Aplicación de técnicas de inteligencia artificial para predecir mora en créditos de una institución financiera en Ecuador

Application of Artificial Intelligence techniques to predict loan defaults in a financial institution in Ecuador

Aplicação de técnicas de Inteligência Artificial para prever a moratória em créditos de uma instituição financeira no Equador

• Libia Johana Sánchez-Espinoza[1]

Fecha de recepción: 03 de junio de 2025
Fecha de aceptación: 07 de agosto de 2025

Doi: 10.33210/ca.v14i1.509
Cienciamérica (2025) | Vol. 14 N°. 2 | pp. 1-19
ISSN 1390-9592 **ISSN**-L 1390-681X

[1]Facultad de Ciencias y Tecnología. Universidad Internacional Isabel I. Ambato-Ecuador. Correo: ingenieriainformaticals@gmail.com

***Cómo citar: L. J. Sánchez-Espinoza, "Aplicación de técnicas de inteligencia artificial para predecir mora en créditos de una institución financiera en Ecuador", CienciAmérica, vol. 14, no. 2, pp. 1–19, Ago. 2025, doi: 10.33210/ca.v14i1.509.**

Resumen

INTRODUCCIÓN. Esta investigación aborda la predicción de morosidad en el pago de créditos mediante la implementación de modelos basados en técnicas de Inteligencia Artificial, específicamente Aprendizaje Automático. La variable dependiente es la morosidad, y las independientes incluyen características demográficas, socioeconómicas y del historial crediticio. **OBJETIVO.** Implementar y entrenar modelos predictivos utilizando técnicas de aprendizaje automático supervisado, con el propósito de anticipar posibles moras en créditos y apoyar la toma de decisiones. **MÉTODO.** Se aplicaron las etapas de la metodología CRISP-DM, iniciando con la extracción, transformación y carga de los datos, seguido de análisis exploratorio, limpieza, verificación de correlaciones, entrenamiento de algoritmos supervisados y evaluación del rendimiento. **RESULTADOS.** El mayor índice de recall 0,68, indicador clave para identificar impagos, se obtuvo con el algoritmo de Regresión Logística utilizando la técnica de balanceo SMOTE. **DISCUSIÓN Y CONCLUSIONES.** El resultado contrasta con otras investigaciones que adoptan el modelo Random Forest en problemas de predicción de mora, en este caso los valores de recall obtenidos no fueron significativos. Una limitación importante fue el desbalance en la variable a predecir, abordado mediante técnicas de balanceo. Finalmente se evidencia la importancia de validar empíricamente los resultados según los datos y el contexto específico de aplicación.

Palabras Clave

Predecir, mora, inteligencia artificial, aprendizaje automático, regresión logística.

Abstract

INTRODUCTION. This research addresses the prediction of credit default through the implementation of models based on Artificial Intelligence techniques, specifically Machine Learning. The dependent variable is default, and the independent variables include demographic, socioeconomic, and credit history characteristics. **OBJECTIVE.** Implement and train predictive models using supervised machine learning techniques, with the aim of anticipating possible loan defaults and supporting decision-making. **METHOD.** The stages of the CRISP-DM methodology were applied, starting with data extraction, transformation, and loading, followed by exploratory analysis, cleaning, correlation verification, supervised algorithm training, and performance evaluation. **RESULTS.** The highest recall rate of 0.68, a key indicator for identifying defaults, was obtained with the Logistic Regression algorithm using the SMOTE balancing technique. **DISCUSSION AND CONCLUSIONS.** The result contrasts with other studies that adopt the Random Forest model in default prediction problems, in which case the recall values obtained were not significant. An important limitation was the imbalance in the variable to be predicted, which was addressed using balancing techniques. Finally, the importance of empirically validating the results according to the data and the specific context of application is evident.

Keywords

Predict, default, artificial intelligence, machine learning, logistic regression.

Resumo

INTRODUÇÃO. Esta investigação aborda a previsão de incumprimento no pagamento de créditos através da implementação de modelos baseados em técnicas de Inteligência Artificial, especificamente Aprendizagem Automática. A variável dependente é o incumprimento, e as independentes incluem características demográficas, socioeconómicas e do histórico de crédito. **OBJETIVO.** Implementar e treinar modelos preditivos utilizando técnicas de aprendizagem automática supervisionada, com o objetivo de antecipar possíveis atrasos nos pagamentos de créditos e apoiar a tomada de decisões. **MÉTODO.** Foram aplicadas as etapas da metodologia CRISP-DM, começando com a extração, transformação e carregamento dos dados, seguidas de análise exploratória, limpeza, verificação de correlações, treino de algoritmos supervisionados e avaliação do desempenho. **RESULTADOS.** O maior índice de recall, 0,68, indicador-chave para identificar inadimplências, foi obtido com o algoritmo de regressão logística utilizando a técnica de balanceamento SMOTE. **DISCUSSÃO E CONCLUSÕES.** O resultado contrasta com outras investigações que adotam o modelo Random Forest em problemas de previsão de inadimplência, neste caso os valores de recall obtidos não foram significativos. Uma limitação importante foi o desequilíbrio na variável a ser prevista, abordado por meio de técnicas de equilíbrio. Finalmente, fica evidente a importância de validar empiricamente os resultados de acordo com os dados e o contexto específico de aplicação.

Palavras-chave

Prever, mora, inteligência artificial, aprendizagem automática, regressão logística.

Introducción

La Cuarta Revolución Industrial es conocida por la era de la digitalización o Industria 4.0 y se hacen referencia con términos como tecnologías disruptivas, tecnologías emergentes y tecnologías habilitadoras [1]. El término más escuchado en los últimos años es Tecnologías Disruptivas y son aquellas que convierten en obsoleta una tecnología existente, no representan pequeñas mejoras sobre algo existente, sino que introducen algo completamente nuevo, dejando lo anterior obsoleto o ineficiente.

Dentro de las tecnologías disruptivas de la Industria 4.0 destaca la Inteligencia Artificial IA, que incluye sistemas capaces de imitar procesos de la inteligencia humana como el reconocimiento de voz, la toma de decisiones y el aprendizaje. Un subconjunto de la IA es el Aprendizaje Automático (Machine Learning ML) que emplea algoritmos basados en datos para tareas de predicción, clasificación y generación de conocimiento. A su vez, el Aprendizaje Profundo Deep Learning es una rama del ML que utiliza redes neuronales profundas para simular el comportamiento del cerebro humano [1].

La Inteligencia Artificial marca una nueva etapa en la transformación digital en todos los sectores económicos y con mayor razón en el sector financiero [2]. La aceleración de la transformación digital se hizo más evidente a raíz de la pandemia mundial COVID-19 y se ha integrado de manera progresiva en la cultura organizacional de las instituciones financieras con la finalidad de ser más innovadoras en la gestión de la tecnología, adoptando un enfoque estratégico [3].

A diferencia de la programación tradicional, el Aprendizaje Automático genera modelos a partir de datos de entrada y salida, descubriendo patrones que permiten hacer predicciones [4]. Un algoritmo de aprendizaje no es exclusivo para dar respuesta a tareas específicas ya que un mismo procedimiento de aprendizaje puede implementarse para resolver diferentes tareas, considerando siempre que se disponga de datos debidamente etiquetados y específicos de cada tarea [5].

