

**MODELO DE SCORE CREDITICIO: UN ENFOQUE BASADO EN
METODOLOGÍAS DE MACHINE LEARNING PARA UNA COOPERATIVA**

Daniel Alejandro Rengifo Aguirre

**Trabajo de grado para optar al título de
Máster en Ciencia de datos**

Director:

Mario Andrés Oviedo Belalcazar



FACULTAD DE INGENIERÍA, DISEÑO Y CIENCIAS APLICADAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI

2024

MODELO DE SCORE CREDITICIO: UN ENFOQUE BASADO EN METODOLOGÍAS DE MACHINE LEARNING PARA UNA COOPERATIVA

INTRODUCCIÓN

Las cooperativas son organizaciones que juegan un papel crucial en el desarrollo social y económico, promoviendo la inclusión y el crecimiento económico sostenible fuera de sus funciones sociales, las cooperativas de ahorro y crédito, permiten a sus asociados, obtener servicios financieros como; ahorros, préstamos, seguros y servicios, bajo condiciones especiales logrando facilitar el acceso a estos recursos, por ello es fundamental brindar herramientas precisas que permitan facilitar la toma de decisiones en el otorgamiento de crédito con el fin de salvaguardar la estabilidad de las entidades. (Naciones Unidas, 2023)

El proyecto, busca aplicar metodologías basadas en analítica, mediante uso de *machine learning*, para determinar de manera rigurosa un *score* o puntaje de crédito que permita evidenciar durante la fase de estudio de crédito, la capacidad y disposición de pago de los asociados aspirante a crédito, con el fin no solo de mejorar las herramientas bajo las cuales actualmente se administra el riesgo de crédito, sino que también contribuya a la sostenibilidad financiera de la cooperativa.

Este estudio se fundamenta en el análisis previo relacionado con los factores que influyen en el comportamiento de crédito, enfocándose en la selección de características más relevantes para el desarrollo del *score*, este análisis incluye la evaluación de los modelos de clasificación usando métricas clave como precisión (Accuracy), sensibilidad (Recall) y puntaje F1 (F1-Score), con el objetivo de garantizar la robustez, precisión y aplicabilidad del modelo en el contexto de la cooperativa.",

Este trabajo cobra validez dado que, en la actualidad, no se cuenta con un *score* interno preciso sino que las decisiones de otorgamiento se realizan en función del criterio experto del evaluador. Con esta propuesta se pretende facilitar el análisis y disminuir los tiempos de respuesta al asociado, así como contribuir a la adopción de metodologías y herramientas que robustecen el área de administración del riesgo crediticio dentro de la organización.

CONTENIDO

1.	CONTEXTO Y ANTECEDENTES.....	5
1.1	FORMULACIÓN DEL PROBLEMA – PREGUNTA DE INVESTIGACIÓN	5
1.2	JUSTIFICACIÓN.....	6
1.3	OBJETIVOS.....	6
1.3.1	Objetivo General	6
1.3.2	Objetivos Específicos	6
2.	METODOLOGÍA	7
2.1	<i>Comprensión del negocio:</i>	8
2.2	Comprensión de los datos:.....	8
2.3	Preparación de los datos:.....	8
2.4	Modelado:.....	9
2.5	Evaluación:	9
3.	MARCOS DE REFERENCIA.....	10
3.1	MARCO TEÓRICO CONCEPTUAL.....	10
3.1.1	Conceptos de Negocio:	10
3.1.2	Conceptos estadísticos:.....	10
3.1.3	Métricas de decisión:.....	11
4.	ESTADO DEL ARTE	11
4.1	Trabajos seleccionados	12
4.1.1	Diseño de un modelo de scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento colombiana (Arango & Restrepo, 2017)	12
4.1.2	Aprendizaje supervisado en la construcción de un modelo de Credit Scoring para cooperativas de ahorro y crédito en Colombia (Cano Bedoya, 2021)	13
4.1.3	Modelo de credit scoring para predecir el otorgamiento de crédito personal en una cooperativa de ahorro y crédito (Rodríguez & Ulfe, 2015)	13
4.1.4	Comparación de modelos de riesgo de crédito: modelos logísticos y redes neuronales (Becerra Ladino, 2014).....	13
4.1.5	Método para evaluar el scoring de crédito de la línea de libranzas en las cooperativas de crédito de Medellín” (Gonzalez Mora, 2023)	14
4.1.6	Eficiencia en los modelos de aprendizaje de máquina para evaluar el riesgo crediticio de personas naturales en una institución financiera de Chiclayo (Tensén Aroyo, 2017)	14
4.1.7	Predicción del Riesgo Crediticio en Colombia Usando Técnicas de Inteligencia Artificial (Borrero Tigreros & Bedoya Leiva, 2020)	14

5.	ANÁLISIS EXPLORATORIO DE DATOS	15
6.	SELECCIÓN DE VARIABLES.....	20
6.1	Selección de variables por conocimiento del negocio	20
6.2	Análisis de importancia de variables con Lasso	21
7.	RESULTADOS.....	21
8.	CONCLUSION.....	24
9.	REFERENCIAS	25

1. CONTEXTO Y ANTECEDENTES

Las cooperativas se crean para satisfacer las necesidades comunes de sus miembros, ya sea en forma de productos, servicios o ambos. Desempeñan un papel fundamental en la sociedad al impactar positivamente a las comunidades donde operan, mediante transferencias sociales como auxilios, beneficios o el otorgamiento de préstamos bajo condiciones especiales. Es importante resaltar que una de sus principales misiones es promover prácticas sociales sostenibles que beneficien a sus miembros y a la comunidad a largo plazo, asegurando el equilibrio entre el impacto social y la sostenibilidad económica.

Para lograr este equilibrio, es esencial que las cooperativas cuenten con metodologías precisas y herramientas avanzadas que respalden la administración del riesgo crediticio. Actualmente, muchas decisiones relacionadas con la concesión de crédito se toman en función del criterio experto, lo que puede dar lugar a inconsistencias y subjetividades. Este proyecto busca abordar este desafío desarrollando un modelo de puntuación crediticia basado en técnicas de ciencia de datos y *machine learning*. Dicho modelo permitirá identificar las variables más influyentes en el rechazo o aprobación de una solicitud de crédito, proporcionando un enfoque más objetivo, estandarizado y eficiente.

El resultado final será un sistema de puntuación crediticia que no solo apoyará la toma de decisiones estratégicas en la cooperativa, sino que también servirá como referencia para implementar metodologías similares en otras organizaciones del sector. Este modelo será capaz de integrarse en los procesos actuales de evaluación de crédito, permitiendo identificar, con un alto grado de confiabilidad, los perfiles más idóneos para recibir financiación y bajo qué condiciones, fortaleciendo así la misión social y la sostenibilidad de la cooperativa.

1.1 FORMULACIÓN DEL PROBLEMA – PREGUNTA DE INVESTIGACIÓN

Si bien la cooperativa desempeña un rol social en la comunidad, también administra recursos que deben salvaguardarse para garantizar su equilibrio financiero. Por lo tanto, el problema se centra en: ¿Cómo puede la cooperativa desarrollar un modelo de puntuación crediticia basado en técnicas de ciencia de datos y *machine learning* que permita una evaluación precisa y estandarizada de la disposición de pago de sus asociados, mejorando así la calidad de la cartera de crédito y la sostenibilidad financiera de la entidad?

