, 2017), que permiten descomponer la predicción en contribuciones aditivas por variable, lo que facilita su aplicación en entornos financieros regulados.

**SVM con Kernels No Lineales**

Las Máquinas de Vectores de Soporte (SVM) son algoritmos discriminativos que buscan un hiperplano que maximice el margen entre dos clases. Su formulación básica con separación no estricta se plantea como un problema de optimización cuadrática convexa:

[4] **Shrinkage,** reducción de la tasa de aprendizaje.

$$\min_{w,b,\xi_i} \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{N}\xi_i$$

(10)

Sujeto a:

$$y_i(w^t x_i + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0,\ \ \forall i = 1, \ldots, N$$

(11)

Donde $W$ es el vector normal al hiperplano, $b$ es el sesgo, $\xi_i$ son variables de holgura y $C > 0$ controla el compromiso entre complejidad y error.

Para problemas no linealmente separables, las SVM utilizan una función kernel $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ que calcula productos internos en un espacio de características de dimensión superior, evitando computar explícitamente la transformación $\phi(x)$.

El kernel Gaussiano o de base radial (RBF), ampliamente utilizado en riesgo crediticio, se define como:

$$K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right), \gamma > 0$$

(12)

Este kernel permite aproximar fronteras de decisión altamente no lineales, adaptándose a relaciones complejas entre variables demográficas, financieras o conductuales.

Las SVM son modelos no paramétricos que no suponen distribuciones específicas para las clases. Su rendimiento depende fuertemente de una correcta selección de hiperparámetros: $C$ es la tolerancia al error, $\gamma$ es el ancho del kernel RBF, y la normalización previa de las variables, ya que la distancia euclidiana es sensible a la escala.

Desde la teoría estadística del aprendizaje, el principio de maximización del margen se asocia a menores cotas de error de generalización (Vapnik, 1998). Además, la penalización implícita $||w||^2$ actúa como regularización L2, útil en contextos de alta dimensionalidad ($p \gg N$).

En aplicaciones reales de credit scoring, las SVM con kernel RBF han demostrado desempeños sobresalientes. Por ejemplo, Amzile et al. (2022) reportaron un AUC = 0.98 frente a 0.717 de la regresión logística usando datos reales de cartera bancaria. Además, estudios como Letkovsky et al. (2023) muestran que, en escenarios bien calibrados, SVM puede rivalizar con ensambles más complejos.

Su principal limitación es la escasa interpretabilidad, pero su capacidad de modelar relaciones no lineales las convierte en una herramienta útil cuando se dispone de conjuntos de datos moderados y se busca precisión predictiva sin recurrir a arquitecturas profundas.

**Métricas de Evaluación del Desempeño Predictivo**

La evaluación rigurosa del desempeño de modelos en riesgo crediticio requiere métricas que vayan más allá de la precisión global, especialmente por el desbalance entre clases. Modelos que aciertan en la mayoría de los casos pueden no detectar casi ningún incumplimiento real. Por eso, se utilizan métricas que capturan desempeño en ambas clases, discriminación entre perfiles y utilidad práctica:

$$\textbf{Precision} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

Donde TP (verdaderos positivos) es la cantidad de incumplimientos predichos correctamente, TN (verdaderos negativos) los no incumplimientos predichos correctamente, FP los falsos positivos (predicciones de default que no ocurrieron) y FN los falsos negativos (casos

de default no predichos). La precisión es una métrica intuitiva que resume el rendimiento general del modelo. Interpretación: indica la proporción de decisiones correctas; p.ej., una precisión de 0.95 significa que el 95% de los clientes fueron clasificados correctamente como buenos o malos.

Métrica global. Útil si las clases están balanceadas, pero puede ser engañosa en datasets con pocos incumplidores (Zhao et al., 2024).

La sensibilidad, también llamada recall o tasa de verdaderos positivos (TPR), mide la proporción de positivos reales (incumplimientos) que el modelo logra capturar. Su fórmula es:

$$\textbf{Sensibilidad} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

Es la fracción de deudores morosos correctamente identificados por el modelo. Una sensibilidad de, digamos, 0.90 (90%) indica que el modelo logra atrapar el 90% de todos los defaults ocurridos, aunque se le puedan escapar el 10%. En gestión de riesgo, la sensibilidad responde a la pregunta: "¿Qué tan bien detectamos a los clientes problemáticos?".

Es crucial en escenarios donde pasar por alto un default (FN) es muy costoso. Un alto recall asegura que pocos malos pagadores "se cuelen" sin ser detectados. Desventajas: Puede venir al costo de muchos falsos positivos (baja precisión), ya que una forma de aumentar sensibilidad es marcar más clientes como arriesgados. Por ello, un modelo con sensibilidad del 100% podría lograrlo clasificando todos los clientes como morosos –lo cual sería impráctico por la cantidad de buenos clientes rechazados.

La especificidad, o tasa de verdaderos negativos (TNR), es la proporción de negativos reales (clientes solventes) que el modelo clasifica correctamente como tales. Se calcula como:

$$\boldsymbol{Especificidad} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

(15)

Representa la fracción de buenos pagadores que el modelo reconoce como buenos. Por ejemplo, una especificidad de 0.80 significa que el 80% de los clientes cumplidores fueron aprobados (no señalados erróneamente como de alto riesgo), mientras que el 20% restante resultaron falsos positivos (rechazados sin necesidad). En un entorno crediticio, la especificidad responde a: "¿Qué tan bien evitamos rechazar clientes que sí pagarán?".

En datos muy desbalanceados con pocos impagos, la especificidad tiende a ser naturalmente alta incluso con modelos mediocres –pues la mayoría son TN. De hecho, en un conjunto donde 97% pagan, un modelo podría tener especificidad ~0.97 casi por default si tiende a clasificar muchos como bueno.

El F1-Score es la métrica que representa el compromiso entre precisión y recall, siendo la media armónica de ambos. Se define como:

$$\boldsymbol{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Sensibilidad}}{\text{Precision} + \text{Sensibilidad}}$$

(16)

La media armónica penaliza valores muy dispares, así que F1 solo será alta si tanto la precisión como la sensibilidad lo son. Un F1 cercano a 1 indica que el modelo logra simultáneamente una alta sensibilidad y precisión en la clase positiva. Por ejemplo, si recall = 0.8 y precision = 0.8, entonces F1 = 0.80. Pero si recall = 0.95 y precision = 0.50, F1 ≈ 0.66 (bastante moderado), reflejando el desequilibrio entre ambas. F1 es especialmente útil en escenarios desbalanceados donde nos importa el desempeño sobre la clase positiva, pero queremos un balance entre detectarla y no abusar de falsos positivos.

La G-Mean es otra métrica agregada útil en contextos desbalanceados, definida como la raíz cuadrada del producto de la sensibilidad y la especificidad:

$$\boldsymbol{G - Mean} = \sqrt{\text{Sensibilidad} \times \text{Especificidad}} \tag{17}$$

La G-Mean refleja el equilibrio del modelo en ambas clases. Será alta solo si tanto la sensibilidad como la especificidad son altas – si una de ellas es baja, la G-Mean se reduce notablemente. Por eso, es vista como un indicador de desempeño balanceado: penaliza que el modelo sacrifique en exceso una clase por la otra.

El AUC-ROC (Área bajo la Curva ROC) es una métrica de desempeño global muy extendida en credit scoring. La curva ROC (Receiver Operating Characteristic) traza la relación entre la Tasa de Verdaderos Positivos (TPR = Sensibilidad) en el eje $y$ y la Tasa de Falsos Positivos (FPR = 1 – Especificidad) en el eje x a medida que varía el umbral de clasificación del modelo. El AUC es literalmente el área bajo esa curva, resumida en un número entre 0 y 1.

El AUC puede interpretarse como la probabilidad de que, tomando al azar un cliente incumplidor y uno cumplidor, el modelo asigne mayor puntaje de riesgo al incumplidor que al cumplido. Así, $\text{AUC} = 0.5$ indica un modelo sin poder discriminante (equivalente a clasificar al azar), mientras $\text{AUC} = 1.0$ sería un modelo perfecto que siempre ordena antes a los malos que a los buenos. En la práctica, los modelos de crédito competentes suelen lograr AUC en torno a 0.7–0.8 o más. Por ejemplo, un AUC = 0.86, como el reportado para un modelo LDA por Reis et al. (2023), denota alta capacidad discriminatoria.

El estadístico KS es otra medida clásica del poder predictivo en credit scoring, particularmente en ámbitos bancarios. Se basa en la máxima diferencia entre las distribuciones acumuladas de puntajes para las dos clases (buenos vs malos). Para calcularlo, se ordenan todos los clientes por su score de riesgo (de mayor a menor riesgo, por ejemplo) y se construyen dos

funciones de distribución acumulada: $F_{malo}(s)$ = % de malos con score <= s y $F_{bueno}(s)$ = % de buenos con score <= s. El estadístico KS es entonces:

$$\boldsymbol{KS} = \max_{s}; |F_{\text{malo}}(s) - F_{\text{bueno}}(s)|, \tag{18}$$

KS indica cuánto separa el modelo las poblaciones de buenos y malos. Un KS = 0 significaría que las distribuciones acumuladas son iguales (el modelo no distingue nada: por cada porcentaje de población con score bajo, vienen mezclados buenos y malos en igual proporción). Un KS = 1 (o 100%) sería separación perfecta.

Más recientemente, algunos autores han propuesto métricas orientadas a la utilidad económica del modelo. Peykani et al. (2025) introdujeron una métrica llamada "Saving" (Ahorro) que estima de forma más realista los costos de clasificación errónea en riesgo crediticio. La idea es cuantificar el beneficio financiero neto que aporta el modelo en comparación con decisiones aleatorias o con una estrategia base.