Los algoritmos de Aprendizaje Automático pueden ser clasificados en supervisados, no supervisados y reforzados. La principal diferencia entre estos tres tipos de algoritmos es la presencia o ausencia de una variable conocida como de resultado, dependiente o explicada [6].

Los algoritmos de Aprendizaje Automático Supervisado se utilizan para tareas de predicción. Una vez entrenados, pueden predecir resultados sobre nuevos datos [6]. Un aspecto clave del Aprendizaje Supervisado es que los datos disponibles incluyen observaciones de la variable objetivo y aportan valor a la variable que se desea explicar o predecir [5].

El objetivo principal de esta investigación es implementar modelos basados en técnicas de Inteligencia Artificial con el fin de predecir la existencia de mora en el pago de créditos de los clientes de una institución financiera en Ecuador. Para cumplir con este propósito, se plantea: revisar la aplicación de la Inteligencia Artificial en el sector financiero, preprocesar los datos históricos de los clientes, realizar un análisis exploratorio de los datos (EDA), desarrollar e implementar modelos de Aprendizaje Automático supervisado y evaluar la precisión y efectividad de los modelos entrenados. La incorporación de esta técnica de predicción en los procesos crediticios permite sustentar decisiones estratégicas basadas en evidencia, lo que contribuye a la mitigación de riesgos financieros y al fortalecimiento de una gestión eficiente y orientada a la sostenibilidad institucional.

Estado del Arte

La banca de Reino unido, Japón, Alemania entre otras, trabajan para minimizar los riesgos como ciberataques, fraudes o impagos en créditos implementando algoritmos de Aprendizaje Automático. Estas tecnologías permiten procesar grandes volúmenes de datos y analizar características del prestatario con el fin de identificar factores asociados al riesgo de impago, lo que contribuye a una toma de decisiones crediticias más precisa y a la reducción de pérdidas financieras [7].

Gimeno y Márquez [8], se refieren al uso de la IA en las finanzas como un nuevo paradigma y

consideran que su principal uso se enfoca en predecir una variable generalmente binaria. Las técnicas de la IA brindan la posibilidad de ir «aprendiendo» y corrigiendo errores en la predicción.

Investigadores de universidades de Egipto y Arabia Saudita, destacan la importancia de evaluar correctamente el otorgamiento de un crédito [9]. En su estudio, se implementan técnicas de Aprendizaje Automático para predecir el comportamiento de pago de los créditos. Empiezan por realizar el análisis de los datos, seguido del procesamiento, aplicación de modelos de Inteligencia Artificial (ML, DL), optimización del modelo y finalizan con la evaluación de los modelos. Las variables que integran el dataset son de tipo demográficas, laborales, socioeconómicas, geográficas y la variable a predecir es el indicador de riesgo. Los algoritmos de Aprendizaje Automático implementados fueron Gaussian NB, Ada Boost, Gradiant Boosting, Regresión logística, Random Forest, Árbol de decisión y K neighbors. La técnica de balanceo de datos utilizada fue SMOTE (Synthetic Minority Oversampling Technique) y SMOTE- Tomek (combinación de SMOTE y Tomek Links) que combinada con el algoritmo K neighbors retorno una precisión de predicción del 95% [9].

Kamil Dawid Grzebień [10], implementa técnicas de preprocesamiento de datos históricos de préstamos con variables relevantes como forma de pago, cantidad prestada, número de cuotas, interés, duración del empleo en años, ingresos anuales, estado de propiedad de vivienda, finalidad del préstamo y la variable a predecir riesgo de impago. Para balancear los datos utiliza técnicas de remuestreo como SMOTE-Tomek. Utiliza los algoritmos Regresión Logística, Árbol de decisión, Random Forest y Red Neuronal. En sus resultados mostraron que Random Forest tiene un mejor rendimiento en la predicción de impagos.

En una institución financiera de Ecuador Juan Freire López elabora un estudio que analiza la información de créditos entre el año 2017 y 2020, implementa el modelo Random Forest para clasificar a los clientes como "buenos" o "malos" pagadores. Compara los resultados de precisión de los algoritmos con otros modelos como Redes Neuronales, Árboles de Decisión, Máquinas de Soporte Vectorial y concluye que los resultados de Random Forest fueron los mejores para el conjunto de datos, obteniendo una precisión del 97,2% y una tasa de error del 2.8% [11].

En Colombia Diego Borrero y Oscar Bedoya [12] proponen técnicas de IA para identificar clientes que podrían incurrir en un estado de mora. Implementan los algoritmos de Aprendizaje Automático supervisado Redes Neuronales, Árboles de decisión y Máquinas de soporte vectorial. Las variables que analizan en el dataset son capital, género, educación, estado civil, historia de pago, importe de estado de cuenta, monto pagado y la variable a predecir es incumplimiento de cuota. Concluyen que el modelo con mayor índice de precisión es Random Forest con 72.72% y el que menos precisión tiene es Red Neuronal con un índice de precisión 43.58%.

Dado el enfoque de este artículo, se seleccionan cuatro modelos de aprendizaje supervisado: Regresión Logística, Árbol de Decisión, Random Forest y XGBoost (Extreme Gradient Boosting).

Regresión logística es un modelo estadístico utilizado para predecir variables dicotómicas (0 o 1), se basa en la regresión lineal, pero adaptado a variables dependientes binarias. Permite evaluar la relación entre esta variable y una o más variables independientes, sean cuantitativas o categóricas (estas últimas transformadas en variables Dummies) [13].

Árbol de decisión es un modelo que clasifica datos mediante una estructura en forma de árbol. Cada nodo representa una decisión basada en una característica, las ramas indican las posibles respuestas, y las hojas muestran el resultado o predicción. Se construye de forma descendente, dividiendo progresivamente el conjunto de datos en subconjuntos más específicos [10].

Random Forest es un algoritmo de ensamblado que combina múltiples árboles de decisión para mejorar la precisión y estabilidad de las predicciones. Cada árbol se entrena

con una muestra aleatoria de los datos mediante la técnica de bagging, lo que reduce la correlación entre árboles. La predicción final se obtiene mediante votación mayoritaria en clasificación o promedio en regresión [10].

XGBoost es una técnica de boosting que utiliza múltiples árboles de decisión como clasificadores, mejorando la velocidad y precisión respecto al algoritmo original Gradient Boosting. Incorpora un mecanismo de regularización que reduce el sobreajuste y mejora la capacidad de generalización del modelo. Además, permite construir modelos más simples y eficientes, utilizando menos variables [14].

Método

La metodología de trabajo implementada es CRISP-DM (CRoss-Industry Standard Process for Data mining) una guía de referencia altamente utilizada en proyectos de minería y ciencia de datos. Esta guía propone 6 fases iterativas [15] y se observan en la Tabla 1.

Tabla 1. Fases CRISP-DM

Núm.	Fases CRISP-DM	Descripción
1	Comprensión del negocio	Objetivos del negocio, evaluación de la situación.
2	Comprensión de los datos	Recolección de datos, descripción, exploración y verificación.
3	Preparación de los datos	Selección, limpieza, estructuración, integración, formateo.
4	Modelado	Selección de técnica de modelado, plan de prueba, entrenamiento y evaluación del modelo.
5	Evaluación	Evaluación de resultados, definición de métricas, proceso de revisión.
6	Implementación	Plan de implementación. Queda fuera del presente proyecto de investigación.

