

# *Modelo de Predicción de Éxito en Recuperación de Préstamos Financieros Utilizando Machine Learning en una Entidad Financiera de Perú*

## *Successful Prediction Model in Financial Loan Recovery Using Machine Learning in a Financial Institution in Peru*

Nombres de Autores de Primera  
Institución

Línea 1 (Institución) Dep.,  
Universidad, Organización  
Línea 2 (Institución)  
Línea 3: Ciudad, País  
Línea 4: Correo electrónico

Nombres de Autores de  
Segunda Institución

Línea 1 (Institución) Dep.,  
Universidad, Organización  
Línea 2 (Institución)  
Línea 3: Ciudad, País  
Línea 4: Correo electrónico

Nombres de Autores de Tercera  
Institución

Línea 1 (Institución) Dep.,  
Universidad, Organización  
Línea 2 (Institución)  
Línea 3: Ciudad, País  
Línea 4: Correo electrónico

**Resumen** — Las entidades Financiera del Perú, en su esfuerzo por mejorar la recuperación de préstamos, ha recurrido a modelos predictivos de Machine Learning para identificar los créditos con mayor probabilidad de pago. Este estudio propone un modelo para predecir la recuperación de préstamos financieros usando 3 algoritmos de Machine Learning: Random Forest, Regresión Logística y Máquinas de Soporte Vectorial. Se utilizó la metodología CRIS-DM considerando los siguientes pasos: recopilación de los datos, preprocesamiento de datos, modelamiento del algoritmo y definición de métricas. El dataset que se utilizó fue recogidos con 438 registros históricos de préstamos de una empresa financiera del Perú, lo cual se identificaron 10 variables. Los resultados revelaron que el algoritmo Random Forest sobresalió por su alta precisión (97.8%) y capacidad de generalización, evidenciando un mejor desempeño en la predicción de la recuperación de préstamos. Estos hallazgos subrayan la viabilidad del algoritmo Random Forest como herramienta para la mejora en la gestión del riesgo crediticio y sugieren su implementación como parte de un sistema integral de evaluación de créditos.

**Palabras Clave:** recuperación de préstamos; aprendizaje automático; modelos de predicción financiera; interpretabilidad de modelos financieros; predicción de impago de préstamos; bosque aleatorio.

**Abstract** — Financial entities in Peru, in their effort to improve loan recovery, have turned to predictive Machine Learning models to identify loans with the highest probability of repayment. This study proposes a model to predict financial loan recovery using three Machine Learning algorithms: Random Forest, Logistic Regression, and Support Vector Machines. The CRISP-DM methodology was used, considering the following steps: data collection, data preprocessing, algorithm modeling, and metric definition. The dataset used consisted of 438 historical loan records from a financial company in Peru, identifying 10 variables. The results revealed that the Random Forest algorithm stood out for its high accuracy

(97.8%) and generalization ability, demonstrating better performance in predicting loan recovery. These findings underscore the feasibility of the Random Forest algorithm as a tool for improving credit risk management and suggest its implementation as part of an integrated credit assessment system.

**Keywords** - loan recovery; machine learning; financial predictive models; interpretability of financial models; loan default prediction; random forest.

### I. INTRODUCTION

La industria financiera ha enfrentado persistentemente el desafío de recuperar préstamos de manera efectiva. A pesar de los esfuerzos basados en análisis demográfico e historiales crediticios, la tasa de préstamos no recuperados sigue siendo alta en muchas regiones [1]. Las técnicas convencionales frecuentemente no logran adaptarse a las cambiantes dinámicas del comportamiento del prestatario [2]. Además, la evolución en la naturaleza del préstamo y recuperación ha llevado a la necesidad de enfoques más sofisticados [3].

A nivel nacional, las instituciones crediticias, como las cajas municipales y cooperativas, enfrentan desafíos significativos en la recuperación de préstamos. La falta de recuperación afecta no solo la solidez financiera de estas instituciones, sino también la economía general de la región [4]. A nivel internacional, la problemática es similar, y muchas instituciones financieras, independientemente de su escala, enfrentan pérdidas sustanciales debido a préstamos no recuperados [5].

Numerosos estudios han abordado la tarea de predecir el riesgo crediticio utilizando técnicas de aprendizaje automático. Por ejemplo, Zhu et al. [1] emplearon el algoritmo de bosques aleatorios para predecir el incumplimiento de préstamos, logrando una precisión del 85%. Sin embargo, su enfoque no consideró las dinámicas socioeconómicas cambiantes, que podrían influir en la

capacidad de pago del prestatario. Por otro lado, Moscato et al. [2] presentaron un benchmark de diferentes técnicas de machine learning para la predicción de scores de crédito. A pesar de que su estudio proporcionó una visión amplia de las técnicas, no exploró en profundidad la adaptabilidad de los modelos en diferentes contextos financieros. Nazemi et al. [5] se centraron en la predicción inter-temporal de recuperaciones de bonos en mora utilizando machine learning, obteniendo resultados prometedores en términos de precisión. Sin embargo, su modelo podría no ser práctico para instituciones con registros limitados. Finalmente, Arora et al. [8] propusieron técnicas de análisis de datos y machine learning para predecir incumplimientos de tarjetas de crédito. A pesar de que su enfoque mostró ser efectivo, la falta de interpretabilidad del modelo podría ser una limitación en entornos regulados.

Dada la importancia de predecir el riesgo crediticio con precisión y eficiencia, esta investigación propone una evaluación comparativa de tres técnicas prominentes en el campo del aprendizaje automático: bosques aleatorios, regresión logística y máquina de soporte vectorial.

Para llevar a cabo esta evaluación, se utilizará un conjunto de datos proporcionado por una institución financiera local, que incluye características demográficas, socioeconómicas y de historial crediticio de los prestatarios. La metodología seguirá los siguientes pasos: (1) Preprocesamiento de los datos, que incluye la limpieza, normalización y división en conjuntos de entrenamiento y prueba; (2) Implementación de los tres modelos de machine learning mencionados (bosques aleatorios, regresión logística y máquina de soporte vectorial); (3) Entrenamiento de los modelos utilizando el conjunto de entrenamiento y evaluación de su desempeño en el conjunto de prueba; (4) Comparación de los modelos basada en métricas como precisión, sensibilidad, especificidad y el área bajo la curva ROC (AUC-ROC).