En esencia, asigna un costo monetario a cada FN (un default no predicho, que implica pérdida de capital, intereses no cobrados, gastos de cobranza, etc.) y un costo a cada FP (un cliente rechazado indebidamente, que implica pérdida del margen de interés que hubiera dado). Con esas asignaciones, se puede calcular el ahorro total que el modelo proporciona respecto a un clasificador sin habilidad.

Por ejemplo, imaginemos que sin modelo se aprueba a todos: se incurre en pérdidas por todos los malos (FN totales) y se obtiene ganancias por todos los buenos. Con el modelo, se rechazan algunos malos (evitando sus pérdidas) pero también se pierden ganancias de algunos buenos rechazados. El "Saving" en porcentaje podría definirse como la proporción de la pérdida

neta evitada gracias al modelo. Si $C_{\mathrm{FN}}$ es el costo de un falso negativo (un default no anticipado) y $C_{\mathrm{FP}}$ el costo de un falso positivo (negocio perdido), entonces una formulación posible es:

$$\textbf{Saving} = \frac{\mathrm{FN}_{\mathrm{aleatorio}} \cdot C_{\mathrm{FN}} + 0 \cdot C_{\mathrm{FP}} - [\mathrm{FN}_{\mathrm{modelo}} \cdot C_{\mathrm{FN}} + \mathrm{FP}_{\mathrm{modelo}} \cdot C_{\mathrm{FP}}]}{\mathrm{FN}_{\mathrm{aleatorio}} \cdot C_{\mathrm{FN}}}, \tag{19}$$

En términos menos formales, la métrica de ahorro contesta: "¿Cuánto dinero (en pérdidas evitadas menos ingresos perdidos) nos ahorra el modelo comparado con no usar modelo?". Peykani et al. (2025) reportan que esta métrica puede ser más informativa para instituciones financieras que las métricas puramente estadísticas, pues incorpora el perfil de costos reales.

Para detallar de mejor manera el uso de estás métricas véase la tabla 1.

**Tabla 1**

*Métricas de evaluación comparativa según tipo de modelo predictivo.*

| Métrica | Aplicable a | ¿Por qué? |
|---|---|---|
| AUC, KS, G-Mean | XGBoost y SVM | Evalúan desempeño de clasificación sin necesidad de umbral; disponibles para cualquier modelo con score. |
| Recall, F1 | XGBoost y SVM | Revelan errores sobre la clase minoritaria; SVM es más sensible al desbalance. |
| Savings | XGBoost y SVM | Mide impacto financiero sin importar el modelo. |

Para concluir con el marco teórico La formulación matemática de los modelos LDA, Regresión Logística, SVM y XGBoost revela los fundamentos estadísticos y algorítmicos que los sustentan. El Análisis Discriminante Lineal (LDA) se basa en la estadística multivariada clásica, optimizando una razón entre la dispersión interclase e intraclase para lograr una separación lineal óptima de clases (Fisher, 1936; Rao, 1948). Por otro lado, la Regresión Logística formula el problema de clasificación binaria desde la perspectiva probabilística, modelando directamente la

probabilidad de pertenecer a una clase mediante la función logit, lo que permite estimar las probabilidades de clase y facilitar la interpretación de los coeficientes como razones de momios (Cox, 1958).

El método de Máquinas de Vectores de Soporte (SVM), propuesto originalmente por Vapnik (1995), formula el problema de clasificación como una optimización convexa en la cual se busca un hiperplano separador óptimo que maximice el margen entre clases. Para problemas no linealmente separables, SVM incorpora funciones kernel y términos de penalización (margen suave), permitiendo resolver tareas complejas al proyectar los datos a espacios de mayor dimensionalidad, con control explícito sobre el compromiso entre margen máximo y errores de clasificación (Cortes & Vapnik, 1995).

El modelo XGBoost, desarrollado por Chen y Guestrin (2016), es una extensión avanzada del algoritmo de árboles potenciados (gradient boosting), que incorpora una optimización mediante derivadas de segundo orden y regularización explícita para controlar la complejidad estructural de los árboles. Estas características proporcionan un rendimiento predictivo superior y gran eficiencia computacional en tareas predictivas, especialmente con grandes volúmenes de datos.

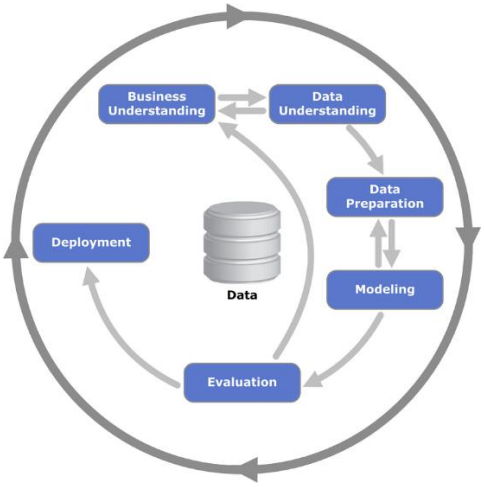
Cada una de las formulaciones expuestas fue presentada originalmente en trabajos pioneros y posteriormente enriquecida por contribuciones teóricas, asegurando un sólido respaldo académico a los fundamentos y procedimientos descritos en este marco teórico.

## Metodología

La investigación se desarrollará siguiendo el modelo estándar CRISP-DM (Cross-Industry Standard Process for Data Mining, véase Figura 1), ampliamente adoptado en aplicaciones analíticas orientadas a la toma de decisiones. Este modelo divide el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación y Despliegue. A continuación, se detalla cada fase adaptada al objetivo de esta tesis: comparar el desempeño predictivo de dos modelos de aprendizaje automático (XGBoost y SVM no lineal) y un modelo tradicional (Regresión Logística) en la predicción del impago en una cooperativa de ahorro y crédito colombiana. Cada fase incluye actividades metodológicas específicas, decisiones técnicas, criterios de validación y referencias normativas pertinentes.

**Figura 1**

*Imagen Referente al Flujo de la metodología CRISP-DM.*




*Tomado de: Health Data Miner*

## Comprensión del Negocio

Esta fase tiene como propósito establecer una comprensión profunda del problema desde la perspectiva institucional, regulatoria y estratégica. En Colombia, el término impago (o incumplimiento de pago) refiere a la situación en que un deudor no cumple sus obligaciones crediticias en los plazos pactados. De acuerdo con la Superintendencia Financiera de Colombia (SFC), un crédito puede considerarse en mora desde el primer día de atraso, pero se clasifica como vencido a partir de los 30 días, y como incumplido (evento de default) al superar los 90 días de vencimiento o ante evidencia clara de irrecuperabilidad (SFC, 2024; Bancóldex, 2024). Esta definición se encuentra alineada con los estándares internacionales de la NIIF 9, que presume impago tras 90 días de mora salvo justificación documentada.

Sobre esta base normativa, se plantea como objetivo de negocio el desarrollo de modelos predictivos que permitan anticipar qué créditos caerán en impago, generando alertas tempranas para reducir la cartera vencida, mejorar la gestión de riesgo y optimizar las decisiones de aprobación y recuperación. Desde el punto de vista estratégico, los criterios de éxito para el negocio incluyen: (i) aumentar la tasa de detección de incumplimientos reales (recall clase impago); (ii) mejorar la precisión general sin deteriorar el trato justo a los clientes; y (iii) proponer modelos que sean implementables bajo las restricciones técnicas y éticas del entorno cooperativo.

### *Definición Formal del Problema de Negocio y Objetivos de la Tesis*

Se formaliza el objetivo general de diseñar un sistema comparativo de predicción de impagos basado en dos enfoques algorítmicos distintos (XGBoost y SVM-RBF), con datos reales de clientes de la cooperativa. El objetivo específico de esta fase es alinear las metas del proyecto analítico con las prioridades y restricciones del negocio cooperativo.

### *Definición Normativa de Impago Respaldada en Fuentes Oficiales*

El evento de impago se define operativamente como aquel en el que el cliente presenta más de 90 días de mora o se tiene evidencia documentada de que no cumplirá su obligación (NIIF 9; Circular Externa 100 de 1995 de la SFC). Para créditos de consumo, el umbral puede reducirse a 60 días conforme al manual de cartera de la entidad financiera.

### *Criterios de Éxito y Métricas de Evaluación Acordadas*

Se establecen como métricas clave: AUC-ROC, Recall, F1-Score, KS Statistic, Savings y Brier Score. Se considera como exitoso un modelo si supera las métricas actuales empleadas por la cooperativa y si puede ser implementado sin transgredir principios de equidad o transparencia.

### *Actividades Metodológicas para la Comprensión del Negocio*

Para realizar los objetivos de esta fase, se realizarán las siguientes actividades: (i) Reuniones Técnicas con Expertos del Área de Riesgo Crediticio y Oficiales de Cumplimiento de la Cooperativa, (ii) Revisión Documental de Políticas Internas de Evaluación y Monitoreo de Riesgo, (iii) Estudio de la Normativa Vigente en Colombia: Circulares de la SFC, Lineamientos de Supersolidaria y Referencias a NIIF 9, (iv) Revisión Ética sobre el Uso de Predicciones Automatizadas y su Impacto en Procesos de Aprobación o Rechazo de Crédito.

## Comprensión de los Datos

Esta fase comprende la recolección, auditoría preliminar y caracterización del conjunto de datos con el que se desarrollará el análisis. El dataset a trabajar es proporcionado por una cooperativa de ahorro y crédito bogotana, la cual pertenece al sector solidario, donde los datos

están anonimizados, no contiene temporalidad en los datos y además son transversales abarcando desde 2022 hasta 2025. La base de datos consta de 10.173 registros y 61 variables.

