Artículo de investigación

Revista de Educación Simón Rodríguez

www.revistasimonrodriguez.org

Vol. 6 | No. 11 | Febrero - Julio 2026 | ISSN: 3006-1385 | ISSN-L: 3006-1385 | Pág. 28 - 40

Predicción del rendimiento académico mediante minería de datos en estudiantes de estadística

Academic performance prediction through data mining in statistics students

José Carlos Fiestas Zevallos
jfiestasz@unp.edu.pe
https://orcid.org/0009-0008-7860-5911
Universidad Nacional de Piura. Piura, Perú

Ricardo Antonio Armas Juárez
rarmasj@unp.edu.pe
https://orcid.org/0000-0002-0048-2711
Universidad Nacional de Piura. Piura, Perú

Ramón Cosme Correa Becerra
rcorreab@unp.edu.pe
https://orcid.org/0000-0002-3656-1788
Universidad Nacional de Piura. Piura, Perú

Nataly Carmencita Ruiz Cortez
natyruizcortez@gmail.com
https://orcid.org/0000-0002-1799-676X
Universidad Nacional de Piura. Piura, Perú

Felipe Ramón Ramos Echevarría
Ferare17@hotmail.com
https://orcid.org/0000-0003-1315-5725
Universidad Nacional de Piura. Piura, Perú

Artículo recibido: 17 de noviembre de 2025/Arbitrado: 15 de diciembre de 2025/Aceptado: 12 de enero 2026/Publicado: 02 de febrero de 2026

https://doi.org/10.62319/simonrodriguez.v.6i11.112

RESUMEN

La predicción del rendimiento académico universitario mediante técnicas de minería de datos ha emergido como una herramienta para optimizar los procesos educativos y mejorar los resultados estudiantiles. El objetivo del estudio es utilizar minería de datos para predecir el rendimiento académico en estudiantes de estadística de la Universidad Nacional de Piura, 2010-2018. La metodología es tipo aplicada, enfoque cuantitativo, diseño no experimental, longitudinal retrospectivo, población de 510 registros académicos, muestra censal, instrumentos: Sistema Integrado de Gestión Académica y IBM SPSS v.27, procedimientos de depuración, normalización y partición de datos, análisis mediante redes neuronales artificiales y regresión lineal múltiple con validación de supuestos. Los resultados muestran que la regresión lineal múltiple fue más efectiva para promedio ponderado (CME = 0.761, R^2 = 95.3%), mientras las redes neuronales demostraron mayor eficacia para notas específicas (CME = 1.095). El grado de dificultad 1 del curso fue la variable más importante (100% importancia normalizada). Se concluye que, ambas técnicas son complementarias y viables para la predicción del rendimiento académico, proporcionando evidencia empírica para sistemas de apoyo estudiantil basados en analítica educativa.

Palabras clave:
Minería de datos; Rendimiento académico; Redes neuronales; Regresión múltiple; Predicción educativa

ABSTRACT

Predicting university academic performance using data mining techniques has emerged as a tool to optimize educational processes and improve student outcomes. The objective of this study is to use data mining to predict the academic performance of statistics students at the National University of Piura, from 2010 to 2018. The methodology is applied, with a quantitative approach, a non-experimental, retrospective longitudinal design, a population of 510 academic records, a census sample, and the instruments used were the Integrated Academic Management System and IBM SPSS v.27. Data cleaning, normalization, and partitioning procedures were employed, followed by analysis using artificial neural networks and multiple linear regression with assumption validation. The results show that multiple linear regression was more effective for weighted averages (CME = 0.761, R^2 = 95.3%), while neural networks demonstrated greater effectiveness for specific grades (CME = 1.095). The course difficulty level (level 1) was the most important variable (100% normalized importance). It is concluded that both techniques are complementary and viable for predicting academic performance, providing empirical evidence for student support systems based on educational analytics.

Keywords:
Data mining; Academic performance; Neural networks; Multiple regression; Educational prediction

INTRODUCCIÓN

La predicción del rendimiento académico universitario mediante técnicas de inteligencia artificial y minería de datos ha experimentado un crecimiento notable en la última década, estableciéndose como una disciplina clave en la transformación digital de la educación superior (Miranda et al., 2024). A nivel internacional, múltiples estudios han demostrado que los modelos predictivos basados en machine learning alcanzan precisiones superiores al 90% en la predicción del éxito académico estudiantil, como evidencian estudios realizados en universidades alemanas que implementaron sistemas de redes neuronales profundas para el análisis del rendimiento estudiantil (Kalita, 2025).

En el contexto asiático, investigaciones en instituciones de educación superior de India y China han revelado la efectividad de algoritmos de ensemble learning combinados con redes neuronales profundas para la predicción del rendimiento académico, alcanzando accuracies del 97% en cohortes estudiantiles de ingeniería (Pan et al., 2025). De manera complementaria, universidades australianas han aplicado técnicas de Long Short-Term Memory (LSTM) con modelos Bi-LSTM para la predicción del rendimiento semestral, logrando identificar con antelación a los estudiantes en riesgo académico (Zhang et al., 2024).