Fuente: [15].

Fase de comprensión del negocio / Muestra

La institución financiera del Ecuador brinda los servicios de ahorros, inversiones, créditos y pagos de servicios básicos. Actualmente se encuentra en el segmento 3 según la Superintendencia de Economía Popular y Solidaria de Ecuador, cuenta con más de 5000 socios y según sus balances sus mayores ingresos se generan por los intereses de cartera de créditos. Dispone de un historial centralizado que incluye datos demográficos de los clientes como edad, género, ubicación, nivel socioeconómico y registros detallados de créditos, incluyendo comportamiento de pago, montos, frecuencia, intereses, saldos y estado de mora. Actualmente esta información es considerada brevemente de forma manual por el personal de créditos y no se aplican técnicas de clasificación ni modelado con Aprendizaje Automático.

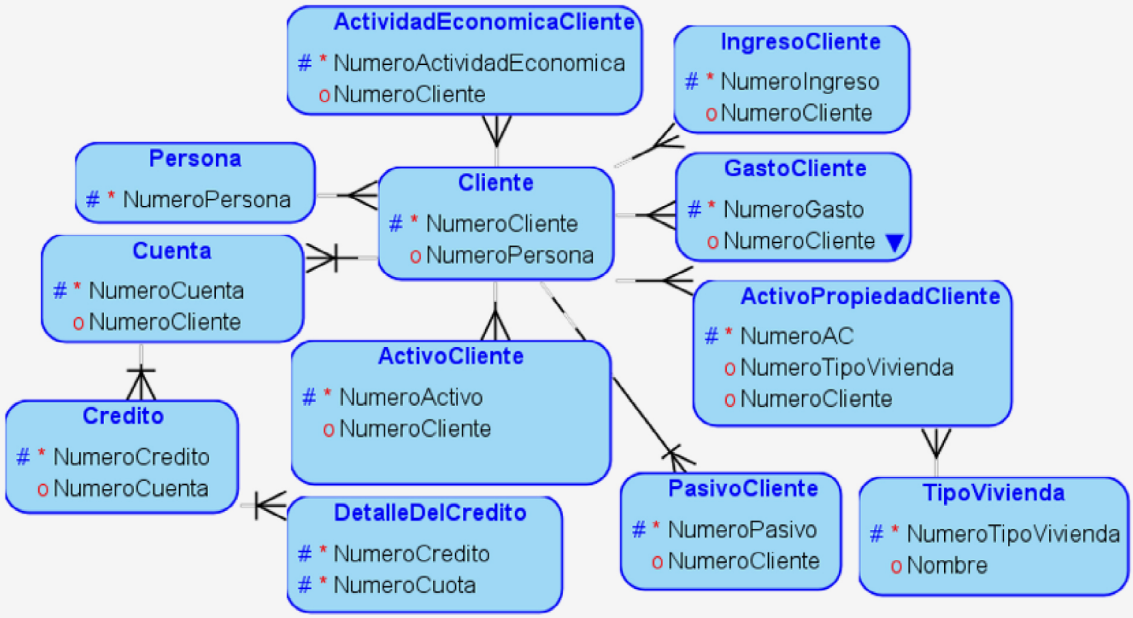
Fase de comprensión de los datos / Fuente de datos

Los datos se encuentran almacenados en una base de datos Microsoft SQL Server, que se encuentra operativa en modo On-premise. Se realiza un análisis de la estructura, contenido y distribución de los datos. En la Figura 1 se observa el esquema de la base de datos relacional con las tablas que contienen información relacionada a la investigación con claves primarias y foráneas.

Fase de preparación de los datos / Técnicas de análisis de datos

Por medio del Lenguaje SQL (Structured Query Language) se crea un script que extrae la información relevante de las tablas, se ejecutan cálculos, se combinan datos y se elimina información redundante. Se obtiene la información en un archivo en formato CSV (Comma Separated Values) con las siguientes

Figura 1. Esquema de la estructura de la información necesaria

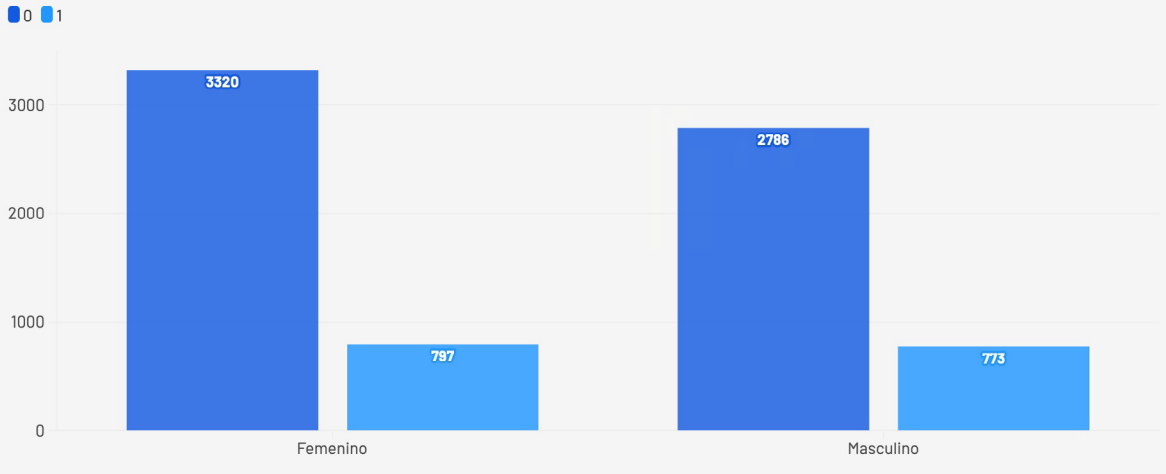


Fuente: Elaboración propia.

variables: "Sexo", "EstadoCivil", "FechaNacimiento", "Instrucción", "CodigoBarrio", "NumeroCargas", "TipoActividad", "NumeroCredito", "Monto", "NumeroCuotas", "ValorCuota", "FechaEntrega", "Tasa", "Calificacion", "encaje", "tipo", "vivienda", "totalActivo", "totalPasivo", "Ingresos", "Gastos", "MoraCredito".