En un primer momento resulta necesario el poder clasificar las variables, categorizarlas conceptualmente y organizarlas según las dimensiones y su naturaleza técnica como se verá en la tabla 2.

**Tabla 2**

*Clasificación de variables por dimensión y tipo.*

| Dimensión | Variables Representativas | Tipo | Observaciones |
|---|---|---|---|
| Demográficas | Edad, Genero, Estado_Civil, Estrato | Numérica / Categórica | Edad: continua. Estrato: ordinal discreto (1–6). |
| Socioeconómicas | Nivel_Educativo, Tipo_Contrato, Ocupación, Nombre_Empresa | Categórica / Ordinal | Nivel educativo incluye: Técnico, Universitario, etc. |
| Financieras | Saldo_Total, Monto_Credito, Cuota, Mora_Histórica | Numérica | Variables clave para inferencia de capacidad de pago. |
| Operativas | Forma de Pago, Riesgo, Puntuación | Categórica / Ordinal | RIESGO: etiquetas como “medio bajo”, “alto medio”. |
| Derivadas | Saldo_Total/Ingresos, Cuota/Ingresos, rangosaldotota/ing | Continua | Calculadas para representar indicadores relativos. |
| Variable Objetivo | Default_Q | Binaria | 0 = pago completo, 1 = impago. |

Para caracterizar a los clientes de esta cooperativa y así poder determinar esta información que ayuda al trabajo a saber a que público se está enfocando el estudio y con el fin de responder al primer objetivo específico se realiza un análisis exploratorio de las variables.

Como primer análisis se realiza una revisión del desbalance en la variable objetivo donde el resultado es un desbalance del 5.64% para la clase minoritaria y del 94.36% para la clase mayoritaria.

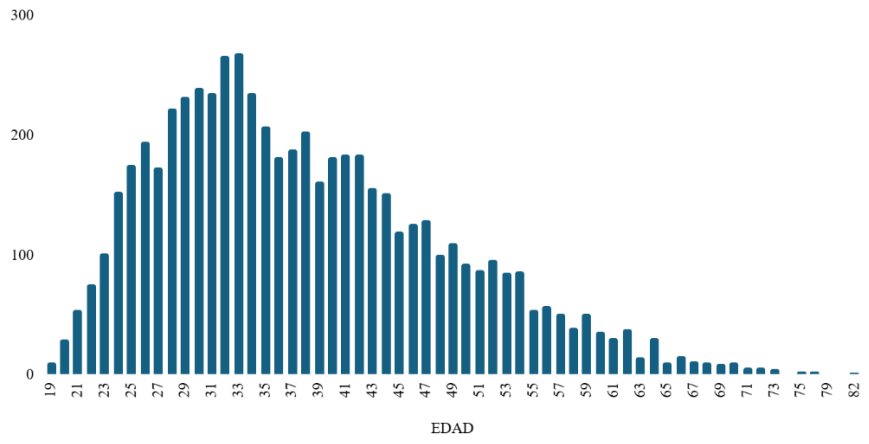
En un segundo momento se revisa la presencia de valores nulos dentro de los datos, el resultado encontrado resalta que en la variable género existen 25 registros nulos, debido a la naturaleza anonimizada de los datos resulta imposible la imputación de estos valores, estos registros comprenden aproximadamente el 0.25% de la totalidad de los datos.

En tercer lugar, se busca hacer la salvedad de que existen variables anómalas que, aunque no contengan nulos en ella todos sus valores son 0 o inclusive en el diccionario de datos se reporta que es una variable que no se recolecta más o la entidad que la comparte dejó de enviarla.

Ahora, para caracterizar la edad de la población de la cooperativa trabajada se hace el gráfico mostrado en la figura 2 donde se ve que la mayor concentración de clientes se encuentra entre los 27 años y los 34 años, aunque en la población hay personas desde los 19 años hasta los 82 años.

**Figura 2**

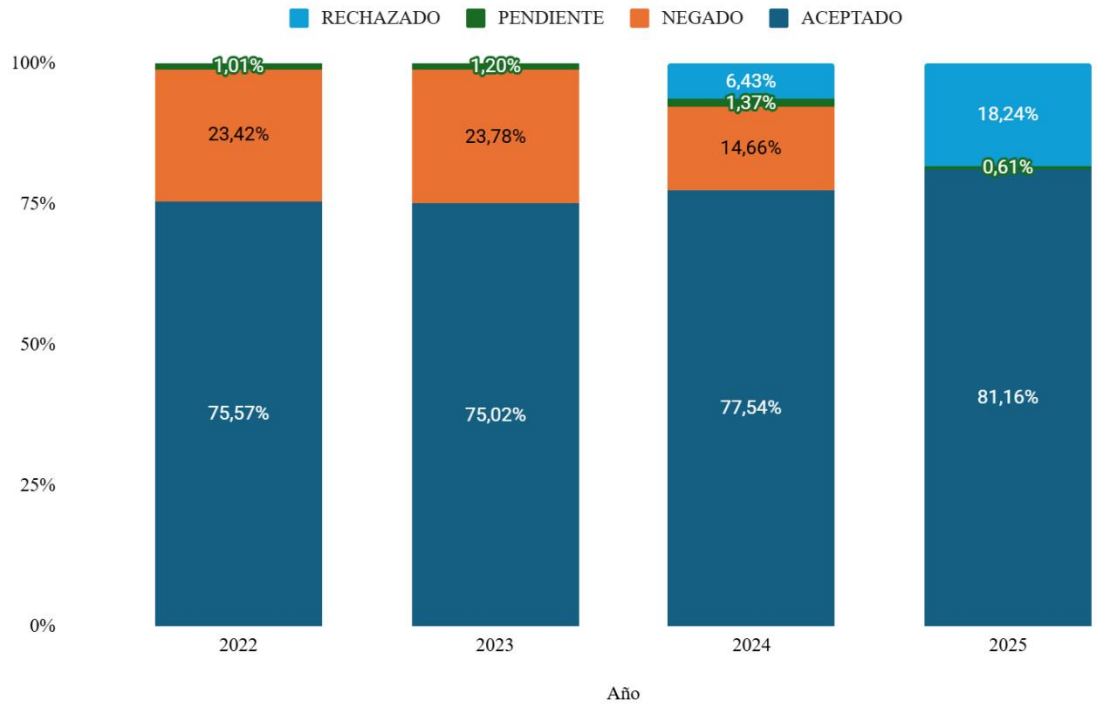
*Imagen Referente a la Distribución de Edad en la Población.*



Dentro de la base de datos existe una variable llamada Resultado Precisión que se interpreta como una evaluación realizada por la Cifin acerca de un cliente y sus posibles valores son: Aceptado, Pendiente, Rechazado y Negado, en la figura 3 se podrá observar la distribución año a año de los solicitantes de crédito de la cooperativa.

**Figura 3**

*Imagen Referente a la Distribución de Resultado Precisión en la Población Año a Año.*



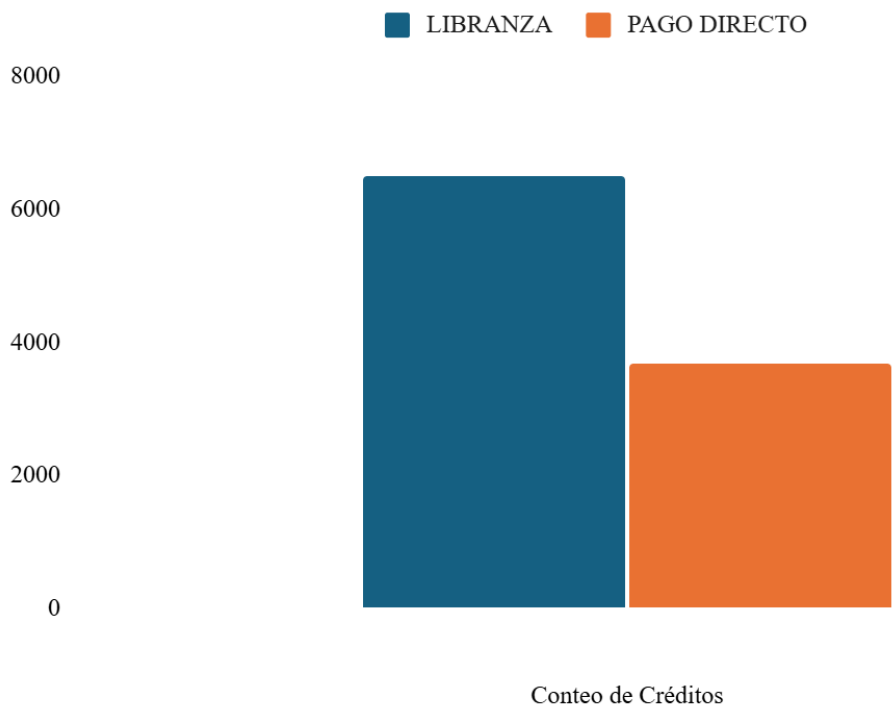

Resulta clave denotar que, las personas que resultan aceptadas de esta decisión han ido aumentando y que la categoría Negado desaparece completamente en 2025 lo que puede indicar un cambio de término lo que puede también orientar con respecto a las dos categorías similares (Negado y Rechazado) y que acá se debería hacer un tratamiento con respecto a esto más aún

cuando se debe hacer encoding para esta variable y el tratamiento incorrecto de esto puede llevar a una variable extra no necesaria.

Ahora para entender los tipos de pagos que se dan en los créditos de esta cooperativa se realiza un análisis que se divide en las dos categorías existentes que son o por libranza o por pago directo donde la figura 4 muestra los resultados.