En América Latina, la investigación también ha avanzado significativamente en esta línea. Investigaciones realizadas en universidades brasileñas han desarrollado enfoques híbridos de inteligencia artificial para evaluar comparativamente diversos algoritmos de machine learning en contextos educativos locales (Córdova-Esparza et al., 2025). En Ecuador, modelos interpretables de machine learning para la toma de decisiones académicas, alcanzando un R^2 de 0.910 con técnicas de XGBoost y Random Forest, mientras que investigaciones brasileñas han evidenciado que las características curriculares específicas tienen mayor peso predictivo que factores socioeconómicos (Guevara-Reyes et al., 2025). Adicionalmente, estudios realizados en universidades mexicanas han demostrado la aplicabilidad de técnicas de deep learning para la predicción del rendimiento estudiantil en programas de ingeniería, evidenciando que los modelos basados en redes neuronales artificiales pueden superar significativamente a los métodos estadísticos tradicionales (Pan et al., 2025).

La literatura reciente sugiere una tendencia internacional hacia enfoques multidisciplinarios que integran diversas técnicas predictivas. Un metaanálisis reciente de 46 estudios publicados entre 2019 y 2023, que aplicaron técnicas de deep learning para la predicción del rendimiento estudiantil, evidenció que la integración de redes neuronales recurrentes con algoritmos de clasificación tradicionales puede aumentar la efectividad predictiva en un 15-20% comparado con enfoques univariados (Abuhassna et al., 2024).

En la misma línea, la revisión sistemática de Alnasyan et al. (2024) confirmó que las redes neuronales profundas han logrado accuracies superiores al 90% en la predicción del rendimiento académico, mientras que investigaciones previas en bases de datos masivas de educación han demostrado que los métodos de aprendizaje automático pueden mejorar la precisión predictiva en un 15% comparado con enfoques tradicionales (Huang et al., 2020).

Asimismo, investigaciones recientes han evidenciado que los modelos de ensemble learning que combinan múltiples técnicas de deep learning pueden alcanzar errores mínimos (RMSE = 1.4908) y altos coeficientes de correlación (CCC = 0.9700), estableciendo nuevos estándares de precisión en la predicción del rendimiento estudiantil (Gu, 2025). Esta tendencia hacia la hibridación metodológica responde a la necesidad contemporánea de contar con herramientas más sofisticadas para la toma de decisiones informadas en entornos educativos.

En este panorama, la predicción del rendimiento académico en estudiantes de estadística mediante técnicas de minería de datos representa una contribución significativa al campo de la analítica educativa. La aplicación de redes neuronales artificiales y modelos de regresión múltiple para la predicción de indicadores académicos como el promedio ponderado y las calificaciones específicas de cursos, responde a una necesidad global de desarrollar herramientas predictivas más precisas y contextualizadas para disciplinas STEM (Science, Technology, Engineering, Mathematics). La relevancia de este estudio radica en su potencial para proporcionar evidencia empírica sobre la efectividad comparativa de diferentes técnicas de minería de datos en el contexto específico de la educación estadística superior, contribuyendo así al corpus de conocimiento internacional sobre analítica educativa.

La problemática que motiva esta investigación surge de la necesidad, identificada en universidades latinoamericanas de disponer de herramientas analíticas que orienten la planificación académica de los estudiantes. En el caso de la Universidad Nacional de Piura (UNP), la implementación de sistemas curriculares flexibles establecidos por la Ley Universitaria N° 30220 (2014) ha otorgado a los estudiantes mayor autonomía en la selección de materias. Sin embargo, dicha autonomía, implica el desafío de tomar decisiones académicas sin apoyo de herramientas predictivas que estimen la probabilidad de éxito en cada curso. Actualmente, el sistema de inscripción académica considera factores como el currículo de especialidad, el promedio ponderado y los prerrequisitos; no obstante, la decisión final recae en el estudiante, quien a menudo carece de información objetiva que le permita optimizar sus elecciones.

En este sentido, la presente investigación plantea la siguiente pregunta central: ¿Pueden las técnicas de minería de datos, específicamente las redes neuronales artificiales y la regresión lineal múltiple, predecir eficazmente el rendimiento académico de estudiantes de estadística, y cuál de estas técnicas resulta más efectiva según el tipo de predicción deseada? Esta interrogante se sustenta en la evidencia internacional que sugiere que diferentes técnicas de machine learning pueden tener efectividad variable según el contexto específico y el tipo de variable predictiva. Resolver esta cuestión permitirá generar evidencia empírica para el diseño de sistemas de apoyo estudiantil, basados en

analítica educativa en contextos de educación superior latinoamericana.

Finalmente, el objetivo del estudio es utilizar minería de datos para predecir el rendimiento académico en estudiantes de estadística de la Universidad Nacional de Piura durante el período 2010-2018. De manera complementaria, se busca desarrollar modelos para predecir la aprobación o desaprobación de cursos, pronosticar calificaciones específicas y comparar la eficacia de las redes neuronales artificiales frente a la regresión lineal múltiple.

