Machine learning models for predicting mortgage payment difficulties in Peru

Geraldo-Campos, Luis Alberto[1], Carreño-Flores, Oscar David[2], and Soria-Quijaite, Juan Jesús[3]
[1,2,3]Universidad Tecnológica del Perú, Perú, *lgeraldo@utp.edu.pe, c28416@utp.edu.pe, c20723@utp.edu.pe*

Abstract*– The objective of this study was to analyze which machine learning models best predict mortgage payment difficulties in Peru. A quantitative method was used with longitudinal data from 2018 to 2022 from the National Household Survey (ENAHO), where a total of 5,716 households with mortgage loans were examined. The input variables considered were geographical area, type of housing, use of credit, and source of financing, with difficulty in meeting the payment schedule as the output variable. The analyses were performed in Google Colab, reporting frequency statistics and exploratory and class balancing analyses to evaluate machine learning models such as Logistic Regression, Random Forest (RF), and Support Vector Machine (SVM) with SMOTE. In the training phase of the classification models, the Scikit-learn, XGBoost, and Keras models were trained and compared. The results showed that, of all the models evaluated, Random Forest without adjustments showed the best performance (F1-score = 0.67; recall = 0.71), although combined Stacking (RF + XGBoost) showed a better balance between classes, but its overall performance was lower. In addition, models such as SVM without adjustments show problems in situations of unbalanced classes, highlighting the need to use techniques such as SMOTE. It is concluded that the Random Forest model is more effective in detecting payment difficulties in mortgage loans.

***Keywords*-- machine learning, mortgage credit, credit risk, longitudinal study.**

Modelos de machine learning para la predicción de la dificultad de pago hipotecario en el Perú

Geraldo-Campos, Luis Alberto[1], Carreño-Flores, Oscar David[2], and Soria-Quijaite, Juan Jesús[3]
[1,2,3]Universidad Tecnológica del Perú, Perú, *lgeraldo@utp.edu.pe, c28416@utp.edu.pe, c20723@utp.edu.pe*

Resumen– El objetivo de este estudio fue analizar qué modelos de machine learning predicen mejor la dificultad de pago hipotecario en el Perú. Se utiliza un método cuantitativo con datos longitudinales del 2018 al 2022 de la Encuesta Nacional de Hogares (ENAHO), donde se examinaron un total de 5716 hogares con créditos hipotecarios. Se consideraron como variables de entrada el dominio geográfico, tipo de vivienda, uso del crédito y fuente de financiamiento, y la dificultad con el cronograma de pago como variable de salida. Los análisis se realizaron en Google Colab, reportándose los estadísticos de frecuencia y los análisis exploratorios y de balanceo de clases para evaluar los modelos de machine learning como Regresión Logística, Random Forest (RF) y Support Vector Machine (SVM) con SMOTE, en la fase de entrenamiento de los modelos de clasificación, se entrenaron y compararon los modelos Scikit-learn, XGBoost y Keras. Los resultados mostraron que, de todos los modelos evaluados, el Random Forest sin ajustes mostró el mejor rendimiento (F1-score = 0.67; recall = 0.71), aunque el Stacking combinado (RF + XGBoost), mostraron un mejor equilibrio entre clases, pero su rendimiento total fue más bajo. Además, modelos como el SVM sin ajustes muestran problemas en situaciones de clases desbalanceadas, resaltando la necesidad de usar técnicas como SMOTE. Se concluye que el modelo de Random Forest es más efectivo para detectar la dificultad de pago en los créditos hipotecarios.

Palabras clave-- aprendizaje automático, crédito hipotecario, riesgo crediticio, estudio longitudinal.

I. Introducción

La estabilidad económica y el bienestar de los hogares en países emergentes como Perú, el acceso al crédito desempeña un rol fundamental, ya que contribuye a la reducción de las desigualdades sociales; sin embargo, depende de la capacidad que tienen los hogares para acceder a un crédito formal o informal, el cual está condicionada a diversos factores como el nivel de ingresos, la educación, la ubicación geográfica y otras variables importantes en su determinación [1], [2]. A pesar de los avances en la bancarización y la expansión de productos financieros, los hogares peruanos continúan enfrentando dificultades para acceder a créditos adecuados a sus necesidades, lo que genera una dependencia significativa de fuentes informales con condiciones menos favorables [3], [4].