**Figura 4**

*Imagen Referente a la Distribución del Tipo de Pago.*



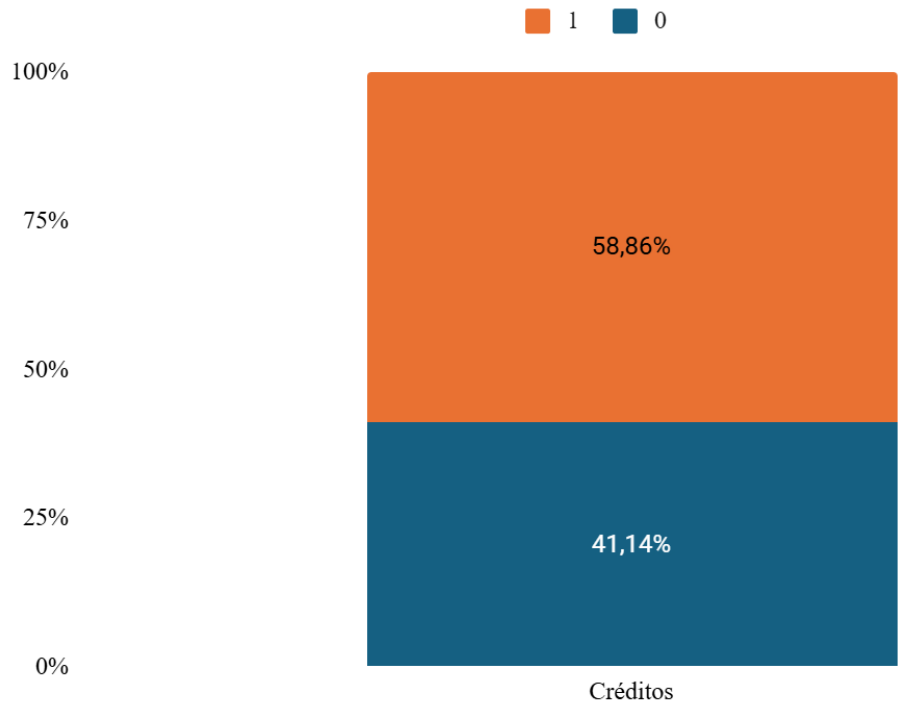

Donde los morosos se distribuyen en 40% por libranza y 60% de pago directo.

Para continuar el análisis descriptivo de los datos se analiza la variable aprobado2 que es aquella que dictamina si el crédito se aprobó, esta variable resulta clave al analizar los datos ya que el objetivo del trabajo es un análisis y detección de la morosidad en la entidad financiera y

solo se conoce si existió morosidad o no en un crédito que se aprobó. La distribución se presenta en la siguiente figura.

**Figura 5**

*Imagen Referente a la Distribución de Créditos Aprobados y Negados en la Cooperativa.*



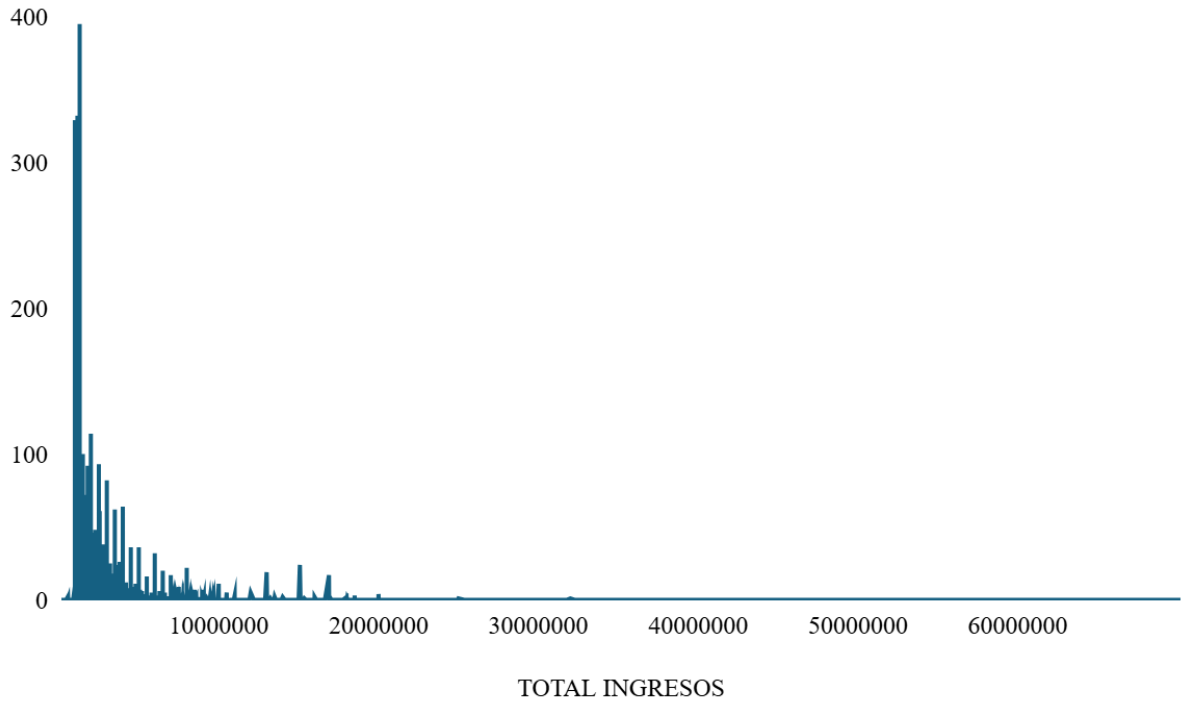

El histograma de la variable Total Ingresos de la figura 6 evidencia una distribución fuertemente asimétrica hacia la derecha, característica típica de variables económicas como salarios o ingresos mensuales. La mayoría de los registros se concentran en valores relativamente bajos, próximos al límite inferior de la distribución, lo que se observa en la barra inicial con frecuencias superiores a 350 observaciones. Conforme aumenta el nivel de ingresos, la frecuencia

disminuye de manera abrupta, mostrando una caída pronunciada durante el primer tramo de la gráfica.

A partir de aproximadamente $5.000.000, la presencia de observaciones se vuelve escasa y dispersa, hasta llegar a valores aislados cercanos a los $30.000.000 y superiores, que corresponden a posibles outliers o ingresos atípicamente altos dentro de la población analizada. La forma de la distribución sugiere que la cooperativa trabaja con una base de asociados cuyo ingreso típico se encuentra en rangos bajos o medios, mientras que los ingresos altos representan una proporción mínima del conjunto total de datos.

**Figura 6**

*Imagen Referente a la Distribución de Ingresos en la Cooperativa.*



La matriz de correlación mostrada en la figura 7 permite identificar la estructura relacional entre las variables financieras, socioeconómicas y operativas incluidas en el estudio, y aporta evidencia sobre los patrones de comportamiento de los solicitantes. En términos generales, se observa que la mayoría de las correlaciones presentan magnitudes bajas o moderadas, lo cual sugiere que las variables aportan información complementaria sin redundancias excesivas. Sin embargo, se identifican algunos grupos de asociaciones fuertes que resultan coherentes con la lógica del riesgo crediticio.

En primer lugar, se destacan las correlaciones elevadas entre variables asociadas al endeudamiento y al uso de productos financieros, particularmente entre cuota_cifin/ingreso y cuotacifin (0.84), así como entre saldo de tarjeta de crédito y cupo disponible (0.74). Estas relaciones reflejan dependencias estructurales: la capacidad de pago relativa se explica directamente por la magnitud de la cuota, y el saldo de tarjetas está acotado por el cupo asignado. De igual manera, se observa una asociación significativa entre monto solicitado y plazo solicitado (0.55), lo cual señala que los créditos de mayor valor suelen requerir plazos más extensos. Este conjunto de patrones confirma la presencia de grupos de variables estrechamente vinculadas, aspecto que debe considerarse para evitar redundancia o inestabilidad en modelos lineales como la regresión logística.

En segundo lugar, variables relacionadas con la capacidad económica del solicitante, tales como total de ingresos, total de activos y saldos consolidados, presentan correlaciones moderadas con varias variables crediticias (entre 0.30 y 0.50). Este comportamiento indica que los clientes con mayor solidez económica tienden a manejar productos de mayor magnitud, con montos y saldos más altos. Desde la perspectiva de modelado, estas variables constituyen indicadores

relevantes del perfil financiero y pueden aportar señales sustanciales al momento de estimar el riesgo de impago.