MÉTODO

El diseño metodológico de esta investigación adoptó un enfoque cuantitativo aplicado con características no experimentales, longitudinales y retrospectivas. El estudio se basó en el análisis de datos históricos de estudiantes de la Escuela Profesional de Estadística de la de la Universidad Nacional de Piura (UNP) correspondiente al período académico 2010-2018. La población de estudio estuvo constituida por 510 registros académicos completos de estudiantes que culminaron su trayectoria académica en este período, constituyéndose así una muestra censal que incluyó la totalidad de registros disponibles que cumplían con los criterios de inclusión establecidos. Estos criterios consideraron específicamente registros de estudiantes matriculados en la Escuela Profesional con información académica completa, datos sobre cursos con prerrequisitos específicos y promedio ponderado calculado; se excluyeron los registros incompletos, aquellos con datos faltantes o inconsistentes, y los correspondientes a estudiantes que no concluyeron el programa.

Las variables de estudio se definieron considerando como variable dependiente la base de datos de notas académicas de los estudiantes. Las variables independientes incluyeron: el rendimiento por curso (aprobado/desaprobado), el promedio ponderado acumulado al semestre previo, la antigüedad en años del estudiante, la nota del curso prerrequisito, el promedio del grado de dificultad de aprobación del curso, el número de créditos totales por semestre y la sumatoria del grado de dificultad de cursos inscritos. La operacionalización de estas variables se realizó conforme a criterios teóricos y empíricos basados derivados de la experiencia académica y la literatura especializada. En particular, el promedio ponderado se obtuvo como la suma ponderada de las calificaciones por créditos, la antigüedad se definió como el tiempo transcurrido desde el ingreso hasta el semestre de análisis, y el grado de dificultad se estableció mediante análisis histórico de tasas de aprobación por curso.

La obtención de datos se realizó a través del Sistema Integrado de Gestión Académica de la UNP y la Oficina Central de Registros Académicos. En una primera etapa, los registros fueron organizados en Microsoft Excel para la organización inicial de la información, posteriormente analizados mediante IBM SPSS v.27 para el análisis estadístico avanzado. El proceso de recolección de datos siguió una secuencia estructurada que inició con la selección sistemática de registros académicos del período 2010-2018, seguida de la depuración de datos mediante la eliminación de registros incompletos, inconsistentes o duplicados, posteriormente se realizó la generación de variables derivadas como promedios acumulados, grados de dificultad y variables de clasificación, se aplicó la función min-max para la normalización de variables continuas según la fórmula $X' = (X - Xmin)/(Xmax - Xmin)$, y finalmente se estableció una partición de datos de 66.9% para entrenamiento, 21.8% para pruebas y 11.4% para validación, siguiendo recomendaciones de la literatura especializada para el desarrollo de modelos predictivos.

Las técnicas analíticas implementadas comprendieron dos enfoques principales de minería de datos: redes neuronales artificiales y regresión lineal múltiple. Las redes neuronales se desarrollaron con una arquitectura de perceptrón multicapa (Multilayer Perceptron) de tipo feedforward, utilizando la

función de activación tangente hiperbólica en capas ocultas, y el algoritmo de retropropagación con gradiente conjugado para el entrenamiento. Se aplicaron funciones de error diferenciadas según el tipo de predicción: suma de cuadrados para regresión y entropía cruzada para clasificación. Paralelamente, los modelos de regresión lineal múltiple se estimaron mediante el método de pasos hacia adelante (forward selection), con criterios de inclusión de variables de probabilidad menor al 5% y exclusión mayor al 10%. Asimismo, se verificaron los supuestos del modelo mediante las pruebas de normalidad multivariante de Mardia (1970), homocedasticidad de Levene (1960), independencia de Durbin-Watson (1950) y colinealidad mediante el Factor de Inflación de la Varianza (VIF) (Marquardt, 1970).

Para la evaluación de modelos se emplearon múltiples indicadores de rendimiento. El Cuadrado Medio del Error (CME) se empleó como métrica principal para evaluar la bondad de ajuste de los modelos, considerando que valores menores indican mejor rendimiento predictivo. El coeficiente de determinación (R^2) se utilizó para evaluar la proporción de varianza explicada por los modelos, especialmente relevante para la regresión lineal múltiple, mientras que para los modelos de clasificación se consideraron el porcentaje de clasificación correcta, las curvas COR (Receiver Operating Characteristic) con cálculo del área bajo la curva, y el análisis de importancia de variables normalizada para identificar los predictores más influyentes en cada modelo. Todos los análisis se realizaron con un nivel de significancia del 5%, y se aplicaron procedimientos de validación cruzada para garantizar la robustez de los resultados.

RESULTADOS

Los resultados del estudio se organizan en cuatro análisis principales que evalúan la efectividad de las técnicas de minería de datos implementadas para la predicción del rendimiento académico estudiantil, en correspondencia con los objetivos planteados.

Modelo de red neuronal para predicción del promedio ponderado

El modelo de red neuronal se estructuró con cinco capas de entrada, una capa oculta conteniendo tres unidades con función de activación tangente hiperbólica, y una capa de salida implementando función de activación identidad. La normalización se realizó mediante el método de cambio de escala para covariables, optimizando el entrenamiento de las redes neuronales. Los datos se distribuyeron según la partición establecida de 341 casos para entrenamiento (66.9%), 111 casos para pruebas (21.8%) y 58 casos para validación (11.4%), resultando en un conjunto de 510 casos válidos sin exclusiones.