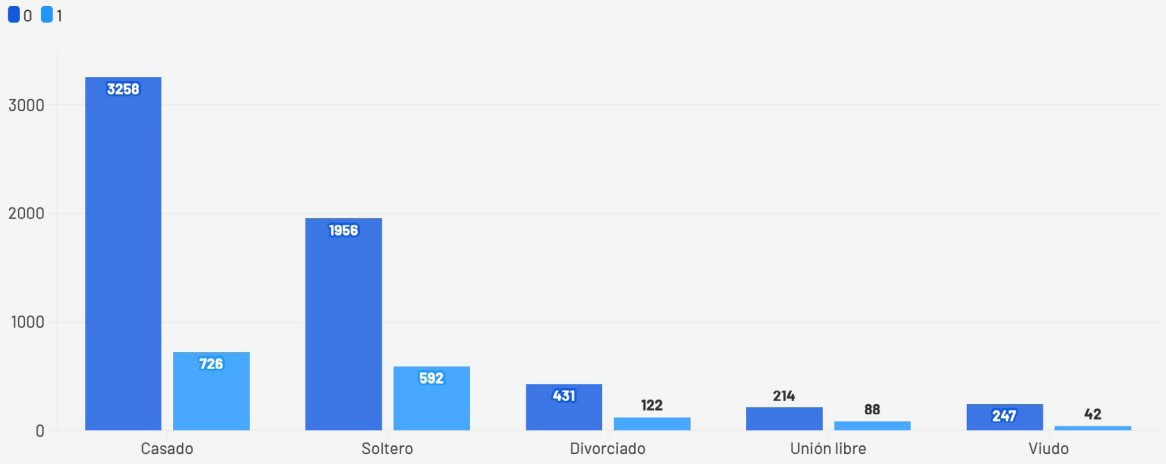
La variable "MoraCredito" es la variable objetivo o variable a predecir y contiene datos binarios donde 0 indica "Puntualidad sin mora" y 1 indica "Impuntualidad con mora". El análisis del archivo CSV se realiza mediante un script en Python utilizando librerías como pandas, Numpy, Matplotlib, Seaborn, Sklearn, Xgboost entre otras, especializadas en análisis de datos, para su edición y ejecución se utiliza el entorno de Google Colab. Se realiza el análisis exploratorio EDA (Exploratory Data Analysis), se visualiza la estructura del dataframe, se obtiene la estadística descriptiva de las columnas con datos numéricos, resumen estadístico de las variables categóricas y se crean los gráficos de barras categóricas para visualizar la distribución de las variables con respecto a la variable "MoraCredito".

Figura 2. Distribución "MoraCrédito" según "Sexo"



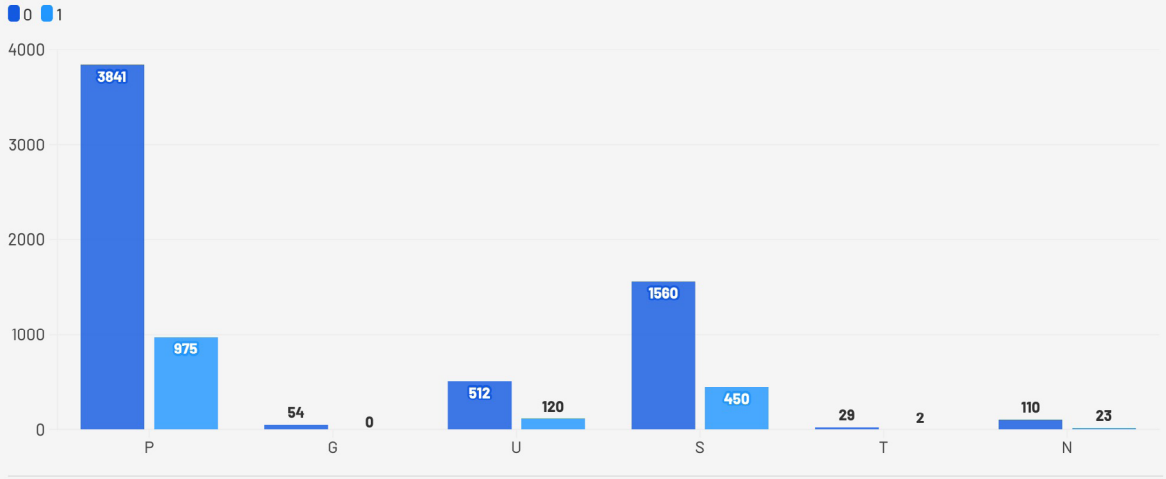
Fuente: Elaboración propia.

Figura 3. Distribución "MoraCrédito" según "EstadoCivil"



Fuente: Elaboración propia.

Figura 4. Distribución "MoraCrédito" según "Instrucción"



Fuente: Elaboración propia.

La Figura 2 muestra que las mujeres concentran un mayor volumen de créditos otorgados en comparación con los hombres. Sin embargo, la incidencia de morosidad presenta valores similares entre ambos grupos, lo que sugiere que, en términos proporcionales, la tasa de morosidad es relativamente más alta en la población masculina.

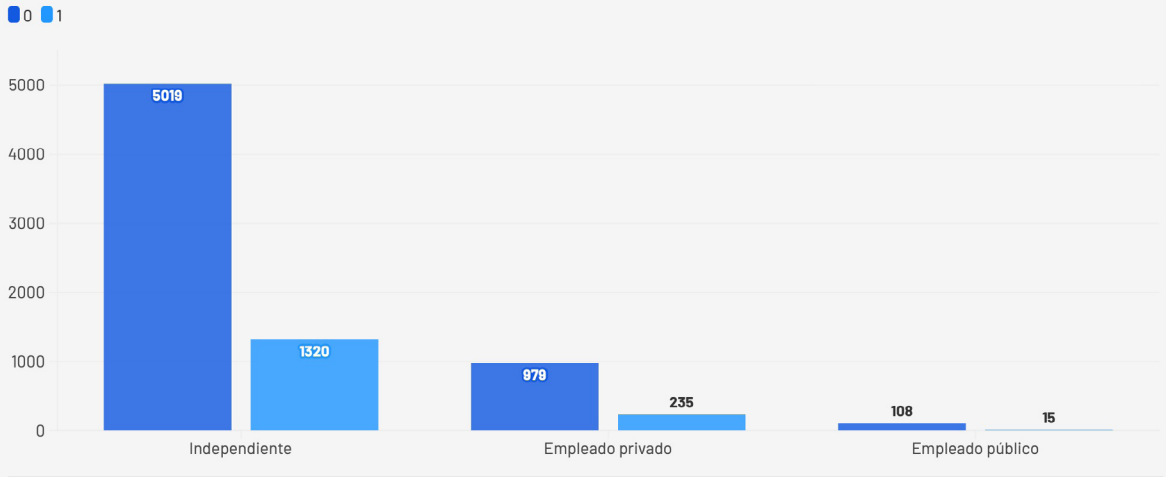
La Figura 3 evidencia que los clientes con estado civil casado concentran una mayor cantidad de créditos en comparación con los clientes solteros. Sin embargo, la incidencia de morosidad entre ambos grupos no presenta diferencias sustanciales, lo que sugiere que el estado civil no constituye un factor determinante en el comportamiento de pago.

La Figura 4 muestra que los niveles de morosidad son más elevados entre los clientes con instrucción secundaria (S) en comparación con aquellos que poseen únicamente instrucción primaria (P), lo que sugiere una posible relación entre el nivel educativo y el riesgo de incumplimiento crediticio.

En la Figura 5, se observa la distribución de "MoraCredito" con respecto a la actividad económica del cliente y se evidencia que cuando el cliente es empleado independiente la morosidad es mayor y si es empleado público la morosidad es menor.