El presente artículo está estructurado de la siguiente manera: La sección 2 presenta una revisión detallada de los trabajos relacionados; La sección 3 describe el modelo propuesto y la metodología empleada; La sección 4 discute los experimentos realizados y los resultados obtenidos; Finalmente, se presentan las conclusiones en la sección 5.

## II. RELATED WORKS

En la literatura se han aplicado diferentes técnicas de Machine Learning para la prevención de riesgos en préstamos financieros, tales como: Random Forest (RF) [1], Support Vector Machine (SVM) [8], Decision Trees (DT)[3] y Logistic Regression (LR)[2] tal como se muestra en la Tabla I.

La técnica RF ha demostrado ser una técnica robusta en la predicción de incumplimientos de préstamos. Zhu et al. [1] alcanzaron una alta precisión con RF en la predicción de préstamos en mora. Kumar et al. [13] confirmaron la eficiencia de RF en clasificación de préstamos, con una precisión del 97.16%. Sigrist y Leuenberger [12] destacaron la superioridad de RF sobre modelos lineales para predicciones a largo plazo. La versatilidad de RF se manifiesta en la integración con otras técnicas, como en la investigación de Li et al. [18], que mejora significativamente la precisión de las predicciones. Wang et al. [7] mostraron cómo el Zhima Credit Score, integrado con RF, mejora la predicción de incumplimientos, con SVM mostrando el mayor incremento en rendimiento. En cuanto a la aplicación en datos bancarios, Wang

et al. [19] reportaron una precisión del 85.16% con RF, subrayando su eficacia en manejar grandes conjuntos de datos y variables relevantes. Estos hallazgos, junto con los estudios de Chen et al. [20] y Dong et al. [21], enfatizan la capacidad de RF para profundizar en el análisis de riesgos en el sector financiero.

Las técnicas de SVM son ampliamente utilizadas en la predicción de incumplimientos financieros, ofreciendo resultados significativos en términos de precisión y rendimiento. Arora et al. [8] reportaron una precisión de SVM del 82% en la predicción de incumplimientos de tarjetas de crédito. En otro contexto, Zhang et al. [11] desarrollaron un modelo basado en SVM que logró altos valores de APR, con un promedio del 3.0872% y 2.1858% en dos conjuntos de datos diferentes, lo que se traduce en un enfoque rentable para los prestamistas. Bellotti et al. [9] aplicaron SVM en la previsión de tasas de recuperación de préstamos no productivos, logrando una proporción explicada de la variación del 18.54% al 19.79%. Estos resultados resaltan la capacidad de SVM para manejar eficientemente la variabilidad y complejidad de los datos de recuperación de activos. Por último, Cela y Cuenca [10] propusieron un modelo de SVM para la evaluación de riesgo de crédito, el cual mostró un buen rendimiento con una precisión del 77.42%, aunque se observó un desbalance entre la sensibilidad, con un 96.069%, y la especificidad, con solo un 4.277%.

Las técnicas DT son reconocidos por su utilidad en la predicción de riesgos crediticios. Sotomayor Alzamora et al. [14] lograron una precisión de 93.06% en la concesión de microcréditos rurales. En cuanto a la evaluación de la concesión de créditos en general, el modelo DT alcanzó una precisión global del 90% en 2018 y 2020 [3]. Adicionalmente, Uddin et al. [15] reportaron una precisión del DT de 0.81 y 0.77 para las categorías 'Sí' y 'No' respectivamente en la aprobación de préstamos bancarios, demostrando su utilidad en entornos financieros urbanos. Sin embargo, al comparar DT con métodos más avanzados, se encontró que RF superó al DT tradicional con una precisión del 79.44% frente a un 67.28% en un estudio de Uddin et al. [15], lo que sugiere que las variantes mejoradas de DT pueden ofrecer resultados más precisos en la predicción de préstamos.

Por otro lado, los modelos de LR son ampliamente reconocidos por su utilidad en la predicción de riesgo crediticio y han sido comparados con enfoques más contemporáneos de aprendizaje automático. Moscato et al. [2] proporcionan un benchmark de diferentes técnicas de aprendizaje automático, incluida la LR, destacando su viabilidad para predecir puntajes crediticios. Srinath y G.H.S. [4] exploran la aplicabilidad de métodos explicables de aprendizaje automático, donde la LR sigue siendo un método de referencia por su interpretabilidad. Liu et al. [16] aplican la LR junto con otros algoritmos de aprendizaje automático para predecir la probabilidad de incumplimiento en el mercado de crédito en línea, demostrando la importancia de la representación de datos y la selección del modelo en la precisión de las predicciones. Machado y Karray [17] evalúan el riesgo de crédito de clientes comerciales utilizando algoritmos híbridos de aprendizaje automático, indicando que la combinación de LR con otros métodos puede mejorar la precisión del modelo. Dong et al. [21] analizan el puntaje de crédito de big tech y el riesgo de incumplimiento, subrayando que características como historiales de puntaje de crédito pueden mejorar las predicciones de modelos híbridos que incluyen LR.

TABLE I. TÉCNICAS MACHINE LEARNING APLICADOS EN RIESGOS DE PRÉSTAMOS FINANCIEROS.

Técnica	Propósito	Reference
RF	Predicción de préstamos en mora	[1], [5], [7], [12], [13], [18]- [21]
SVM	Evaluación de riesgo crediticio	[8], [9], [10], [11]
DT	Predicción de tasas de recuperación en préstamos no productivos	[3], [6], [14], [15]
LR	Identificación de incumplimientos de tarjetas de crédito	[2], [4], [16], [17], [21]

Estas investigaciones destacan la variedad y eficacia de las técnicas empleadas en diferentes contextos del ámbito financiero. Es esencial considerar estos estudios y sus conclusiones al abordar los desafíos específicos de la Entidad Financiera Peruana.