1.2 JUSTIFICACIÓN

La importancia de este proyecto radica en que contar con un sistema robusto de administración del riesgo crediticio es indispensable para asegurar una gestión administrativa y financiera eficiente en una entidad dedicada a la colocación de recursos. Una gestión deficiente puede provocar un aumento en la morosidad de la cartera, afectando la liquidez y, por ende, la continuidad de la organización en el sector. En este contexto, la ausencia de un sistema estandarizado puede conducir a decisiones crediticias inapropiadas. Por ello, la aplicación de técnicas de *machine learning* en la construcción de modelos de puntuación crediticia ha demostrado ser más precisa y eficiente en la identificación de patrones y elementos clave que podrían pasar desapercibidos mediante métodos tradicionales.

Este enfoque permitirá a la entidad conocer de manera anticipada el perfil de riesgo del asociado al que puede ofrecer sus productos financieros, y establecer relaciones que identifiquen el potencial de clientes óptimos, asegurando el sano crecimiento de la cartera. Además, servirá como insumo para la toma de decisiones en el otorgamiento, no solo identificando de a que asociados se les puede prestar sino también bajo qué condiciones. De igual manera, se espera que con este estudio se genere un precedente en la aplicación de este tipo de metodologías en el sector, sirviendo de base a investigaciones futuras en la región y sector.

1.3 OBJETIVOS

1.3.1 Objetivo General

- Desarrollar un modelo de score crediticio para una cooperativa del valle del cauca a partir de las variables clave que influyen en la capacidad y decisión de pago de los asociados por medio del uso de la ciencia de datos y el machine learning.

1.3.2 Objetivos Específicos

- Determinar la relación entre las variables independientes y el comportamiento de la variable objetivo.
- Mejorar la calidad y adecuación de los datos para asegurar su idoneidad en el desarrollo del modelo de score de crédito.
- Identificar las variables más relevantes que influyen el modelo de score de crédito, optimizando su capacidad predictiva.
- Evaluar el modelo predictivo más efectivo para el score de crédito, asegurando su alta precisión y confiabilidad.

2. METODOLOGÍA

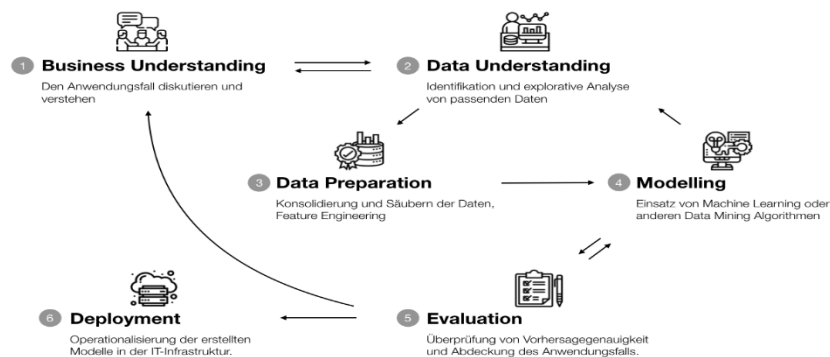
La Dirección de Planeación y Riesgo de la Cooperativa proporcionó un archivo CSV con información sobre los asociados con créditos vigentes en la entidad. Este archivo contiene 6,628 registros y 113 variables, tanto cualitativas como cuantitativas. Incluye información actualizada de los créditos que forman parte de la cartera activa de la organización y proporciona, entre otros campos de interés, el saldo actual de las obligaciones, el comportamiento de pago y elementos sociodemográficos de los asociados relacionados con las deudas registradas. El propósito de este estudio es analizar las variables que tienen mayor incidencia en la determinación del otorgamiento de crédito a los asociados y, con ello, responder a los objetivos de análisis planteados en el documento.

Esta información se extrajo directamente de la base de datos de la entidad, para lo cual fue necesario construir un Query que permitiera la exportación de la totalidad de las obligaciones existentes así como de la información relacionada con el titular de la deuda, dentro de esta información se encuentra la variable objetivo del estudio “Score de Crédito” que en la actualidad se asigna de manera manual y que se constituirá como la base de este estudio. Es importante mencionar, que durante la exportación inicial de la tabla a la base de datos se eliminó variables redundantes o duplicadas dentro del dataset que va a ser usado en el proyecto.

El proyecto emplea la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), que es ampliamente utilizada en proyectos de ciencia de datos y analítica. Esta metodología sigue un enfoque estructurado que guía el desarrollo de soluciones desde la comprensión del negocio y los datos, hasta el despliegue de los modelos predictivos, asegurando un proceso integral y repetible.(Chapman et al., 2000).

DATADRIVENCOMPANY.DE

CRISP DM



Fuente: **Data Driven Company**. (s.f.). *CRISP DM: Das Data Mining Modell einfach erklärt*.

Recuperado de <https://datadrivencompany.de>

2.1 Comprensión del negocio:

La fase de comprensión del negocio tiene como objetivo alinear el desarrollo del modelo de score de crédito con los objetivos estratégicos de la cooperativa. En este contexto, es crucial entender que la cooperativa no solo desempeña un rol social al brindar servicios financieros a sus asociados, sino que también tiene la responsabilidad de garantizar la sostenibilidad económica y la adecuada gestión de los riesgos asociados a su cartera crediticia.

2.2 Comprensión de los datos:

Se realizó la extracción y consolidación de datos relevantes para el análisis, en este caso relacionado con las variables asociadas a las obligaciones vigentes por cada asociado. En esta etapa también se incluye el análisis exploratorio de los datos para comprender su distribución y la calidad de los mismos.

2.3 Preparación de los datos:

En la fase de preparación de los datos, se realizaron diversas tareas para garantizar la calidad y adecuación del conjunto de datos utilizado en el desarrollo del modelo de score. Inicialmente, se llevó a cabo la extracción de los datos desde la base de la cooperativa, consolidando un conjunto de 6,628 registros con 113 variables cualitativas y cuantitativas relevantes. Durante esta etapa, se eliminaron registros duplicados y valores faltantes mediante estrategias de imputación, como el uso de la moda para variables categóricas y métodos estadísticos para datos numéricos. Además, se gestionaron valores atípicos detectados en variables críticas, como ingresos y saldos de deuda, para minimizar su impacto en los análisis posteriores.

Posteriormente, se transformaron las variables categóricas en un formato numérico utilizando técnicas como *one-hot encoding*, lo que facilitó su incorporación en los algoritmos de *machine learning*. También se aplicaron técnicas de normalización y escalado a las variables numéricas para garantizar que todas estuvieran en un rango comparable, optimizando así el rendimiento de los modelos. Estas tareas aseguraron que el conjunto de datos final estuviera limpio, estructurado y listo para el análisis, reduciendo riesgos de errores y mejorando la eficiencia en las siguientes fases del proyecto.

Para reducir la dimensionalidad y optimizar el conjunto de datos, se utilizó dos técnicas complementarias: Lasso y PCA. El método de selección de variables Lasso permitió identificar aquellas características más relevantes para el modelo penalizando los coeficientes de las variables menos significativas, eliminando

así aquellas que no aportaban valor predictivo y ayudando a evitar el sobreajuste. Por su parte, el Análisis de Componentes Principales (PCA) transformó las variables originales en un conjunto reducido de componentes principales que explicaban la mayor parte de la varianza. Este enfoque dual no solo mejoró la interpretabilidad y eficiencia del modelo, sino que también garantizó que los datos finales fueran representativos y adecuados para capturar los patrones clave necesarios para el desarrollo del score crediticio.

