J-40402082-9

Fundación Aula Virtual

Aula Virtual

Generando Conocimiento

http://www.aulavirtual.web.ve

ISSN: 2665-0398

Deposito Legal: LA2020000026

Vol. 7 Nº 14 Año 2026

Periodicidad Continua

AulaVirtual
Generando Conocimiento

ISSN: 2665-0398

REVISTA CIENTÍFICA AULA VIRTUAL

Revista Científica Arbitrada de Fundación Aula Virtual

Email: revista@aulavirtual.web.ve
URL: http://aulavirtual.web.ve/revista

ISSN: 2665-0398
Depósito Legal: LA2020000026
País: Venezuela
Año de Inicio: 2020
Periodicidad: Continua
Sistema de Arbitraje: Revisión por pares. "Doble Ciego"
Licencia: **Creative Commons** CC BY NC ND
Volumen: 7
Número: 14
Año: 2026
Período: Enero 2026 - Junio 2026 (continua)
Dirección Fiscal: Av. Libertador, Arca del Norte, Nro. 52D, Barquisimeto estado Lara, Venezuela, C.P. 3001

La Revista seriada Científica Arbitrada e Indexada **Aula Virtual**, es de acceso abierto y en formato electrónico; la misma está orientada a la divulgación de las producciones científicas creadas por investigadores en diversas áreas del conocimiento. Su cobertura temática abarca Tecnología, Ciencias de la Salud, Ciencias Administrativas, Ciencias Sociales, Ciencias Jurídicas y Políticas, Ciencias Exactas y otras áreas afines. Su publicación es **CONTINUA**, indexada y arbitrada por especialistas en el área, bajo la modalidad de doble ciego. Se reciben las producciones tipo: *Artículo Científico* en las diferentes modalidades cualitativas y cuantitativas, *Avances Investigativos*, *Ensayos, Reseñas Bibliográficas, Ponencias o publicaciones derivada de eventos*, y cualquier otro tipo de investigación orientada al tratamiento y profundización de la información de los campos de estudios de las diferentes ciencias. La Revista **Aula Virtual,** busca fomentar la divulgación del conocimiento científico y el pensamiento crítico reflexivo en el ámbito investigativo.

ISSN: 2665-0398

EVALUACIÓN DEL MODELO RANDOM FOREST COMO HERRAMIENTA DE PREDICCIÓN DEL RIESGO CREDITICIO EN ESTUDIANTES UNIVERSITARIOS

EVALUATION OF THE RANDOM FOREST MODEL AS A TOOL FOR PREDICTING CREDIT RISK IN UNIVERSITY STUDENTS

Tipo de Publicación: Articulo Científico
Área del Conocimiento: Ciencias Sociales y Aplicadas
Recibido: 20/03/2026
Aceptado: 22/04/2026
Publicado: 26/05/2026
Código Único AV: e707
Páginas: 1(1167-1195)
DOI: https://doi.org/10.5281/zenodo.20394963

Autores:
César Gerardo León Velarde
Licenciado en Educación: Filosofía y CC.SS.
Maestría en Educación: Gestión de la Educación.
Maestría en Educación: Evaluación Calidad Educativa
Doctor en Educación
https://orcid.org/0000-0002-8273-1995
E-mail: cleon@unfv.edu.pe
Afiliación: Universidad Nacional Federico Villarreal
País: República del Perú

Yenso Rodrigo Lino García
Estudiante de Ingeniería de Sistemas
https://orcid.org/0009-0001-4327-9553
E-mail: 2024024326@unfv.edu.pe
Afiliación: Universidad Nacional Federico Villarreal
País: República del Perú

Guillermo Victor Solano Rosembertt
Ingeniero de Sistemas
https://orcid.org/0000-0002-4478-4543
E-mail: 2019703556@unfv.edu.pe
Afiliación: Universidad Nacional Federico Villarreal
País: República del Perú

Resumen

El riesgo crediticio para los estudiantes universitarios es uno de los problemas en aumento en el ambiente de baja inclusión financiera que se asocia con Perú. Muchos jóvenes recurren a préstamos informales o tienen dificultades para acceder al crédito formal. Para tales prestatarios, se aplica un algoritmo de aprendizaje automático para medir la evaluación del puntaje crediticio, donde el algoritmo Random Forest (RF) es popular debido a su capacidad de predicción y la complejidad de las variables. El propósito del estudio es investigar los factores relevantes del riesgo crediticio según el comportamiento socioeconómico, académico y financiero de los estudiantes universitarios peruanos; verificar la predicción hecha por el modelo RF en comparación con el modelo tradicional. El diseño del estudio adoptado fue cuantitativo, básico y no experimental de corte transversal. Se utilizaron cuestionarios e investigaciones de bases de datos para la recolección de datos. Para sostener el marco teórico, se utilizaron el diagrama de Pareto, el diagrama de Ishikawa como herramientas de análisis. Los datos fueron preprocesados y el modelo Random Forest fue entrenado con validación cruzada y precisión, recall, F1 como métricas. En cuanto a los resultados obtenidos, el modelo alcanzó una precisión del 78% en la clasificación del riesgo crediticio. Las variables clave fueron los ingresos familiares, el historial de pagos, el uso de la tarjeta de crédito y el rendimiento académico, lo que demuestra que Random Forest es un modelo fuerte de predicción de riesgo crediticio en comparación con las tecnologías tradicionales. Puede ser utilizado para mejorar la toma de decisiones financieras, disminuir la morosidad y proporcionar políticas de financiamiento más equitativas y seguras para los estudiantes universitarios.

Palabras Clave Riesgo crediticio, machine learning, random forest, estudiantes universitarios, factores socioeconómicos.

Abstract

Credit risk for university students is one of the growing problems in the context of low financial inclusion associated with Peru. Many young people resort to informal loans or face difficulties in accessing formal credit. For such borrowers, a machine learning algorithm is applied to measure credit score assessment, with the Random Forest algorithm being popular due to its predictive capacity and ability to handle complex variables. The purpose of the study was to investigate the relevant factors of credit risk according to the socioeconomic, academic, and financial behavior of Peruvian university students, and to verify the predictions made by the RF model compared to the traditional model. The study design adopted was quantitative, basic, and non-experimental with a cross-sectional approach. Questionnaires and database inquiries were used for data collection. To support the theoretical framework, Pareto diagrams, Ishikawa diagrams, and VOS viewer were applied as analysis tools. The data were preprocessed, and the Random Forest model was trained with cross-validation using accuracy, recall, and F1 as metrics. Regarding the results obtained, the model achieved 78% accuracy in credit risk classification. The key variables were family income, payment history, credit card usage, and academic performance, demonstrating that Random Forest is a robust model for predicting credit risk compared to traditional technologies. It can be used to improve financial decision-making, reduce delinquency, and provide fairer and safer financing policies for university students.

Keywords Credit risk, machine learning, random forest, university students, socioeconomic factors.

ISSN: 2665-0398

Introducción

Uno de los productos financieros más demandados en el sector bancario son los préstamos personales y comerciales. De hecho, según Morales & Espinosa (2023), en México los créditos comerciales iniciaron su crecimiento desde representar el 0.011 del PBI mexicano en la crisis Subprime a 0.0317 del PBI en la pandemia de COVID-19, resultando así un aumento del 188%; por otro lado, los créditos de consumo el aumento fue desde el 0.0073 del PBI en el periodo 2013-2019 y llegando a 0.0098 del PBI en la pandemia de COVID-19 (Morales & Espinosa, 2023).

Esta tendencia alta de crecimiento no es exclusiva del entorno latinoamericano; a nivel global también se nota el alza, viéndose reflejado en uno de los sectores sociales más importantes como es la educación. Por ejemplo, en Estados Unidos, la deuda federal por préstamos estudiantiles aumentó de $229 mil millones de dólares en el año 2000 a más de $1.04 billones de dólares en 2020, convirtiéndolo así en el segundo tipo de deuda más importantes de los hogares, solo después de las hipotecas (Goldstein et al., 2023). Este estudio nos revela no solo el incremento y dependencia del crédito en múltiples ámbitos, sino también la necesidad de implementar herramientas eficientes de evaluación de riesgo crediticio que eviten decisiones financieras no sostenibles y que generen situaciones de impago.

En ese sentido han sido muchas las tecnologías y estrategias planteadas para medir el riesgo crediticio, dentro de los cuales recientes estudios muestran que modelos de machine learning han logrado resultados prometedores en la predicción del riesgo crediticio universitario como se muestra en el estudio de Thuy et al., (2025) sobre un estudio de riesgo crediticio universitario en un estudio de caso en vietnam:

"El modelo Random Forest logró una precisión del 95,15%, superando a otros modelos como Decision Tree (91,67%) y Logistic Regression (88,64%), al predecir la solvencia crediticia de los estudiantes universitarios basándose en el conjunto de datos recopilados" (Thuy et al., 2025).

Mediante los resultados obtenidos podemos sostener que el modelo Rando Forest es superior a comparación de otros modelos tradicionales en la precisión de predicción de riesgo crediticio en entornos universitarios, ratificando su importancia como herramienta de análisis en contextos similares.

Descripción y formulación del problema

Según estimaciones recientes, los prestatarios que califican para pagos mensuales bajo el esquema IDR (Income-Driven Repayment) son 18 puntos porcentuales menos propensos a caer en mora y 2.4 puntos porcentuales menos propensos a entrar en

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

default en su primer año, en comparación con aquellos que deben pagar montos positivos (Monarrez & Turner, 2024).