Se ha demostrado que los hogares con menores ingresos y baja escolaridad tienden a recurrir a préstamos informales, mientras que aquellos con mayor nivel educativo y estabilidad laboral tienen mayores probabilidades de acceder a productos financieros formales [5], [6]. No obstante, la creciente digitalización y la proliferación del comercio electrónico han transformado las dinámicas de acceso al crédito, favoreciendo la adopción de créditos digitales como complemento o sustituto del crédito tradicional [7]. A pesar de estos avances, la falta de alfabetización financiera y la desconfianza en los servicios formales siguen siendo barreras que limitan la inclusión financiera, especialmente en los segmentos socioeconómicos más vulnerables [8], [9]. Esta problemática evidencia la necesidad de identificar los patrones de comportamiento crediticio de los hogares peruanos basado en su ubicación geográfica, el acceso al crédito mediante la fuente de financiamiento, uso del crédito, el monto del crédito accedido y la dificultad con el cronograma de pago.

En este marco estudios previos han demostrado que el uso de técnicas de Machine Learning (ML) son eficaces para evaluar los de Riesgos Crediticios. Técnicas como la de Random Forest (RF) suelen ser más efectivas para la predicción de riesgos crediticios, reportando una precisión del 93% en predicciones a corto plazo y del 95% a largo plazo [10]; asimismo, se ha reportado una precisión del 88.3%, destacando su capacidad para manejar patrones de datos complejos y características de alta dimensionalidad, demostrando robustes frente a datos no lineales y con múltiples características [11]. Otra de las técnicas usualmente utilizadas en este marco es la regresión logística (LR), la cual es una técnica estándar en la evaluación de riesgos crediticios. Su desempeño suele ser inferior a técnicas avanzadas como la RF y la support vector machine (SVM), reportando precisión de hasta el 78.4%, la cual la hace más efectiva en datos lineales y menos efectiva en datos complejos y no lineales [11]; sin embargo, a pesar de sus limitaciones sigue siendo utilizada por su simplicidad y facilidad de interpretación [10], [11]. En esa misma línea, la técnica SVM también ha demostrado un buen desempeño, principalmente en el manejo de datos de alta dimensionalidad; se ha reportado una precisión del 84.5%, teniendo como desventaja el tiempo, el cual puede ser largo y presentar cuellos de botella computacionales con grandes volúmenes de datos; sin embargo [11], es eficaz en la clasificación de datos no lineales mediante el uso de funciones como kernel [10], [11]. En este marco, el uso de la técnica de sobremuestreo sintético (SMOTE) es ampliamente utilizada para abordar el problema del desbalanceo de datos en la predicción de riesgos crediticios, por la que su uso ha demostrado mejorar significativamente el desempeño de los modelos. Por ejemplo, se encontró que el uso de SMOTE combinado con técnicas como RF y SVM mejoró la precisión y otras métricas de desempeño [12], [13], siendo ser efectiva en la combinación con Edited Nearest Neighbours (ENNs) alcanzando una precisión del 90.49% y mejorando las métricas de precisión y recall [12]

Estas técnicas han sido de gran beneficio para evaluar el riesgo crediticio, sin embargo, existe escaza evidencia del uso de estas técnicas en la evaluación del incumplimiento de pago en el tipo de créditos hipotecarios. Por ello, este estudio tiene como propósito evaluar cuál de los modelos de machine learning previamente identificados permiten una mejor predicción de la dificultad de pago hipotecario en el Perú. Por lo tanto, se tomaron datos registrados mediante la Encuesta Nacional de Hogares (ENAHO) de los periodos del 2018 al 2022. Los resultados ofrecen una visión integral que contribuye a la formulación de políticas públicas más eficaces para mejorar la inclusión financiera y el bienestar de los hogares peruanos. Además, los resultados permiten tener una comprensión más profunda de las dinámicas crediticias en los hogares peruanos, en cuanto a su ubicación geográfica, el tipo de acceso, uso del crédito hipotecario y su dificultad de pago basado respecto al monto accedido.

Este estudio se organiza de la siguiente manera. La sección 2 presenta la estrategia metodológica utilizada, dando detalles de las características de la población encuestada y las variables y técnicas utilizadas, así como el proceso realizado para llevar a cabo el presente estudio. La sección 3 muestra los resultados de las tres técnicas utilizadas, comparándolas para demostrar la más efectiva con los datos y variables utilizadas. La Sección 4 presenta la discusión que emana de los resultados obtenidos en comparación a estudios previos identificados. Finalmente, en la sección 5 se exponen las conclusiones principales del estudio.