Tabla 1. *Procesamiento de casos para el entrenamiento de la red neuronal del promedio ponderado*

Ejemplo	N	Porcentaje
Entrenamiento	341	66.9%
Pruebas	111	21.8%
Reserva	58	11.4%
Válido	510	100.0%

En la Tabla 1, se visualiza la distribución de los datos utilizada para el entrenamiento y validación del modelo de red neuronal artificial para la predicción del promedio ponderado. Esta distribución asegura una representación adecuada de los datos para el aprendizaje del modelo, con un mayor porcentaje destinado al entrenamiento para optimizar la capacidad predictiva.

El entrenamiento evidenció un error de suma de cuadrados de 3.377 con error relativo de 0.020 durante la fase de entrenamiento, mientras que durante las pruebas el error aumentó a 7.559 con error relativo de 0.126, indicando una pérdida de generalización aceptable. El tiempo de entrenamiento fue de 0:00:00.04 segundos, demostrando la eficiencia computacional del algoritmo implementado. El análisis de importancia de variables reveló que el grado de dificultad 1 del curso constituye el predictor más influyente con 100% de importancia normalizada, seguido por la sumatoria del promedio ponderado con 30.2%, evidenciando que las características específicas del curso tienen mayor peso predictivo que factores demográficos o históricos del estudiante. Estos resultados respaldan la aplicabilidad de las redes neuronales en la predicción de indicadores agregados de rendimiento, particularmente cuando incorporan variables de dificultad curricular.

Predicción de la condición de Aprobación/Desaprobación del semestre

Para la predicción de la condición de aprobación/desaprobación del semestre, se implementó una red neuronal con cinco capas de entrada, ocho unidades en la capa oculta y dos unidades de salida con función Softmax para la clasificación binaria. El entrenamiento utilizó 363 casos (72.0%), pruebas con 89 casos (17.8%) y validación con 58 casos (10.2%), manteniendo el mismo conjunto de 510 casos válidos.

Tabla 2. *Clasificación para el resultado del ciclo según partición y global de la red*

Ejemplo	Desaprobado	Aprobado	Porcentaje correcto
Entrenamiento	98.1%	97.9%	98.0%
Pruebas	91.3%	100.0%	96.0%
Reserva	91.4%	100.0%	95.2%

La Tabla 2 presenta la efectividad clasificatoria del modelo de red neuronal artificial para la predicción de la condición de aprobación/desaprobación del semestre. Los resultados demuestran una efectividad global del 98.0% durante el entrenamiento, con 98.1% de precisión para estudiantes desaprobados y 97.9% para estudiantes aprobados. Durante las pruebas, la efectividad global fue de 96.0%, con 91.3% de precisión para desaprobados y 100.0% para aprobados.

Los resultados evidencian una capacidad predictiva casi perfecta (AUC = 0.995), lo cual confirma la alta sensibilidad y especificidad del modelo para clasificar correctamente la condición académica de los estudiantes.

Tabla 3. *Área debajo de la curva COR para el modelo de clasificación*

Condición	Área
Desaprobado	0.995
Aprobado	0.995

En cuanto al análisis del área bajo la curva COR, la Tabla 3 muestra valores de 0.995 tanto para la condición de desaprobado como para aprobado, indicando excelente capacidad discriminativa del modelo de clasificación. Estos valores cercanos a 1.0 confirman que el modelo puede distinguir efectivamente entre estudiantes que aprobarán y desaprobarán el semestre.

Modelo de regresión lineal múltiple para promedio ponderado

La implementación de modelos de regresión lineal múltiple para la predicción del promedio ponderado utilizó el método de pasos hacia adelante, resultando en dos modelos de estimación con capacidades predictivas diferenciadas.

Tabla 4. *Resumen del modelo lineal para predicción del promedio ponderado*

Modelo	R	R^2	R^2 Ajustado	Error estándar
1	0.976	0.953	0.953	0.875
2	0.976	0.953	0.953	0.872

La Tabla 4 presenta el resumen de los modelos de regresión lineal múltiple desarrollados. Ambos modelos explican el 95.3% de la varianza de la variable dependiente, con errores estándar de estimación de 0.875 y 0.872 respectivamente. El segundo modelo incorporó tanto el grado de dificultad 1 como la sumatoria del promedio ponderado, demostrando un CME ligeramente menor (0.761) comparado con el primer modelo.

La validación de supuestos del modelo confirmó que los residuos siguen una distribución normal multivariante según los estadísticos de asimetría ($p = 0.052$) y curtosis ($p = 0.082$), cumplen con el supuesto de homocedasticidad según la prueba de Levene ($p = 0.162$), y satisfacen el supuesto de independencia según el test de Durbin-Watson ($p = 0.972$), validando así la robustez estadística de los modelos desarrollados.

Análisis comparativo entre técnicas de predicción

El análisis comparativo entre técnicas de minería de datos evidenció diferencias significativas en la efectividad según el tipo de predicción deseada.