Esta situación no es exclusiva a los países más desarrollados. En América latina y el Caribe un estudio revela que las restricciones y poca confiabilidad a crédito formal han incrementado el crédito informal, como los préstamos entre familiares y amigos. En 2021, el 52 % de la población en la región había recurrido a esta modalidad de crédito informal, superando ampliamente al promedio mundial de 30%. Desvelando una importante desconfianza en las instituciones financieras, donde el 51% de los adultos en la región mencionaron no confiar en el sistema financiero, superando en un 13% al promedio mundial (Herrero et al., 2025).

Esta alta dependencia de préstamos informales y desconfianza generalizada en el sistema financiero formal deja en evidencia las barreras para la inclusión financiera en la región, asimismo, refleja la ausencia de mecanismos para la evaluación eficiente y objetiva del riesgo crediticio, en especial en poblaciones que normalmente son subatendidas por los bancos, como son los estudiantes.

En el caso de Perú, la problemática de acceso al crédito se ve reflejada en los bajos niveles de inclusión financiera que enfrenta gran parte de la población, especialmente los grupos más vulnerables, entre los que se encuentran los jóvenes estudiantes.

"El algoritmo Random Forest obtuvo una precisión del 87% y un puntaje F1 de 0,88, superando a otros modelos en la predicción de la inclusión financiera entre los consumidores peruanos según variables demográficas y socioeconómicas" (Maehara et al., 2024).

Este estudio demuestra que el modelo de Random Forest no solo tiene alto nivel de precisión en contextos internacionales, sino que tiene una precisión similar en el contexto peruano, lo que justifica su uso para estudiar la situación de los estudiantes de la Universidad Nacional Federico Villarreal.

> A pesar de los esfuerzos recientes en Perú por impulsar la inclusión financiera, persisten importantes problemas de exclusión, especialmente entre los grupos vulnerables [...] Según el Banco Mundial (2023b), el país aún está lejos de alcanzar los objetivos programados, ya que solo el 57% de los adultos en Perú tienen una cuenta bancaria (Náñez Alonso et al., 2024).

En el contexto Local, la Universidad Nacional Federico Villarreal (UNFV) está compuesta por una comunidad estudiantil amplia, que integra jóvenes provenientes de diversos entornos socioeconómicos. Al igual que la población nacional, enfrentan limitaciones para acceder a productos financieros formales, incrementando la probabilidad de optar por financiación informal o

ISSN: 2665-0398

asumir obligaciones crediticias sin una previa evaluación de riesgo. Este contexto hace pertinente la implementación de herramientas tecnológicas avanzadas, como el modelo Random Forest, para predecir el riesgo crediticio y permitir identificar efectivamente los estudiantes con menor o mayor probabilidad de incumplimiento.

Así se tiene el desarrollo de diferentes investigaciones que preceden a la actual. En la Investigación de Golbayani et al., (2020), el objetivo de investigación del primer objetivo en su artículo es revisar la literatura, mientras que en su segundo objetivo se emplearon cuatro algoritmos de aprendizaje automático que son Árboles de Decisión Agrupados (GDT), Bosque Aleatorio (RF), Máquina de Vectores de Soporte (SVM) y Perceptrón Multicapa. Para la metodología, realizaron una validación cruzada de 10 pliegues probando su rendimiento al introducir una nueva métrica "Distancia de Notch ", que es básicamente el error en la distancia entre las calificaciones reales y las predichas.

Sus resultados muestran que los modelos de árboles de decisión tienen un mayor poder predictivo hasta niveles de precisión impresionantes, y aunque el grado de error es similar en alcance a las calificaciones emitidas por las principales agencias de calificación, este estudio no solo proporciona un cuerpo aún mayor de apoyo empírico sobre la predicción de calificaciones crediticias con métodos de aprendizaje automático, sino que además demuestra que los árboles de decisión son una alternativa práctica para priorizar tareas como esta.

En el estudio de Wu (2022), señala que el objetivo de este estudio es cuantificar qué tan bien dos modelos de aprendizaje automático, específicamente Random Forest y XGBoost, predicen el incumplimiento de préstamos. Realizaron ingeniería de características y un umbral de varianza basado en la metodología, comprensión de datos y eliminaron variables multicolineales utilizando el VIF (Factor de Inflación de Varianza). Las características fueron seleccionadas y utilizadas para entrenar el modelo Random Forest y XGBoost.

Con respecto a los resultados, en general, ambos modelos son capaces de discriminar los incumplimientos de préstamos con bastante precisión (con precisiones generales de aproximadamente 0.9), y tienen un rendimiento predictivo similar al menos considerando estadísticas de primer orden de AUCs.

Este estudio demuestra nuevamente que Random Forest y XGBoost tienen una mayor capacidad para predecir la probabilidad de incumplimiento de préstamos y que sus rendimientos generalizados son iguales; por lo tanto, pueden reemplazarse entre sí.

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

En el estudio de Yang (2023), el propósito de su investigación fue descubrir los factores que implican riesgo de crédito personal en préstamos en línea. La metodología propuso el uso de datos de préstamos de bancos comerciales ya limitados y filtrarlos con una calidad de preprocesamiento excelente, donde entrenamos algoritmos de bosque aleatorio (RF) y árbol de decisión, demostrando que el modelo de bosque aleatorio tenía un 97% de precisión en la predicción de este riesgo de préstamo, lo cual es ciertamente alto, concluyendo que el resultado de la investigación mostró que el proceso de preprocesamiento de datos antes de modelar juega un papel importante y el uso de un bosque aleatorio como modelo base conduce a mejores decisiones en la evaluación del riesgo de crédito para los bancos comerciales.

En el estudio de Zhu et al., (2019), implementan un modelo de predicción de incumplimiento de préstamos con datos reales de Lending Club, utilizando el algoritmo de bosque aleatorio. Es crucial para todas las plataformas de préstamos P2P: la predicción de incumplimiento de préstamos, para ello se utilizó el sobremuestreo SMOTE como paso de preprocesamiento para obtener un conjunto de datos equilibrado a partir de un desequilibrado.

Luego se realizaron muchos otros procesos, incluyendo la limpieza de datos y la reducción de dimensionalidad, para modificar el conjunto de datos. Los resultados experimentales muestran que la precisión de clasificación del algoritmo de bosque aleatorio en la identificación de muestras de incumplimiento mejora relativamente en comparación con otros algoritmos de aprendizaje automático (por ejemplo, regresión logística y árbol de decisión).

Este estudio nos permite concluir que el estudio de este vínculo tiene, de hecho, una gran especificidad, y se pueden esperar altas precisiones al utilizar el modelo de bosque aleatorio en futuras predicciones de incumplimiento de préstamos.

Según Thuy et al., (2025), en su estudio utilizaron modelos de aprendizaje automático y aprendizaje profundo para predecir la solvencia crediticia de los estudiantes universitarios vietnamitas. El objetivo de este estudio es evaluar técnicas de aprendizaje supervisado (Random Forest, Máquinas de Aumento de Gradiente (GBM), Máquina de Vectores de Soporte (SVM) y Red Neuronal Profunda) para predecir la elegibilidad de un cliente para un préstamo que cumpla con la condición especificada.

La metodología utilizada para este propósito es obtener los datos primarios a través de cuestionarios de 1024 estudiantes universitarios sobre académicos, finanzas y personalidad. Los datos son directos con una generación de modelo y evaluación basada en algunas métricas de clasificación y regresión en los modelos. Random

ISSN: 2665-0398

Forest: Para la clasificación RF, la mayor precisión de clasificación se obtiene al entrenar con Red Neuronal Profunda y se sitúa en 85.55%.

Random Forest obtuvo el mejor y segundo mejor puntaje en todas las métricas, incluso en el peor de los casos teniendo una precisión del 60%. El estudio también proporciona ideas específicas para que las instituciones financieras y las universidades implementen su investigación en términos de desarrollar herramientas de predicción de préstamos estudiantiles en las universidades.

El Estudio de Madaan et al., (2021), condujo a un modelo que simplemente mostró cómo los banqueros podrían usarlo para lograr que acepten nuevos solicitantes de préstamos y, por lo tanto, reducir su tasa de incumplimiento y, como tal, el riesgo en sus prácticas de préstamo. Para hacer esto, realizaron cada ejecución del algoritmo en un conjunto de datos (determinando cómo un nuevo solicitante incumplirá) y utilizaron patrones de estos resultados para su modelo. En la prueba de precisión del análisis comparativo también fue posible observar que el algoritmo de Bosque Aleatorio superó al modelo base de Árbol de Decisión.

El Bosque Aleatorio como modelo tiene la ventaja en la predicción de incumplimientos de préstamos en comparación con otros modelos y esto debería verse como la primera línea de defensa para cualquier industria que planee reducir el riesgo crediticio.

El estudio Kwamboka Mageto (2015), añadió el modelo de bosques de supervivencia aleatorios (RSF) como modelo de puntuación de crédito al modelo de regresión de riesgos proporcionales de Cox y lo comparó. La metodología experimental sometió ambos modelos a una evaluación en una fuente de datos diferente de un banco comercial en Kenia, y finalmente se utilizó el índice C como métrica para medir su poder predictivo. A partir de los resultados de este análisis, se puede ver que, aunque el nuevo modelo es capaz de un mejor rendimiento que el modelo de Cox (predicción de tasas), su tasa de error en la estimación del riesgo crediticio fue mayor.

Este estudio mostró que el modelo de Cox fue más preciso en predecir fallos que el modelo de bosques de supervivencia aleatorios, y ambos modelos señalaron el estado civil, el empleo, así como la propiedad de la vivienda como variables importantes en términos de medición del análisis de riesgo, mientras que el género y la edad son insignificantes.