### III. MODELO PROPUESTO

Esta sección presenta el modelo para la predicción de éxito en la recuperación de préstamos financieros utilizando Machine Learning en la Entidad Financiera y aplicando la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [22]. La fase (1) se centra en la selección y recopilación de datos financieros relevantes para la recuperación de préstamos. En la fase (2) se realiza el preprocesamiento de los datos para garantizar su integridad y utilidad en el modelo. La fase (3) comprende el entrenamiento de algoritmos de Machine Learning, específicamente Random Forest, Regresión Logística y Máquinas de Soporte Vectorial. Finalmente, la fase (4) involucra la evaluación de los modelos mediante métricas adecuadas para identificar la técnica más efectiva. (ver Figura. 1).

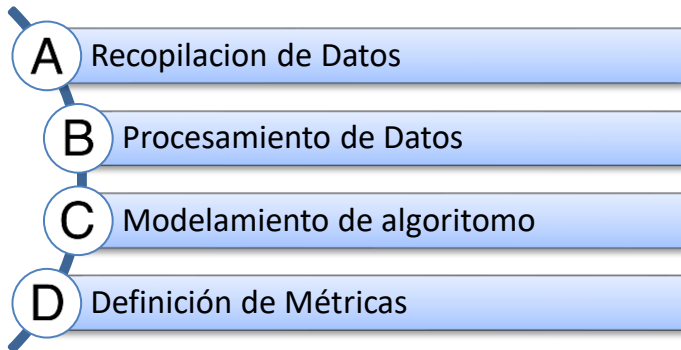


Figura. 1. Modelo Propuesto con CRISP-DM [10].

#### A. Fase 1: Recopilación de Datos

Para la recopilación de datos, se han seleccionado datos de préstamos de Consumo Personal y PYME de la Entidad Financiera Peruana desde 2020 hasta 2023 con un total de 6,814 registros con 52 variables clave exportadas en formato XLS, permitiendo una integración directa con herramientas de análisis como Python. Este rango abarca el impacto económico de la pandemia y es de suma importancia para adaptar el modelo a los patrones de pago cambiantes y mejorar la precisión de las predicciones de recuperación de préstamos. La tendencia creciente en el número de préstamos, detallada en la Tabla II, subraya la importancia de estos datos en la estrategia analítica para nuestro estudio que aborda técnicas de Machine Learning.

TABLE II. EVOLUCIÓN ANUAL DE LA CARTERA DE PRÉSTAMOS DE LA ENTIDAD FINANCIERA DE PERÚ.

Año	Cantidad de Préstamos
2020	43,587
2021	33,062
2022	71,341
2023	189,708

#### B. Fase 2: Preprocesamiento de Datos

Siguiendo la metodología CRISP-DM y basándonos en las 52 variables financieras proporcionadas para el DataSet de la Entidad Financiera de Perú, se enfocará en preparar los datos para que sean adecuados para el modelado predictivo. Aquí está una descripción general del proceso y cómo cada paso se ajusta a la metodología CRISP-DM:

**Identificación de Variables.** Para poder determinar las variables más importantes que intervendrán en el modelo para la recuperación de préstamos que luego serán exportados en formato XLS para facilitar el trabajo con código Python, se trabajó en conjunto con el equipo de cobranzas y recuperaciones pertenecientes al área de riesgos identificando así las variables (ver Tabla III).

TABLE III. VARIABLES PARA LA GENERACIÓN DEL MODELO

Variables	Descripción
Crédito	Número del préstamo
Tipo de Moneda	Tipo de moneda del préstamo
Producto	Producto del Préstamo
Fecha de Pago	Fecha que realizo el Pago
Fecha de Vencimiento	Fecha de vencimiento de cada cuota
Fecha Desembolso	Fecha que se solicitó el préstamo
Número de Calendario	Número de calendario del crédito
Número de cuota	Número de Cuota actual del crédito
Días de Atraso	Días de atraso después de la fecha de vencimiento
Estado	Estado contable del crédito
Monto	Monto por cada cuota
Monto Pendiente	Monto pendiente de pago por cuota
Monto Total Pendiente	Monto total pendiente del préstamo
Saldo Capital	Monto de capital restante

Fecha de Castigo	Fecha que se castigó crédito
------------------	------------------------------

**Limpieza de Datos.** Esta etapa se enfocó en corregir o eliminar los registros incorrectos, incompletos, inexactos o irrelevantes del conjunto de datos. Las fechas fueron convertidas de formatos de texto a valores ordinales para permitir su análisis cuantitativo, y se manejaron adecuadamente los valores faltantes mediante la imputación de un valor predeterminado, lo cual preserva la integridad de los datos para el análisis sin introducir sesgos indebidos.

**Transformación de Datos.** Durante la preparación del conjunto de datos, utilizamos StandardScaler de scikit-learn para estandarizar las características numéricas. Este proceso ajusta las variables para que tengan una media de cero y una desviación estándar de uno, nivelando su importancia durante el

entrenamiento de modelos como SVM y Regresión Logística, que son sensibles a la escala de los datos. Es crucial aplicar esta estandarización únicamente a los datos de entrenamiento para evitar la fuga de información y asegurar que el aprendizaje se base en patrones reales.

**Integración de Datos.** Al procesar los datos desde un único archivo .xls, enfocamos atención especial en mantener la coherencia de la información a través de cada etapa de transformación. Los identificadores únicos de los registros, como los números de crédito, sirvieron de anclaje para que cada dato procesado conservara su vinculación con el historial original del préstamo. Este cuidadoso manejo asegura la fidelidad de los datos para análisis y validación futura del modelo predictivo (ver Tabla IV).