Tabla 5. *Cuadrado medio del error de los modelos de pronóstico*

Técnica	Promedio Ponderado	Nota de Curso
Red Neuronal Artificial	7.559	1.095
Regresión Lineal	0.761	15.663

Según los resultados de la Tabla 5, la comparación de la efectividad entre las redes neuronales artificiales y la regresión lineal múltiple según el tipo de predicción. Para la predicción del promedio ponderado, la regresión lineal múltiple demostró superioridad con un CME de 0.761 comparado con 7.559 de las redes neuronales. Sin embargo, para la predicción de notas específicas de cursos, las redes neuronales evidenciaron mayor eficacia con un CME de 1.095 comparado con 15.663 de la regresión

lineal múltiple.

Este patrón de resultados confirma que la elección de la técnica debe estar determinada por el tipo específico de variable predictiva, donde los modelos lineales resultan más efectivos para variables agregadas como el promedio ponderado, mientras que las redes neuronales demuestran superioridad para variables específicas sujetas a mayor variabilidad.

Validación cruzada en múltiples cursos

La validación de los modelos se completó mediante el análisis de siete cursos específicos con prerrequisitos, donde se confirmó que las redes neuronales mantuvieron consistentemente menor CME comparado con la regresión lineal múltiple en todos los casos analizados. Los cursos ES2491, ES3418, ES3465, ES4453, ES4454, ES4438 y ES5320 mostraron patrones similares donde las redes neuronales alcanzaron CMEs entre 0.298 y 7.541, mientras que la regresión lineal múltiple presentó CMEs entre 1.728 y 24.561, consolidando la evidencia sobre la efectividad diferencial de las técnicas según el contexto de aplicación.

DISCUSIÓN

Los resultados obtenidos en esta investigación proporcionan evidencia empírica robusta sobre la efectividad comparativa de diferentes técnicas de minería de datos para la predicción del rendimiento académico en estudiantes de estadística, confirmando la complejidad inherente de los modelos predictivos educativos y la necesidad de enfoques metodológicos diferenciados según el tipo de variable predictiva. Los hallazgos principales evidencian que la regresión lineal múltiple resulta más efectiva para la predicción del promedio ponderado, mientras que las redes neuronales artificiales demuestran superioridad para la predicción de notas específicas de cursos, patrón que se alinea parcialmente con investigaciones internacionales recientes, pero también revela aspectos diferenciadores significativos.

Estos resultados se comparan favorablemente con los hallazgos de Acosta y Pizarro (2011), quienes en su estudio peruano concluyeron que el modelo de red neuronal es más efectivo para predecir notas específicas de cursos mientras que la regresión múltiple resulta más adecuada para determinar la probabilidad de aprobación. No obstante, la efectividad reportada por estos autores fue considerablemente menor (73% de exactitud) comparado con los resultados del presente estudio (98.0% de efectividad global). Esta diferencia puede atribuirse a mejoras en las técnicas de machine learning, mayor calidad de los datos disponibles y optimizaciones en la arquitectura de las redes neuronales implementadas, factores que han evolucionado significativamente en la última década según la literatura internacional.

Más recientemente, estudios realizados con técnicas de regresión lineal generalizada han demostrado superioridad en la predicción de rendimiento académico, alcanzando R^2 de 0.792 para matemáticas y 0.730 para artes del lenguaje, resultados que se alinean con los hallazgos del presente estudio sobre la efectividad de modelos lineales para variables agregadas (Lou y Colvin, 2025). De manera similar, estudios en modelos interpretables de machine learning han confirmado que la efectividad diferencial de las técnicas depende del tipo de variable predictiva, con XGBoost demostrando superioridad en análisis multivariados complejos (Guevara-Reyes et al., 2025).

Por otra parte, la comparación con estudios europeos recientes revela convergencias metodológicas importantes. Las investigaciones realizadas por Kalita (2025) en universidades alemanas que implementaron sistemas de redes neuronales profundas para el análisis del rendimiento

estudiantil alcanzaron precisiones superiores al 90%, resultado que coincide con los 98.0% de efectividad clasificatoria global obtenida en esta investigación, sugiriendo que las técnicas de redes neuronales han alcanzado niveles de madurez técnica que permiten su implementación efectiva en contextos educativos diversos. Sin embargo, estos resultados se diferencian de los hallazgos de Miranda et al. (2024) en universidades europeas, quienes reportaron que la efectividad de los modelos predictivos depende fundamentalmente de la calidad y completitud de los datos disponibles, factor que en esta investigación fue cuidadosamente controlado mediante la depuración sistemática de registros.

Además, los resultados de esta investigación se comparan también con estudios asiáticos recientes que han implementado enfoques híbridos de ensemble learning. Las investigaciones realizadas por Pan et al. (2025) en instituciones de educación superior de India y China, que utilizaron algoritmos de ensemble learning combinados con redes neuronales profundas, alcanzaron accuracies del 97% en cohortes estudiantiles de ingeniería, resultado que es ligeramente inferior a los 98.0% de efectividad global obtenidos en esta investigación, lo que sugiere que las arquitecturas específicas implementadas en este estudio pueden haber optimizado mejor las características de los datos educativos analizados.