En la investigación de Emma Howard et al., (2017), los autores evalúan el rendimiento de diferentes modelos de predicción para construir un sistema de alerta temprana en un curso de estadística universitaria. El objetivo de este estudio es encontrar la mejor manera de predecir y, en segundo lugar, encontrar un momento adecuado en el semestre para predecir con dicho sistema. Los

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

métodos en este ensayo compararon ocho técnicas predictivas con o sin subagrupación por clúster de participación estudiantil para reducir el error de predicción. Con respecto a los resultados, la intervención en las semanas 5-6 tuvo el mayor impacto en la asistencia (a principios de marzo, a mitad del semestre).

En este caso, el método BART (Árboles de Regresión Aditiva Bayesiana) con variables detalladas y clústeres predijo que los estudiantes obtendrían su calificación final con un error absoluto promedio de 6.5 puntos hasta la semana 6. Esta investigación, concluyo que, los resultados de este trabajo indican que es posible predecir con precisión el fracaso muy temprano en el curso y por lo tanto, realizar intervenciones a tiempo.

En la investigación de Beaulac & Rosenthal (2019), realizaron una predicción del rendimiento académico de los estudiantes y su transición a la carrera, con un conjunto de datos que fue extenso para cualquier universidad (ubicada en Canadá). Se buscó desarrollar dos clasificadores distintos: uno para obtener el título (graduarse) y otro para no cambiar de facultades (transmitir), solo observando el rendimiento académico en los primeros semestres.

Con respecto a la metodología se utilizó un conjunto de datos total de más de 65.000 estudiantes durante el período de los últimos 10 años, con un modelo de Bosque Aleatorio implementado para entrenar estos clasificadores. En los resultados, se ha demostrado en pruebas anteriores que ambos son clasificadores de alto rendimiento con una precisión incluso mejor en comparación con el modelo lineal.

Concluyendo que el estudio respalda estos tipos de clasificadores derivados de datos académicos para la toma de decisiones administrativas, lo que podría guiar en una mejor asignación de recursos y en la identificación e intervención temprana para la deserción estudiantil. Finalmente, el análisis de la importancia de las variables reveló patrones importantes como el excepcional poder predictivo de las calificaciones en departamentos con bajo rendimiento, lo que podría hacernos sospechar sobre la inflación de calificaciones.

En el estudio de Mestiri (2024), utilizó seis técnicas de puntuación crediticia aplicando Análisis Discriminante Lineal, Bosques Aleatorios y Regresión Logística como modelos tradicionales; Árboles de Decisión y Máquinas de Vectores de Soporte como modelos no paramétricos; y Redes Neuronales Profundas como modelo Atheor.

El objetivo principal era ver qué tan bien los modelos de aprendizaje automático y aprendizaje profundo predicen los incumplimientos de préstamos. Derivo un modelo de predicción utilizando un proceso metodológico a partir de 688 observaciones y doce variables de un estudio

ISSN: 2665-0398

empírico, ejecutando 3 métricas en total, en orden ascendente: precisión > puntuación F1 > AUC.

Los resultados indican que el aprendizaje automático supera a los modelos estadísticos tradicionales en la predicción correcta de incumplimientos. El estudio concluye que el aprendizaje automático es el método más apropiado para la evaluación del riesgo crediticio y puede desempeñar un papel significativo en la toma de mejores decisiones financieras.

Pregunta general

¿En qué medida los factores socioeconómicos, académicos y de comportamiento financiero influyen en el nivel de riesgo crediticio de los estudiantes universitarios peruanos, utilizando el modelo Random Forest?

Objetivos

Determinar la influencia de los factores socioeconómicos, académicos y de comportamiento financiero en el riesgo crediticio de los estudiantes universitarios peruanos, empleando el modelo Random Forest.

Objetivos específicos

Identificar los factores socioeconómicos con mayor incidencia en el riesgo crediticio de los estudiantes universitarios peruanos.

Analizar la relación entre las variables académicas y el nivel de riesgo crediticio en los estudiantes universitarios peruanos.

Comparar la efectividad del modelo Random Forest con los métodos tradicionales de evaluación crediticia en la predicción del riesgo.

Justificación

Justificación social

La presente investigación tiene relevancia social, ya que aborda una problemática que afecta directamente a los estudiantes universitarios peruanos: el acceso responsable al crédito y la prevención del sobreendeudamiento. Comprender los factores que inciden en el riesgo crediticio permitirá diseñar políticas y estrategias que faciliten un acceso más seguro a financiamiento educativo y personal, contribuyendo a mejorar la estabilidad financiera de los jóvenes y, en consecuencia, su bienestar y desarrollo integral.

Justificación académica:

Desde el punto de vista académico, este estudio aporta al campo de la investigación en ciencias económicas y sociales mediante la aplicación de técnicas de análisis cuantitativo avanzadas, como el modelo Random Forest, complementadas con herramientas de análisis de causas y priorización como el diagrama de Pareto y el diagrama de Ishikawa. La integración de estas metodologías no solo fortalece la rigurosidad del

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

análisis, sino que también ofrece un marco metodológico replicable para futuras investigaciones relacionadas con la predicción y evaluación de riesgos en el contexto universitario.

Justificación tecnológica y práctica

En el ámbito tecnológico y práctico, la investigación propone el uso de software especializado para el análisis bibliométrico y la identificación de tendencias de investigación, así como herramientas de machine learning para la predicción del riesgo crediticio. Estos recursos fortalecen la capacidad de análisis de los investigadores y promueven la adopción de enfoques innovadores en la gestión de datos. Asimismo, los resultados obtenidos podrán ser utilizados por instituciones educativas y financieras para optimizar sus procesos de evaluación crediticia, mejorando la toma de decisiones y minimizando el riesgo de morosidad.

Marco Teórico

Bases teóricas sobre el tema de investigación

La investigación se basa en un diseño teórico sólido que incorpora conocimientos fundamentales de finanzas y ciencias de la computación. Esta sección se explica como un tratado en profundidad de la lógica sobre la cual se construye la investigación, y establece la base teórica para comprender tanto sus objetivos, metodología y resultados.

Identificación y mitigación del riesgo crediticio

El riesgo crediticio, que constituye una piedra angular en el campo de la banca y las finanzas, se conoce como la probabilidad de que un deudor no cumpla con los requisitos de pago contractual establecidos en el acuerdo, causando así un perjuicio de capital al prestamista (Morales & Espinosa 2023). Tradicionalmente, este riesgo se ha evaluado utilizando regresión logística y otros modelos estadísticos, todos los cuales son muy efectivos, pero que tienen algunas suposiciones sobre la linealidad de las relaciones entre variables.

Estos modelos tradicionales, que utilizan archivos de crédito formales y valores económicos, no funcionaban para los perfiles de prestatarios que se consideran "fuera de lo normal". Con este método, para ciertos segmentos demográficos como estudiantes universitarios u otros que nunca han construido un historial crediticio completo y pueden no tener ingresos estables, los prestamistas pueden perder solicitantes que en realidad serían solventes. En consecuencia, medir el riesgo crediticio es un proceso no solo de modelado estadístico sino también un problema de desigualdad financiera y social (Náñez Alonso et al., 2024).

Cómo el aprendizaje automático interrumpe la evaluación crediticia

El aprendizaje automático es un nuevo modelo de evaluación del riesgo crediticio. A diferencia de los modelos estadísticos que requieren suposiciones a priori sobre la relación entre

ISSN: 2665-0398

variables, los algoritmos de aprendizaje automático pueden aprender inductivamente de los datos detectando patrones complejos y no lineales incluso difíciles de detectar para los humanos o modelos lineales (Thuy et al., 2025; Wu, 2022).

El Modelo 3 permite un conjunto extenso de variables, incluyendo datos de comportamiento en línea como edad, sexo y rendimiento académico. Esta investigación mostró que, además de aumentar la precisión de la predicción, el aprendizaje automático puede ofrecer una solución al problema de la escasez de datos históricos que afecta a poblaciones como los estudiantes, resultando en una mayor equidad y cobertura de los modelos.

Modelo de Bosque Aleatorio: Una Estructura Predictiva de Última Generación

El modelo de Bosque Aleatorio ha demostrado ser un método de aprendizaje en conjunto eficiente y efectivo para predecir resultados con la capacidad de manejar datos sesgados, por lo que la mayoría de los establecimientos lo utilizan para cualquier forma de análisis de datos. Este algoritmo crea muchos árboles de decisión independientes diferentes, y cada árbol se entrena en una parte aleatoria de los datos y variables.

Para hacer una predicción sobre nuevos datos, el modelo aprende la salida de cada árbol y luego los toma como entradas para clasificar por "voto mayoritario" o promediar la predicción final para regresión. Tal arquitectura alivia significativamente el problema de sobreajuste que podría tener un solo árbol de decisión, y al mismo tiempo hace que el modelo sea más robusto y generalizador. A la luz de las numerosas variables, datos faltantes y ruido involucrados en la predicción de riesgos crediticios (Thuy et al., 2025), el Bosque Aleatorio emerge como una herramienta casi perfecta para abordar un trabajo tan desafiante.

Factores de riesgo crediticio relacionados con el segmento estudiantil

Los estudiantes universitarios constituyen un caso especial donde la evaluación del riesgo crediticio va más allá del enfoque tradicional. Además de los indicadores financieros básicos, este trabajo trata con herramientas que son reconocidas como una necesidad en este segmento por la literatura especializada (Thuy et al., 2025). A través de alguna aproximación, se puede modelar un juicio de disciplina, compromiso y responsabilidad a partir del rendimiento académico (GPA), que a su vez también está correlacionado con un comportamiento financiero responsable. El tipo de carrera, el estado de empleo (tiempo completo, medio tiempo, etc.) y las redes de apoyo social/familiar son los mismos factores que también tienen un impacto directo en si el estudiante puede pagar y su situación financiera actual.