2.4 Modelado:

En la etapa de modelado, se evaluaron y compararon múltiples algoritmos de clasificación para identificar el modelo más adecuado para predecir el score crediticio. Los datos fueron divididos mediante validación cruzada estratificada, lo que aseguró una representación equilibrada de las clases en los conjuntos de entrenamiento y validación. Esto garantizó que los resultados fueran consistentes y representativos del comportamiento real de los datos.

Los modelos evaluados:

- Regresión Logística: Utilizada como modelo base debido a su simplicidad y capacidad para interpretar los coeficientes asociados a las variables predictoras.
- Máquinas de Soporte Vectorial (SVM): Elegidas por su habilidad para manejar datos linealmente no separables mediante el uso de funciones kernel.
- Árboles de Decisión: Evaluados por su facilidad de interpretación y su capacidad para manejar interacciones entre variables.
- Modelos Basados en Ensamblados: Se exploraron enfoques como *Random Forest* y *Gradient Boosting*, conocidos por su robustez y precisión en tareas de clasificación.

2.5 Evaluación:

Durante la validación, se utilizaron métricas como F1-score, AUC-ROC y precisión para evaluar el desempeño de cada modelo. Finalmente, tras un análisis comparativo, se seleccionó el modelo con mejor desempeño, alineado con las necesidades operativas y estratégicas de la cooperativa. Este modelo será la base para la implementación y despliegue del sistema de score crediticio.

3. MARCOS DE REFERENCIA

3.1 MARCO TEÓRICO CONCEPTUAL

3.1.1 Conceptos de Negocio:

Score de Crédito: Es una medida numérica utilizada para evaluar la solvencia de un individuo en función a su comportamiento de pago, este puntaje revela la probabilidad de que una persona cumpla con sus obligaciones crediticias. (Fieldman & Schmidt, 2016)

Asociado: Persona que se une voluntariamente a una cooperativa para satisfacer sus necesidades económicas, sociales o culturales, por medio de una empresa asociada en la cual los miembros de la cooperativa participan en la toma de decisiones en la entidad. (International Cooperative Alliance, 2024)

Capacidad de Pago: Se refiere a la habilidad de una entidad para cumplir con sus obligaciones financieras, considerando sus ingresos, gastos y otras responsabilidades financieras.

Incumplimiento: Se refiere al hecho de que una contrapartida de una operación financiera, incumpla con sus obligaciones contractuales. En sentido estricto, es el retraso en el pago de sus compromisos financieros. (Garcia, 2007)

3.1.2 Conceptos estadísticos:

Modelo de Clasificación: Un modelo de clasificación es un tipo de algoritmo de machine learning que se utiliza para predecir la clase o categoría a la que pertenece una nueva observación, basándose en datos de entrenamiento con clases conocidas. (Goodfellow, Bengio, & Courville, 2016)

Regresión Logística: La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. (Chirarroni, 2002)

Máquinas de Soporte Vectorial: Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) son un conjunto de algoritmos de aprendizaje supervisado, se utilizan para problemas de clasificación y regresión. En la clasificación, las SVM buscan encontrar el hiperplano que mejor separa las clases en el espacio de características, maximizando la distancia entre las muestras más cercanas de cada clase (los vectores de soporte). (Vapnik & Cortes, 1995)

Árboles de Decisión: Los árboles de decisión para clasificación y regresión son una metodología no paramétrica que consiste en dividir iterativamente el espacio de características para predecir un valor de salida. Los árboles de decisión se utilizan tanto para problemas de clasificación, donde el objetivo es asignar etiquetas a las observaciones, como para problemas de regresión, donde el objetivo es predecir valores continuos. (Morgan & Jerome, 1963)

3.1.3 Métricas de decisión:

El F1-score: es la media armónica entre la precisión y el recall. Proporciona una medida equilibrada que es útil cuando existe un desbalance entre las clases, ya que toma en cuenta tanto los falsos positivos como los falsos negativos. (Marina Sokolova, 2009)

Precisión: es la proporción de instancias correctamente clasificadas como positivas frente al total de predicciones positivas realizadas por el modelo. Evalúa la capacidad del modelo para evitar falsos positivos (Powers, 2020)

Recall: mide la proporción de instancias positivas correctamente identificadas por el modelo frente al total de instancias positivas existentes en el conjunto de datos. Evalúa la capacidad del modelo para detectar todos los verdaderos positivos. (Christopher D. Manning, 2008)

4. ESTADO DEL ARTE

El score de crédito es el término utilizado para describir los métodos estadísticos formales empleados para clasificar a los clientes potenciales de una entidad que realiza actividades de intermediación financiera, así como para segmentar a aquellos que ya son clientes y poseen productos de la misma (Hand & Henley, 1997).

Los modelos de score, son hoy en día una de las herramientas más utilizadas en la administración del riesgo crediticio, de acuerdo con Gutierrez Girault (2007), la

utilización de esta técnica comenzó desde los años 70's y aumentó su uso con el desarrollo computacional de los 90's, lo cual permitió la implementación de modelos eficientes que permitieran la correcta evaluación del riesgo en un portafolio, pese a estos avances el juicio humano aún continúa siendo empleado en el otorgamiento de crédito.

La implementación de un score busca relacionar la probabilidad de incumplimiento de un cliente basándose en el análisis de características y su ponderación entre elementos de tipo cualitativo y cuantitativo. Aunque, existe un gran número de publicaciones relacionadas al proyecto, a continuación se detallan algunas investigaciones que han sido seleccionadas por su relevancia con el problema planteado en la investigación:

4.1 Trabajos seleccionados

El desarrollo de modelos de score crediticio ha experimentado una evolución significativa en los últimos años, impulsada tanto por los avances en algoritmos de aprendizaje automático como por el mayor acceso a datos financieros. Inicialmente, los modelos clásicos, como la regresión logística, destacaron por su simplicidad y capacidad de interpretación. Sin embargo, investigaciones recientes han incorporado técnicas más avanzadas, como máquinas de soporte vectorial, redes neuronales y árboles de decisión, que han demostrado su eficacia en escenarios complejos. En este contexto, se presenta una recopilación de trabajos seleccionados que han servido como referencia para la elaboración del presente estudio.

4.1.1 Diseño de un modelo de scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento colombiana (Arango & Restrepo, 2017)

En este trabajo de grado, emplea cuatro técnicas para modelos de scoring las cuales son: análisis discriminante, modelo probabilístico, modelo logístico y redes neuronales artificiales, utilizando métricas de decisión tales como matrices de confusión, prueba de Kolmogorov-Smirnov, Finalmente, se escogió el modelo de regresión logística, no solo por ser el utilizado como referencia por la Superintendencia Financiera de Colombia, sino también porque presentó el mejor ajuste en las métricas empleadas

4.1.2 Aprendizaje supervisado en la construcción de un modelo de Credit Scoring para cooperativas de ahorro y crédito en Colombia (Cano Bedoya, 2021)

Este trabajo de grado, utiliza métodos Biplot, análisis de correspondencias y componentes principales para la reducción de dimensionalidad, técnicas de árboles de decisión, modelos de regresión probabilísticos, K-Vecinos más cercanos, máquinas de soporte vectorial y redes neuronales, con el fin de asignar el puntaje de crédito y seleccionar las variables que son significativas en el estudio. El modelo de regresión logística es empleado para el Score de crédito y se compara con las demás técnicas supervisadas mediante métricas de rendimiento.