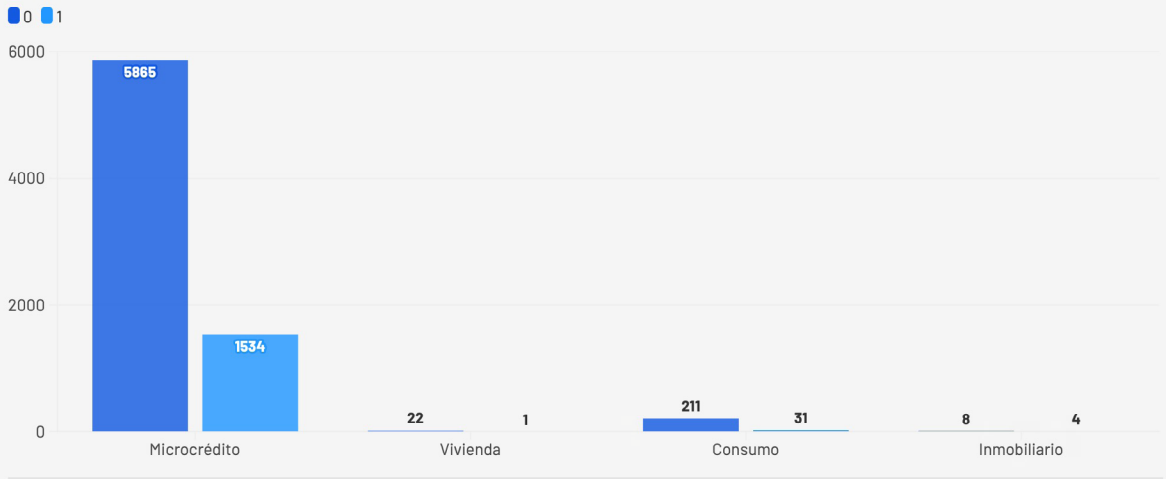
La Figura 6 presenta la distribución de la variable "MoraCredito" en función del destino del crédito. Se observa que, aunque los cré-

Figura 5. Distribución "MoraCrédito" según "TipoActividad"



Fuente: Elaboración propia.

Figura 6. Distribución "MoraCrédito" según "Calificación"



Fuente: Elaboración propia.

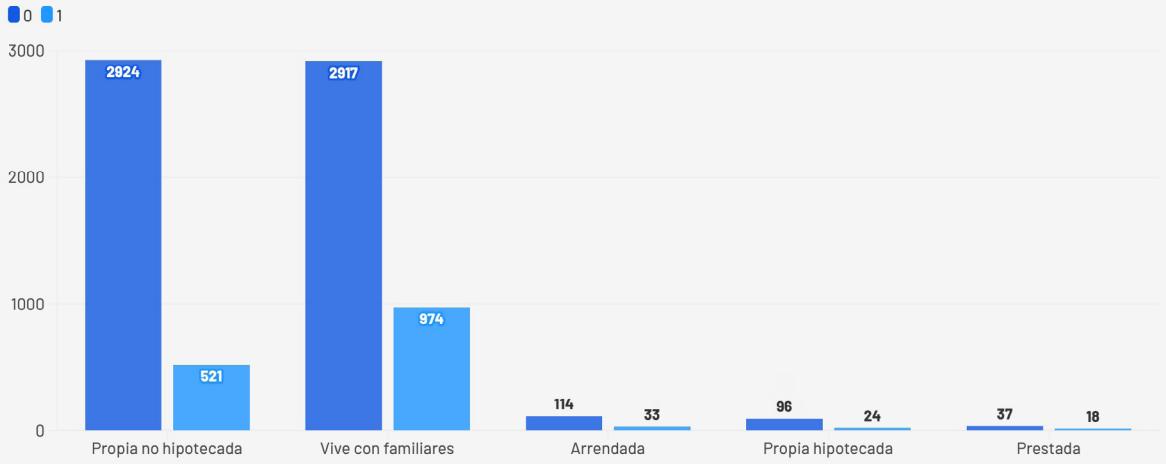
ditos destinados al sector inmobiliario representan una proporción reducida del total, registran una tasa de morosidad considerablemente superior en comparación con los microcréditos.

En la Figura 7, se visualiza que los clientes con vivienda propia no hipotecada tienen menor morosidad, mientras aquellos que viven con familiares o en condiciones menos estables como prestada o hipotecada, tienden a tener mayor riesgo de mora.

Se aplica el test Chi Cuadrado para evaluar la asociación con un nivel de significancia de 0.05, entre las variables y "MoraCredito", sus valores se observan en la Tabla 2.

Para la limpieza de datos y transformación se corrige valores redundantes, se tratan valores ausentes, se transforman las variables "Sexo", "EstadoCivil" y "vivienda" en variables dummies, se calcula la "EdadAlCredito", se crean grupos con las edades Juvenil, Adulto y Senior para facilitar su distribución, se define un mapeo para convertir la variable "Instrucción" de categórica a numérica, se convierte la variable "TipoActividad", "Calificacion" y "Tipo" en numérica con el método Label Encoding, se eliminan variables que no aportan valor como "NumeroCredito, se calcula la diferencia entre "Activos – Pasivos" y se identifica como "Patrimonio", se calcula la diferencia entre "Ingresos – Gastos" y se identifica como "Disponible".

Figura 7. Distribución "MoraCrédito" según "Vivienda"



Fuente: Elaboración propia.

Tabla 2. Test Chi Cuadrado para evaluar la asociación con "MoraCredito" con un nivel de significancia de 0.05

Variable	p-valor	Condición	Interpretación
"Sexo"	0.0114	0.0114 < 0.05	Se rechaza la hipótesis nula e indica que el género influye en la probabilidad de caer en mora.
"EstadoCivil"	3.28e-09	3.28e-09 < 0.05	Se rechaza la hipótesis nula e indica que el estado civil influye considerablemente en la probabilidad de caer en mora.
"Instruccion"	0.0002	0.0002 < 0.05	Se rechaza la hipótesis nula e indica que la instrucción influye significativamente en la probabilidad de caer en mora.
"Calificacion"	0.003	0.003 < 0.05	Se rechaza la hipótesis nula e indica que la calificación influye en la probabilidad de caer en mora.
"Tipo" (cuota fija o variable)	0.173	0.173 > 0.05	No se rechaza la hipótesis nula e indica que el tipo de crédito no afecta la probabilidad de caer en mora.
"Vivienda"	1.52e-24	1.52e-24 < 0.05	Se rechaza la hipótesis nula e indica que la calificación influye significativamente en la probabilidad de caer en mora.

Fuente: Elaboración propia.

Se calcula la matriz de correlación y se grafica el mapa de calor con el método de Pearson de todas las variables con respecto a "MoraCredito". Se elige este método considerando que se requiere observar la correlación de las variables explicativas cuantitativas con la variable "MoraCredito" que es dicotómica. Sus resultados muestran que la variable "NumeroCuotas" tiene la mayor correlación positiva con "MoraCredito" (0.139), sugiriendo que más cuotas están ligeramente asociadas con un mayor riesgo de mora. La variable "vivienda_PROPIA NO HIPOTECADA" muestra la correlación negativa más significativa (-0.122), indicando que tener una vivienda propia sin hipotecar está asociado con un menor riesgo de mora. Las correlaciones para variables como "Monto" y "encaje" son moderadas, mientras que las correlaciones negativas con "GrupoEdad", "Patrimonio" y "totalActivo" sugieren que personas mayores y aquellos con más activos tienen menor probabilidad de incurrir en mora. Las variables como "Ingresos", "Gastos" y "EstadoCivil" tienen correlaciones débiles con la mora. En la Figura 8 se observan las variables seleccionadas.