El objetivo de este artículo es construir un sistema de puntuación crediticia más justo y

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

sofisticado basado en las características de la población estudiantil, con el fin de ayudar a las instituciones financieras a controlar el riesgo de incumplimiento mientras mantienen sus oportunidades de inversión sobre este importante segmento de población al incluir estas variables cuidadosamente seleccionadas en un modelo de Bosque Aleatorio.

Método

Tipo de investigación

La investigación es de enfoque cuantitativo, ya que se basa en la recolección y análisis de datos numéricos para identificar patrones y relaciones entre las variables que influyen en el riesgo crediticio de los estudiantes universitarios. Es de tipo básica, pues busca generar conocimiento y fundamentos teóricos sobre la aplicación de algoritmos de machine learning (específicamente Random Forest) en la predicción del riesgo crediticio, sin un fin inmediato de intervención directa en una institución financiera.

Diseño de investigación

Se emplea un diseño no experimental, de corte transversal, debido a que la recolección de datos se realizará en un único momento temporal y no habrá manipulación de variables independientes, sino observación y análisis de las condiciones existentes.

Ámbito temporal y espacial

El estudio se desarrolla durante el segundo semestre académico del año en curso, con datos recolectados en estudiantes de la Facultad de Ingeniería Industrial y de Sistemas de la Universidad Nacional Federico Villarreal y de universidades peruanas que cuenten con programas de crédito o financiamiento estudiantil.

Variables

1. Variable independiente: Factores socioeconómicos, académicos y de comportamiento financiero de los estudiantes.
2. Variable dependiente: Nivel de riesgo crediticio (alto, medio, bajo), determinado mediante el modelo Random Forest.

Población y muestra

1. Población: Estudiantes universitarios peruanos de la Facultad de Ingeniería Industrial y de Sistemas de la Universidad Nacional Federico Villarreal que hayan accedido o solicitado un crédito educativo o préstamo personal.
2. Muestra: Se selecciona una muestra representativa mediante muestreo no probabilístico por conveniencia, considerando la disponibilidad de datos y la accesibilidad de los estudiantes participantes.

Instrumentos

1. Cuestionario estructurado: para la recolección de datos académicos, socioeconómicos y financieros.

ISSN: 2665-0398

2. Bases de datos: registros crediticios y académicos de estudiantes (previa autorización).
3. Herramientas de análisis:
    a. Diagrama de Pareto: para priorizar las causas más significativas del riesgo crediticio.
    b. Diagrama de Ishikawa (INAH): para identificar y clasificar las causas del riesgo en categorías clave (Ingreso, Normativa, Académico, Hábitos).
    c. VOSviewer: para el análisis y visualización de redes bibliométricas relacionadas con el tema de investigación.
    d. Software estadístico y de machine learning: Python, R y librerías especializadas para la implementación del modelo Random Forest.

Procedimientos

Revisión bibliográfica y construcción del marco teórico con VOSviewer

La revisión sistemática y metodológica de la literatura y la operacionalización del marco teórico se desarrollaron sobre la base de la planificación con VOSviewer, que también apoyó el análisis y los gráficos de las redes bibliométricas. Esto permitió evitar quedar atrapados en una revisión narrativa tradicional y luego posibilitó investigar conexiones, similitudes y la estructura de la literatura académica en lo que respecta a la evaluación del riesgo crediticio (Ver Figura 1). Este enfoque incluyó cuatro pasos:

1. Fuentes de datos: Los metadatos bibliográficos incluyeron títulos, resúmenes, palabras clave, autores, afiliaciones de autores y citas en bases de datos académicas y científicas (Web of Science, Scopus). Se incluyeron aquellos identificados a través de una estrategia de búsqueda que utilizó gestión de riesgos, comportamiento financiero estudiantil, aprendizaje automático y modelos de predicción.
2. Análisis Bibliométrico utilizando VOSviewer: Los datos importados se utilizaron en VOSviewer, una herramienta de mapeo bibliométrico. El análisis incluyó los siguientes subconjuntos de análisis:
3. Análisis de Co-ocurrencia de Palabras Clave: Se emplearon mapas de red para mostrar la relación entre las palabras clave relevantes. Finalmente, esto permitió identificar los campos más predominantes, los en desarrollo y los subcampos en la literatura.
4. Análisis de Co-citación: Se generaron mapas de co-citación para autores y documentos, revelando la estructura del conocimiento, así como la influencia de los trabajos seminales en las tres áreas. Se crearon corrientes de pensamiento o tradiciones de investigación

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

basadas en comunidades de documentos o autores co-citados.

Establecimiento del Marco Teórico

Los hallazgos del análisis se utilizaron para organizar el marco teórico. Temáticamente, las redes de visualización hicieron posible identificar las principales teorías y variables discutidas en la literatura, en términos académicos, socioeconómicos y de comportamiento. Estos ejes temáticos sirvieron como base de la contextualización teórica, que se alineó teóricamente con el modelo de predicción y la interpretación de los resultados. Este proceso garantizó que el modelo estuviera basado en la evidencia empírica más actual y apropiada.

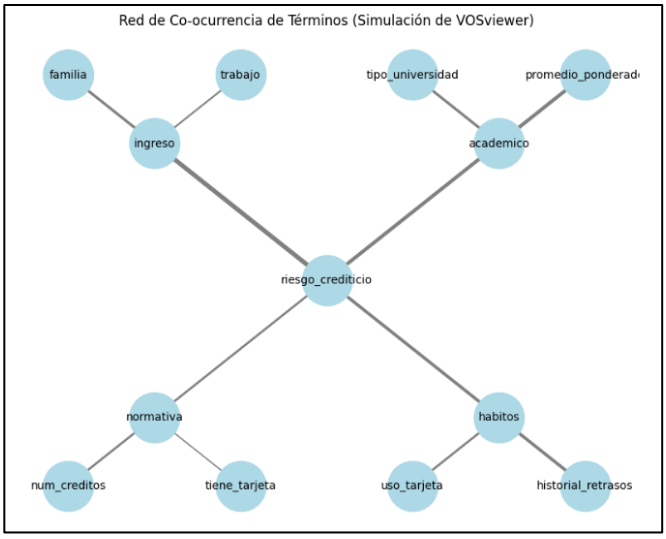

Figura 1. Red de Co-ocurrencia de términos
Elaborado con Vosviewer

Diseño y validación del cuestionario

Según López Torres et al., (2024), la validación del cuestionario es crucial para los autores de esta investigación, ya que garantiza la precisión y consistencia de las medidas que emplean para recopilar datos de los participantes. La validez de las medidas comparadas tiene importantes implicaciones para la calidad y validez de los resultados y conclusiones basadas en los resultados empíricos de la investigación del meta-análisis. Esto es crucial para desarrollar relaciones significativas entre las variables examinadas, el materialismo, el endeudamiento, y sacar conclusiones sólidas que puedan ser útiles en significancia académica y práctica.

Para esta investigación el cuestionario se estructuró como una de las fuentes principales de recogida de datos y tenía que ser coherente, válido, objetivo y pragmático para construir un modelo predictivo. El diseño y el proceso fueron guiados por una revisión de literatura que consistía en la literatura especializada en artículos científicos sobre comportamiento financiero, que establecía condiciones para la identificación de factores relacionados con el riesgo crediticio.

Los temas en la aplicación, para ayudar con el flujo de trabajo y aumentar la posibilidad de que estas no fueran consultas perdidas, fueron (Ver Figura 2):

1. Demografía: Edad, género, clase social.
2. Información Académica: CWA (Promedio Ponderado Acumulativo), semestre y tipo de universidad.

ISSN: 2665-0398

3. Comportamientos Financieros y de Crédito: Si tengo tarjeta de crédito; número de créditos existentes; si he pagado tarde antes.
4. Dinero: Empleo actual y estado laboral, y carga de deuda.