4.1.3 Modelo de credit scoring para predecir el otorgamiento de crédito personal en una cooperativa de ahorro y crédito (Rodríguez & Ulfe, 2015)

Este trabajo de grado, realizó la construcción de un modelo scoring para predecir el otorgamiento de crédito personal, con la finalidad de clasificar a los clientes a partir de la probabilidad de default. Metodológicamente la investigación es aplicada, con propósito predictivo y explicativo, basada en el proceso CRISP-DM. Las técnicas utilizadas fueron Regresión Logística, Árboles de Clasificación y Redes Neuronales; la comparación de los modelos se realizó considerando las capacidades de clasificación y predicción, eligiendo como mejor modelo el de regresión logística por agrupación interactiva.

4.1.4 Comparación de modelos de riesgo de crédito: modelos logísticos y redes neuronales (Becerra Ladino, 2014)

Este Trabajo de grado, comparó el poder predictivo de las regresiones logísticas y las redes neuronales utilizando varias funciones de costos. Dicha comparación se llevó a cabo en un portafolio de tarjetas de crédito de una institución financiera colombiana. Inicialmente se estimaron modelos de redes neuronales feedforward con las mismas variables del modelo Logístico, utilizando una función de activación arco tangente (tanh) y en la última capa (output layer) una función de activación logística, buscando que el resultado tenga un rango (0,1), y variando el número de capas ocultas y unidades ocultas en cada capa.

4.1.5 Método para evaluar el scoring de crédito de la línea de libranzas en las cooperativas de crédito de Medellín” (Gonzalez Mora, 2023)

En el presente proyecto de grado, se desarrolla una investigación empírica a modo de profundización sobre algunos modelos de clasificación de riesgo, sustentados en la metodología de la curva “ROC”, a partir de la cual se selecciona el modelo que mejor área bajo la curva “AUC” tenga, y así, predecir de una mejor forma el comportamiento del buen y mal hábito de pago de los clientes de las Cooperativas Financieras de Crédito de Medellín, en su línea de libranzas. Dicha investigación utilizó clasificadores de riesgo como Generalized Linear Models (GLMs)-Logit, Deep Learning (Neural Networks)-Redes Neuronales y Árboles Binomiales.

4.1.6 Eficiencia en los modelos de aprendizaje de máquina para evaluar el riesgo crediticio de personas naturales en una institución financiera de Chiclayo (Tensén Aroyo, 2017)

En el presente proyecto, se analizan diferentes modelos de aprendizaje de máquina evaluando eficazmente el riesgo crediticio de personas naturales de una institución financiera de Chiclayo, la investigación es de tipo descriptivo, explicativo y predictivo para lo cual se trabajó bajo el esquema CRISP-DM aplicado al uso de herramientas como arboles de clasificación, redes neuronales, máquinas de soporte vectorial y el modelo clásico de regresión logística.

4.1.7 Predicción del Riesgo Crediticio en Colombia Usando Técnicas de Inteligencia Artificial (Borrero Tigreros & Bedoya Leiva, 2020)

Este trabajo se centra en la aplicación de técnicas de inteligencia artificial para la predicción del riesgo crediticio en Colombia. Los autores implementaron y compararon diversos modelos, incluyendo regresión logística, árboles de decisión y máquinas de soporte vectorial (SVM), para evaluar su eficacia en la clasificación de clientes según su probabilidad de incumplimiento. Utilizando un conjunto de datos proporcionado por una entidad financiera colombiana, se llevaron a cabo pruebas de validación cruzada y se analizaron métricas como la precisión, el recall y el F1-score para determinar el desempeño de cada modelo. Los resultados indicaron que, aunque la regresión logística ofreció una interpretación clara de las variables predictoras, los modelos basados en árboles de decisión y SVM presentaron una mayor precisión en la clasificación de clientes. El estudio concluye que la integración de

estas técnicas puede mejorar significativamente los sistemas de evaluación de riesgo crediticio en el sector financiero colombiano.

Tabla No 1: Resumen de la comparación de los trabajos relacionados para este proyecto de grado

Estado del Arte	4.1.1.	4.1.2.	4.1.3.	4.1.4.	4.1.5.	4.1.6.	4.1.7
Fecha de Publicación	2017	2021	2015	2014	2023	2017	2020
País	Colombia	Colombia	Perú	Colombia	Colombia	Ecuador	Colombia
Idioma	Español	Español	Español	Español	Español	Español	Español
Algoritmo de Clasificación Usado	Arboles de Decisión - Redes Neuronales	Arboles de Decisión - SVM - Redes Neuronales	Regresión Logística - Arboles de Clasificación - Redes Neuronales	Regresión Logística - Redes Neuronales	Redes Neuronales	Redes Neuronales - SVM - Regresión Logística	Regresión Logística, SVM, Arboles de Decisión

La tabla anterior, compara siete investigaciones sobre modelos de Scoring crediticio, destacando su evolución desde 2014 hasta 2023. La mayoría de los estudios fueron realizados en Colombia, lo que refleja el interés del país en el desarrollo de técnicas avanzadas para evaluar el riesgo crediticio. Los estudios están escritos en español, facilitando su acceso y comprensión en la región. Temporalmente, se observa un avance en la complejidad de los modelos empleados, desde enfoques tradicionales como la regresión logística hasta técnicas más modernas como redes neuronales y máquinas de soporte vectorial (SVM).

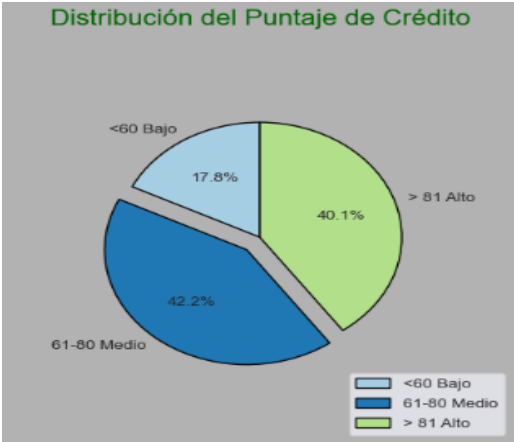
En cuanto a los algoritmos de clasificación, la regresión logística se mantiene como un método ampliamente utilizado debido a su simplicidad e interpretabilidad. Sin embargo, modelos más avanzados, como redes neuronales y árboles de decisión, están presentes en la mayoría de los trabajos, mientras que enfoques como SVM destacan en investigaciones recientes. Esta tendencia refleja un interés creciente en metodologías más sofisticadas que combinan la capacidad predictiva con un análisis profundo de las variables, alineándose con las necesidades del sector financiero para mejorar la precisión y robustez de los sistemas de Scoring crediticio.

5. ANÁLISIS EXPLORATORIO DE DATOS

El Análisis exploratorio es una parte fundamental de la fase inicial en un proyecto de ciencia de datos. Su relevancia radica en que permite comprender y establecer relaciones entre la variable dependiente y las independientes relacionadas con esta. En el caso de estudio, se busca evidenciar el comportamiento de la variable “Scoring” frente a las que se tienen consolidadas en el dataset como el nivel educativo, cantidad de ingreso, edad, entre otras.