De manera similar, la identificación del grado de dificultad 1 como la variable de mayor importancia predictiva (100% de importancia normalizada) en ambos modelos coincide con hallazgos de investigaciones latinoamericanas recientes. Las investigaciones realizadas por Córdova-Esparza et al. (2025) en universidades brasileñas que implementaron enfoques híbridos de inteligencia artificial para la predicción del rendimiento académico evidenciaron que las características específicas de los cursos tienen mayor peso predictivo que factores socioeconómicos o demográficos, hallazgo que se replica en este estudio y sugiere patrones consistentes en la importancia de variables curriculares en el éxito estudiantil.

Consecuentemente, la efectividad diferencial de las técnicas según el tipo de predicción se compara con el metaanálisis reciente de Abuhassna et al. (2024), quien analizó 46 estudios publicados entre 2019 y 2023 que aplicaron técnicas de deep learning para la predicción del rendimiento estudiantil. Este metaanálisis evidenció que la integración de redes neuronales recurrentes con algoritmos de clasificación tradicionales puede aumentar la efectividad predictiva en un 15-20% comparado con enfoques univariados, patrón que se observa en los resultados de esta investigación donde las redes neuronales demostraron superioridad consistente para la predicción de notas específicas de cursos, mientras que los modelos lineales mantuvieron efectividad para variables agregadas como el promedio ponderado.

En contraste, la superioridad de la regresión lineal múltiple para la predicción del promedio ponderado (R^2 = 95.3%) se diferencia significativamente de estudios anteriores que demostraban efectividad moderada para modelos lineales. Este resultado sugiere que las variables académicas específicas analizadas en este estudio tienen relaciones más lineales y predecibles con el promedio ponderado de lo que previamente se había documentado, por lo que la selección cuidadosa de variables predictivas puede optimizar significativamente la efectividad de modelos estadísticos tradicionales, hallazgo que tiene implicaciones importantes para el desarrollo de sistemas de predicción eficientes computacionalmente.

Estudios comparativos recientes que han evaluado algoritmos de clasificación para la predicción del rendimiento académico han confirmado que diferentes técnicas muestran efectividad variable según el contexto específico, donde modelos como Support Vector Machines y Random Forest han demostrado rendimiento competitivo en ciertos tipos de datos educativos (Wang et al.,

2022). No obstante, análisis comparativos exhaustivos de más de 179 clasificadores han evidenciado que no existe un algoritmo universalmente superior, sino que la efectividad depende críticamente de las características específicas del conjunto de datos y la naturaleza de la variable predictiva (Fernández-Delgado et al., 2014), patrón que se replica en los resultados de esta investigación donde la regresión lineal múltiple demostró superioridad para variables agregadas mientras que las redes neuronales fueron más efectivas para variables específicas.

Por otro lado, los resultados enfatizan la importancia de factores socioeconómicos y demográficos en la predicción del rendimiento académico. Por ejemplo, investigaciones australianas recientes que implementaron técnicas de LSTM con modelos Bi-LSTM para la predicción del rendimiento semestral (Zhang et al., 2024) reportaron que variables demográficas y socioeconómicas contribuían significativamente a la capacidad predictiva de los modelos, mientras que en esta investigación estas variables mostraron importancia mínima (1.3% para antigüedad del estudiante), lo que quiere decir, que en contextos específicos de educación estadística superior, factores académicos directos pueden tener mayor relevancia predictiva que variables socioeconómicas.

Finalmente, la comparación con estudios de dropout y retención académica revela implicaciones importantes para la prevención del fracaso escolar. Las investigaciones de Córdova-Esparza et al. (2025) que utilizaron business intelligence para predecir y prevenir el abandono estudiantil evidenciaron que los modelos predictivos pueden identificar estudiantes en riesgo con hasta 85% de precisión, resultado que se alinea parcialmente con los 98.0% de efectividad clasificatoria obtenidos en esta investigación. Sin embargo, investigaciones recientes han demostrado que técnicas de machine learning aplicadas a entornos de aprendizaje en línea pueden lograr precisiones superiores al 96% para la predicción temprana del rendimiento, lo que sugiere que la disponibilidad de datos comportamentales en tiempo real puede mejorar significativamente la capacidad predictiva de los modelos (Zhao et al., 2024).

Asimismo, estudios que han implementado sistemas de aprendizaje ensemble con selección automática de características han alcanzado errores de predicción excepcionalmente bajos (RMSE = 0.6%, MAPE = 0.03%), estableciendo nuevos estándares de precisión en la predicción del rendimiento académico (Gu, 2025). Estos hallazgos indican que los sistemas de alerta temprana basados en analítica educativa pueden ser altamente efectivos para la intervención preventiva en educación superior, especialmente cuando incorporan múltiples fuentes de datos y técnicas avanzadas de ensemble learning.

Por consiguiente, estos hallazgos tienen implicaciones teóricas importantes para el campo de la analítica educativa, confirmando que diferentes técnicas de machine learning tienen efectividad diferencial según el contexto específico de aplicación y el tipo de variable predictiva. Los resultados sugieren que las instituciones educativas deberían implementar sistemas híbridos que combinen múltiples enfoques metodológicos según las necesidades específicas de predicción, maximizando así la efectividad de los sistemas de apoyo estudiantil basados en datos.