II. Metodología

Esta investigación parte de una estrategia metodológica cuantitativa de tipo longitudinal basado en datos de la Encuesta Nacional de Hogares (ENAHO) del Perú, centrando sus esfuerzos en las encuestas de los periodos del 2018 al 2022, específicamente de los módulos relacionados con las variables de estudio dominio geográfico, tipo de vivienda, uso del crédito, fuente de financiamiento evaluándose como acceso al crédito y dificultad con el cronograma de pago, considerado como dificultad de pago.

El procesamiento y análisis se realizó íntegramente en el entorno Google Colab, utilizando bibliotecas especializadas de Python para ciencia de datos. Para ello, se realizó la preparación y preprocesamiento de los datos, en el cual se cargó y se cumplió con la limpieza de los valores perdidos y codificación de variables categóricas mediante. Además, al seleccionar registros de hogares que informaron de que habían accedido a algún tipo de crédito hipotecario, se logró identificar una muestra de 5716 hogares.

Antes de la ejecución de los análisis se definió la variable binaria como dependiente, dificultad de pago, basada en respuestas sobre dificultades reportadas en el cronograma de pago del crédito (Sí = 1, No = 0). Como variables independientes se incluyeron el dominio geográfico, el tipo de vivienda, uso del crédito, fuente financiera del crédito y monto total del crédito hipotecario en soles.

Con estas variables se realizaron los análisis descriptivos de frecuencias y se realizaron los análisis exploratorio y balanceo de clases. Se tuvo a bien explorar la distribución de la variable objetivo, evidenciándose un fuerte desbalance de clases (~25% con dificultad de pago). Para mitigar este problema, se aplicaron dos técnicas de balanceo supervisado: class_weight='balanced' para modelos como Regresión Logística [14], Random Forest [15] y Support Vector Machine (SVM) [16], y SMOTE (Synthetic Minority Oversampling Technique) usando ImbLearn para generar observaciones sintéticas de la clase minoritaria.

En la fase de entrenamiento de modelos de clasificación, se entrenaron y compararon los modelos Scikit-learn [17], [18], XGBoost y Keras. Para la regresión logística se utilizó con y sin balanceo de clases, y con SMOTE; en cambio, para random forest se utilizó la versión estándar, con balanceo, con SMOTE y versión optimizada mediante GridSearchCV [19], [20], [21]. En esa misma línea, la técnica de XGBoost [22] se utilizó el modelo base, versión optimizada y combinación con Random Forest mediante StackingClassifier. Por su parte, la técnica de SVM se utilizó la versión básica, sin ajuste por desbalance. Finalmente, la técnica de MLPClassifier (Red Neuronal) se utilizó el modelo sin ajuste y modelo ajustado con SMOTE y estandarización (StandardScaler).

En la fase de visualización y evaluación, se utilizó una validación cruzada estratificada (5-fold) para evaluar el desempeño, considerándose las métricas de Precisión (Accuracy; VP / (VP + FP)), Recall (Sensibilidad: VP / (VP + FN)), F1-score (promedio macro y ponderado=2 × (Precisión × Recall) / (Precisión + Recall)) y AUC-ROC (Área bajo la curva ROC: $\int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(x))\,dx$, donde: TPR = VP / (VP + FN); FPR = FP / (FP + VN)), todas expresadas con base en los valores de la matriz de confusión: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP), falsos negativos (FN) [23], [24]

III. Resultados

La TABLA I muestra que la mayor concentración de solicitantes de crédito se ubica en la Costa Norte (19.52%) y la Sierra Sur (13.09%), seguidos por la Selva (12.73%) y Lima Metropolitana (11.51%). Esto evidencia una distribución territorial diversa, aunque con predominancia en zonas urbanizadas y de desarrollo económico intermedio. Además, una mayoría de los encuestados reside en casas independientes (91.29%), lo que refleja la tipología de vivienda dominante en el Perú, en cambio, las viviendas en edificios solo están representado por el 4.02%, lo que podría vincularse a un mercado inmobiliario vertical aún incipiente.

La TABLA I también revela que el 74.66% de los encuestados indica que el crédito fue destinado a mejoramiento y ampliación de vivienda, seguido por compra de casa o departamento (9.07%). Este dato es crítico, pues evidencia que el crédito hipotecario en Perú no se utiliza predominantemente para adquisición de vivienda nueva, sino para mejorar condiciones habitacionales existentes. Por su

parte, los bancos privados (36.39%) y las cajas municipales (30.14%) constituyen las principales fuentes de financiamiento; en cambio, con un porcentaje menor, las cooperativas de ahorro y crédito (5.58%) también cumplen un rol importante, particularmente en regiones rurales o con menor infraestructura financiera.