TABLE IV. CARACTERÍSTICAS ESTANDARIZADAS PARA EL MODELO

Código	Campo	Descripción	Tipo dato
F1	DiasDesdeUltimoPago	Días transcurridos desde el último día de pago	Numérico
F2	PromedioDiasAtraso	Promedio de días de atraso total del crédito	Numérico
F3	DesviacionDiasAtraso	desviación de días de atraso	Numérico
F4	CantidadPagosRealizados	Cantidad de Pagos realizados	Numérico
F5	Cuota	Cuota de pago	Numérico
F6	CuotaActual	Monto o importe	Numérico
F7	MontoPendiente	Monto Pendiente	Numérico
F8	MontoCuotaActual	Monto Pendiente de la cuota actual	Numérico
F9	MontoTotPendiente	Monto total pendiente del crédito	Numérico
F10	SaldoCapital	Saldo Capital del crédito	Numérico

**Formateo Final.** El conjunto de datos fue meticulosamente organizado para asegurar la lógica y funcionalidad en el análisis predictivo. Se ordenaron las columnas para reflejar el historial crediticio de los clientes, con información relevante como montos de créditos, saldos pendientes y el historial de pagos. Este ordenamiento estratégico facilita la interpretación y el uso eficiente en los algoritmos predictivos. En la secuencia final de las columnas, se posicionó la variable F2 que actúa como el indicador objetivo del modelo, reflejando la probabilidad de que un préstamo sea pagado a tiempo según la función analítica aplicada.

En la Figura 2 se destacan correlaciones significativas como la alta relación positiva entre F6 y F4 (0.99) y entre F7 y F8 (0.85), así como una correlación negativa notable entre F2 y F9 (-0.26). Esta matriz es fundamental para identificar relaciones entre variables, detectar colinealidad y mejorar modelos predictivos, proporcionando una base sólida para decisiones informadas en el análisis de datos y modelado

En la Figura 3 se tuvo el siguiente nivel de importancia de las variables, donde F2 ocupa el primer lugar para las características que se aplicarán al entrenamiento de este modelo predictivo.

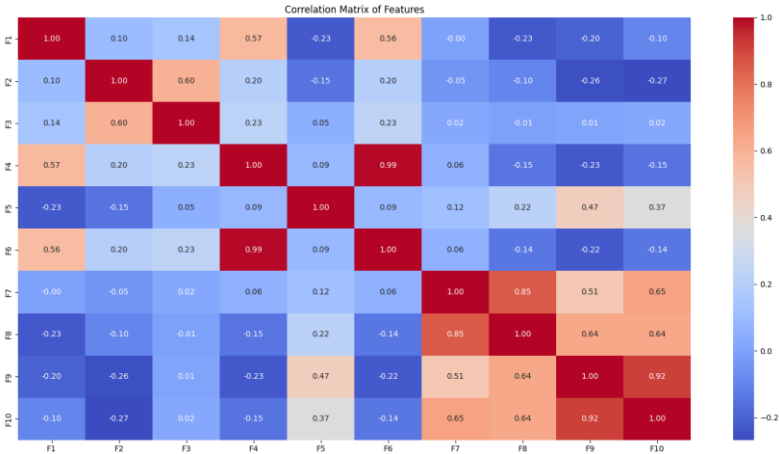


Figura. 2. Matriz de Correlación .

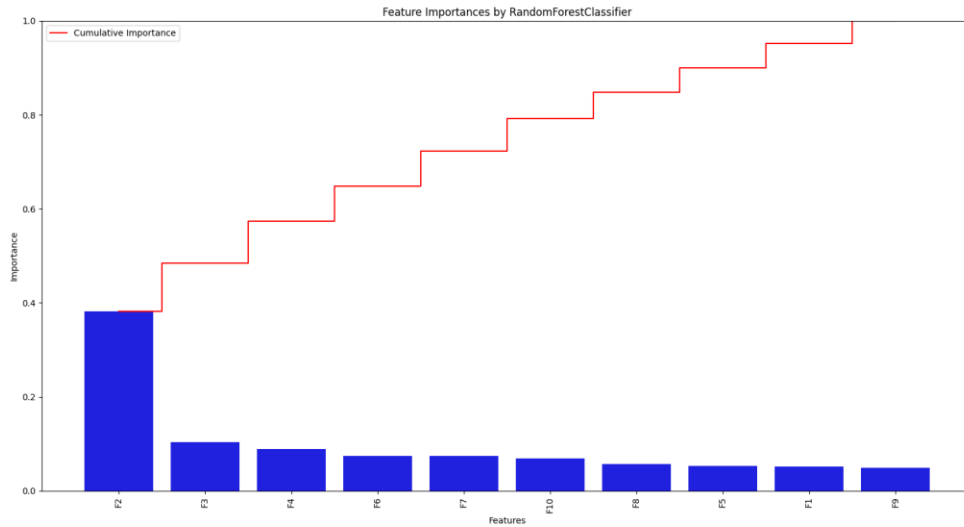


Figura. 3. Clasificación de importancia de variables.

### C. Fase 3: Modelamiento de Algoritmo

La Tabla V muestra el script del proceso que se realizó para el modelamiento de algoritmos RF, para ello se utilizó Python como lenguaje de programación y Visual Studio Code como entorno de desarrollo integrado (IDE). Además, empleamos diversas bibliotecas para llevar a cabo el análisis y modelado de datos. El proceso de análisis y predicción de pagos de préstamos se realizó en varios pasos utilizando además librerías de Machine Learning, los paseos se presenta continuación:

- 1) Primero, se importaron las librerías necesarias, tales como: *pandas*, *numpy* y *sklearn*, para la manipulación de datos, la normalización, imputación de valores faltantes, construcción del modelo y evaluación de su desempeño.
- 2) En este paso se cargaron los datos desde un archivo *Excel*.
- 3) Luego, se realizaron transformaciones en los datos, convirtiendo las fechas en formato *datetime* y creando nuevas características basadas en el comportamiento histórico de los créditos, tales como: el número de días hasta el pago más reciente, el promedio y desviación estándar de días de atraso, el número de pagos realizados, entre otros.
- 4) En este paso los datos se dividieron en conjuntos de entrenamiento y prueba, y se aplicaron técnicas de imputación y escalado para manejar valores faltantes y normalizar las características.
- 5) Posteriormente, se entrenó el modelo de clasificación *Random Forest* con los datos preprocesados.
- 6) Finalmente, se evaluó el desempeño del modelo en el conjunto de prueba utilizando una matriz de confusión y un reporte de clasificación, los cuales proporcionan información sobre la precisión, sensibilidad, especificidad y otras métricas clave.