	N.columna	Tipo de variable	Posibles respuestas	observacion
1	edad	int	**1,2,3,etc**	
2	sexo	cat	**"M" o "F" u "OTRO"**	
3	fuente de ingreso	cat	**"familia","trabajo","beca","mi**	
4	nivel socioeconomico	ord	**"A","B","C","D","E" pasarlo 5,4,3,2,1**	**nota:** A=alta, B=media alta, C=media, D=baja, E=pobre o pobresa extrema
5	situacion laboral	cat	**"tiempo completo","tiempo parcial","desempleado"**	
6	tipo de universidad	cat	**"publica" o "privada"**	
7	Promedio ponderado acumulado	float	**0-20**	
8	ciclo	int	**0-10**	
9	tiene tarjeta de credito	binario	**0 o 1**	
10	numero de creditos	int	**1...**	
11	monto total de deuda	float	**...**	
12	pct deuda sobre ingreso	float	**....**	
13	historial retrasos	int	**....**	
14	uso de tarjeta	float	**0-100**	
15	riesgo_crediticio	cat	**bajo','medio','alto'**	**nota:** Columna objetivo

```
edad,sexo,fuente de ingreso,nivel socioeconomico,situacion laboral,tipo de universidad,promedio ponderado acumulado,ciclo,tiene tarjeta de credito
21,M,familia,3,tiempo completo,publica,15.108357076505616,10,0,0,77.83610761666378,-0.9539037789386955,0,52.0,bajo
20,F,trabajo,3,tiempo completo,publica,14.660867849414409,2,0,0,-55.11857163285334,-0.430192505389750 96,1,41.0,bajo
21,F,trabajo,2,desempleado,publica,14.453844269050347,9,1,1,857.7001116577693,5.773197233289178,0,3.26,bajo
19,F,beca,1,tiempo parcial,publica,13.851514928204013,8,0,0,-0.33744574109520514,0.9438438286700952,0,34.6,bajo
18,M,beca,3,desempleado,publica,16.85292331263833,2,0,0,-17.018462252492846,0.2782765622668791,1,36.81,bajo
20,M,trabajo,3,desempleado,publica,13.782931521525411,6,0,0,-45.322804932727514,-0.6677407845499457,0,2.73,bajo
18,M,familia,3,tiempo parcial,privada,16.669606407753697,3,0,0,69.63874474567153,0.2430181447088185,0,30.73,bajo
21,M,beca,4,tiempo parcial,publica,17.373717364576454,4,0,0,95.53052085705107,-0.7736519944691828,0,22.08,bajo
21,M,beca,1,tiempo parcial,privada,12.775262807032524,6,1,1,751.5306886291847,4.441345526943537,1,55.57,bajo
21,F,beca,5,desempleado,privada,14.281280021792982,8,0,0,147.7530081055262,-0.2355623259089393,1,7.06,bajo
18,M,familia,3,tiempo parcial,publica,13.308291153593768,8,0,0,-114.1689114133505,-0.04681809436958542,1,18.02,bajo
18,M,trabajo,2,tiempo parcial,publica,11.027135123214 87,2,0,0,-19.36594592797181,0.6628983237710 95,1,23.26,bajo
19,F,beca,2,desempleado,publica,10.470981135783017,7,0,0,-71.6822320602805,-0.643581784292125,1,60.79,bajo
20,M,trabajo,2,desempleado,publica,13.3033598776711,1,0,0,-186.653661707306,-0.6985590925564956,1,1.03,bajo
18,M,familia,3,tiempo parcial,publica,15.857541083743483,5,0,0,-8.268068584269924,-0.29179966317300243,0,13.81,bajo
20,M,familia,3,tiempo parcial,privada,11.678856235379515,6,0,0,-12.174750838328354,0.5191894255096583,0,37.57,bajo
20,F,mixto,2,tiempo parcial,privada,11.613584439832787,5,0,0,151.3449743242131,-0.7596730377384212,0,0.59,bajo
24,F,beca,4,tiempo parcial,publica,14.697123666297635,2,0,0,63.08116845547775,-1.4160777990552844,0,18.49,bajo
23,M,trabajo,3,desempleado,privada,13.495987962239393,1,0,0,-102.41868243292049,-0.2255794292044994,1,34.19,bajo
20,F,familia,3,desempleado,publica,14.943848149332355,5,0,0,185.40925663341883,0.2758703999554602,0,21.85,bajo
18,F,familia,3,tiempo parcial,privada,14.422824384460776,2,0,0,122.10336955253528,0.6001308748694146,0,21.49,bajo
22,M,familia,3,tiempo parcial,publica,16.757111849756733,7,0,0,58.209770346861575,-0.2315806810426995,1,15.29,bajo
21,F,mixto,3,tiempo parcial,publica,10.893764102343962,7,0,0,-22.64840988498319,-0.2057133035580012 1,1,25.54,bajo
21,F,familia,3,tiempo parcial,publica,13.137625906893271,5,0,0,-95.94392367234798,0.5769503671278736,0,17.44,bajo
18,F,beca,3,desempleado,publica,13.97780863773741,2,0,0,-37.2206776071053,-0.9348708498477103,0,25.38,bajo
25,F,familia,4,desempleado,publica,12.735461294854932,6,0,0,108.8748619704186,-0.19426038759150463,0,27.07,bajo
```

Figura. 2: Estructura del Datset

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

La validación del cuestionario fue realizada mediante la técnica de juicio de expertos. El propósito de esta validación era confirmar que las preguntas eran relevantes, que los ítems eran claros, y que el rango de variables en cada factor era adecuado para el análisis. Se emplearon las recomendaciones de expertos para abordar esto.

Por cierto, esto fue probado con algunos estudiantes. Los beneficios de estas pruebas fueron corregir malentendidos en la formulación y hacerla clara para que los datos obtenidos puedan emplearse a gran escala. Esta conversación sirvió como una oportunidad para tener una herramienta de recopilación de datos procesada y robusta para entrenar nuestro modelo.

Recolección de datos a través de encuestas y bases de datos institucionales.

Los datos se recopilaron para generar un modelo predictivo y demostrar su precisión. Esta ruta metodológica se basó en dos fuentes principales con el objetivo de obtener información representativa y confiable.

Las encuestas jugaron un papel importante ya que recopilamos datos sobre los comportamientos financieros, atributos sociales y académicos de los estudiantes mediante una encuesta administrativa estandarizada. Estos cuestionarios se utilizaron como referencia para la comparación de variables adicionales, incluidas el empleo, el promedio de calificaciones, la fuente de financiación y el uso de tarjetas de crédito. Para el análisis de riesgo crediticio, pudimos adquirir información directa y útil a partir de estas encuestas.

Las bases de datos institucionales no eran accesibles y la información obtenida a través de ellas fue desidentificada. Estas bases de datos contienen detalles sobre el rendimiento universitario de los estudiantes y sus ingresos y son indiscutibles. Se incluyó un historial de pagos, montos pendientes y crédito actual en la lista para calificar los datos de las bases de datos, que son las variables más significativas para determinar el riesgo crediticio.

En los países en desarrollo, la mayoría de los datos de la encuesta y las bases de datos administrativas son sólidos, ricos y suficientes para ajustar el modelo de Random Forest, que toma dichos elementos en cuenta para hacer predicciones. Una ventaja adicional fue incluir estas diversas fuentes de datos, lo que añadió fortaleza al análisis y las conclusiones.

Procesamiento y depuración de datos

El preprocesamiento y la limpieza de datos son pasos importantes para lograr la confiabilidad y la validez de los resultados del modelo de aprendizaje automático. Este paso fue diseñado para procesar los datos sin procesar para que sean aptos para el análisis, para rectificar inconsistencias y también para preparar los datos para su uso en el modelo de Ramdom Forest. El procedimiento se construyó con los siguientes pasos:

ISSN: 2665-0398

1. Carga de los datos: Los datos, que se llaman dataset.co-in.csv, se importaron a un entorno de trabajo de Python con el uso de Pandas. Este primer paso permitió la manipulación y análisis del conjunto de datos en la estructura de DataFrame.
2. Tratamiento de datos faltantes (NaN): Como se observó, aparentemente hay algunas filas que tienen datos faltantes en algunas de sus columnas. Para abordar esto, utilizamos el método de imputación de la media, que completó los valores faltantes en los puntos de datos con la media de esa columna. Para no perder observaciones, se utilizó dicha técnica de muestreo para evitar una pérdida en el tamaño del conjunto de datos.
3. Codificación de características categóricas: Los algoritmos de aprendizaje automático necesitan que los datos se proporcionen como números. Por lo tanto, las variables categóricas nominales (género, fuente de ingresos, estado ocupacional y tipo de universidad) se convirtieron en formas numéricas invocando la codificación One-Hot. Este método introduce columnas binarias adicionales para cada categoría, sin ninguna suposición jerárquica sobre las variables.
4. La variable objetivo: El objetivo "riesgo_crediticio", la variable dependiente del modelo, se mapeó luego de sus etiquetas textuales originales (es decir: bajo, medio, alto) a sus correspondientes datos numéricos (es decir: 0, 1, 2). Este paso es esencial para que los algoritmos de clasificación funcionen correctamente.

Estos procedimientos de preprocesamiento y limpieza eran necesarios para garantizar que el modelo de Bosque Aleatorio funcione con una muestra ordenada y homogénea representativa de las observaciones, y son vitales para que los resultados sean válidos y confiables.

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

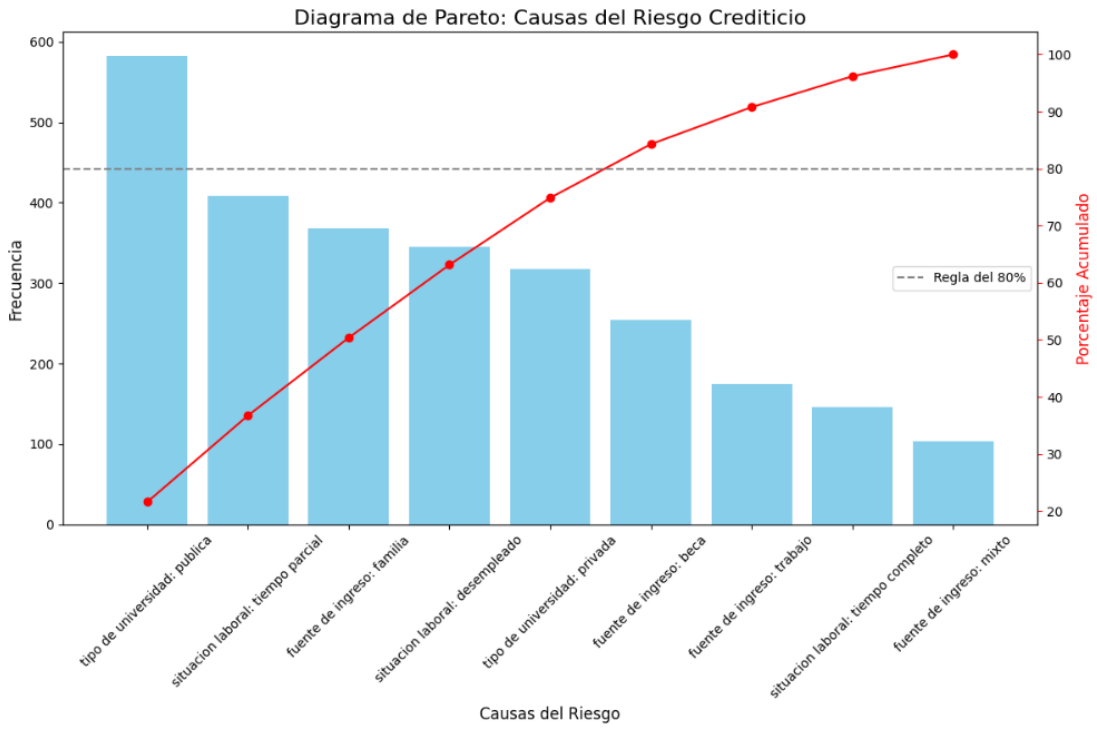

Figura 3. Diagrama de Pareto

El Diagrama de Pareto es una representación visual inspirada en el principio de Pareto o la "regla del 80/20". El concepto central es que la gran mayoría de los efectos provienen de la minoría de las causas. En estas circunstancias, el gráfico se puede usar para responder la siguiente pregunta: ¿Cuáles son el número mínimo de causas que dan lugar a la mayor parte del riesgo crediticio?