En primer lugar, se analiza la distribución de la variable dependiente (puntaje de crédito) en el dataset, que contiene 6.628 registros con corte abril 2024 que corresponden a obligaciones vigentes con la entidad. Del total, el 17% de las obligaciones es decir 1.127 registros tiene un score bajo, en el rango medio se encuentra el 42% con 2.784 y en alto 2.717 obligaciones.

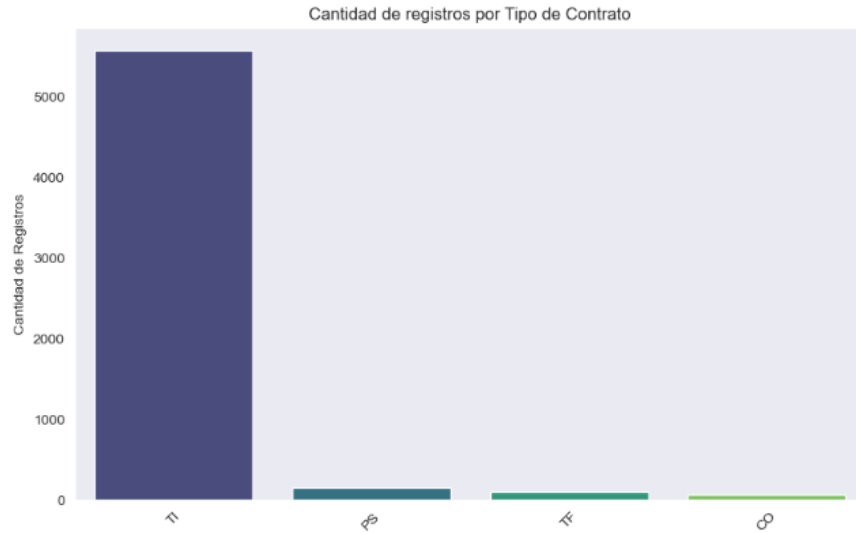
Ilustración 1. Distribución del puntaje de crédito



Fuente: Cooperativa – Elaboración Propia

La mayoría de los asociados con créditos vigentes se concentran bajo el tipo de contrato indefinido, esto es una característica dado el tipo de relación que tiene la cooperativa con las empresas vinculadas al grupo, se encontró que 5.577 asociados tienen este tipo de contrato, 152 se encuentran vinculados mediante prestación de servicios y 104 con término fijo. Finalmente, 795 asociados cuentan con contrato por obra.

Ilustración 2. Cantidad de registros por tipo de contrato

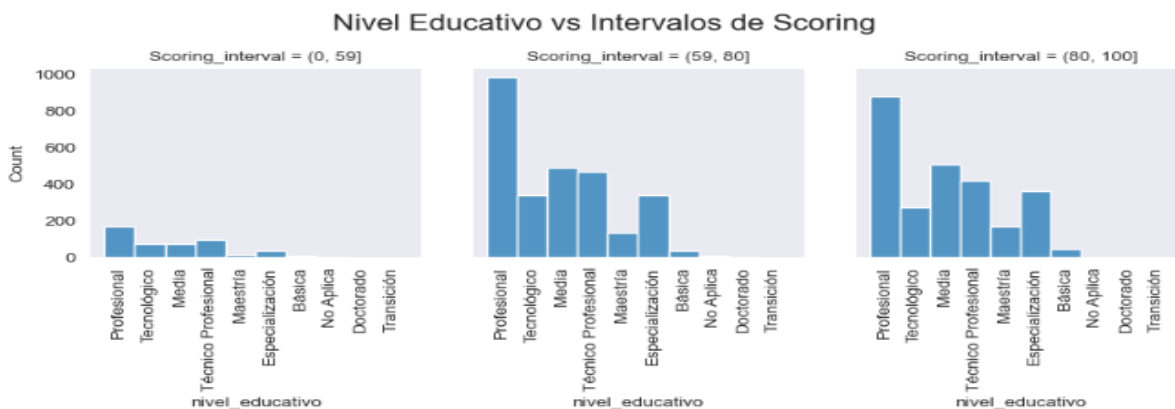


Fuente: Cooperativa – Elaboración Propia

Es importante resaltar que existe un sesgo hacia los contratos indefinidos, lo que crea un precedente inicial: el tipo de Scoring desarrollado en este ejercicio académico será aplicable específicamente al contexto de la cooperativa, limitando su generalización a otros entornos. Este sesgo constituye un reto para la adopción de algoritmos que puedan aprender y adaptarse a este contexto particular.

Ilustración 3. Nivel educativo vs Scoring

De acuerdo con la ilustración, se observa que en los rangos con score mayor a 60 existe mayor presencia de asociados con nivel educativo superior, entre profesional, maestría y doctorado frente al intervalo menor a 60.

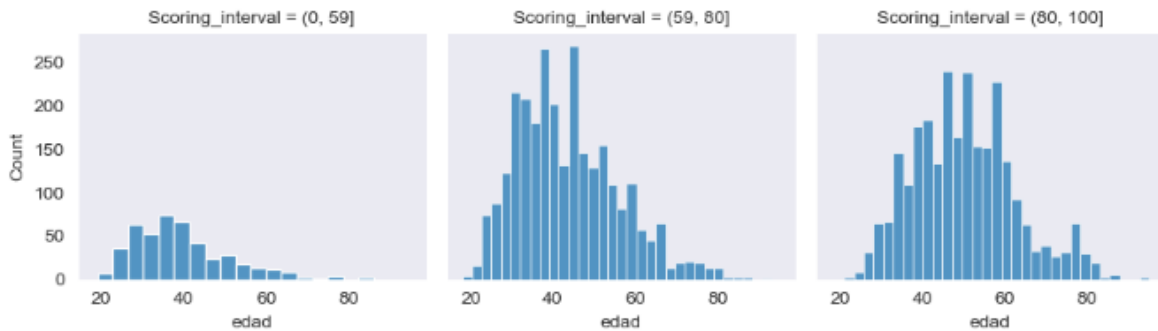


Fuente: Cooperativa – Elaboración Propia

Ilustración 4. Edad del asociado vs Scoring

En el intervalo de score menor a 60, se evidencia que existe una población con edad entre los 25 a 50 años, mientras que para el score con rangos mayores a 60 existe mayor presencia de asociados con edades mayores a 60 años.

Edad del Asociado vs Intervalos de Scoring

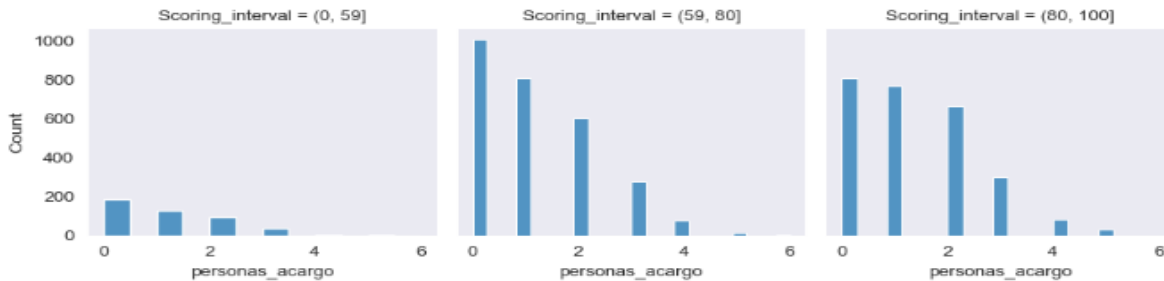


Fuente: Cooperativa – Elaboración Propia

Ilustración 5. Personas a cargo vs Scoring

Para un puntaje de score superior a 60, existe una mayor población de asociados con mayor cantidad de personas a cargo.