TABLE V. SCRIPT PARA ENTRENAMIENTO CON RF.

1	<pre>import pandas as pd, numpy as np from sklearn.model_selection import train_test_split from sklearn.preprocessing import StandardScaler from sklearn.impute import SimpleImputer from sklearn.ensemble import RandomForestClassifier from sklearn.metrics import confusion_matrix, classification_report</pre>
2	<pre>data = pd.read_excel("../Recuperaciones.xls")</pre>
3	<pre>data['FechaPago'] = pd.to_datetime(data['FechaPago'], errors='coerce') data['FechaVencimiento'] = pd.to_datetime(data['FechaVencimiento'], errors='coerce') data['PagoATiempo'] = (data['nDiasAtraso'] &lt;= 0).astype(int) data = data.assign( F1=data.groupby('Credito')['FechaPago'].transform(lambda x: (x.max() - x).dt.days), F2=data.groupby('Credito')['nDiasAtraso'].transform('mean'), F3=data.groupby('Credito')['nDiasAtraso'].transform('std'), F4=data.groupby('Credito')['FechaPago'].transform('count'), F5=data['nCuota'], F6=data['nCuotaActual'], F7=data['nMonto'], F8=data['nMontoPendiente'], F9=data['nMontoTotPendiente'], F10=data['nSaldoCap']) features = ['F1', 'F2', 'F3', 'F4', 'F5', 'F6', 'F7', 'F8', 'F9', 'F10']</pre>
4	<pre>train = data[data['nCuota'] != data['nCuotaActual']] test = data[data['nCuota'] == data['nCuotaActual']] X_train, y_train = train[features], train['PagoATiempo'] X_test, y_test = test[features], test['PagoATiempo'] imputer, scaler = SimpleImputer(), StandardScaler() X_train = scaler.fit_transform(imputer.fit_transform(X_train)) X_test = scaler.transform(imputer.transform(X_test))</pre>
5	<pre>model = RandomForestClassifier(random_state=42) model.fit(X_train, y_train)</pre>
6	<pre>y_pred = model.predict(X_test) print(confusion_matrix(y_test, y_pred)) print(classification_report(y_test, y_pred))</pre>

#### D. Fase 4: Definición de Métricas

La evaluación del rendimiento de los algoritmos se lleva a cabo mediante métricas ampliamente utilizadas en ML: Precisión (Ecuación (1)), Exactitud (Ecuación (2)), Sensibilidad o Recall (Ecuación (3)), y la Puntuación F1 (Ecuación (4)). Además, se utilizó la Característica Operativa del Receptor Micro-Promedio (ROC) para un análisis más detallado:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 - Score} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

Donde:

- Verdaderos Positivos (TP): Predicción correcta de cuotas que pagaran a tiempo.
- Verdaderos Negativos (TN): Predicción correcta de cuotas que no pagaran a tiempo.
- Falsos Positivos (FP): Predicción incorrecta de cuotas que pagaran a tiempo.
- Falsos Negativos (FN): Predicción incorrecta de cuotas que no pagaran a tiempo.

Las métricas se definen como:

**Accuracy:** Representa la proporción de predicciones correctas (tanto pagos como no pagos) entre todas las realizadas por el modelo [24].

**Recall:** Mide la habilidad del modelo para identificar correctamente las cuotas que fueron pagadas a tiempo [24].

**Precision:** Indica qué tan precisas son las predicciones del modelo cuando identifica una cuota como pagada a tiempo [24].

**F1 - Score:** Es el promedio armónico de la precisión y la sensibilidad, proporcionando una medida única de la precisión del modelo en la clasificación de cuotas pagadas a tiempo [24].

Estas métricas proporcionan una comprensión integral de la efectividad del modelo en la tarea específica de predecir la puntualidad de los pagos de las cuotas. La precisión mide qué tan bien el modelo identifica las cuotas que se pagarán a tiempo, mientras que la exactitud mide el rendimiento general del modelo en todas las clases. La sensibilidad indica la habilidad del modelo para detectar todas las instancias positivas reales (pagos a tiempo), y la puntuación **F1** proporciona un balance entre precisión y sensibilidad, siendo particularmente útil en situaciones donde las clases están desbalanceadas.

## IV. RESULTADOS Y DISCUSIÓN

Para validar el rendimiento de nuestros modelos, se utilizaron muestras de 438 créditos, clasificados en pago realizado a tiempo y pago no realizado a tiempo. Luego, al calcular las métricas utilizando las matrices de confusión (Figura.5), identificamos los algoritmos que lograron los mejores resultados en predicciones correctas e incorrectas en la puntualidad de los pagos. La Tabla VI muestra el número de predicciones correctas e incorrectas por resultado y algoritmo. En este sentido, el modelo que utilizó el algoritmo RF logró el mayor número de predicciones correctas, seguido por SVM y RL con cantidades respectivamente inferiores.

TABLE VI. PREDICCIONES CORRECTAS E INCORRECTAS DE PAGO A TIEMPO

Algoritmo	Resultado	Predicciones Correctas	Predicciones Incorrectas
RF	Pagará a tiempo	278	23
	No pagara a tiempo	136	1
LR	Pagará a tiempo	279	71
	No pagara a tiempo	88	0
SVM	Pagará a tiempo	278	50
	No pagara a tiempo	109	1

La Figura 4 muestra la comparación del resultado de la curva ROC de los tres algoritmos: RF, LR y SVM.