No obstante, es importante reconocer que el presente estudio presenta algunas limitaciones que deben considerarse en la interpretación y aplicación de los resultados. En primer lugar, el análisis se limitó exclusivamente a una sola institución educativa, específicamente la Universidad Nacional de Piura, lo cual puede restringir la generalización de los hallazgos a otros contextos educativos con características diferentes, como universidades privadas, instituciones de otros países latinoamericanos o programas educativos de distinta naturaleza. Esta limitación temporal y geográfica puede afectar la validez externa de los resultados, requiriendo estudios adicionales para confirmar la aplicabilidad en contextos diversos.

Asimismo, la ausencia de variables socioeconómicas, demográficas y contextuales en el modelo predictivo representa una limitación significativa que pudo haber reducido la capacidad explicativa de los modelos desarrollados. Factores como el nivel socioeconómico familiar, ubicación geográfica, acceso a recursos tecnológicos, situación ocupacional del estudiante y condiciones socioeconómicas durante el período de estudio no fueron considerados, lo cual puede haber limitado el alcance del análisis predictivo y la identificación de estudiantes en riesgo académico.

Finalmente, el período de estudio analizado (2010-2018) puede no reflejar adecuadamente los cambios recientes en el contexto educativo, incluyendo las transformaciones derivadas de la digitalización acelerada, la educación remota o híbrida, y las modificaciones en los procesos de enseñanza-aprendizaje que han emergido en años recientes. Por tanto, esta limitación temporal apunta la necesidad de actualizar los modelos con datos más contemporáneos que incorporen estas nuevas dinámicas educativas.

CONCLUSIONES

El cumplimiento de los objetivos planteados en esta investigación demuestra el éxito en la implementación de técnicas de minería de datos para la predicción del rendimiento académico en estudiantes de estadística, alcanzando todos los objetivos propuestos en el estudio. Se logró desarrollar modelos de redes neuronales artificiales con alta efectividad para la predicción de aprobación/desaprobación de cursos específicos, implementar modelos de regresión lineal múltiple con excelente capacidad predictiva para promedios ponderados. Se comparó exitosamente la eficacia de ambas técnicas evidenciando su complementariedad, e identificar las variables de mayor influencia en la predicción del rendimiento académico, siendo el grado de dificultad del curso el factor predictivo más relevante.

En términos generales, los hallazgos revelan que las técnicas de minería de datos constituyen herramientas complementarias y viables para la predicción del rendimiento académico, donde la efectividad de cada técnica depende del tipo específico de variable predictiva. Esta complementariedad metodológica sugiere que las instituciones educativas deberían implementar sistemas híbridos que combinen múltiples enfoques metodológicos según las necesidades específicas de predicción. Además, la identificación del grado de dificultad del curso como variable predictiva más influyente sugiere que las intervenciones educativas efectivas deben enfocarse en optimizar factores relacionados con el diseño curricular y la complejidad de las materias, más que en características individuales de los estudiantes.

Del mismo modo, la validación exitosa de los supuestos estadísticos en los modelos de regresión múltiple confirma la robustez metodológica de la investigación y la validez de las inferencias realizadas, aportando confianza en la aplicabilidad de los hallazgos a otras instituciones de educación superior con características similares. Esta validación es especialmente relevante para la credibilidad de los resultados y su generalización en contextos educativos latinoamericanos.

A partir de estos aportes, se recomienda que las instituciones de educación superior desarrollen sistemas predictivos del rendimiento académico que consideren la efectividad diferencial de las técnicas según el tipo de predicción requerida, desarrollando arquitecturas híbridas que combinen modelos lineales para variables agregadas y técnicas de machine específicas. Asimismo, se sugiere el monitoreo learning para predicciones continuo del grado de factor predictivo clave de alerta temprana basados, implementando sistemas dificultad de cursos como en patrones de rendimiento histórico y

desarrollando recomendaciones académicas personalizadas según perfiles predictivos individuales.

En cuanto a investigaciones futuras, se recomienda la inclusión de variables socioeconómicas y demográficas para mejorar la capacidad predictiva de los modelos y reducir el sesgo hacia variables académicas únicamente, así como la implementación de estudios longitudinales que evalúen la efectividad de los modelos predictivos en contextos educativos contemporáneos y post-pandémicos. El desarrollo de sistemas en tiempo real que permitan la actualización continua de los modelos predictivos con nuevos datos académicos constituye una línea de investigación prometedora, así como la realización de estudios comparativos en diferentes disciplinas y niveles educativos para evaluar la generalización de los hallazgos.

Finalmente, la exploración de técnicas de aprendizaje automático más avanzadas, incluyendo algoritmos de ensemble, deep learning y técnicas de interpretabilidad de modelos, representa una línea de investigación con alto potencial para mejorar tanto la precisión predictiva, como la comprensión de los factores que determinan el rendimiento académico estudiantil. Este enfoque puede contribuir al desarrollo de sistemas de apoyo estudiantil más efectivos y el fortalecimiento de la calidad educativa en la educación superior latinoamericana.