Figura 8. Información del DataFrame final con las variables definitivas

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Instruccion                     7676 non-null   int64
 1   NumeroCargas                    7676 non-null   int64
 2   TipoActividad                   7676 non-null   int64
 3   Monto                           7676 non-null   float64
 4   NumeroCuotas                    7676 non-null   int64
 5   ValorCuota                      7676 non-null   float64
 6   Tasa                            7676 non-null   float64
 7   Calificacion                    7676 non-null   int64
 8   encaje                          7676 non-null   float64
 9   tipo                            7676 non-null   int64
 10  totalActivo                     7676 non-null   float64
 11  totalPasivo                     7676 non-null   float64
 12  Ingresos                        7676 non-null   float64
 13  Gastos                          7676 non-null   float64
 14  MoraCredito                     7676 non-null   int64
 15  Sexo_Femenino                   7676 non-null   int64
 16  Sexo_Masculino                  7676 non-null   int64
 17  EstadoCivil_Casado              7676 non-null   int64
 18  EstadoCivil_Divorciado          7676 non-null   int64
 19  EstadoCivil_Soltero             7676 non-null   int64
 20  EstadoCivil_Union Libre         7676 non-null   int64
 21  EstadoCivil_Viudo               7676 non-null   int64
 22  GrupoEdad                       7676 non-null   int64
 23  vivienda_ARRENDADA              7676 non-null   int64
 24  vivienda_PRESTADA               7676 non-null   int64
 25  vivienda_PROPIA HIPOTECADA      7676 non-null   int64
 26  vivienda_PROPIA NO HIPOTECADA   7676 non-null   int64
 27  vivienda_VIVE CON FAMILIARES    7676 non-null   int64
 28  Patrimonio                      7676 non-null   float64
 29  Disponible                      7676 non-null   float64
dtypes: float64(10), int64(20)
memory usage: 1.8 MB
```

Fuente: Elaboración propia.

Fase de modelado / Procedimiento

Se cuenta con la variable explicada en formato binario "MoraCredito" y las variables explicativas en formato numérico. Se separa los datos un 30% para pruebas y el 70% de datos para entrenamiento y como es de esperarse existe un desbalance. En datos de entrenamiento se observa "MoraCredito" = 0 existen 4274 registros y "MoraCredito" = 1 existen 1099 registros.

```
# Se aplica sobremuestreo aleatorio sobre los datos de entrenamiento
ROS = RandomOverSampler(sampling_strategy = 0.8, random_state = 1)
balance_training_x, balance_training_y = ROS.fit_resample(x_val, y_val)
```

Para balancear los datos se aplica el sobre muestreo aleatorio ROS Random Over-Sampling sobre los datos de entrenamiento. La estrategia de esta técnica es aumentar la cantidad de la clase minoritaria hasta alcanzar el 80% de la cantidad de la clase mayoritaria.

```
########## Regresión Logística #############
# Instanciamos el modelo
modelo_regresion_log = LogisticRegression(random_state=42, max_iter=1000,
solver='liblinear')
# Entrenamos el modelo con los datos de entrenamiento (train)
modelo_regresion_log = modelo_regresion_log.fit(balance_training_x,
balance_training_y)
# Hacemos la predicción del modelo entrenado sobre los datos de prueba (test)
y_pred_rl = modelo_regresion_log.predict(x_test_)
```

Modelo Regresión Logística con balanceo ROS, clasifica correctamente a 1435(TN) clientes sin mora y 202(TP) con mora, pero comete 397(FP) falsos positivos, afectando la precisión en clientes sin mora, y 269(FN) falsos negativos, lo que representa un riesgo al no identificar clientes con mora.

```
####### Árbol de decision ##########
# Instanciamos el modelo
modelo_arbol = DecisionTreeClassifier(max_depth = 5, random_state = 42)
# Entrenamos el modelo con los datos de entrenamiento (train)
modelo_arbol = modelo_arbol.fit(balance_training_x, balance_training_y)
# Hacemos la predicción del modelo entrenado sobre los datos de prueba (test)
y pred ad = modelo arbol.predict(x test )
```

Modelo Árbol de decisión con balanceo ROS, clasifica correctamente a 1246 clientes sin mora y 281 con mora, pero comete 586 falsos positivos, afectando significativamente la confianza en clientes sin mora, y 190 falsos negativos, lo que implica riesgos al no identificar algunos clientes con mora.

```
########## Random Forest ############
# Instanciamos el modelo
modelo bosque = RandomForestClassifier(n estimators = 500, random state = 42,
max_depth = 5)
# Entrenamos el modelo con los datos de entrenamiento (train)
modelo_bosque.fit(balance_training_x,balance_training_y)
# Hacemos la predicción del modelo entrenado sobre los datos de prueba (test)
y_pred_ba = modelo_bosque.predict(x_test_)
```

Modelo Random Forest con balanceo ROS, clasifica correctamente a 1428 clientes sin mora, pero genera 404 falsos positivos, lo que puede afectar la confianza de clientes clasificados erróneamente. Los 254 falsos negativos que representan clientes con mora que no son identificados, puede ser un riesgo financiero importante y detecta 217 verdaderos positivos clientes con mora correctamente clasificados.

```
######### XGBoost ##########
# Instanciamos el modelo
xgb_model = xgb.XGBClassifier(
    n_estimators=100,       # Número de árboles
    max_depth=5,            # Profundidad máxima del árbol
    learning_rate=0.1,      # Tasa de aprendizaje
    random_state=42,      # Asegura la reproducibilidad
    use_label_encoder=False # Desactivar la codificación de etiquetas
)
# Entrenamos el modelo con los datos de entrenamiento (train)
xgb_model.fit(balance_training_x, balance_training_y)
# Hacemos la predicción del modelo entrenado sobre los datos de prueba (test)
y_pred_xgboost1 = xgb_model.predict(x_test_)
```

Modelo XGBoost con balanceo ROS, clasifica correctamente 1393 clientes sin mora, pero genera un número significativo de 439 falsos positivos, lo que puede afectar significativamente la confianza de clientes clasificados erróneamente. Los 207 falsos negativos que representan clientes con mora que no son identificados, puede ser un riesgo financiero importante. Aunque detecta 264 verdaderos positivos clientes con mora correctamente clasificados.