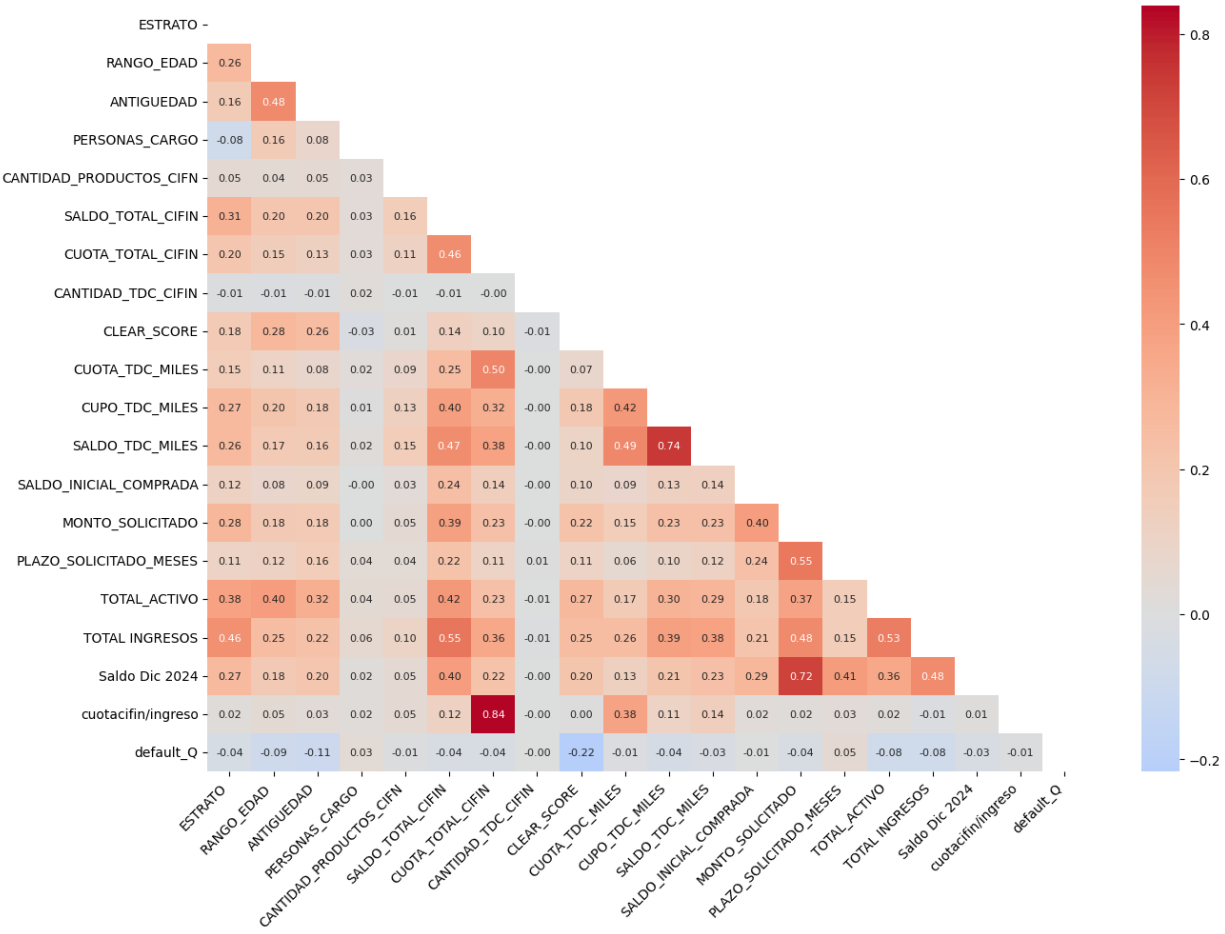
Un elemento central de esta matriz es la relación entre las variables explicativas y la variable objetivo (default_Q). En este caso, las correlaciones con el default son extremadamente bajas, rondando valores entre −0.05 y 0.05. Este comportamiento es habitual en problemas de riesgo crediticio, dado que el impago no depende linealmente de un único factor sino de interacciones complejas entre múltiples variables. La ausencia de correlaciones fuertes con el default no implica falta de información predictiva; por el contrario, confirma que los modelos lineales simples son insuficientes para capturar adecuadamente el fenómeno, y que el uso de técnicas no lineales como XGBoost y SVM resulta pertinente para identificar patrones más sofisticados que no son visibles mediante relaciones lineales bivariadas.

Finalmente, variables socioeconómicas como estrato, personas a cargo o rango de edad presentan correlaciones muy bajas con el resto de los predictores. Esto sugiere que su contribución directa al modelo podría ser limitada cuando se analizan de forma aislada; no obstante, estas variables pueden adquirir relevancia en combinación con otras características, especialmente en modelos capaces de capturar interacciones y efectos no lineales.

En conjunto, la matriz de correlación muestra que la base de datos tiene una estructura heterogénea, con señales distribuidas entre múltiples variables y sin un predictor dominante del impago. Este hallazgo refuerza la necesidad de utilizar modelos avanzados y técnicas de interpretabilidad que permitan identificar relaciones complejas dentro del conjunto de datos.

**Figura 6**

*Imagen Referente a la Matriz de Correlación de las Variables.*




## Preparación de los Datos

Esta fase comprende la transformación del dataset bruto en una versión lista para modelado, cuidando la calidad, representatividad y reproducibilidad del conjunto de entrenamiento. Se implementa un pipeline documentado de preprocesamiento que se separa en dos tipos de tratamientos de datos diferentes, debido a la realización de un modelo logit como base comparativa con los de ML se segmenta el procesamiento en una rama par XGBoost y SVM y otra rama para el Logit debido a los requerimientos a cumplir en los datos:

A continuación, se muestra el primer flujo trabajado, realizado para modelos de ML siendo estos XGBoost y SVM no lineal con kernel RBF.

En un primer momento se cuentan con 10.173 datos los cuales son reducidos a 5988 por la filtración a solo los créditos aprobados en la cooperativa, aquellos créditos de los cuales se tiene conocimiento de un estado en mora o al día.

***Tratamiento de Valores Faltantes***

Primero se tratan los valores faltantes de la base de datos, esto para poder realizar demás procesos que requieren de no contener nulos en la composición de los datos, haciendo un análisis se haya que solo existen valores nulos en la variable género y esos valores no registrados son 25 siendo esto el 0.25% del dataset completo, es por eso que la decisión tomada es eliminar los registros por la poca significancia de los registros para el trabajo; habiendo solucionado este hallazgo se tiene una base de datos sin nulos en ella.

***Eliminación de Variables***

En la comprensión de la base de datos es posible ver variables que caben dentro de alguno de los siguientes 3 casos: (i) Valores Erróneamente Recolectados; (ii) Variables Deprecadas; (iii) Variables Compuestas. Algunos ejemplos de estas eliminaciones son: VALOR_INICIAL_TOTAL que es completamente 0 y en el diccionario se comenta que la cifin como entidad que comparte este campo ya no lo reporta, o existen casos como la variable rangocuota/ing que es una variable aritméticamente compuesta con base a otras variables contenidas en el dataset, donde la única composición conservada fue la capacidad de

endeudamiento dada por (cuotacifin/ingreso); es a partir de ese proceso que se dejan 29 variables incluyendo la objetivo.

### *Corrección de Multicolinealidad*

A las variables Año y tasaea presentan una multicolinealidad bastante alta llegada a 25 con el VIF, lo que lleva a eliminar las variables y deja el VIF más alto en 13 siendo este el del Score siguiendo en 12 el estrato y 10 la cuota mensual en créditos, variables que se dejan por la importancia de estas dentro del estudio y de la decisión de los modelos, sin embargo, los valores del VIF si presentaron una disminución al eliminar las dos variables con el índice más alto.

### *Identificación y Tratamiento de Outliers*

Para el tratamiento de outliers se toma la determinación de hacer uso de DBSCAN como el algoritmo a usar donde lo inicial al implementarlo fue configurar los parámetros de este, los parámetros finalmente ajustados fueron 31 como la cantidad mínima de vecinos y 4 como el eps que sería la distancia mínima para ajustar el grupo; el resultado visto a partir de la implementación de este algoritmo constó de 265 registros eliminados de la clase mayoritaria.

### *División Entrenamiento y Prueba*

Se dividieron los datos en subconjuntos de entrenamiento (70%), validación (10%) y prueba (20%) mediante muestreo estratificado, asegurando la preservación de la proporción de las clases de la variable objetivo (default_Q). Esta división permitirá evaluar el desempeño del modelo sobre datos no vistos y prevenir sobreajuste, además al realizar la partición en este momento evitamos cualquier tipo de figa de información del dataset de prueba.

***Codificación, Escalamiento y PCA***

Las variables categóricas se transformarán según su naturaleza: ordinales mediante codificación por ranking, y nominales mediante codificación one-hot. Las variables numéricas fueron escaladas mediante estandarización z-score. Se realizó un PCA para agrupar variables que puedan compartir su comportamiento bajando así la cantidad de variables a 7 recogiendo 0.7 de la varianza acumulada. Esto con el fin de probar diferentes combinaciones de datasets buscando aquellas que logren los mejores resultados, para los modelos de ML se tendrían particiones de un dataset con PCA y sin PCA y otros con escalamiento y sin escalamiento dejando 4 tipos de particiones a probar y determinar la mejor para este contexto.

Para el desarrollo del modelo de regresión logística se utilizaron las particiones originales del conjunto de datos y se identificaron las variables numéricas y categóricas en la base de entrenamiento utilizando sus tipos de dato nativos. Dado que el supuesto de linealidad del Logit en el logaritmo de las variables independientes se puede evaluar mediante el procedimiento Box–Tidwell, se trabajó únicamente con variables numéricas estrictamente positivas. Los valores iguales a cero se reemplazaron por un valor insignificante para permitir la aplicación del logaritmo sin alterar la estructura de la variable. Para cada variable numérica positiva se estimó un modelo auxiliar de la forma $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \beta_2(xln(x))$, el estadístico asociado a $\beta_2$ permite evaluar si la relación entre la variable y el logit es lineal. Las variables con p-valor $\leq 0.05$ fueron consideradas como violadoras del supuesto de linealidad y, por tanto, candidatas a transformación.

Las variables categóricas fueron transformadas mediante OHE con eliminación de la primera categoría (drop-first), reduciendo la colinealidad y manteniendo la interpretabilidad.

Solo se aplicó OHE a variables categóricas verdaderas, evitando la explosión dimensional propia de los modelos no lineales.