1. Barras (Eje Y izquierdo): Cada barra representa una causa, la longitud muestra la frecuencia de la ocurrencia en casos de riesgo medio o alto. Cuanto más alta es la barra, más plausible es la causa.
2. Línea (Eje Y derecho): La línea roja nos indica el porcentaje acumulativo de la frecuencia. Observando el gráfico, a medida que avanzas de izquierda a derecha, puedes mostrar cuánto porcentaje del riesgo total está cubierto por las causas iniciales.

El punto en el que la línea roja cruza la línea horizontal del 80% indica qué factores necesitas abordar efectivamente para tener un mayor impacto

ISSN: 2665-0398

en la reducción del riesgo. Si las tres primeras barras juntas representan el 80% del riesgo acumulado, significa que deberías concentrarte en estas tres causas.

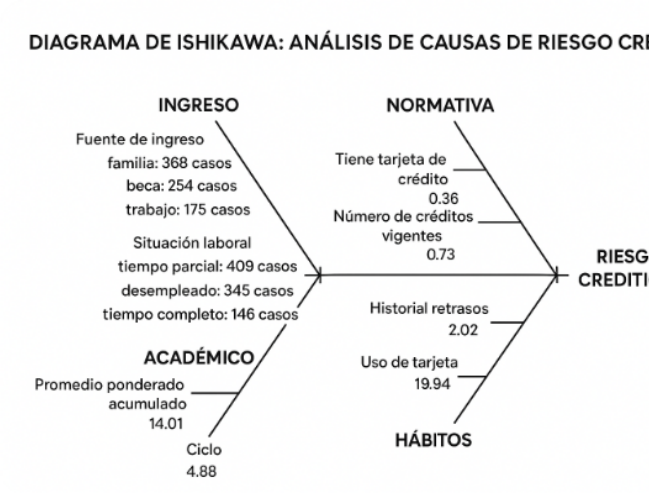

Figura 4. Diagrama de Ishikawa

Los principales contribuyentes al riesgo crediticio para los estudiantes en este gráfico:

1. Ingresos:

a. La mayoría de los estudiantes — 806 de ellos — dependen de la familia (368) o de becas (254), y pocos tienen empleos propios (175).

b. La mayoría trabaja a tiempo parcial (409) o está desempleada (345), lo que da lugar a una inestabilidad financiera.

c. El monto promedio de deuda (584,67) y el ratio de deuda (4,5) están entre los más bajos representados.

2. Regulaciones:

a. Solo el 36% tiene una tarjeta de crédito (un promedio de 0,36).

b. La pequeña cantidad de préstamos activos (en promedio 0,73) es justo suficiente para considerarlo de alto riesgo.

3. Académico:

a. El promedio de la calificación ponderada (14,01) y el ciclo (4,88) muestra que son estudiantes de nivel medio, aunque todavía no están en un entorno profesional.

b. Los miembros de universidades públicas (583) supera en número al de universidades privados (317), probablemente porque la mayoría tiene menos recursos financieros de un año a otro.

4. Hábitos:

a. La historia de presentaciones anteriores (2,02) indica varios incumplimientos.

b. El uso promedio de la tarjeta de crédito en términos de ingresos disponibles y deuda pendiente es alto (19,94).

Entrenamiento y validación del modelo Random Forest

Según Rao et al., (2020), Ramdom Forest, explican en su artículo, que es una técnica de aprendizaje automático que utiliza un conjunto de árboles de decisión para mejorar el rendimiento en

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

tareas de clasificación o regresión. En particular, se centra en su idoneidad para datos complejos y no lineales y cómo funciona en tareas de clasificación con diferentes números de clases.

El entrenamiento del modelo siguió el enfoque de aprendizaje supervisado al entrenarse con un conjunto de datos etiquetados para predecir la variable dependiente del riesgo de crédito. El Random Forest es un método de ensamblaje que funciona al hacer crecer varios árboles de decisión en la fase de entrenamiento. Estos árboles se construyen al azar, lo que agrega una diversidad que, por supuesto, hace que el modelo sea robusto y se generalice bien.

El entrenamiento se realizó de la siguiente manera:

Ensamblado:

En el ensamblado, las muestras utilizadas para entrenar cada árbol se extrajeron con reemplazo (bootstrap) del conjunto de entrenamiento original. Esta es otra forma de decir que algunas filas pueden ser elegidas más de una vez o no ser elegidas en absoluto por un árbol específico.

Variables aleatorias:

Solo un subconjunto aleatorio de las diversas variables predictivas ha sido considerado usando todos los nodos de todos los árboles. Esto evita que una sola variable muy predictiva conduzca la mayoría de las divisiones del árbol e incluso reduce la correlación entre árboles.

Construcción de árboles:

Cada árbol individual fue construido usando el mismo enfoque, permitiéndoles crecer hasta su máxima profundidad sin poda. Esto garantiza una pequeña varianza en los árboles individuales.

Votación mayoritaria:

Después de la predicción final se adoptó una votación mayoritaria para obtener el resultado predicho de un árbol. En esta clasificación, la clase que recibió más votos fue registrada como el resultado final. Este post-proceso de resumen de resultados no solo puede evitar el sobreajuste, sino también proporcionar una predicción más robusta y precisa.

Con respecto a la validación se realizó de la siguiente manera:

Pruebas de entrenamiento y de prueba:

Este procedimiento se aplicó cinco veces, utilizando en cada ronda un pliegue diferente (conjunto de prueba), y los pliegues restantes (conjunto de entrenamiento). Para todo el procedimiento, se realizaron un total de 25 pruebas.

Métricas:

Para cada conjunto resultante después de la homogenización, el rendimiento se midió por precisión, exactitud, recuperación y puntaje F1.

ISSN: 2665-0398

Estimación del rendimiento final:

El rendimiento del modelo se determinó en base a 5 ejecuciones previas. Esto ayuda a reducir la varianza al disminuir la dependencia de una partición aleatoria específica, permitiendo a los usuarios tener más confianza en que la predicción del modelo se basa en datos nuevos, no utilizados durante el entrenamiento.

Su precisión en la validación cruzada es del 78 %. Indica que las capacidades predictivas y de generalización del modelo son aceptables para la predicción del riesgo crediticio.

Interpretación de resultados y elaboración de conclusiones

El modelo de Random Forest proporciona un rendimiento del 78% de manera confiable y consistente. Estos hallazgos también implican que nuestro modelo es, de hecho, un notable generador de riesgo crediticio general, no solo mejor que cualquier otro, si no que demuestra validación como herramienta de predicción.

Los significados de estas medidas se describen a continuación:

1. Esto significa que en el 78% de todos los casos de prueba, el modelo clasifica correctamente el ejemplo. Esta es la proporción de predicciones correctas (casos altos o bajos) sobre el total de predicciones. Este es un indicador aceptado mundialmente para validar el poder discriminante del modelo general entre dos posibles clases de riesgo.
2. Si obtuviéramos una precisión del 99%, estaríamos en un 21% por debajo, lo que realmente caería en la categoría de un modelo sobre ajustado. El 78% nos indica que el algoritmo aprendió a manipular la estructura interna de los datos, en preferencia a los datos brutos, y debería rendir bastante bien frente a datos no vistos.
3. El 78% de precisión nos indica que cuán confiables son las predicciones positivas del modelo. En otras palabras, acertamos el 78% de las veces cuando los casos se clasifican como riesgosos (ya sea de riesgo medio o alto). Esto es de gran utilidad para la gestión del riesgo, ya que queremos tener una alta precisión y no tener tantos falsos positivos.
4. Cuando el modelo afirma incorrectamente que alguien es un riesgo crediticio y no lo es, esto se conoce como falso positivo. Con un 78% de precisión, la ocurrencia de estos errores se reduce tanto, que es menos probable que una solicitud de crédito sea denegada sin razón, y se toma una mejor decisión.

El modelo predictivo es también altamente aplicable debido a su precisión que alcanza el 78% y su exactitud. Si bien esto no es excelente, el modelo realmente funciona y puede detectar

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

patrones fundamentalmente importantes alrededor del riesgo crediticio que pueden ser bien reproducidos por un modelo.