```
# Se implementa la técnica de sobremuestreo SMOTE
smote = SMOTE()
# Ajustar el predictor y la variable objetivo
balance_smote_training_x, balance_smote_training_y = smote.fit_resample(x_val,
y_val)
```

Con la intención de mejorar la precisión de la predicción se aplica la técnica de sobre muestreo SMOTE (Synthetic Minority Over-sampling Technique) [10] para balancear los datos. Esta técnica aumenta la cantidad de muestras en la clase minoritaria de forma conjunta tanto en variables explicativas como explicadas sobre los datos de entrenamiento.

Se aplican los mismos algoritmos con la técnica de balanceo SMOTE y se obtienen los siguientes resultados:

Modelo Regresión Logística con balanceo SMOTE, clasifica correctamente a 1304 clientes sin mora y 232 clientes con mora. Sin embargo, comete 528 falsos positivos (clientes sin mora clasificados incorrectamente como morosos), lo cual afecta la precisión en la detección de clientes sin mora. Además, presenta 239 falsos negativos (clientes con mora no detectados), lo que representa un riesgo importante al no identificar adecuadamente a los clientes con probabilidad de incumplimiento.

Modelo Árbol de decisión con balanceo SMOTE, clasifica correctamente a 1195 clientes sin mora y 272 clientes con mora. Sin embargo, comete 637 falsos positivos (clientes sin mora que fueron clasificados incorrectamente como morosos), lo cual disminuye la precisión al identificar clientes confiables. Además, presenta 199 falsos negativos (clientes con mora no detectados por el modelo), lo que representa un riesgo significativo, ya que estos casos podrían implicar pérdidas para la institución financiera al no anticipar el incumplimiento.

Modelo Random Forest con balanceo SMOTE, clasifica correctamente a 1353 clientes sin mora y a 244 clientes con mora. Sin embargo, comete 479 falsos positivos (clientes sin mora clasificados erróneamente como morosos), lo cual puede afectar la experiencia del cliente al etiquetar injustamente a personas confiables. Por otro lado, presenta 227 falsos negativos (clientes con mora que no fueron detectados), lo que representa un riesgo financiero para la institución, ya que estos casos podrían derivar en incumplimientos no previstos.

Modelo XGBoost con balanceo SMOTE, clasifica correctamente a 1683 clientes sin mora y a 119 clientes con mora. No obstante, comete 149 falsos positivos (clientes sin mora clasificados erróneamente como morosos), lo que puede afectar la precisión en la identifica-

Tabla 3. Fórmulas de métricas implementadas

Métrica	Fórmula	Descripción
Precisión (Accuracy)	(TP + TN) / (TP + TN + FP + FN)	Proporción de aciertos globales.
Exactitud (Precisión)	TP / (TP + FP)	Confiabilidad de los positivos predichos.
Exhaustividad (Recall)	TP / (TP + FN)	Positivos reales identificados.
F1-Score	2 * (Exactitud * Recall) / Exactitud	Media armónica entre precisión y Recall), útil para conjuntos desbalanceados.

Fuente: Elaboración propia.

ción de buenos clientes. Además, presenta 352 falsos negativos (clientes con mora que no fueron detectados), lo cual implica un riesgo importante para la institución, ya que representa una cantidad considerable de morosos no identificados a tiempo.

Fase de evaluación

Las métricas de evaluación se derivan directamente de la matriz de confusión generada por cada modelo. Esta matriz permite identificar y clasificar los resultados de las predicciones en cuatro categorías: verdaderos positivos (TP), falsos negativos (FN), falsos positivos (FP) y verdaderos negativos (TN). En la Tabla 3 se observan las fórmulas de las métricas y su descripción para evaluar los modelos.

Normas éticas de investigación

La presente investigación tiene un carácter tecnológico y no involucra la participación directa de seres humanos ni animales, ya que se basa en el análisis de datos secundarios provenientes de un historial de créditos de una entidad financiera. Por lo tanto, no fue necesario aplicar protocolos de consentimiento informado ni aprobación de comités de ética para la recolección de datos. No obstante, se garantizó la confidencialidad y anonimato de la información utilizada, cumpliendo con las normativas internas de protección de datos de la institución y la legislación vigente en Ecuador sobre manejo responsable de información sensible.

Resultados

La tabla 4 presenta la matriz comparativa de métricas de evaluación para cada combinación modelo-balanceo. Entre las métricas evaluadas, se otorga especial énfasis al recall, dado que este indicador mide la capacidad del modelo para identificar correctamente a los clientes morosos.

Mejor desempeño en detección de morosos, el modelo de Regresión Logística con SMOTE alcanzó el valor más alto de recall 0.68, aunque su precisión general (accuracy) fue más baja 0.61. Esta combinación resulta particularmente útil para el propósito de esta investigación, ya que privilegia la detección de clientes en riesgo de mora.

Modelos con mejor equilibrio general, XGBoost con ROS mostró el mejor balance global entre métricas, con un accuracy de 0.72 y un recall de 0.56, lo que representa un com-

Tabla 4. Matriz comparativa de métricas de evaluación de los modelos con balanceo ROS y SMOTE

Modelo (balanceo)	Precisión (Accuracy)	Exactitud (Precisión)	Recall	F1 Score	Interpretación
Regresión Logística (ROS)	0.71	0.34	0.43	0.38	Recall medio, captura limitada de morosos.
Árbol de Decisión (ROS)	0.66	0.32	0.60	0.42	Buen recall, útil para detectar morosos.
Random Forest (ROS)	0.71	0.35	0.46	0.40	Recall aceptable, balance moderado.
XGBoost (ROS)	0.72	0.38	0.56	0.45	Buen balance, buen desempeño en recall.
Regresión Logística (SMOTE)	**0.61**	**0.30**	**0.68**	**0.42**	**Mejor recall, excelente para predecir morosos.**
Árbol de Decisión (SMOTE)	0.69	0.32	0.46	0.38	Recall medio, desempeño moderado.
Random Forest (SMOTE)	0.69	0.34	0.51	0.41	Buen recall, modelo equilibrado.
XGBoost(SMOTE)	0.78	0.44	0.25	0.32	Alto accuracy pero bajo recall, no ideal para detectar morosos.

Fuente: Elaboración propia.

promiso razonable entre rendimiento general y capacidad de detección.

Modelos con desempeño robusto y estable, Random Forest, tanto con ROS como con SMOTE, se comportó de manera consistente con métricas equilibradas (recall de 0.46 y 0.51, respectivamente), posicionándose como una opción confiable para escenarios donde se busca un equilibrio entre sensibilidad y precisión.

Modelos con alto accuracy pero baja sensibilidad, XGBoost con SMOTE obtuvo el mayor accuracy de 0.78, pero con un recall de apenas 0.25. Este resultado indica que el modelo clasifica correctamente a la mayoría (no morosos), pero penaliza fuertemente a la clase minoritaria, haciéndolo poco adecuado para el objetivo de esta investigación. Similar comportamiento se observa con el Árbol de Decisión con SMOTE (recall de 0.46).

Impacto del balanceo en el rendimiento, los resultados evidencian el impacto de la técnica de balanceo en el rendimiento de los modelos. Se revisa detenidamente el modelo de Regresión Logística, el cual alcanzó un valor de recall de 0.68 al utilizar el balanceo con SMOTE, que es considerablemente más alto en comparación con el recall de 0.43 obtenido con el balanceo ROS.

El modelo de Regresión Logística con SMOTE emerge como una opción apropiada para predecir mora en créditos en este contexto.

Discusión y Conclusiones

En este proyecto de investigación, se implementaron modelos basados en técnicas de Inteligencia Artificial, específicamente de aprendizaje automático supervisado, con el objetivo de predecir mora en créditos de una institución financiera en Ecuador y sustentar decisiones estratégicas basadas en evidencia.