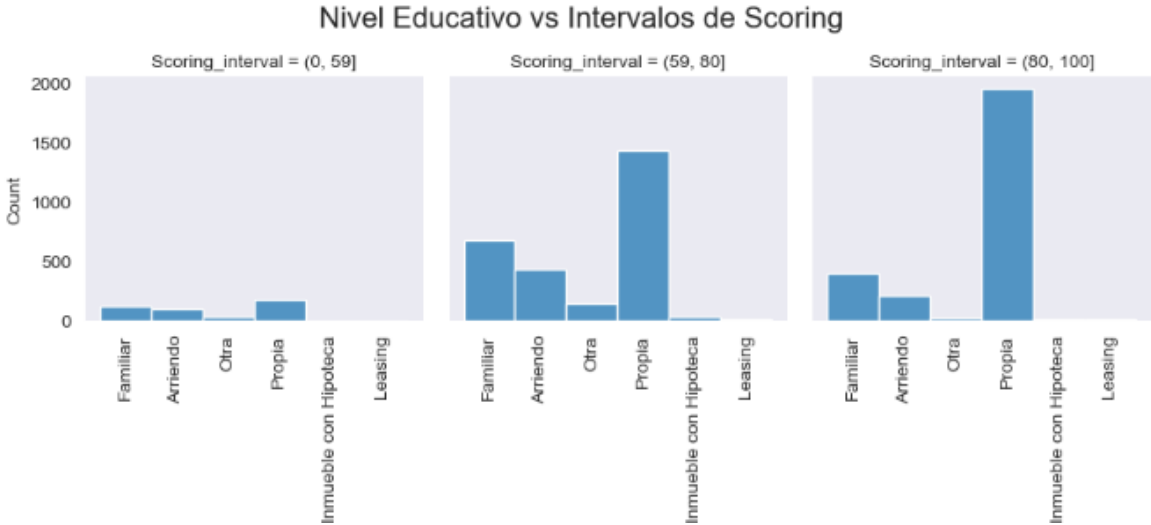
Personas a cargo vs Intervalos de Scoring



Fuente: Cooperativa – Elaboración Propia

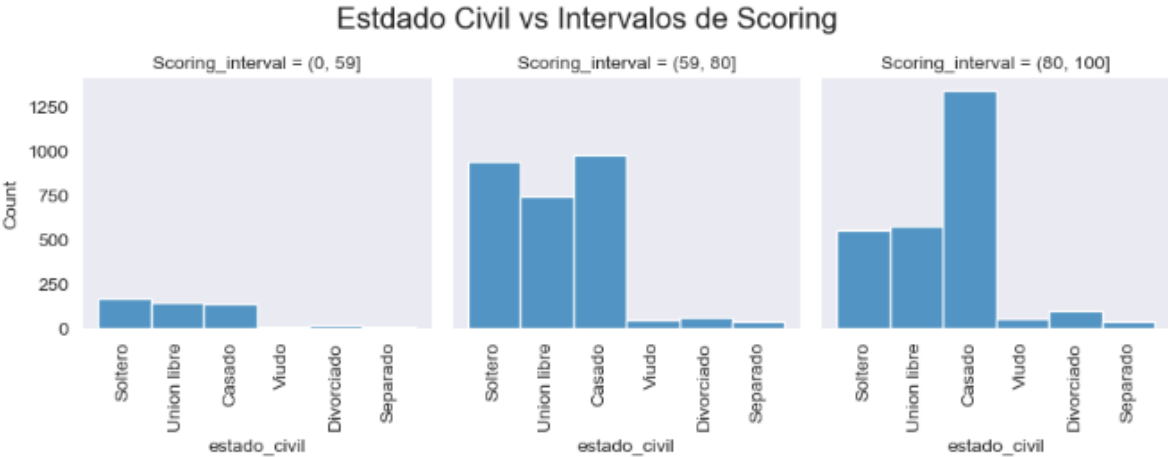
Ilustración 6. Tipo de vivienda vs Scoring

La mayoría de los asociados, con vivienda propia se encuentran en el puntaje de score mayor a 60.



Fuente: Cooperativa – Elaboración Propia

Ilustración 7. Estado civil vs Scoring



Fuente: Cooperativa – Elaboración Propia

En conclusión, el EDA para el caso de estudio refleja el punto de inicio en la identificación de relaciones y patrones entre las distintas variables relacionadas y el score de crédito. La presencia de sesgo, como la predominancia de contratos indefinidos, debe ser considerada en las fases posteriores de implementación de los algoritmos para clasificación. Este tipo de hallazgos no solo ayudan a entender mejor el comportamiento del score dentro de la cooperativa sino que constituye la base para la implementación de los modelos a usar, por lo cual en adelante será crucial comprender estas relaciones y el sesgo para encontrar modelos que generalicen bajo este contexto.

6. SELECCIÓN DE VARIABLES

En los modelos de clasificación es importante la selección de aquellas variables de mayor trascendencia, lo anterior mejora la precisión del modelo y aporta a mejorar la eficiencia en la máquina, en este ejercicio se llevó a cabo dos fases para la selección y la eliminación de variables, el primero tiene que ver con el descarte basado en el entendimiento del contexto o el negocio que a criterio experto se puede tener, y el segundo por medio del análisis de la importancia de variables utilizando el método Lasso.

6.1 Selección de variables por conocimiento del negocio

Se utilizó el coeficiente de correlación entre las variables independientes y la variable objetivo para identificar aquellas con mayor relevancia en el modelo. Además, se aplicó el criterio experto, asignando una puntuación que evaluó la importancia de cada variable en relación con el objetivo, lo que permitió una selección fundamentada tanto desde un enfoque estadístico como desde la perspectiva del negocio.

Adicionalmente, se revisaron los campos relacionados directamente con los datos personales de los asociados, tales como teléfono, correo electrónico, número de identificación, fecha de nacimiento, tipo de documento, dirección de correspondencia y ciudad de correspondencia, asegurando un manejo adecuado de esta información sensible.

También se eliminaron variables relacionadas con la descripción del crédito otorgado, ya que estas comprenden características meramente informativas y no aportan valor predictivo directo al modelo. Entre las variables descartadas se encuentran fechas específicas del proceso crediticio, como Fecha. Aprobación, Fecha. Préstamo, Fecha. Liquidación, Fecha. Inicio, Fecha. Vcto Final, Fecha. Vencto. y Fecha. Histórico. Asimismo, se excluyeron variables relacionadas con

calificaciones previas, como Calificación. Antes Ley Arras., Calificación. Aplicada y Calificación. Mes Ante., así como el campo Pagaré.