Según se observa en la TABLA I, el 25.02% de los encuestados reportó haber experimentado dificultad de pago. Este valor es significativo, pues sugiere que 1 de cada 4 hogares enfrenta algún nivel de vulnerabilidad financiera posterior al otorgamiento del crédito. Esta proporción debe ser monitoreada cuidadosamente, especialmente ante escenarios macroeconómicos adversos (inflación, desempleo, tasas de interés elevadas, etc.), que podrían aumentar el riesgo de morosidad.

TABLA I
DESCRIPTIVOS SOBRE EL PERFIL HIPOTECARIO EN HOGARES PERUANOS

Variables	Opciones de respuesta	f	%
Dominio geográfico	Costa Norte	1116	19.52
	Costa Centro	561	9.82
	Costa Sur	478	8.36
	Sierra Norte	277	4.85
	Sierra Centro	636	11.13
	Sierra Sur	748	13.09
	Selva	1242	21.73
	Lima Metropolitana	658	11.51
Tipo de Vivienda	Casa independiente	5218	91.29
	Departamento en edificio	230	4.02
	Vivienda en quinta	39	0.68
	Vivienda en casa de vecindad (Callejón, solar o corralón)	182	3.18
	Choza o cabaña	4	0.07
	Local no destinado para habitación humana	43	0.75
Uso de Crédito	Comprar casa departamento	267	4.67
	Comprar terreno para vivienda	519	9.08
	Mejoramiento y ampliación de la vivienda	4256	74.46
	Construcción de vivienda nueva	664	11.62
	Más de un uso	10	0.18
Acceso al Crédito	Banco privado	2080	36.39
	Banco de la Nación	384	6.72
	Caja Municipal	1723	30.14
	Persona Particular	181	3.17
	Techo propio	38	0.67
	Financiera de Ahorro y Crédito	688	12.04
	Cooperativa de ahorro y crédito	336	5.88
	Derrama Magisterial	44	0.77
	Más de una fuente	242	4.23
Dificultad de pago	Sí	1430	25.02
	No	4286	74.98

La TABLA II resume el rendimiento comparativo de los modelos evaluados para predecir la dificultad de pago de créditos hipotecarios en el contexto peruano. Las métricas como Precisión, Recall, F1-Score y AUC-ROC permiten evaluar tanto la sensibilidad del modelo hacia la clase minoritaria (dificultad de pago), como su precisión general en la clasificación.

TABLA II
COMPARACIÓN DE MODELOS PARA LA PREDICCIÓN DE DIFICULTAD DE PAGO

Modelo	Precisión	Recall	F1-Score	AUC-ROC
Regresión Logística	0.62	0.48	0.54	0.512
Regresión Logística + SMOTE	0.66	0.58	0.62	0.561
Random Forest	0.65	0.71	0.67	0.554
Random Forest + SMOTE	0.66	0.54	0.58	0.529
Random Forest Optimizado	0.60	0.60	0.61	0.513
XGBoost	0.74	0.00	0.47	0.560
XGBoost Optimizado	0.56	0.33	0.50	0.540
Modelo Combinado (RF + XGBoost)	0.58	0.32	0.51	0.540
SVM	0.00	0.00	0.00	0.490
MLP	0.00	0.00	0.00	0.520
MLP + SMOTE + Estandarización	0.25	0.56	0.45	0.510

La TABLA II y la comparación visual que se muestra en la Fig. 1 permite señalar que, entre todos los modelos evaluados, el Random Forest sin ajustes demostró el mejor desempeño global con un F1-score de 0.67 y un recall de 0.71, siendo altamente sensible a la detección de hogares en riesgo. Aunque algunos modelos como el Stacking combinado (RF + XGBoost) mostraron un balance más equitativo entre clases, el rendimiento general fue inferior. Además, modelos como SVM y MLP sin ajustes evidencian limitaciones en contextos de clases desbalanceadas, subrayando la necesidad de aplicar técnicas como SMOTE o ajustes de ponderación para mejorar la sensibilidad hacia la clase minoritaria. Esta comparación permite concluir que, en entornos de riesgo crediticio como el crédito hipotecario, no basta con métricas globales como la precisión: es necesario optimizar la capacidad de detección específica de los hogares más vulnerables.