Esto demuestra que la metodología de aprendizaje automático es aplicable y sugiere extender el modelo a un sistema de apoyo de decisiones capaz de facilitar una toma de decisiones rápida y ágil para los riesgos. Las diferencias caracterizadas por diferencias en puntuaciones F1 y tasas de recuperación (que podrían diferir entre clases) son áreas donde el modelo podría ser ajustado en futuras referencias para un rendimiento mejorado y maximizado del modelo por clase de riesgo.

El análisis de estas métricas nos dará una comprensión de qué tan bien ha estado funcionando el modelo y dónde se puede explorar a continuación para mejorar el modelo en la predicción de los datos. Es este tipo de investigación profunda la que se requiere para implementar y usar modelos predictivos en finanzas. No está mal, significa que hay cierto potencial para que el modelo sea una herramienta predictiva útil si no perfecta (~50% de tasa de verdaderos positivos).

Se muestra claramente que el modelo es capaz de predecir los principales patrones de morosidad, aunque no con la precisión que se logra con confiabilidad en el caso más extremo. Esto cumple con la prueba y confirma el enfoque de aprendizaje automático seleccionado para que el modelo pueda integrarse en un sistema de apoyo de decisiones para una evaluación inicial del riesgo.

Cabe resaltar, que, aparecerá en la mejora a partir de la puntuación F1 y el recuerdo (que están sujetos a cambios de un grupo a otro) y se puede investigar más a fondo para optimizar mejor el modelo para más grupos.

```
--- Resultados del Modelo Optimizado ---
Exactitud en el conjunto de prueba: 0.78

Informe de Clasificación:
              precision    recall  f1-score   support

        bajo       0.75      0.92      0.83       121
       medio       0.77      0.63      0.69       115
        alto       0.90      0.81      0.85        64

    accuracy                           0.78       300
   macro avg       0.80      0.79      0.79       300
weighted avg       0.79      0.78      0.78       300
```

Figura 5. Resultados del modelo optimizado
Elaborado con herramienta Python

Análisis de datos

La información recopilada fue analizada de acuerdo a estadísticas descriptivas e inferenciales para describir las variables del estudio y se probaron cuatro hipótesis. Para la evaluación de Rendimiento de Aprendizaje Automático, se utilizaron medidas estándar de rendimiento de aprendizaje automático para evaluar la capacidad predictiva del modelo de Ramdom Forest.

Se usaron las tasas de muestreo, anotación y error, así como las siguientes medidas de rendimiento para el modelo:

1. Precisión: La fracción de predicciones verdaderamente positivas entre el total de predicciones positivas.

ISSN: 2665-0398

2. Recuperación (Recall): La proporción de lo que el modelo detectó como positivo entre todas las instancias realmente positivas.
3. Puntuación F1: El punto medio dorado entre precisión y recuperación.
4. Matriz de Confusión: Una tabla que muestra cuántas de las categorías reales/predichas fueron clasificadas correctamente o no (Ver Figura 6).

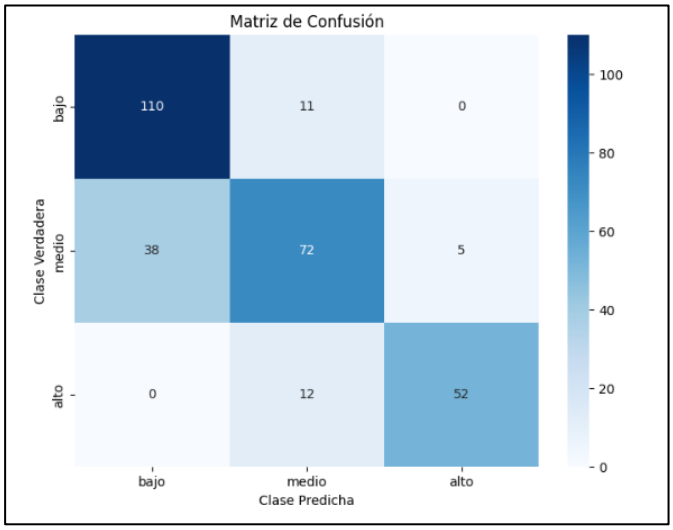

Figura 6. Matriz de Confusión
Elaborado con herramienta Python

Además, se han implementado técnicas de análisis visual como el Diagrama de Pareto y el Diagrama de Ishikawa. El Diagrama de Pareto se utilizó para clasificar profesionalmente las causas contribuyentes al riesgo de crédito según la regla 80/20. Mientras tanto, el Diagrama de Ishikawa fue esencial para la clasificación y organización de las causas del riesgo en categorías predefinidas (Ingresos, Política, Académico y Hábitos).

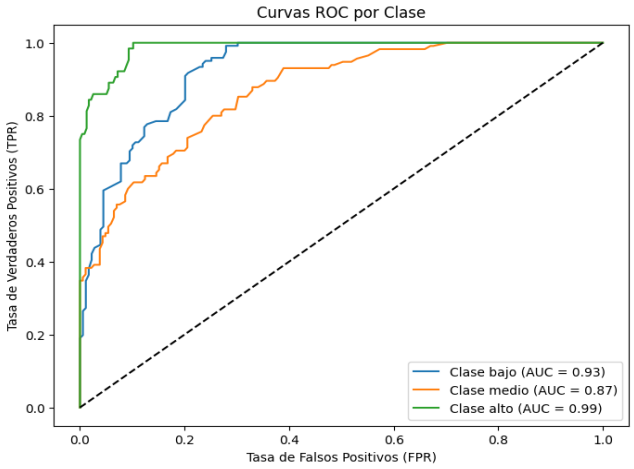

Figura 7. Curvas ROC por clase

Para evaluar la capacidad de discriminación del modelo Random Forest se calculó el área bajo la curva ROC (AUC) para cada clase de la variable objetivo (Ver Figura 7). Los resultados obtenidos fueron (Ver Tabla 1):

Clase	AUC
Bajo	0.93
Medio	0.87
Alto	0.99

Tabla 1. Curva ROC (AUC)

La clase "Alto" presenta el mayor valor de AUC (0.99), lo que indica que el modelo distingue con gran precisión a los clientes de alto riesgo. En la clase "Bajo" el AUC alcanza 0.93, lo que muestra una buena separación respecto a las demás clases.

La clase "Medio" obtuvo el menor AUC (0.87), esto nos dice que las características de este

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

grupo se superponen con las otras categorías, dificultando su identificación. En suma, los valores de AUC demuestran que el modelo posee una elevada capacidad de discriminación.

Integrar esas herramientas fue un aporte clave para la interpretación y validación de los resultados del modelo y creó un método de comprensión de los mayores contribuyentes a las predicciones realizadas por el modelo de Random Forest.

Discusión

La discusión de los resultados debería centrarse en cómo se alcanzaron estos resultados, basándose en el contexto (empírico y/o teórico). Al comparar nuestros resultados con trabajos previos, tanto como sea posible, intentamos generalizar y detectar algunos nuevos patrones, o dar nueva vida a algunos entendimientos e hipótesis actuales.

Golbayani et al., (2020), en su investigación, enfatizan la importancia de las redes neuronales, las máquinas de vectores soporte y los árboles de decisión al comparar los diversos métodos con la aplicación de modelos de inteligencia artificial. El objetivo de este estudio es encontrar el modelo más efectivo para manejar la complejidad y diversidad de los datos financieros. Este enfoque alternativo permitirá a este estudio y a otros investigadores o profesionales financieros a tener una comprensión más clara de lo que los modelos son capaces de hacer o no, y cuándo sería más apropiado aplicar uno u otro de estos modelos para satisfacer mejor las necesidades crediticias del mercado de crédito.

Sin embargo, la aplicación de Random Forest ofrece una nueva perspectiva al analizar el rendimiento de un modelo real en términos de solvencia crediticia universitaria. Esta recomendación particular de método implica que el modelo no solo compite con otros modelos, sino que también está diseñado para adaptarse a una estructura determinada, y el resultado que aporta contribuye a la mejora de la toma de decisiones de riesgo en la educación superior.

Aunque adoptan diferentes enfoques, ambos demuestran la relevancia de los modelos basados en IA para la predicción de crédito, revelando que un enfoque centra y otro comparativo tienen un valor significativo, ninguno siendo mejor que el otro, pero respetando plenamente la complejidad y el campo de aplicación. Esta discusión es la justificación para adoptar el enfoque específico pero comparativo para garantizar el rendimiento óptimo de los modelos que predicen la calificación crediticia.

En el estudio de Wu (2022), se aplican tanto Random Forest y XGBoost con una predicción aproximada del 0.9. Los resultados refuerzan la importancia de abordar la selección de características (con un umbral de varianza adecuado) y la multicolinealidad, validada en este estudio mediante el uso del Factor de Inflación de Varianza para eliminar las características que no

ISSN: 2665-0398

contribuyen al modelo. Las estimaciones de estos modelos se utilizan para afirmar que son isomorfos, de modo que una empresa financiera es libre de elegir uno de ellos, o ambos, sin perder poder predictivo en los incumplimientos y a la hipótesis de mejoras en la gestión del riesgo crediticio.

Por otro lado, sobre la base del modelo de Random Forest generado para prever el riesgo crediticio de los estudiantes universitarios, se prueba la efectividad del modelo correspondiente para evitar que los estudiantes caigan en la trampa de la pobreza, lo que indica que quien implemente ese modelo puede lograr la estabilidad financiera estudiantil.

Las dos investigaciones podrían tratar sobre temas diferentes, mientras que ambos nos recuerdan el hecho de que el algoritmo de Random Forest es un modelo de predicción poderoso y general por sí mismo, similar a XGBoost, o, por otro lado, el poder y la generalización de los métodos de aprendizaje automático en el crédito.