6.2 Análisis de importancia de variables con Lasso

Continuando con el paso anterior, para las demás variables se aplicó la regresión Lasso (*Least Absolute Shrinkage and Selection Operator*), una técnica de análisis de regresión que combina la selección de variables y la regularización para mejorar tanto la precisión predictiva como la interpretabilidad de los modelos estadísticos (Hastie, Tibshirani, & Friedman, 2009). Este enfoque no solo contribuye a simplificar el modelo, sino que también mejora su interpretación y capacidad de generalización en los datos. Tras aplicar este método, el conjunto de datos pasó de 80 características iniciales a 55, identificando 25 variables como no relevantes para el análisis

7. RESULTADOS

La selección de variables basada en el conocimiento del negocio y en el coeficiente de correlación de las variables, permitió identificar aquellas variables más relevantes para el modelo que se pretende construir, aquí el criterio experto fue fundamental donde se evaluó la importancia de cada una mediante una puntuación donde se reflejó su importancia y su contribución potencial al desempeño del modelo. Como resultado se seleccionaron variables clave como; el nivel educativo, el tipo de contrato, la edad, el estado civil, la cantidad de personas a cargo y el tipo de vivienda, que mostraron una correlación significativa con el score de crédito, esto hace que la optimización de la selección no guarde solo un componente estadístico sino que además se aplicable al contexto de la cooperativa.

El análisis de estas variables mostró relaciones importantes. Por ejemplo, los asociados con niveles educativos superiores tendieron a tener scores de crédito más altos, al igual que aquellos con contrato indefinido y vivienda propia. Así mismo se observó que la edad y el estado civil tienen relación con el umbral en el que se puntuó el score de crédito. Estos resultados son las primeras aproximaciones del modelo que se pretende construir y permiten tener una clara comprensión de los factores que limitan la capacidad y disposición al pago de los asociados. Los resultados a futuro contemplan la implementación del modelo la evaluación del mismo y elección del mejor modelo para el contexto de la cooperativa.

Durante el desarrollo del proyecto, se implementaron y evaluaron diferentes modelos de clasificación para predecir la variable objetivo "SCORING_CAT", los modelos probados son; regresión logística, linear support vector classifier (Linear

SVC), arboles de decisión y random forest. Cada modelo fue seleccionado y configurado cuidadosamente, considerando tanto su rendimiento en métricas como su aplicación en un entorno productivo.

El modelo de regresión logística destacó como la opción final para el proyecto debido a su simplicidad, facilidad de implementación y desempeño. Este modelo utiliza una función lineal combinada con probabilidades logísticas para clasificar instancias en las categorías bajo, medio y alto. Adicionalmente, la regularización L2 se aplicó para evitar problemas de sobreajuste, y el balanceo automático de clases permitió abordar la desproporción en la distribución de las categorías. Por lo cual, dichas características hacen que la regresión logística sea una herramienta robusta y fácilmente interpretable, ideal para el uso en el contexto productivo.

Además de la regresión logística, se probaron otros modelos como el LinearSVC, que se fundamenta en la optimización de márgenes para crear un clasificador lineal. Aunque mostró un rendimiento adecuado su complejidad y menor interpretabilidad lo hicieron menos adecuado para el objetivo del proyecto. Los arboles de decisión y el Random Forest también se evaluaron debido a su capacidad para modelar relaciones no lineales. No obstante, el Random Forest aunque muy preciso, resultó más complejo y menos interpretable y los arboles de decisión mostraron propensión al sobre ajuste, lo que disminuyó la utilidad para este caso.

Para evaluar el rendimiento de los modelos, se utilizaron métricas como la precisión, recall, el F1-score y la matriz de confusión. El modelo de regresión logística presentó un desempeño sobresaliente, logrando una precisión global del 98,89% y un F1-score de 99% en el conjunto de prueba. La matriz de confusión mostró que el modelo tiene un buen balance entre precisión y recall en las tres categorías, por ejemplo de las 399 observaciones para el nivel de score alto 397 se clasificaron correctamente y dos quedaron como medio, y en el score bajo de 177 observaciones 167 fueron clasificadas correctamente y 9 quedaron como medo. Lo anterior, confirma la capacidad para diferenciar de manera efectiva los niveles de riesgo crediticio.

Tabla 1: Comparación del rendimiento de modelos de aprendizaje automático en la clasificación del score de crédito.

Modelo	Accuracy	Precisión	Recall	F1-Score
Linear SVC	0.99	0.99	0.99	0.99
Regresión Logística	0.99	0.99	0.99	0.99
Árbol de Decisión	1.00	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00	1.00

Fuente: Datos Cooperativa – Elaboración propia

Como parte del proceso de evaluación y definición del modelo final, se realizó validación cruzada con el objetivo de medir de manera más robusta el desempeño de los modelos y reducir el riesgo de sobreajuste. Esta técnica permitió evidenciar como los modelos generalizan y proporciono métricas importantes al promediar resultados a partir de las particiones de los datos.

Se utilizó validación cruzada estratificada para garantizar que cada partición mantuviera la proporción original de las clases de la variable objetivo, se configuró el número de particiones en K=5 dividiendo los datos en cinco subconjuntos (folds). Para cada iteración: cuatro de los subconjuntos se usaron para entrenar el modelo, el quinto subconjunto se utilizó para validar el modelo, este proceso se repitió cinco veces, asegurando que cada subconjunto se usara como conjunto de validación exactamente una vez. Las métricas finales se calcularon como el promedio de las métricas obtenidas en cada iteración, proporcionando una visión más equilibrada del desempeño.

Tabla 2: Comparación del rendimiento de modelos de aprendizaje automático en la clasificación del score de crédito con validación cruzada.

Modelo	Accuracy	Precisión	Recall	F1-Score
Linear SVC	0.98	0.98	0.98	0.98
Regresión Logística	0.98	0.98	0.98	0.98
Árbol de Decisión	0.97	0.97	0.97	0.97
Random Forest	0.99	0.99	0.99	0.99

Fuente: Datos Cooperativa – Elaboración propia

Por último, la elección de la regresión logística como modelo final respondió a un equilibrio entre rendimiento, simplicidad y aplicabilidad en el entorno de la cooperativa. Aunque modelos como Random Forest, Linear SVC mostraron métricas de desempeño comparables, la regresión logística destacó por su facilidad de interpretación, lo que permite a los responsables de la toma de decisiones comprender claramente la influencia de cada variable en el score crediticio, además su menor complejidad computacional facilita la implementación en los sistemas actuales de la cooperativa.

8. CONCLUSIÓN

El desarrollo del modelo de score basado en metodologías de machine learning ha demostrado ser una estrategia efectiva para el desarrollo de modelos que permitan y aseguren una correcta evaluación del perfil de riesgo crediticio, Este enfoque no solo mejora la precisión de las decisiones crediticias, sino que también reduce la subjetividad inherente al uso exclusivo del criterio experto. Esto contribuye a mejorar la eficiencia operativa y la transparencia en la gestión crediticia dentro de la cooperativa.

La metodología empleada para la selección de las características combinó criterios estadísticos, como el método Lasso, y el conocimiento del negocio, lo que permitió identificar factores clave como el nivel educativo, el tipo de contrato y el estado civil de los asociados. Estas variables, se encontraron estrechamente relacionadas con el score de crédito y proporcionan una base sólida para garantizar la relevancia y aplicabilidad del modelo en el contexto específico de la cooperativa

Entre los modelos evaluados, la regresión logística sobresalió por su simplicidad, interpretabilidad y destacado desempeño, alcanzando métricas relevantes como una precisión global del 98.89% y un F1-Score del 99%. Estas cualidades la convierten en una opción ideal para su implementación en un entorno productivo. Sin embargo, el despliegue de este modelo se contempla como una etapa futura, fuera del alcance principal del presente estudio.