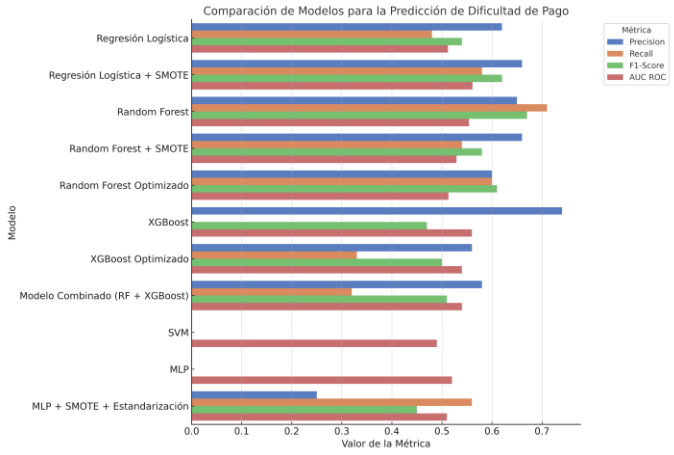

Fig. 1 Comparación visual de las métricas de desempeño por modelo

IV. DISCUSIÓN

Nuestros resultados revelan que, entre los modelos evaluados, Random Forest clásico fue el que logró mejor equilibrio entre recall (0.71) y F1-score (0.67). Este hallazgo es consistente con estudios previos que destacan la capacidad del Random Forest para generalizar adecuadamente en contextos de datos estructurados y desequilibrados, particularmente en tareas de clasificación binaria relacionadas con el riesgo crediticio [25] y en este caso en particular con la dificultad de pago de los créditos hipotecarios. Aunque XGBoost mostró la mayor precisión (0.74), su recall fue considerablemente más bajo (0.40), lo cual limita su utilidad

práctica para la detección temprana de hogares en riesgo, una prioridad crítica para instituciones financieras [26], [27].

El énfasis en recall como métrica clave se alinea con una investigación reciente [28] que aborda la predicción de impago como una tarea de identificación de casos minoritarios en conjunto de datos desbalanceados. Modelos como MetaNTM (metadata-guided variational neural topic model) han sido propuestos para mejorar la discriminación de señales débiles en escenarios donde la precisión por sí sola no es suficiente [28]. Así, aunque modelos como XGBoost pueden alcanzar altos niveles de precisión, su limitada capacidad para capturar positivos reales afecta su aplicabilidad en dominios donde las consecuencias de un falso negativo son significativas [29].

Modelos optimizados como XGBoost optimizado y el modelo combinado Random Forest más XGBoost no superaron al Random Forest clásico en recall ni en AUC. Este comportamiento difiere de lo reportado en un estudio anterior [26], donde modelos optimizados con Optuna mejoraron significativamente el rendimiento de XGBoost. La diferencia podría atribuirse a factores como el tamaño de muestra, la calidad de las variables explicativas o la heterogeneidad geográfica y socioeconómica de los hogares peruanos que accedieron a los créditos hipotecarios, lo que sugiere que la generalización de modelos debe ser cuidadosamente evaluada en función del contexto de aplicación [30], [31].

El uso de SMOTE mostró mejoras modestas en el rendimiento, especialmente al ser combinado con regresión logística (F1-score = 0.62, recall = 0.58), lo que reafirma hallazgos previos sobre su utilidad para tratar desequilibrios de clase [25]. Sin embargo, en modelos más complejos como Random Forest o XGBoost, el beneficio marginal del uso de SMOTE fue limitado, lo que indica que, en ciertos escenarios, los modelos son capaces de manejar desequilibrios sin necesidad de técnicas adicionales de sobremuestreo [32].

Aunque los modelos avanzados comprobados en este estudio ofrecen alta capacidad predictiva, se evidenció que modelos como SVM y MLP no lograron clasificar correctamente los casos de dificultad de pago, arrojando métricas nulas. Esta ineficacia puede relacionarse con la falta de ajuste adecuado, el requerimiento de grandes volúmenes de datos o la sensibilidad a la escala de las variables, aspectos que han sido resaltados como desafíos en estudios sobre redes neuronales aplicadas a problemas financieros [33]. Incluso estudios exitosos con redes neuronales profundas [29], reconocen limitaciones en cuanto a interpretabilidad y necesidad de conjuntos de datos más diversificados. Además, la precisión depende de la calidad, representatividad y actualización del conjunto de datos utilizados, en este caso se entrenó con datos poblacionales, lo cual puede no capturar características especificas de los segmentos financieros como las microempresas o créditos comerciales, ya que se centra en créditos hipotecarios. Por lo tanto, en contextos financieros, no es suficiente saber que modelo tiene mayor precisión, sino que es fundamental entender por qué ciertos factores predicen el riesgo y como estas pueden ayudar en la toma de decisiones estratégicas.