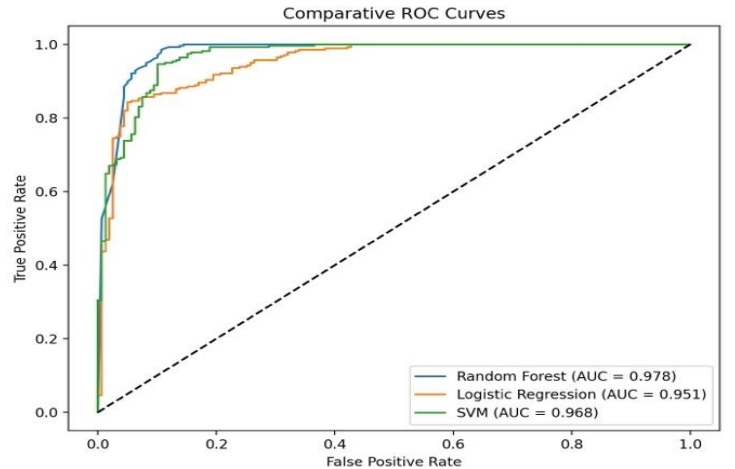
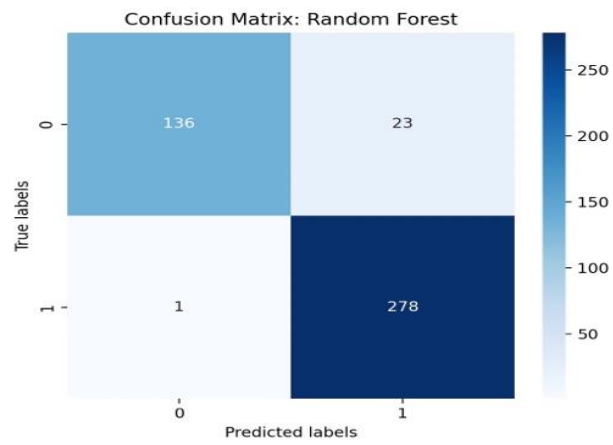
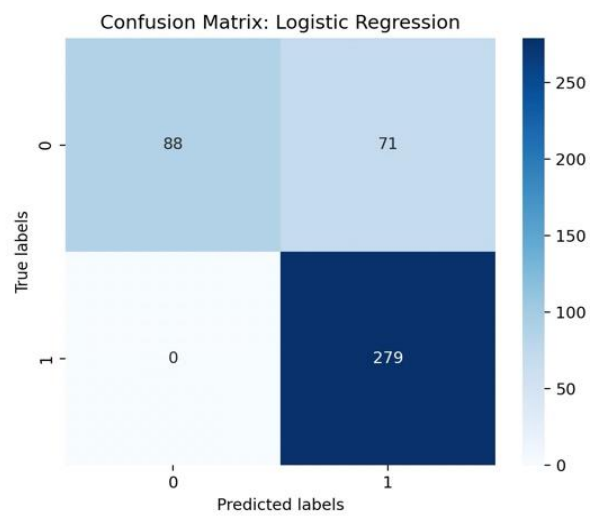


Figura. 4. Comparación de rendimiento ROC de las 3 técnicas

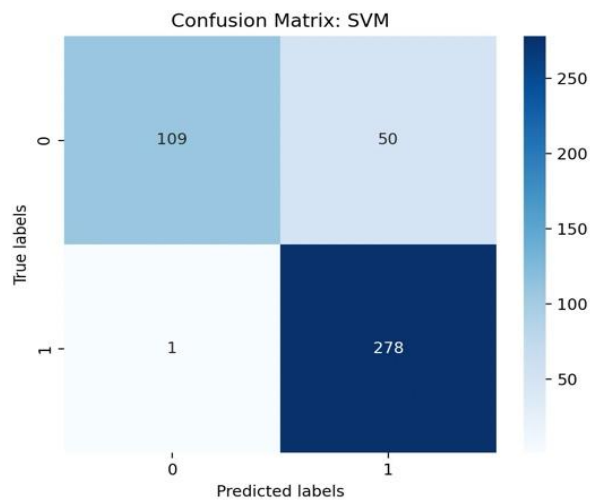
Se observa que el algoritmo que obtuvo mejor resultado de AUC es RF con un valor de 0.978. Los valores recopilados de cada algoritmo se consideran buenos para capturar valores positivos, sin embargo, se observa que los algoritmos SVM y RF generan más predicciones falsos positivos, lo que reduce la precisión del rendimiento predictivo final. Según el porcentaje de falsos positivos, el algoritmo RF tiene pocos fallos en predicciones incorrectas. Además, la Figura 6 muestra la curva ROC obtenida en cada algoritmo.



(a)



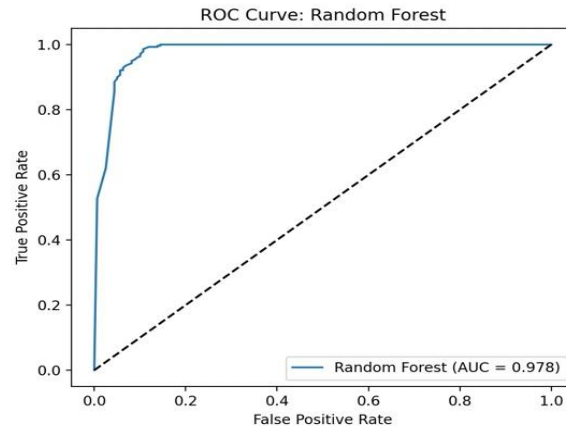
(b)



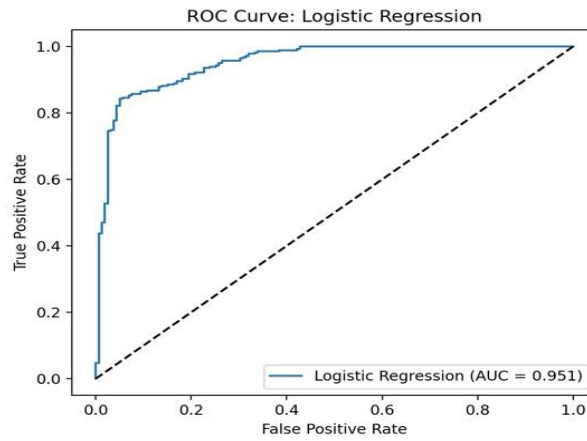
(c)