Por otra parte, el uso de la metodología CRISP-DM proporcionó un marco estructurado que guió el desarrollo del proyecto desde la comprensión inicial del negocio hasta la evaluación final del modelo. Este enfoque asegura no solo la robustez del proceso, sino también su replicabilidad para futuras aplicaciones en otros contextos del sector cooperativo.

Finalmente, este proyecto establece un precedente para la incorporación de herramientas analíticas avanzadas en el ámbito cooperativo, contribuyendo a la sostenibilidad financiera y a la innovación del sector. No obstante, el modelo enfrenta desafíos en términos de generalización, dado los sesgos específicos de los datos utilizados, por lo cual se hace relevante la adopción de futuros estudios orientados a ampliar su capacidad de generalización en otros entornos

9. REFERENCIAS

- Arango, L., & Restrepo, D. (2017). Diseño de un modelo de scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento colombiana. *Universidad Eafit*, 46-74.
- Becerra Ladino, C. I. (2014). Comparación de modelos de riesgo de crédito, Modelos Logísticos y Redes Neuronales. *Universidad Javeriana*, 9-24.
- Borrero Tigreros, D., & Bedoya Leiva, O. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. *Revista UIS Ingenierías*, 16.
- Cano Bedoya, J. (2021). 4.1.2. Aprendizaje supervisado en la construcción de un modelo de Credit Scoring paracooperativas de ahorro y crédito en Colombia. *Universidad Nacional*, 54-113.
- Chirarroni, H. (2002). La regresión Logística. *Area Empleo y Población*, 18.
- Christopher D. Manning, P. R. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University.
- Fieldman, R., & Schmidt, M. (2016). *El análisis del crédito: Evaluación y gestión del riesgo*. Mexico: Finacredit.
- García, J. C. (2007). Nuevas técnicas de medición del riesgo de crédito. *Revista económica financiera BBVA*, 29.
- Gonzalez Mora, V. (2023). 4.1.5. Método para evaluar el scoring de crédito de la línea de libranzas en las cooperativas de crédito de Medellín. *Universidad Nacional*, 30-60.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Estados Unidos: MIT Press.
- Gutierrez Girault, M. A., & Gutierrez Girault, M. A. (2007). Credit scoring models: what, how, when. *Banco Central de la República Argentina*, 1-5.
- Hand, J., & Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 523-541.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- International Cooperative Alliance. (2024). *International Cooperative Alliance*. Obtenido de <https://ica.coop/>
- Lecuona, C. M. (2021). *Machine Learning en su Búsqueda de la Responsabilidad Ética*.
- Marina Sokolova, G. L. (2009). *A systematic analysis of performance measures for classification tasks*. *Information Processing & Management*.
- Morgan, J., & Jerome, S. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 415-434.

- Muñoz, M. A. (31 de 07 de 2019). *Blogs.sas.com*. Obtenido de <https://blogs.sas.com/content/sasla/2019/07/31/la-inteligencia-artificial-y-su-impacto-en-el-scoring-credificio/>
- Naciones Unidas. (2023). *Naciones Unidas*. Obtenido de <https://www.un.org/es/observances/cooperatives-day>
- Oscar, O. (26 de 09 de 2024). *Dataprotected*. Obtenido de <https://dataprotected.com.co/blog-proteccion-de-datos/datos-personales/como-afecta-la-ley-1581-de-2012-la-gestion-de-datos-en-las-empresas/>
- Powers, D. (11 de 10 de 2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies*. Obtenido de Cornell University: <https://arxiv.org/abs/2010.16061>
- Rodriguez, M. d., & Ulfe, H. (2015). 4.1.3. Modelo de credit scoring para predecir el otorgamiento de crédito personal en una cooperativa de ahorro y crédito. *Universidad Nacional Pedro Ruiz Gallo*, 37-102.
- Tensén Aroyo, A. (2017). Eficiencia en los modelos de aprendizaje de maquina para evaluar el riesgo crediticio de personas naturales en una institución financiera de ciclayo. *Universidad Nacional del Santa*, 103.
- Vapnik, V., & Cortes, C. (1995). Support-Vector Networks. *Machine Learning*, 273-297.

ANEXOS

Nombre del Campo	Descripción
SUCURSAL	Código de la Sucursal a la que pertenece el Asociado (Cali, Medellín, Bogotá o Barranquilla)
ESTADO	Estado Activo, Sancionado o No Activo del Asociado Medio de pago: Nombre de la pagaduría asociada a la obligación
EDAD	
TIPO_DE_ASOCIADO	Descripción del Asociado, Empleado actual, Ex-Empleado, Pensionado
SUELDO	Salario Mensual
COD_DEPEND	Código de la dependencia o empresa asociada a Carvajal donde labora el asociado
SEXO	Masculino / Femenino

ANTIGUEDAD	Número de meses desde la última afiliación del asociado
ESTADO_CIVIL	Soltero, Casado, Unión Libre
NRO_HIJOS	Cantidad de Hijos
PERSONAS_ACARGO	Número de personas a cargo
ESTRATO	Estrato Socio-Económico
NIVEL_EDUCATIVO	Bachiller, Profesional, Posgrado
VALOR_INGRESO_MENSUAL	Valor del Ingreso Mensual
VALOR_EGRESO_MENSUAL	Valor de Egreso Mensual
DESC_TIPO_VIVIENDA	Familiar, Propia, Alquilada, Hipotecada
AHORRO_PERMANENTE	Saldo del ahorro permanente
SLD_APORTES	Saldo de los aportes
MONTO_SOLICITADO	Monto Solicitado del Crédito
SALDO_CAPITAL	Saldo Capital al corte
SLD_INT	Saldo Interés
SLD_MORA	Saldo en mora
DIAS_VENCIDOS	Saldo vencido capital
VENCCAPITAL	Valor vencido de capital adeudado
VLR_GARANTIA	Valor de la Garantía
VLRCOBERTURA_DISPONIBLE	Cobertura de la Garantía
VLR_CUOTA	Valor Cuota Mensual
NOCUOTAS	Numero de Cuotas pactadas
ALTURA	Altura de Mora
INT_CTE_ORDEN	Intereses de cuentas de orden
INT_MORA_ORDEN	Intereses Mora
SLD_SEG_VIDA	Valor Seguro de Vida
SUCCREDITO	Sucursal donde se realizó desembolso
CANTIDAD_GARANTÍAS	Numero de Garantías
VLRAPLICADO_GARANT	Valor aplicado a la garantía
CODEUDORES	Numero de codeudores
GARANTÍA_REAL	Valor garantía real
DIAS_VENCIDOS_INT	Días Vencidos Interés
VENCIDO_INT	Vencidos Interés
DIAS_VENCIDOS_CAPITAL	Días Vencidos Capital
VALOR_PERDIDA_ESPERADA_APLICADA	Valor Perdida Esperada
VALOR_COMERCIAL_ACTIVOS	Valor Activos
VALOR_SALDO_PASIVOS	Valor Pasivos
SCORING	Puntaje de Score interno
ANTIGÜEDAD_AÑOS	Antigüedad como asociado en años
CANAL	Canal de pago de la obligación
SCORING_CAT	Variable Objetivo Alto-Medio-Bajo