El presente estudio refuerza la utilidad de modelos robustos y fácilmente interpretables como Random Forest para escenarios de crédito hipotecario en contextos emergentes. A pesar de las tendencias recientes hacia arquitecturas más complejas, los resultados sugieren que la adecuación al contexto local y la calidad de las variables explicativas pueden ser más decisivas que la complejidad del modelo. Es por ello, que en futuras investigaciones, se recomienda, explorar modelos híbridos explicables que combinen interpretabilidad y rendimiento, como modelos SHapley Additive exPlanations (SHAP) más Random Forest (RF) [25]. Además, se debe evaluar el impacto de fuentes de datos adicionales como comportamiento financiero longitudinal, historial de pagos, nivel educativo, ingresos o integración de variables no estructuradas como las opiniones de calificación crediticia. Además, aplicar algoritmos de optimización avanzados y técnicas de aprendizaje semi-supervisado para mejorar la sensibilidad sin comprometer precisión.

Estos modelos evaluados tienen implicaciones prácticas en las entidades financieras, al ser herramientas predictivas valiosas que pueden ser utilizadas por los bancos, cajas municipales, cooperativas de ahorro y crédito, entre otros como las Fintech, dado que se podrían utilizar para gestionar el riesgo crediticio, clasificándolos a los solicitantes según su probabilidad de dificultad de pago antes de otorgarles un préstamo. Por lo tanto, al identificar perfiles con alto o bajo riesgo de impago, estas pueden diseñar productos financieros personalizados, estableciendo tasas, plazos y requisitos ajustados a sus perfiles de nivel de riesgo.

Sin embargo, su aplicabilidad puede verse opacadas por las barreras para su implementación en el sistema financiero. Primero, la disponibilidad y calidad de datos, pues, muchas veces no se integran datos socioeconómicos no financieros en sus procesos, como los identificados en la encuesta de ENAHO. Segundo, falta de infraestructura tecnológica, ya que implementar modelos de ML suelen requerir sistemas robustos, gran capacidad de procesamiento y almacenamiento, así como herramientas de automatización seguras. Otra de las barreras a considerar es la capacidad técnica, dado que existe escasez de personal con formación en ciencia de datos y analítica predictiva dentro del sector financiero tradicional, como es el caso peruano. Esto implica tener conocimiento sobre el cumplimiento de normativas y protección de datos, lo cual puede ser complejo. Finalmente, la barrera de la aceptación organizacional, ya que el uso de modelos complejos y desconocimiento de los mismo puede generar resistencia por parte de los directivos y analistas acostumbrados a metodologías clásicas.

Estas barreras permiten establecer implicancias para políticas públicas, regulación y estrategias de inclusión financiera, en el Perú, fortalecer el marco regulatorio debe ser prioridad de la Superintendencia de Banca, Seguros y AFP (SBS), la cual podría promover lineamientos técnicos y estándares éticos para su uso de modelos predictivos en la

evaluación crediticia. Además, los resultados permiten sugerir la integración de datos públicos y privados, ya que la inclusión de datos socioeconómicos como los de la ENAHO pueden contribuir en la evaluación de riesgos, lo cual permite ampliar la inclusión financiera, particularmente en hogares rurales o informales.

V. Conclusión

Se concluye que Random Forest clásico constituye una herramienta eficaz para la predicción de riesgo de impago hipotecario en hogares peruanos, superando incluso a versiones optimizadas de modelos más complejos. Además, la aplicación de técnicas de aprendizaje automático debe considerar no solo las métricas globales, sino también la adaptabilidad al entorno socioeconómico local y la capacidad del modelo para identificar correctamente a los grupos más vulnerables.

Por otro lado, los datos muestran que el crédito hipotecario en el Perú tiene un uso predominantemente orientado al mejoramiento de vivienda, con una presencia destacada de las cajas municipales y bancos privados como principales fuentes de financiamiento. No obstante, la distribución geográfica sugiere un avance en el acceso descentralizado al crédito, aunque persisten brechas regionales. Finalmente, el hecho de que un cuarto de los hogares reporte dificultades de pago demanda acciones regulatorias y políticas focalizadas en educación financiera y evaluación de riesgos.

Agradecimiento/Reconocimiento

Esta investigación es parte de un proyecto financiado por la Universidad Tecnológica del Perú mediante el código: P-2024-LIM-44, por ello, nuestro agradecimiento a la institución y quienes la dirigen.