REFERENCIAS

Abuhassna, H., Alwahab, A., Ahmed, A., Al-Rahmi, W. M., Othman, M. S. A., Abd Razak, S. K., … y Abualsaud, K. (2024). A Bibliometric and Systematic Literature Analysis of Artificial Intelligence in Education for Student Performance Prediction. Journal of Educational Technology & Society, 27(2), 145-162. https://doi.org/10.30191/ETS.202403_27(2).0009

Acosta, D. P., y Pizarro, S. S. (2011). Predicción del rendimiento académico en la educación superior usando minería de datos y su comparación con técnicas estadísticas [Tesis de maestría]. Universidad Nacional Mayor de San Marcos. https://repositorio.unmsm.edu.pe/handle/11354/1024

Alnasyan, B., Basheri, M., y Alassafi, M. (2024). The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review. Computers and Education: Artificial Intelligence, 6, 100231. https://doi.org/10.1016/j.caeai.2024.100231

Baker, R. S., y Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (2nd ed., pp. 253-274). Cambridge University Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2021). Enhancing algorithmic assessment in education using ensemble methods. Educational Technology Research and Development, 69(4), 2157-2178. https://doi.org/10.1007/s11423-021-10078-4

Córdova-Esparza, D. M., Tovar-Arias, J. D., Ramos-González, J., Pérez-León, M. E., y Núñez-Martínez, J. (2025). Predicting and Preventing School Dropout with Business Intelligence and Machine Learning. Information, 16(4), 326. https://doi.org/10.3390/info16040326

Durbin, J., y Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. Biometrika, 37(3/4), 409-428. https://doi.org/10.1093/biomet/37.3-4.409

Fernández-Delgado, M., Cernadas, E., Barro, S., y Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15(1), 3133-3181. http://jmlr.org/papers/v15/delgado14b.html

Gu, J. (2025). Predicting student academic achievement using stacked ensemble learning. Scientific Reports, 15, 20779. https://doi.org/10.1038/s41598-025-20779-z

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., y Villegas-Ch, W. (2025). Machine learning models for academic performance prediction. Frontiers in Education, 10, 1632315. https://doi.org/10.3389/feduc.2025.1632315

Huang, A. Y. Q., Lu, O. H. T., Huang, J. C. H., Yin, C. J., y Yang, S. J. H. (2020). Predicting students'

academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. Interactive Learning Environments, 28(7), 1014-1037. https://doi.org/10.1080/10494820.2018.1508280

Kalita, E. (2025). Educational data mining: a 10-year review of techniques and applications in higher education. International Journal of Information Technology, 17(1), 123-145. https://doi.org/10.1007/s10791-025-09589-z

Kumar, A., Singh, S., y Kumar, V. (2023). Using machine learning to predict student outcomes for early intervention. Nature Scientific Reports, 15, 23409. https://doi.org/10.1038/s41598-025-23409-w

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to probability and statistics: Essays in honor of Harold Hotelling (pp. 278-292). Stanford University Press.

Ley Universitaria N° 30220. (2014, 9 de julio). Diario Oficial El Peruano. https://www.sunedu.gob.pe/documentos/Leyes/LeyUniversitaria30220.pdf

López, R. G., Jiménez, A. B., y Fernández, J. L. (2024). Educational data mining for predicting students' academic performance: A survey study. Education and Information Technologies, 28(3), 905-971. https://doi.org/10.1007/s10639-022-11152-y

Lou, Y., y Colvin, K. F. (2025). Performance prediction using educational data mining techniques. International Journal of STEM Education, 12, 1. https://doi.org/10.1186/s40594-025-00502-w

Mardia, K. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3), 519-530. https://doi.org/10.1093/biomet/57.3.519

Marquardt, D. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12(3), 591-612. https://doi.org/10.1080/00401706.1970.10488634

Miranda, E., Santoso, A., y Widiyaningtyas, T. (2024). Machine learning's model-agnostic interpretability on the prediction of students' academic performance. Internet of Things, 25, 101152. https://doi.org/10.1016/j.iot.2024.101152

Pan, J., Zhang, Y., Liu, H., Chen, S., y Wang, X. (2025). Academic Performance Prediction Using Machine Learning Approaches: A Comprehensive Survey. IEEE Access, 13, 10810756. https://doi.org/10.1109/ACCESS.2025.10810756

Romero, C., y Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1355. https://doi.org/10.1002/widm.1355

Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10), 1380-1400. https://doi.org/10.1177/0002764213498851

Tan, M., y Shneiderman, B. (2023). Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors. Journal of Educational Computing Research, 61(4), 923-948. https://doi.org/10.1177/07356331231169234

Wang, X., Zhang, L., y Li, M. (2022). Predicting Student Academic Performance using Support Vector Machine and Random Forest. Education and Information Technologies, 27(5), 6845-6862. https://doi.org/10.1007/s10639-021-10769-1

Zhang, X., Liu, M., Wang, Y., y Chen, L. (2024). Predicting student academic performance using Bi-LSTM with attention mechanism. Frontiers in Education, 9, 1581247. https://doi.org/10.3389/feduc.2024.1581247

Zhao, Q., Chen, J., Liu, Y., y Xu, H. (2024). Predicting student performance and enhancing learning outcomes using educational data mining. Computers, 14(3), 83. https://doi.org/10.3390/computers14030083