**Modelado**

En esta fase se seleccionan, entrenan y evalúan los modelos predictivos con base en los datos preparados. Se implementarán tres algoritmos con enfoques distintos: un modelo de XGBoost (Extreme Gradient Boosting), una Máquina de Vectores de Soporte (SVM) con núcleo RBF y un modelo Logit como base de comparación tradicional. A continuación, se presenta una tabla con los requerimientos, características técnicas, métricas y técnicas de interpretabilidad para cada modelo.

**Tabla 3**

*Modelos, requerimientos técnicos y estrategias de validación.*

| **Modelo** | **Requerimientos Previos** | **Hiperparámetros Claves** | **Consideraciones Técnicas** | **Métricas** | **Técnicas de Interpretabilidad** |
|---|---|---|---|---|---|
| XGBoost | No requiere normalización. Puede manejar valores faltantes. | n_estimators, max_depth, learning_rate, lambda, alpha. | Sensible al overfitting. Permite control de regularización. Puede manejar features categóricas mediante encoding. | AUC, F1, Recall, Precisión, Especificidad, G-Mean, Índice KS, Índice Lift, Brier Score, Saving. | SHAP (TreeSHAP), Permutation Importance. |
| SVM (RBF) | Requiere normalización estricta. Sensible a outliers y desbalance. | C, gamma, kernel. | Kernel no lineal (RBF); alto costo computacional; requiere | AUC, F1, Recall, Precisión, Especificidad, G- | SHAP (Kernel), Permutation Importance. |

| | | | | | |
|---|---|---|---|---|---|
| | | | escalamiento y balanceo de clase. | Mean, Índice KS, Brier Score, Curvas ROC, Saving. | |
| Logit | Linealidad del logit. Eliminación de multicolinealidad. Depuración de outliers. | Se ajusta el treshold. | Linealidad de predictores y logit. Sensible a multicolinealidad. Predicción dependiente del treshold. | AUC, F1, Recall, Precisión, Especificidad, G-Mean, Índice KS, Brier Score, Curvas ROC, Saving. | - |

Ya habiendo abordado el proceso de modelado se procede a mostrar las primeras métricas extraídas de los modelos (véase Tabla 5).

**Resultados**

**Tabla 5**

*Reporte de métricas de evaluación de modelos.*

| MODELO | AUC PR | AUC-ROC | F1 SCORE | Balanced Accuracy |
|---|---|---|---|---|
| XGB SCALED | 0.85 | 0.98 | 0.74 | 0.86 |
| XGB PCA | 0.83 | 0.98 | 0.73 | 0.87 |
| SVM PCA | 0.81 | 0.98 | 0.70 | 0.83 |
| SVM SCALED | 0.78 | 0.98 | 0.70 | 0.88 |

Las métricas que se logran extraer del modelo resultan prometedoras, aún más cuando al parecer superan aquel indicador de cartera bajo de la entidad trabajada, sin embargo, resulta extraño que en primeros intentos de desarrollar modelos se vea un desempeño tan eficiente en test,

para analizar estas métricas mejor resulta conveniente realizar el análisis SHAP para determinar la importancia de cada variable en las decisiones y dar explicabilidad al modelo.

**Tabla 6**

*SHAP para modelo XGBoost únicamente escalado.*

| Característica | Valor absoluto SHAP |
|---|---|
| Rango Mora | 5.22 |
| Clear Score | 0.15 |
| Estado Civil Soltero | 0.13 |
| Saldo Dic 2024 | 0.12 |
| Forma de Pago Directo | 0.11 |
| Tasaea | 0.05 |
| Total Activo | 0.04 |
| Antigüedad | 0.03 |

Al analizar el resultado del análisis y al poder ver la importancia de las variables para el modelo se ve que aquella que tiene el mayor peso en la decisión del modelo es la variable categórica Rango Mora, esta variable tiene a manera de intervalos la cantidad de días de atraso en pagos de créditos anteriores y ha sido una variable clave en la cooperativa para decidir si se otorga un crédito, es por eso que la correlación de esta columna con respecto a la variable objetivo es casi perfecta, alza las métricas de los modelos cuando se está limitando a aprender del comportamiento de esta variable para clasificar y eso impide que el modelo identifique relaciones entre otra información disponible en el dataset. El problema con respecto a esta variable radica principalmente que aunque su presencia está afectando el modelo es importante para el ámbito de negocio conocer la información de la mora de una persona al mes de diciembre con otras entidades ya que eso refleja en un plazo cercano cómo está siendo su comportamiento crediticio; la solución referente a esto es no tomar una variable que indique tiempo que se lleva de atraso y que tome el

dinero que está atrasado, bajando esto la correlación y aun así indicando si se encuentra en incumplimiento de pago.

Con este cambio efectuado se vuelven a realizar los cuatro modelos principales y el base tradicional, dos de XGBoost, dos de SVM RBF y un logit, los resultados de estos modelos se reportan en la siguiente tabla.

**Tabla 7**

*Reporte de métricas de evaluación de modelos sin Rango Mora.*

| MODELO | AUC PR | AUC-ROC | F1 SCORE | Balanced Accuracy |
|---|---|---|---|---|
| XGB SCALED | 0.45 | 0.87 | 0.39 | 0.63 |
| SVM PCA | 0.33 | 0.82 | 0.05 | 0.51 |
| LOGIT | 0.31 | 0.83 | 0.37 | 0.65 |
| SVM SCALED | 0.23 | 0.79 | 0 (Aprox) | 0.50 |
| XGB PCA | 0.16 | 0.79 | 0.07 | 0.51 |

La inclusión del modelo Logit como línea base responde al objetivo de evaluar si los modelos de aprendizaje automático ofrecen una mejora significativa frente a métodos tradicionales utilizados históricamente en entidades financieras. Aunque su desempeño resulta inferior al de XGBoost en métricas como AUC-ROC y F1, el Logit cumple una función metodológica clave: permite cuantificar la ganancia marginal asociada a la flexibilidad no lineal y a la capacidad de interacción implícita en los modelos de boosting y SVM.

Esta comparación evidencia que, bajo las mismas particiones de entrenamiento y prueba, los modelos no lineales logran capturar relaciones más complejas entre variables, obteniendo

mayores niveles de discriminación entre buenos y malos pagadores. Así, el análisis estadístico confirma que la elección de modelos avanzados no responde únicamente a una preferencia técnica, sino a una mejora observable frente al enfoque tradicional, respaldando la pertinencia de adoptarlos como herramientas complementarias en la evaluación del riesgo crediticio.

Al ver estos resultados es claro que el modelo presenta un problema a la hora de distinguir la clase minoritaria (morosos) ya que aunque el AUC puede parecer alto para el mejor modelo en 0.87 presenta un problema y radica en que la sensibilidad sobre la clase positiva es de 0.28 lo que indica que solo clasifica bien el 28% de los casos minoritarios en test generando un problema ya que el costó ligado a la aceptación de tantas personas morosas es alto debido a que se pierde un porcentaje alto del total del préstamo y se tiene que incurrir en procesos costosos para recuperar la cartera; es debido al bajo rendimiento para determinar los morosos que se opta por ajustar el enfoque de optimización de métricas cambiando el AUC PR por el Recall para ello se procede a usar el mejor modelo reportado de momento (XGBoost escalado) para priorizar la métrica mencionada.

Los dos mejores modelos en test cambiando el enfoque a Recall son reportados en las tablas a continuación:

**Tabla 8**

*Reporte de métricas de evaluación de XGBoost Scaled Hiperparametros: n_estimators = 200, max_depth = 5, learning_rate = 0.3, subsample = 0.8, colsample_bytree = 0.8, min_child_weight = 5.*

| Métrica | Valor |
|---|---|
| ROC AUC | 0.82 |
| AUC PR | 0.34 |
| Gini | 0.65 |

| | |
|---|---|
| KS | 0.53 |
| Precisión | 0.16 |
| Sensibilidad | 0.79 |
| Especificidad | 0.74 |
| F1 | 0.26 |
| G-mean | 0.76 |
| Accuracy | 0.76 |
| Lift | 2.75 |
| Matriz de confusión | |

| | V Real | F Real |
|---|---|---|
| V Predicho | 53 | 275 |
| F Predicho | 14 | 801 |

**Tabla 9**

*Reporte de métricas de evaluación de XGBoost Scaled Hiperparametros: n_estimators = 200, max_depth = 3, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 1, min_child_weight = 5.*

| **Métrica** | **Valor** |
|---|---|
| ROC AUC | 0.90 |
| AUC PR | 0.40 |
| Gini | 0.80 |
| KS | 0.67 |
| Precisión | 0.29 |
| Sensibilidad | 0.67 |
| Especificidad | 0.90 |
| F1 | 0.41 |
| G-mean | 0.77 |
| Accuracy | 0.78 |
| Lift | 5.03 |
| Matriz de confusión | |

| | V Real | F Real |
|---|---|---|
| V Predicho | 45 | 107 |
| F Predicho | 22 | 965 |

En métricas el modelo presentado en la tabla 9 resulta mejor en general, sin embargo, es necesario conocer cuánto cuesta cada decisión errónea, se redujeron los falsos negativos pero la cantidad debe ser una mejora económica para la cooperativa.