Referencias

[1] M. S. Camelo, J. S. Amaya, y J. F. Parra, «Determinantes del uso del crédito de vivienda Por Parte de los hogares bogotanos», *Ecos de Economía*, vol. 22, n.º 47, pp. 38-57, dic. 2018, doi: 10.17230/ecos.2018.47.2.

[2] C. Dharmadasa y M. M. Gunatilake, «Determinants of household credit behavior of low-income households in Sri Lanka», *Asian Economic and Financial Review*, vol. 13, n.º 10, Art. n.º 10, ago. 2023, doi: 10.55493/5002.v13i10.4851.

[3] Instituto Peruano de Economía, «El 49% de créditos informales tienen intereses de 500% o más», *El Comercio*, p. 14, 4 de junio de 2023. Accedido: 27 de octubre de 2024. [En línea]. Disponible en: https://www.ipe.org.pe/portal/el-49-de-creditos-informales-tienen-intereses-de-500-o-mas/

[4] M. Zurita, «Crédito informal: 9,3% de los hogares peruanos optó por préstamos fuera del sistema financiero en los últimos 12 meses», *Forbes Perú*, Perú, 21 de octubre de 2024. Accedido: 27 de octubre de 2024. [En línea]. Disponible en: https://forbes.pe/economia-y-finanzas/2024-10-21/credito-informal-93-de-los-hogares-peruanos-opto-por-prestamos-fuera-del-sistema-financiero-en-los-ultimos-12-meses/

[5] M. Bryx, J. Sobieraj, D. Metelski, y I. Rudzka, «Buying vs. Renting a Home in View of Young Adults in Poland», *Land*, vol. 10, n.º 11, Art. n.º 11, nov. 2021, doi: 10.3390/land10111183.

[6] L. Ping, S. Xiaosong, y L. Jinzhao, «Research on farmers' households credit behavior and social capital acquisition», *Front. Psychol.*, vol. 13, nov. 2022, doi: 10.3389/fpsyg.2022.961862.

[7] G. Yu y H. Xiang, «Rural E-commerce development and farmers' digital credit behavior: Evidence from China family panel studies», *PLOS ONE*, vol. 16, n.º 10, p. e0258162, oct. 2021, doi: 10.1371/journal.pone.0258162.

[8] A. Alfageme y N. R. Ramírez, «Acceso a servicios financieros de los hogares en el Perú», *Banco Central de Reserva del Perú*, pp. 1-19, 2016, [En línea]. Disponible en: https://www.bcrp.gob.pe/docs/Publicaciones/Documentos-de-Trabajo/2016/documento-de-trabajo-15-2016.pdf

[9] Instituto Nacional de Estadística e Informática, «El 57,4% de la población tiene una cuenta en el sistema financiero en el primer trimestre del año 2023», *Instituto Nacional de Estadística e Informática*, 2023. Accedido: 17 de septiembre de 2024. [En línea]. Disponible en: https://www.gob.pe/institucion/inei/noticias/780703-el-57-4-de-la-poblacion-tiene-una-cuenta-en-el-sistema-financiero-en-el-primer-trimestre-del-ano-2023

[10] A. Malakauskas y A. Lakstutiene, «The Application of Artificial Intelligence Tools in Creditworthiness Modelling for SME Entities», en *2021 IEEE International Conference on Technology and Entrepreneurship (ICTE)*, ago. 2021, pp. 1-6. doi: 10.1109/ICTE51655.2021.9584528.

[11] C. Li y J. Zhang, «Research on Credit Risk Prediction Models Based on Machine Learning», en *2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, nov. 2024, pp. 1-4. doi: 10.1109/MLBDBI63974.2024.10824021.

[12] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, y B. Ogunleye, «Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction», *Mathematics*, vol. 12, n.º 21, Art. n.º 21, ene. 2024, doi: 10.3390/math12213423.

[13] N. Uddin, Md. K. Md. Khabir Uddin Ahamed, M. A. Uddin, Md. M. Islam, Md. A. Talukder, y S. Aryal, «An ensemble machine learning based bank loan approval predictions system with a smart application», *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 327-339, jun. 2023, doi: 10.1016/j.ijcce.2023.09.001.

[14] D. W. Hosmer, S. Lemeshow, y R. X. Sturdivant, *Applied Logistic Regression*, 1.ª ed. en Wiley Series in Probability and Statistics. Wiley, 2013. doi: 10.1002/9781118548387.

[15] L. Breiman, «Random Forests», *Machine Learning*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[16] C. Cortes y V. Vapnik, «Support-vector networks», *Mach Learn*, vol. 20, n.º 3, pp. 273-297, sep. 1995, doi: 10.1007/BF00994018.