Figura.5. Matrices de confusi3n sobre las predicciones realizadas por (a) RF, (b) LR, (c) SVM

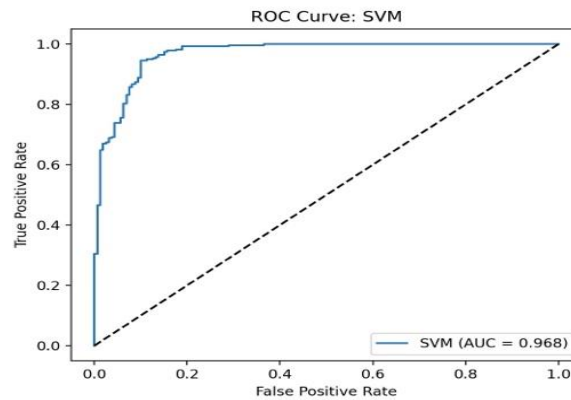




(a)



(b)



(c)

Figura.6. Curvas ROC de algoritmos utilizados en los modelos (a) RF, (b) LR, (c) SVM.

Por otro lado, la Curva ROC es una herramienta gráfica que evalúa la habilidad de un modelo de clasificación para distinguir entre clases. El Área Bajo la Curva ROC (AUC) es una métrica integral que resume la curva completa y mide la probabilidad de que el modelo clasifique una cuota positiva (pagada a tiempo) más alta que una cuota negativa (no pagada a tiempo) al azar. Un AUC más alto indica un mejor rendimiento del modelo. En nuestro estudio, utilizamos la implementación de sklearn para calcular el

AUC, proporcionando así una medida estándar y confiable de la capacidad de discriminación del modelo. Podemos identificar en la Figura.6 la curva ROC obtenida en cada algoritmo.

En la Tabla VII, resume la efectividad de tres modelos de clasificación avanzados evaluados en este estudio. El modelo Random Forest RF lidera el rango con una precisión general del 97.8%, seguido de SVM con una precisión del 96.8%, y la



Regresión Logística LR con una precisión del 95.1%. El valor de recall para RF es de 0.85, lo que indica una alta tasa de verdaderos positivos, mientras que para SVM y RL es de 0.68 y 0.55 respectivamente, lo que muestra una competencia notable en la identificación correcta de las instancias positivas. El F1-score, que combina precisión y recall en una única métrica, es consistentemente alto en todos los modelos, con RF a la cabeza con un puntaje de 0.918, seguido muy de cerca por SVM y LR con puntajes de 0.810 y 0.712, respectivamente. Estas métricas colectivamente subrayan la robustez de los modelos en la tarea de clasificación.

TABLE VII. COMPARACIÓN DE MÉTRICAS DE EVALUACIÓN DE LAS 3 TÉCNICAS DE MACHINE LEARNING.

Rank	Model	Clase	Recall	F1-score	Accuracy (%)
1	RF	1: Pagará a tiempo	1.00	0.92	94.98
		0: No pagara a tiempo	0.86	0.92	
2	SVM	1: Pagará a tiempo	1.00	0.92	88.81
		0: No pagara a tiempo	0.69	0.81	
3	LR	1: Pagará a tiempo	1.00	0.89	85.62
		0: No pagara a tiempo	0.55	0.71	

### V. CONCLUSIONES

En la presente investigación, se desarrolló un estudio comparativo de tres algoritmos de Machine Learning: Random Forest RF, Regresión Logística RL y Máquinas de Soporte Vectorial SVM, con el objetivo de predecir la recuperación de créditos en la Entidad Financiera Peruana. Los resultados han evidenciado que el modelo RF exhibió una superioridad notable en términos de precisión y capacidad para discernir entre los pagos realizados puntualmente y los atrasados.

Durante el proceso de validación, se emplearon métodos rigurosos como la validación cruzada y el análisis de curvas ROC, que reafirmaron la fortaleza del modelo RF frente a los otros algoritmos. Este modelo no solo alcanzó un AUC destacado de 0.978 sino que también mantuvo un balance ideal entre la sensibilidad y la especificidad, demostrando así su aplicabilidad en el ámbito financiero.

Las técnicas de SVM y RL, aunque no superaron al RF, presentaron resultados competentes que sugieren su potencial para ser aplicados en circunstancias donde se requieren modelos con características particulares, como mayor interpretabilidad o adaptabilidad a diferentes tipos de datos financieros.

Asi mismo en la Tabla VII realizamos una comparación con los resultados obtenidos por otros estudios, teniendo en cuenta los siguientes campos que mencionan al algoritmo con mayor precisión acorde al objeto de estudio para el cual fue utilizado, luego cantidad de clases y variables utilizadas y por último el AUC obtenido en los diferentes estudios que se tuvieron en cuenta como base de este estudio referenciados en la parte inferior de este documento.

TABLE VIII. RESULTADO Y OTROS ESTUDIOS RELACIONADOS

Trabajo	Objetivos	Clases	Variables	AUC
RF*	Predicción en recuperación de prestamos	2	10	0.978
RF [2]	predicción de préstamos a reembolsar	4	7	0.717
RF [1]	Predicción de incumplimiento de los préstamos	4	14	0.983
RF [4]	Predicción de incumplimiento de préstamos personales	5	--	0.970
SVM [8]	Predicción de impagos de tarjetas de crédito	5	23	0.820

### ACKNOWLEDGMENT

Agradecemos profundamente a la institución financiera por participar en esta investigación y al Departamento de Investigación de la Universidad XYZ por el apoyo brindado para la realización de este trabajo de investigación a través del incentivo xxx.