Para el modelo de la tabla 8 teniendo en cuenta que el crédito promedio está en $14.000.000 (catorce millones de pesos), la tasa efectiva anual en 14% y el plazo promedio en 43 meses, bajo el supuesto de que el crédito en mora se pierde desde el primer mes y sin tener en cuenta gastos de recuperación de cartera se puede hacer el siguiente cálculo (Siendo FP = Falsos Positivos, $C_{FP}$ = Costo de los Falsos Positivos, FN = Falsos Negativos, $C_{FN}$ = Costo Falsos Negativos, EA = Efectivo Anual y EM = Efectivo Mensual):

$$Costo_{modelo} = FP \times C_{FP} + FN \times C_{FN}$$

$$(1 + i_{EA}) = (1 + i_{EM})^{12} \rightarrow i_{EM} = (1 + i_{EA})^{\frac{1}{12}} - 1$$

$$i_{EM} = (1 + 0.14)^{\frac{1}{12}} - 1 = 1.14^{\frac{1}{12}} - 1 \approx 0.01096 \rightarrow 1 + 0.01096 = 1.01096$$

$$\left(275 \times \left((14.000.000 \times 1.01096^{43}) - 14.000.000\right)\right) + (14 \times 14.000.000)$$

$$= (275 \times 8.246.000) + (14 * 14.000.000) = 2.463.650.000$$

(20)

Viendo este resultado se ve que el modelo puede tener un costo alto producto mayormente del alto rechazo de personas no morosas, el siguiente cálculo corresponde al costo de las decisiones tomadas por el modelo de la tabla 9:

$$Costo_{modelo} = FP \times C_{FP} + FN \times C_{FN}$$

$$y = \left(107 \times \left((14.000.000 \times 1.01096^{43}) - 14.000.000\right)\right) + (22 \times 14.000.000)$$

$$= (107 \times 8.246.000) + (22 * 14.000.000) = 1.190.322.000$$

(21)

El segundo modelo que se planteó resultó ser el mejor teniendo en cuenta que los errores cometidos en la clasificación son menos costosos que los del primer modelo, sin embargo, cabe aclarar que este valor de costo puede ser nocivo para una cooperativa.

A partir de la comparación entre ambos modelos de XGBoost, se observa una mejora consistente en el desempeño del modelo de la tabla 9. Este segundo modelo presenta incrementos notables en métricas globales como ROC AUC (0.90 vs. 0.82), AUC PR (0.40 vs. 0.34) y Gini (0.80 vs. 0.65), lo que indica una mayor capacidad discriminativa. Asimismo, mejora la separación entre clases reflejada en el KS (0.67 vs. 0.53). En cuanto al rendimiento operativo, se evidencia un aumento en Precisión (0.29 vs. 0.16) y F1 (0.41 vs. 0.26), además de un incremento significativo en el Lift (5.03 vs. 2.75), indicando una mayor eficiencia del modelo para identificar casos positivos. Aunque la Sensibilidad disminuye levemente (0.67 vs. 0.79), la Especificidad aumenta de forma importante (0.90 vs. 0.74), lo cual sugiere un mejor control de falsos positivos. En conjunto, las métricas muestran que el segundo modelo ofrece un equilibrio más favorable entre discriminación y estabilidad, superando de manera global al primer modelo y al modelo base tradicional (Logit).

### *Implicaciones Éticas*

El contexto en el que se desarrolla este trabajo corresponde a un sector regulado por entidades como la Superintendencia Financiera y la Superintendencia de la Economía Solidaria. En consecuencia, los modelos utilizados para la evaluación crediticia deben ser transparentes, auditables y explicables tanto para fines de supervisión como para brindar claridad a los solicitantes respecto a las razones de la aprobación o rechazo de su crédito. Para cumplir con estos principios, se incorporaron técnicas de explicabilidad SHAP en sus dimensiones global y local, lo cual permite identificar la importancia relativa de cada variable en el modelo y comprender cómo interactúan para generar una decisión específica.

El uso de aprendizaje de máquina en el ámbito crediticio implica considerar riesgos asociados a sesgos o posibles discriminaciones. Si bien los modelos se entrenan con base en el comportamiento histórico de la cooperativa, esta información puede reflejar patrones que, de no monitorearse adecuadamente, podrían replicar inequidades preexistentes. Aunque el modelo final de XGBoost muestra que el género tiene cierta influencia en la predicción, ello no significa que dicha variable determine por sí sola la aprobación o el rechazo de un crédito, ya que la decisión surge de la combinación de múltiples características. No obstante, la presencia de una variable sensible con relevancia predictiva obliga a implementar mecanismos continuos de vigilancia.

En entornos regulados, atributos como género, raza o edad deben analizarse cuidadosamente para garantizar que no generen efectos discriminatorios indirectos. Por ello, se recomienda complementar la operación del modelo con procesos periódicos de auditoría basados en métricas de equidad —como disparate impact, equal opportunity y comparaciones de tasas de aprobación entre grupos— que permitan identificar si el modelo está reproduciendo desigualdades no deseadas. De esta manera, la cooperativa asegura que las decisiones crediticias se fundamenten en variables financieras y operativas objetivas, evitando que características personales no modificables influyan indebidamente en la evaluación del riesgo.

### *Consideraciones Adicionales*

La aplicación de modelos predictivos en cooperativas de ahorro y crédito tiene un impacto directo sobre la inclusión financiera, especialmente para personas con información limitada en Cifin o TransUnion, o para aquellas en etapas tempranas de su vida crediticia. Los resultados evidencian que, cuando las variables más influyentes provienen de fuentes externas, existe el riesgo de reproducir las mismas barreras estructurales que históricamente han excluido a ciertos

grupos poblacionales. Por ello, el modelo no debe funcionar como un mecanismo de exclusión automática, sino como una herramienta que complemente el análisis del asesor financiero.

Variables internas como el ingreso, la antigüedad en la cooperativa, el tipo de contrato o el comportamiento histórico con la entidad adquieren un papel relevante al permitir que solicitantes sin un puntaje crediticio alto puedan acceder a un crédito de libre inversión. Esta perspectiva fortalece la equidad interna del proceso crediticio y reduce la dependencia de factores externos que podrían introducir sesgos estructurales en la toma de decisiones.

Respecto al despliegue del modelo en un contexto operativo real, no se requieren cambios significativos en la infraestructura actual, dado que la mayoría de los datos utilizados ya se encuentran disponibles durante la solicitud o forman parte del historial de los asociados. Sin embargo, la adopción de modelos predictivos abre la posibilidad de avanzar hacia procesos más automatizados, lo cual demanda una gestión rigurosa de la calidad y consistencia de los datos, así como protocolos de actualización periódica del modelo. Este tipo de automatización solo es viable si el desempeño del modelo mejora de manera sostenida y si se cuenta con mecanismos adecuados de supervisión humana.

Un aspecto crítico en este proceso es la comunicación del rechazo de una solicitud. Aunque SHAP permite explicar con precisión los factores que influyeron en la decisión, esta información debe comunicarse al cliente de manera clara, respetuosa y enfocada en variables modificables, tales como la relación cuota–ingreso, el nivel de endeudamiento o el comportamiento crediticio reciente. Asimismo, para avanzar hacia una mayor automatización, el modelo debe alcanzar niveles más altos de precisión, lo cual depende directamente de mejoras en la calidad de los datos utilizados para su entrenamiento. Con su desempeño actual, el modelo funciona como un apoyo al analista de riesgo, pero no puede reemplazar su criterio experto.

Finalmente, el portafolio analizado corresponde exclusivamente a créditos de libre inversión. En caso de que la cooperativa manejara diferentes modalidades de crédito, sería necesario un análisis estratificado, dado que los perfiles de riesgo y las políticas de evaluación varían significativamente entre productos como consumo, libranza o hipotecarios.

## Conclusiones

El presente estudio tuvo como propósito comparar el desempeño predictivo de modelos de aprendizaje automático —XGBoost y SVM con kernel RBF— frente a un modelo tradicional de regresión logística en la predicción del impago crediticio en una cooperativa de ahorro y crédito colombiana. En conjunto, los resultados permiten responder de manera explícita a los objetivos planteados y evidencian la pertinencia de adoptar modelos avanzados acompañados de mecanismos robustos de explicabilidad y control ético.

En primer lugar, el análisis exploratorio permitió caracterizar el perfil financiero, económico y demográfico de los solicitantes, mostrando una distribución altamente asimétrica en variables como ingresos y activos, así como estructuras de correlación moderadas entre indicadores de endeudamiento. Estos patrones confirman que el riesgo crediticio depende de relaciones complejas, no capturables mediante correlaciones lineales simples, lo que justifica el empleo de modelos no lineales.

En cuanto al desempeño predictivo, tanto XGBoost como SVM mostraron inicialmente métricas elevadas; sin embargo, el análisis de explicabilidad reveló que dichos resultados estaban influenciados por una variable con fuga de información (Rango Mora). Tras su eliminación, los modelos presentaron una caída sustancial en sensibilidad, especialmente en la clase minoritaria. Esta situación llevó a replantear la estrategia de evaluación, priorizando el recall y ajustando el umbral de decisión. Bajo este nuevo enfoque, XGBoost escalado fue el modelo con mejor equilibrio entre precisión y sensibilidad, superando al modelo Logit y demostrando una mayor capacidad para capturar patrones no lineales. Esto confirma que los modelos avanzados ofrecen una mejora sustantiva cuando la información disponible es heterogénea y cuando el evento de interés es raro, como ocurre con el impago.

Desde una perspectiva de inclusión financiera, el estudio muestra que la predicción del riesgo no debe depender exclusivamente de variables externas como puntajes de centrales de riesgo. Aunque estas variables aportan información, su peso desproporcionado podría reproducir barreras estructurales para personas con escasa vida crediticia. El modelo propuesto incorpora variables internas —ingresos, antigüedad, tipo de contrato, entre otras— que permiten ampliar las oportunidades de acceso al crédito para asociados que tradicionalmente serían excluidos por metodologías estrictamente basadas en score.