[17] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011, [En línea]. Disponible en: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

[18] O. Kramer, «Scikit-Learn», en *Machine Learning for Evolution Strategies*, O. Kramer, Ed., Cham: Springer International Publishing, 2016, pp. 45-53. doi: 10.1007/978-3-319-33383-0_5.

[19] Y. Shan, L. Hui, H. Zheng, Y. Qiaoling, y W. Yali, «Research on Imbalanced Classification Problem Based on Optimal Random Forest Algorithm», en *Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing*, PMLR, jul. 2024, pp. 383-392. Accedido: 3 de junio de 2025. [En línea]. Disponible en: https://proceedings.mlr.press/v245/shan24a.html

[20] H. Zhao, C. Zhang, y X. Huang, «Research on Improvement of Random Forest Algorithm Based on Oversampling and Feature Reduction», en *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, jul. 2024, pp. 48-51. doi: 10.1109/CISAT62382.2024.10695367.

[21] Y. Liu, Y. Wang, y J. Zhang, «New Machine Learning Algorithm: Random Forest», en *Information Computing and Applications*, B. Liu, M. Ma, y J. Chang, Eds., Berlin, Heidelberg: Springer, 2012, pp. 246-252. doi: 10.1007/978-3-642-34062-8_32.

[22] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, en KDD '16. New York,

NY, USA: Association for Computing Machinery, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.
[23] D. M. W. Powers, «Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation», *Journal of Machine Learning Technologies*, vol. 2, n.º 1, pp. 37-63, 2011. Accedido: 3 de junio de 2025. [En línea]. Disponible en: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf
[24] T. Fawcett, «An introduction to ROC analysis», *Pattern Recognition Letters*, vol. 27, n.º 8, pp. 861-874, jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[25] R. Fujinuma y Y. Asahi, «Default Factors in Motorcycle Sales in Developing Countries», en *Human Interface and the Management of Information: Visual and Information Design*, S. Yamamoto y H. Mori, Eds., Cham: Springer International Publishing, 2022, pp. 324-336. doi: 10.1007/978-3-031-06424-1_24.
[26] Y. Yang, M. He, y J. Zhang, «Comparative study of Optuna-optimized LightGBM, XGBoost and their hybrid stacking models in personal loan default prediction», en *Proceedings of the 5th International Conference on Artificial Intelligence and Computer Engineering*, en ICAICE '24. New York, NY, USA: Association for Computing Machinery, mar. 2025, pp. 705-710. doi: 10.1145/3716895.3717020.
[27] H. Huang, J. Li, C. Zheng, S. Chen, X. Wang, y X. Chen, «Advanced Default Risk Prediction in Small and Medum-Sized Enterprises Using Large Language Models», *Applied Sciences*, vol. 15, n.º 5, Art. n.º 5, ene. 2025, doi: 10.3390/app15052733.
[28] L. Wang, J. Cuiqing, W. Zhao, D. Yong, y N. Xiaoya, «Incorporating metadata: A novel variational neural topic model for bond default prediction», *Information Sciences*, vol. 715, p. 122219, oct. 2025, doi: 10.1016/j.ins.2025.122219.
[29] W. Lin y Y. Liu, «Deep Learning-Based Attention Mechanism Algorithm for Blockchain Credit Default Prediction», *International Journal of Advanced Computer Science and Applications (ijacsa)*, vol. 16, n.º 2, Art. n.º 2, 28 2025, doi: 10.14569/IJACSA.2025.0160241.
[30] N. Nguyen y D. and Ngo, «Comparative analysis of boosting algorithms for predicting personal default», *Cogent Economics & Finance*, vol. 13, n.º 1, p. 2465971, dic. 2025, doi: 10.1080/23322039.2025.2465971.
[31] Y. Zheng, «A Default Prediction Method using XGBoost and LightGBM», en *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, oct. 2022, pp. 210-213. doi: 10.1109/ICICML57342.2022.10009823.
[32] X. Zhu, Q. Chu, X. Song, P. Hu, y L. Peng, «Explainable prediction of loan default based on machine learning models», *Data Science and Management*, vol. 6, n.º 3, pp. 123-133, sep. 2023, doi: 10.1016/j.dsm.2023.04.003.
[33] Z. Xueping, S. M. Samuri, y M. H. M. Adnan, «Exploring the Performance of XGBOOST and Artificial Neural Network in Personal Credit Default Prediction: An Empirical Study», en *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, jul. 2023, pp. 1-6. doi: 10.1109/ICECCME57830.2023.10252429.