### REFERENCIAS

- [1] J L. Zhu, D. Qiu, D. Ergu and C. Ying, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503-513, 2019.
- [2] V. Moscato, A. Picariello and G. Sperlí, "A benchmark of machine learning approaches for credit score prediction," *Expert Systems with Applications*, vol. 165, p. 113986, 2021.
- [3] V. Lagasio, F. Pampurini, A. Pezzola y A.G. Quaranta, "Assessing bank default determinants via machine learning," *Information Sciences*, vol. 618, pp. 87-97, 2022.
- [4] T. Srinath y G. H.S., "Explainable machine learning in identifying credit card defaulters," *Global Transitions Proceedings*, vol. 3, pp. 119-126, 2022.
- [5] A. Nazemi, F. Baumann y F. J. Fabozzi, "Intertemporal defaulted bond recoveries prediction via machine learning," *European Journal of Operational Research*, vol. 297, pp. 1162-1177, 2022.
- [6] J.-A. Castro, E. A. Cerletti, J. Fernández B., y F. Vásquez L., "IEF2 2019-Minutas citadas-IEF segundo semestre 2019," División de Política Financiera, Banco Central de Chile, 2019.
- [7] H. Wang, W. Chen y F. Da, «Zhima Credit Score in Default Prediction for Personal Loans,» *Procedia Computer Science*, vol. 199, pp. 1478-1482, 2022.
- [8] S. Arora, S. Bindra, S. Singh y V. K. Nassa, «Prediction of credit card defaults through data analysis and machine learning techniques,» *Materials Today: Proceedings*, vol. 51, pp. 110-117, 2022.
- [9] Bellotti, D. Brigo, P. Gambetti, y F. Vrina, "Forecasting recovery rates on non-performing loans with machine learning," *Int. J. Forecast.*, vol. 37, 2021.
- [10] G. Cela y J. P. Cuenca, "Propuesta de modelo de machine learning para la evaluación de riesgo de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito La Merced Ltda. Cuenca," *Maestría en Tecnologías de la Información, Subdirección de Posgrados, Univ. Católica de Cuenca*, 2019.
- [11] L. Zhang, J. Wang, y Z. Liu, "What should lenders be more concerned about? Developing a profit-driven loan default prediction model," *Expert Systems With Applications*, vol. 213, p. 118938, 2023.
- [12] F. Sigríst and N. Leuenberger, "Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail

- probabilities," *European Journal of Operational Research*, vol. 305, pp. 1390–1406, 2023.
- [13] L. Sathish Kumar, V. Pandimurugan, D. Usha, M. Nageswara Guptha, and M.S. Hema, "Random forest tree classification algorithm for predicating loan," in *Materials Today: Proceedings*, vol. 57, pp. 2216–2222, 2022. Available: <https://doi.org/10.1016/j.matpr.2021.12.322>.
  - [14] G. Sotomayor Alzamora, M. R. Aceituno-Rojo, y H. I. Condori-Alejo, "An assertive machine learning model for rural micro credit assessment in Peru," en *Procedia Computer Science*, vol. 202, pp. 301–306, 2022. Available: <https://doi.org/10.1016/j.procs.2022.04.040>.
  - [15] N. Uddin, M. Khabir Uddin Ahamed, M.A. Uddin, M. Manwarul Islam, M. Alamin Talukder et al., "An ensemble machine learning based bank loan approval predictions system with a smart application," en *International Journal of Cognitive Computing in Engineering*, vol. 89, 2023. Available: <https://doi.org/10.1016/j.ijcce.2023.09.001>.
  - [16] Y. Liu, M. Yang, Y. Wang, Y. Li, T. Xiong, y A. Li, "Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China," en *International Review of Financial Analysis*, vol. 79, 2022. Available: <https://doi.org/10.1016/j.irfa.2021.101971>.
  - [17] M. R. Machado y S. Karray, "Assessing credit risk of commercial customers using hybrid machine learning algorithms," en *Expert Systems With Applications*, vol. 200, 2022. Available: <https://doi.org/10.1016/j.eswa.2022.116889>.
  - [18] X. Li, D. Ergu, D. Zhang, D. Qiu, Y. Cai, y B. Ma, "Prediction of loan default based on multi-model fusion," en *Procedia Computer Science*, presentado en The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021), vol. 199, pp. 757–764, 2022. Available: <https://doi.org/10.1016/j.procs.2022.01.094>.
  - [19] Y. Wang, Y. Zhang, Y. Lu, y X. Yu, "A comparative assessment of credit risk model based on machine learning: A case study of bank loan data," en *Procedia Computer Science*, presentado en 2019 International Conference on Identification Information and Knowledge in the Internet of Things (IIKI2019), vol. 174, pp. 141–149, 2020. Available: <https://doi.org/10.1016/j.procs.2020.06.069>.
  - [20] T.-K. Chen, H.-H. Liao, G.-D. Chen, W.-H. Kang, y Y.-C. Lin, "Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports," en *Expert Systems With Applications*, vol. 233, 2023. Available: <https://doi.org/10.1016/j.eswa.2023.120714>.
  - [21] Y. Dong, Q. Gou, y H. Qiu, "Big tech credit score and default risk — Evidence from loan-level data of a representative microfinance company in China," en *China Economic Review*, vol. 81, 2023. Available: <https://doi.org/10.1016/j.chieco.2023.102010>.
  - [22] Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri Ramírez, C.; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramírez Quintana, MJ.... (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*. 33(8):3048-3061. <https://doi.org/10.1109/TKDE.2019.2962680>
  - [23] IBM Corporation, "Conceptos básicos de ayuda de CRISP-DM - Documentación de IBM," Carbon for IBM.com, agosto 17, 2021. [Online]. Available: <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dms-crisp-help-overview>
  - [24] O. Jiménez, A. Jesús and L. Wong, "Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine," *2023 33rd Conference of Open Innovations Association (FRUCT)*, Zilina, Slovakia, 2023, pp. 116–124, doi: 10.23919/FRUCT58615.2023.10143068.