En materia ética y regulatoria, la incorporación de SHAP permitió dotar al modelo de transparencia tanto a nivel global como individual, facilitando auditorías por parte de entidades como la Superintendencia Financiera o Supersolidaria. Adicionalmente, el análisis mostró que ciertas variables sensibles, como el género, tenían influencia en la predicción. Esto evidencia la necesidad de implementar métricas de equidad, monitoreo de sesgos y protocolos de comunicación clara hacia los solicitantes, de manera que las decisiones crediticias puedan ser justificadas sin vulnerar principios de no discriminación.

En síntesis, los resultados demuestran que XGBoost constituye una alternativa viable para el análisis de riesgo crediticio en cooperativas, siempre y cuando su implementación esté acompañada de procesos adecuados de explicabilidad, ajuste de umbrales, monitoreo de sesgos y actualización periódica de los datos. Asimismo, el estudio resalta la importancia de complementar el criterio automatizado con juicio experto, especialmente en contextos donde la inclusión financiera es un objetivo estratégico.

Finalmente, aunque el rendimiento obtenido es adecuado para apoyar la toma de decisiones, el modelo no debe concebirse como un reemplazo del analista de riesgo, sino como una herramienta que potencializa su labor. Se recomienda profundizar en técnicas de calibración,

incorporar variables longitudinales y evaluar el impacto de condiciones macroeconómicas sobre el impago para fortalecer la precisión y robustez del sistema predictivo en futuros desarrollos.

## Referencias

Banca de las Oportunidades. (2024). Desembolso de créditos de consumo. https://www.bancadelasoportunidades.gov.co/es/estadisticas/credito-de-consumo

Banco de la República. (2024). Informe de Inflación y Política Monetaria. https://www.banrep.gov.co

Superintendencia Financiera de Colombia. (2024). Tasas y desembolsos por la modalidad de crédito. https://www.superfinanciera.gov.co/powerbi/reportes/536/504/

MinTIC. (2024). Observatorio de Inteligencia Artificial: Aplicaciones en el sector financiero colombiano. Ministerio de Tecnologías de la Información y las Comunicaciones. Recuperado de https://www.mintic.gov.co

Superintendencia Financiera de Colombia. (2024). Informe de Estabilidad Financiera – Primer semestre de 2024. Recuperado de https://www.superfinanciera.gov.co

Bhatt, T. K., Ahmed, N., Iqbal, M. B., & Ullah, M. (2023). Examining the Determinants of Credit Risk Management and Their Relationship with the Performance of Commercial Banks in Nepal. Journal of Risk and Financial Management, 16(4), 235. https://doi.org/10.3390/jrfm16040235

Shi, Si & Tse, Rita & Luo, Wuman & D'Addona, Stefano & Pau, Giovanni. (2022). Machine learning-driven credit risk: a systemic review. Neural Computing and Applications. 34. 10.1007/s00521-022-07472-2.

Roy, J. K., & Vasa, L. (2025). Transforming credit risk assessment: A systematic review of AI and machine learning applications. Journal of Infrastructure Policy and Development, 9(1), 9652. https://doi.org/10.24294/jipd9652

D. Borrero, O. Bedoya, Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial, Rev. UIS Ing., vol. 19, no. 4, pp. 37-52, 2020, https://doi.org/10.18273/revuin.v19n4-2020004

Salazar Vergara, J. G. (2022). Diseño de un modelo predictivo para otorgar créditos. Semestre Económico, 24(57), 320–347. https://doi.org/10.22395/seec.v24n57a15

Zhao, Z., Cui, T., Ding, S., Li, J., & Bellotti, A.G. (2024, March 07). Class Imbalance Problem in Credit Risk Prediction. In Encyclopedia. https://encyclopedia.pub/entry/55951

Reis, B., & Quintino, A. (2023). Evaluating Classical and Artificial Intelligence Methods for Credit Risk Analysis. Journal of Economic Analysis, 2(3), 35. doi:10.58567/jea02030006

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: A systematic review and overall comparison. arXiv.org. https://arxiv.org/abs/1602.02137v1

El-Qadi, A., Trocan, M., Frossard, T., & Díaz-Rodríguez, N. (2022). Credit Risk Scoring Forecasting Using a Time Series Approach. Physical Sciences Forum, 5(1), 16. https://doi.org/10.3390/psf2022005016

Bhandary, R., & Ghosh, B. K. (2025). Credit Card Default Prediction: An Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods. Journal of Risk and Financial Management, 18(1), 23. https://doi.org/10.3390/jrfm18010023

Karim Amzile and Mohamed Habachi (2022). Assessment of Support Vector Machine performance for default prediction and credit rating. Banks and Bank Systems, 17(1), 161-175. doi:10.21511/bbs.17(1).2022.14

Letkovsky, S., Jencova, S., Vasanicova, P., Gavura, S., & Bacik, R. (2023). Predicting bankruptcy using artificial intelligence: The case of the engineering industry. Economics and Sociology, 16(4), 178-190. doi:10.14254/2071- 789X.2023/16-4/8

Kochański, B. (2024). The shape of an ROC curve in the evaluation of credit scoring models. Statistics in Transition new series. https://www.semanticscholar.org/paper/The-shape-of-an-ROC-curve-in-the-evaluation-of-Kochański/eb51b85cc269db71c2b5d10103aca54b02b36f7f

Suryanto H, Mahidadia A, Bain M, Guan C and Guan A (2022) Credit Risk Modeling Using Transfer Learning and Domain Adaptation. Front. Artif. Intell. 5:868232. doi: 10.3389/frai.2022.868232

Peykani, P., Peymany Foroushany, M., Tanasescu, C., Sargolzaei, M., & Kamyabfar, H. (2025). Evaluation of Cost-Sensitive Learning Models in Forecasting Business Failure of Capital Market Firms. Mathematics, 13(3), 368. https://doi.org/10.3390/math13030368

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Science Direct. Expert Systems with Applications, 83, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer genomics & proteomics, 15(1), 41–51. https://doi.org/10.21873/cgp.20063

Abdou, H., & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature, Intelligent Systems in Accounting, Finance & Management. Manchester: University of Salford. https://doi.org/10.1002/isaf.325

Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, 179-188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

Altman, E.I. (1968), FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. The Journal of Finance, 23: 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Isidore, N., Etoua Magloire, J. C., Mengue Mvondo Jenner, Ngatom Stephane, & Nkague Leontine. (2021). Discriminant analysis and logistic regression applying to credit risk management. Journal of Advances in Applied Mathematics, 6(3). https://doi.org/10.22606/jaam.2021.63003

Copenhagen Business School (2021). Empirical implications of the Ohlson's O-score in relation to stock returns. CBS Research Portal.

https://research.cbs.dk/en/studentProjects/empirical-implications-of-the-ohlsons-o-score-in-relation-to-stoc

Bellotti, T., & Crook, J. (2008). Support vector machines for credit scoring and discovery of significant features. Expert Systems With Applications, 36(2), 3302–3308. https://doi.org/10.1016/j.eswa.2008.01.005

Rao, C.R. (1948), The Utilization of Multiple Measurements in Problems of Biological Classification†. Journal of the Royal Statistical Society: Series B (Methodological), 10: 159-193. https://doi.org/10.1111/j.2517-6161.1948.tb00008.x

Chen, T., & Guestrin, C. (2016). XGBoost. arXiv, 785–794. https://doi.org/10.1145/2939672.2939785

McLachlan, G. J. (1992). Discriminant analysis and statistical pattern recognition. In Wiley series in probability and statistics. https://doi.org/10.1002/0471725293

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/b94608

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.

Brownlee, J. (2016). XGBoost with Python: Gradient boosting and hyperparameter tuning. Machine Learning Mastery.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. (SHAP)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. (Permutation Importance)

IBM. (s.f.). IBM SPSS Modeler CRISP-DM Guide [Manual de usuario]. Recuperado de https://www.ibm.com/spss

ICEF. (2017, October 27). Superfinanciera explicó concepto de mora temprana y riesgo potencial. https://www.icef.com.co/component/k2/itemlist/tag/Superfinanciera?start=360#:~:text=La%20norma%20no%20modifica%20ninguno,y%20no%20exclusivamente%20de%20mora

Bancóldex. (2024). Banco de Comercio Exterior de Colombia S.A. – Bancóldex y subsidiaria. https://www.bancoldex.com/sites/default/files/estados_financieros_consolidado_31_de_diciembre_2024_definitivos.pdf#:~:text=Independientemente%20del%20análisis%20anterior%2C%20el,de%20vencimiento%20es%20más%20apropiado

Emmanuel, T., Maupong, T., Mpoeleng, D. et al. A survey on missing data in machine learning. J Big Data 8, 140 (2021). https://doi.org/10.1186/s40537-021-00516-9

Support vector machines. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/svm.html

Carvalho, M., Pinho, A.J. & Brás, S. Resampling approaches to handle class imbalance: a review from a data perspective. J Big Data 12, 71 (2025). https://doi.org/10.1186/s40537-025-01119-4

Fisher, A., Rudin, C., & Dominici, F. (2018, January 4). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. arXiv.org. https://arxiv.org/abs/1801.01489

Lundberg, S., & Lee, S. (2017, May 22). A unified approach to interpreting model predictions. arXiv.org. https://arxiv.org/abs/1705.07874

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). "Why should I trust you?": explaining the predictions of any classifier. arXiv.org. https://arxiv.org/abs/1602.04938

Breed, D. G., Hurter, J., Marimo, M., Raletjene, M., Raubenheimer, H., Tomar, V., & Verster, T. (2023). A Forward-Looking IFRS 9 Methodology, Focussing on the Incorporation of Macroeconomic and Macroprudential Information into Expected Credit Loss Calculation. Risks, 11(3), 59. https://doi.org/10.3390/risks11030059

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit scoring and its applications (2nd ed.). SIAM. https://doi.org/10.1137/1.9781611974560