De acuerdo con la evaluación de los resultados, se identifica que el modelo de Regresión Logística, combinado con la técnica de balanceo SMOTE, presenta un desempeño superior en términos de recall, es decir, en la correcta identificación de los casos positivos de mora. Este hallazgo difiere de lo reportado en estudios como los de Grzebień [10], Juan Freire [11], Borrero y Bedoya [12], quienes concluyen que el modelo Random Forest ofrece el mejor rendimiento en la predicción de impagos crediticios en sus respectivos contextos. Esta divergencia pone de manifiesto la importancia de validar empíricamente los modelos predictivos, considerando las características específicas del conjunto de datos, el contexto de aplicación y los objetivos del análisis. En este sentido, la selección del modelo más adecuado no debe fundamentarse únicamente en resultados obtenidos en investigaciones previas, sino en una evaluación rigurosa y contextualizada que atienda las particularidades del problema abordado.

Durante la fase de modelado se identificó un desbalance significativo en la variable objetivo "MoraCredito", lo cual representó una limitación en el proceso de entrenamiento de los modelos predictivos. Para mitigar este inconveniente y mejorar el desempeño de los algoritmos, se aplicaron técnicas de balanceo de datos, específicamente ROS y SMOTE. No obstante, estas técnicas pueden modificar la distribución original de los datos, por lo que su aplicación se restringió exclusivamente al conjunto de entrenamiento, con el fin de preservar la representatividad del conjunto de prueba y garantizar una evaluación más objetiva del rendimiento de los modelos.

En futuras investigaciones se recomienda considerar el uso de modelos más avanzados, como redes neuronales, técnicas de aprendizaje profundo o enfoques híbridos, que podrían capturar patrones más complejos en los datos y mejorar la capacidad predictiva del modelo. Asimismo, sería pertinente explorar estrategias adicionales de balanceo de clases y la incorporación de nuevas variables que puedan aportar valor a la variable objetivo y enriquecer la entrada de los modelos.

Fuentes de Financiamiento

La investigación llevada a cabo, no ha sido financiada económicamente por ningún organismo.

Declaración de Conflicto de Interés

El autor declara la no existencia de conflicto de interés alguno..

Aporte del Artículo en la Línea de Investigación

Este artículo aporta a la línea de investigación sobre la aplicación de Inteligencia Artificial en el sector financiero, al proponer una solución práctica y replicable para la predicción de impagos crediticios, que fortalece la toma de decisiones basada en datos y a la reducción de riesgos financieros.

Declaración de Contribución de Cada Autor

Libia Johana Sánchez Espinoza: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Administración del proyecto, Recursos, Software, Supervisión, Validación, Visualización, Redacción del borrador y original del artículo, Redacción-revisión y edición del artículo.

Agradecimientos

Agradezco a la Institución Financiera por abrirme las puertas y facilitarme el acceso a la información, de manera especial, extiendo mi reconocimiento al señor Gerente General de la institución, por su colaboración y apoyo durante todo el proceso.

Referencias

[1] F. Rozo-García, «Revisión de las tecnologías presentes en la industria 4.0», revuin, vol. 19, n° 2, pp. 177-191, may 2020, doi: 10.18273/revuin.v19n2-2020019.

[2] A. Fernández, «Inteligencia artificial en los servicios financieros», BOLETÍN ECONÓMICO 2/2019, n° 2, 29 de marzo de 2019. Accedido: 5 de marzo de 2025. [En línea]. Disponible en: https://repositorio.bde.es/bitstream/123456789/8448/1/be1902-art7.pdf

[3] E. Sauñe-Villalobos y V. Ramos, «Estudio comparativo de cultura organizacional actual y requerida 2019-2021 en una institución bancaria», CienciAmérica, vol. 12, n° 1, pp. 33-46, feb. 2023, doi: 10.33210/ca.v12i1.403.

[4] C. M. Pineda Pertuz, Aprendizaje automático y profundo en Python: una mirada hacia la inteligencia artificial. Bogotá, Colombia, Paracuellos de Jarama, Madrid: Ediciones de la U ; Ra-Ma, 2022. [En línea]. Disponible en: https://www.google.com.ec/books/edition/Aprendizaje_autom%C3%A1tico_y_profundo_en_Py/NEi9EAAAQBAJ?hl=es&gbpv=1&printsec=frontcover

[5] J. M. Robles, Big data para científicos sociales. Madrid: CIS - Centro de Investigaciones Sociológicas, 2020. [En línea]. Disponible en: https://www.google.com.ec/books/edition/Big_data_para_cient%C3%ADficos_sociales_Una/Ir8MEAAAQBAJ?hl=es&gbpv=1&pg=PA3&printsec=frontcover

[6] V. Pedrero, K. Reynaldos-Grandón, J. Ureta-Achurra, y E. Cortez-Pinto, «Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia», Rev. méd. Chile, vol. 149, n° 2, pp. 248-254, feb. 2021, doi: 10.4067/s0034-98872021000200248.

[7] M. Mwangi, «The Role of Machine Learning in Enhancing Risk Management Strategies in Financial Institutions», IJMRM, vol. 2, n° 1, pp. 44-53, jun. 2024, doi: 10.47604/ijmrm.2643.

[8] R. Gimeno y J. M. Marqués, «Tradición e inteligencia artificial: oportunidades y retos del machine learning para los servicios financieros», Revista ICE, n° 926, jun. 2022, doi: 10.32796/ice.2022.926.7403.

[9] E. Hussein Sayed, A. Alabrah, K. Hussein Rahouma, M. Zohaib, y R. M. Badry, «Machine Learning and Deep Learning for Loan Prediction in Banking: Exploring Ensemble Methods and Data Balancing», IEEE Access, vol. 12, pp. 193997-194019, 2024, doi: 10.1109/ACCESS.2024.3509774.

[10] K. D. Grzebień y M. J. Segovia-Vargas, «Predicción del riesgo de impago en los préstamos P2P», REyF, vol. 1, n° 2, pp. 175-188, nov. 2023, doi: 10.32826/reyf.v1i2.352.

[11] J. Freire López, «Modelo de Clasifi-

cación de Riesgo Crediticio Utilizando Random Forest en financiera del Ecuador». agosto de 2021. Accedido: 3 de abril de 2025. [En línea]. Disponible en: https://repositorio.uisek.edu.ec/bitstream/123456789/4256/1/Freire%20L%C3%B3pez%2C%20Juan.pdf

[12] D. Borrero-Tigreros y O. Bedoya-Leiva, «Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial», revuin, vol. 19, n° 4, pp. 37-52, jun. 2020, doi: 10.18273/revuin.v19n4-2020004.

[13] H. Chitarroni, «La regresión logística». Instituto de Investigación en Ciencias Sociales, 1 de diciembre de 2002. Accedido: 8 de abril de 2025. [En línea]. Disponible en: https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf

[14] M. Romero Martínez, P. Carmona Ibáñez, y J. Pozuelo Campillo, «La predicción del fracaso empresarial de las cooperativas españolas. Aplicación del Algoritmo Extreme Gradient Boosting», Ciriec-España, n° 101, pp. 255-288, mar. 2021, doi: 10.7203/CIRIEC-E.101.15572.

[15] C. Schröer, F. Kruse, y J. M. Gómez, «A Systematic Literature Review on Applying CRISP-DM Process Model», Procedia Computer Science, vol. 181, pp. 526-534, 2021, doi: 10.1016/j.procs.2021.01.199.

Libia Johana Sánchez Espinoza

Nota biográfica del autor

https://orcid.org/0009-0009-1329-155X

Investigadora independiente. Máster en Análisis Inteligente de Datos (Big Data) por la Universidad Isabel I (UI1-España). Ingeniera en Sistemas Computacionales e Informáticos por la Universidad Técnica de Ambato (UTA-Ecuador). Cuenta con dos Diplomados de docencia por la Universidad de Las Américas (UDLA-Ecuador) y varios cursos de pedagogía. Sus líneas de investigación se centran en el uso de técnicas de Inteligencia Artificial para el análisis de datos, sistemas de apoyo a la toma de decisiones y la innovación educativa con tecnologías emergentes.