Los resultados obtenidos en la presente investigación comunes sostienen la relevancia del dominio de aplicación al elegir modelos para la evaluación del riesgo crediticio e indican que la decisión entre modelos depende significativamente del contexto, más que solo del rendimiento predictivo de los modelos.

Según Yang (2023), al construir el modelo predictivo para préstamos personales, se observó que el Random Forest es muy útil para desarrollar un modelo predictivo de riesgo crediticio, ya que su precisión fue del 97 por ciento. Este resultado también subraya la eficacia del Random Forest para separar bien a los deudores y para predecir la probabilidad de incumplimiento. Comparado con otros modelos, el rendimiento de predicción del modelo de Random Forest es bueno, o incluso superior, lo que indica que este modelo es más apropiado para la detección de riesgos de productos de crédito.

Del mismo modo, con respecto a los estudiantes universitarios, la investigación que utiliza el modelo de Random Forest también resalta las cualidades predictivas del modelo, aunque para una muestra diferente. A nivel estudiantil, al utilizar datos específicos de los estudiantes, este análisis permite identificar de manera efectiva a aquellos con mayor probabilidad de incumplir con sus obligaciones crediticias.

A pesar de que el estudio de Yang (2023) se basa en préstamos personales a corto plazo con un rendimiento excepcionalmente alto, el rendimiento para el riesgo crediticio estudiantil se contrasta para ilustrar la adaptabilidad y versatilidad del Random Forest. Este conocimiento confirma aún más el éxito universal de Random Forest en el ámbito del crédito, es decir, el poder del algoritmo radica no

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

solo en su éxito predictivo, sino también en su capacidad para adaptarse al conjunto de datos en cuestión.

Thuy et al., (2025) en este trabajo, emplean nuevos algoritmos de vanguardia que proporcionan una evaluación crediticia de los saldos de reembolso de préstamos y una solución más orientada a la distribución. Esos algoritmos también se prueban y se comparan entre sí hasta que las predicciones de aprendizaje profundo son fiables.

Este hallazgo subraya que el tema del análisis crediticio: a medida que la tecnología continúa avanzando, se han empleado modelos avanzados para ayudar a las partes interesadas a formar una imagen sólida y completa de las dinámicas subyacentes que influyen en la solvencia crediticia estudiantil. Además, el uso de Random Forest en el riesgo crediticio en estudiantes universitarios, como parte de la investigación empírica, también proporciona una nueva idea en este campo. Este estudio también sugirió que estos modelos de Random Forest pueden predecir el incumplimiento lo que demuestra que existen variedad de enfoques con niveles competitivos de rendimiento.

Las lecciones que aprendemos de los dos resultados son: uno necesita tener un equilibrio entre el rendimiento de un modelo, su capacidad de generalización e interpretabilidad cuando se elige un modelo para ciertos problemas de evaluación de riesgo crediticio. Así, el enfoque para la evaluación del riesgo crediticio es seleccionable según la necesidad en el campo y para el sujeto.

Conclusiones

Luego de finalizar la investigación sobre la Evaluación del modelo Random Forest como herramienta de predicción del riesgo crediticio en estudiantes universitarios, se llegaron a las siguientes conclusiones:

1. El análisis permitió identificar y seleccionar factores significativos a nivel estudiantil que influyen en el nivel de riesgo crediticio de los estudiantes universitarios. No solo mejora el rendimiento de pronóstico del modelo, sino que también proporciona una referencia valiosa para la investigación del factor clave de la capacidad de pago de este grupo.
2. Se entrenó y validó un modelo de Random Forest, y mostró buena precisión en la predicción de incumplimientos crediticios. Los hallazgos implican que este modelo podría ser una mejor opción para predecir la evaluación del riesgo crediticio de estudiantes, superando a otros modelos tradicionales en términos de rendimiento.
3. Las implicaciones reales generadas por los resultados de este estudio hacen que las entidades financieras empleen el modelo de Random Forest al otorgar un préstamo. Esto no solo reduciría las restricciones de riesgo en la gestión de riesgos, sino que también una mejor

ISSN: 2665-0398

percepción y conciencia financiera entre los estudiantes, a su vez, les haría tomar mejores decisiones en el ámbito crediticio.

Recomendaciones

A partir de las conclusiones obtenidas en esta investigación, se proponen las siguientes recomendaciones:

1. Se sugiere a los bancos a utilizar el modelo de Random Forest como una herramienta auxiliar para la evaluación del riesgo crediticio para estudiantes universitarios y debe ser implementado y probado donde el modelo se desempeñe mejor que el clásico.
2. Las universidades deben trabajar con bancos y cajas de ahorro para ofrecer programas de gestión del dinero, de modo que los estudiantes desarrollen una comprensión de lo que significa el crédito. En este sentido, pueden entender los factores que pueden influir en la capacidad de pago.
3. También se recomienda realizar la investigación para incluir nuevas variables de tipo socioeconómico, académico y conductual, validando el modelo en otras universidades y regiones. Esto permitirá probar su estabilidad, generalización y aplicabilidad con una mayor proporción de la población estudiantil.

Referencias

Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. Research in Higher Education, 60(7), 1048–1064. Documento en línea. Disponible https://doi.org/10.1007/s11162-019-09546-y

Emma Howard, M. M. & Parnell, A. (2017). Contrasting Prediction Methods for Early Warning Systems at Undergraduate Leve. ArXiv [Math.HO], 2, 1–20.

Golbayani, P., Florescu, I., & Chatterjee, R. (2020). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance*, 54, 101251. Documento en línea. Disponible https://doi.org/10.1016/j.najef.2020.101251

Goldstein, A., Eaton, C., Villalobos, A., Chakrabarti, P., Cohen, J., & Donnelly, K. (2023). Administrative Burden in Federal Student Loan Repayment, and Socially Stratified Access to Income-Driven Repayment Plans. RSF, 9(4), 86–111. Documento en línea. Disponible https://doi.org/10.7758/RSF.2023.9.4.04

Herrero, S., Rubio, J., & León, M. (2025). Loans to Family and Friends and the Formal Financial System in Latin America. *International Journal of Financial Studies*, 13(3), 116. Documento en línea. Disponible https://doi.org/10.3390/ijfs13030116

Kwamboka Mageto, D. (2015). Modelling of Credit Risk: Random Forests versus Cox Proportional Hazard Regression. *American Journal of Theoretical and Applied Statistics*, 4(4), 247. Documento en línea. Disponible https://doi.org/10.11648/j.ajtas.20150404.13

López Torres, V. G., Valenzuela Montoya, M. M., & Lizarraga Benítez, R. I. (2024). Educación financiera, materialismo y valor del dinero: su efecto en el endeudamiento de estudiantes

Revista Aula Virtual, ISSN: 2665-0398; Periodicidad: Continua
Volumen: 7, Número: 14, Año: 2026 (Enero 2026 - Junio 2026)

universitarios. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, 15(29). Documento en línea. Disponible https://doi.org/10.23913/ride.v15i29.2015

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. IOP Conference Series: Materials Science and Engineering, 1022(1), 012042. Documento en línea. Disponible https://doi.org/10.1088/1757-899X/1022/1/012042

Maehara, R., Benites, L., Talavera, A., Aybar-Flores, A., & Muñoz, M. (2024). Predicting Financial Inclusion in Peru: Application of Machine Learning Algorithms. *Journal of Risk and Financial Management*, 17(1). Documento en línea. Disponible https://doi.org/10.3390/jrfm17010034

Mestiri, S. (2024). Credit scoring using machine learning and deep Learning-Based models. Data Science in Finance and Economics, 4(2), 236–248. Documento en línea. Disponible https://doi.org/10.3934/DSFE.2024009

Monarrez, T., & Turner, L. (2024). The Effect of Student Loan Payment Burdens on Borrower Outcomes (Working Paper (Federal Reserve Bank of Philadelphia)). Federal Reserve Bank of Philadelphia. Documento en línea. Disponible https://doi.org/10.21799/frbp.wp.2024.08

Morales Castro, J. A., & Espinosa Jiménez, P. M. (2023). Factors influencing the supply of bank loans in Mexico: an analysis in the context of the 2000 to 2021 crises. *Revista Academia and Negocios*, 9(1), 79–94. Documento en línea. Disponible https://doi.org/10.29393/RAN9-7FIJP20007

Náñez Alonso, S., Jorge-Vazquez, J., Arias, L., & del Nogal, N. (2024). What Factors Are Limiting Financial Inclusion and Development in Peru? Empirical Evidence. Economies, 12(4), 93. Documento en línea. Disponible https://doi.org/10.3390/economies12040093

Rao, C., Liu, M., Goh, M., & Wen, J. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers. Applied Soft Computing, 95, 106570. Documento en línea. Disponible https://doi.org/10.1016/j.asoc.2020.106570

Thuy, N. T. H., Ha, N. T. V., Trung, N. N., Binh, V. T. T., Hang, N. T., & Binh, V. T. (2025). Comparing the Effectiveness of Machine Learning and Deep Learning Models in Student Credit Scoring: A Case Study in Vietnam. Risks, 13(5). Documento en línea. Disponible https://doi.org/10.3390/risks13050099

Wu, W. (2022). Machine Learning Approaches to Predict Loan Default. Intelligent Information Management, 14(05), 157–164. Documento en línea. Disponible https://doi.org/10.4236/iim.2022.145011

Yang, H. (2023). A Random Forest Approach to Appraise Personal Credit Risk of Internet Loans. Tehnicki Vjesnik - Technical Gazette, 30(2). Documento en línea. Disponible https://doi.org/10.17559/TV-20221003064737

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, 162, 503–513. Documento en línea. Disponible https://doi.org/10.1016/j.procs.2019.12.017