Revista Amazonía Digital

Vol. 3 Núm. 1: e292 (2024)
https://doi.org/10.55873/rad.v3i1.292
e-ISSN: 2810-8701
Universidad Nacional Amazónica de Madre de Dios

Artículo original / Original article

Predicción del rendimiento académico estudiantil usando algoritmos de aprendizaje supervisado en una universidad de la selva peruana

Prediction of student academic performance using supervised learning algorithms in a university from the Peruvian rainforest

Alex Ali Vargas-Quispe [1]; Jaime Cesar Prieto-Luna [1*]

[1]Universidad Nacional Amazónica de Madre de Dios, Puerto Maldonado, Perú

Recibido: 21/08/2023
Aceptado: 26/10/2023
Publicado: 25/01/2024

*Autor de correspondencia: jprieto@unamad.edu.pe

Resumen: El rendimiento académico es crucial para la gestión educativa, y la predicción de este puede optimizar la toma de decisiones. Este estudio tuvo como objetivo predecir el rendimiento académico de estudiantes universitarios mediante tres algoritmos de aprendizaje supervisado: K-Vecinos más cercanos (KNN), Naive Bayes (NB) y Árbol de Decisión (AD). Se utilizaron datos de 813 estudiantes, incluyendo variables socioeconómicas y académicas. Los datos fueron preprocesados y evaluados a través de métricas como precisión, recuperación, exactitud y AUC-ROC. El modelo KNN fue el más efectivo, alcanzando una exactitud del 81.97%, lo que lo convierte en la mejor opción para predecir el rendimiento académico. Aunque mostró una recuperación moderada, su precisión fue la más alta, demostrando un buen equilibrio en la clasificación de los estudiantes. En conclusión, KNN es una herramienta prometedora para que las instituciones educativas identifiquen estudiantes en riesgo y mejoren sus estrategias de intervención, ayudando a elevar el rendimiento académico y reducir la deserción.

Palabras clave: análisis de datos; aprendizaje automático; minería de datos; predicción educativa

Abstract: Academic performance is crucial for educational management and predicting it can optimize decision-making processes. This study aimed to predict the academic performance of university students using three supervised learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Tree (AD). Data from 813 students, including socioeconomic and academic variables, were processed and evaluated using metrics such as precision, recall, accuracy, and AUC-ROC. The KNN model was the most effective, achieving an accuracy of 81.97%, making it the best choice for predicting academic performance. Although it showed moderate recall, its precision was the highest, demonstrating a good balance in student classification. In conclusion, KNN is a promising tool for educational institutions to identify at-risk students and improve intervention strategies, helping to boost academic performance and reduce dropout rates.

Keywords: data analysis; educational prediction; machine learning; data mining

Cómo citar / Citation: Vargas-Quispe, A. A. & Prieto-Luna, J. C. (2024). Predicción del rendimiento académico estudiantil usando algoritmos de aprendizaje supervisado en una universidad de la selva peruana. *Revista Amazonía Digital*, 3(1), e292. https://doi.org/10.55873/rad.v3i1.292

1. Introducción

El rendimiento académico es un tema central en la educación superior, ya que mide el nivel de aprendizaje alcanzado por los estudiantes y permite evaluar la calidad de los procesos educativos (Goss, 2022). Este concepto ha cobrado mayor relevancia en los últimos años, debido a la creciente demanda de egresados con competencias que respondan a las necesidades del mercado laboral y al desarrollo de habilidades críticas para su inserción en la sociedad (Abelha et al., 2020). A nivel institucional, el rendimiento académico se convierte en un indicador clave no solo para valorar la eficacia educativa, sino también para establecer estrategias que mejoren la experiencia de los estudiantes y aumenten la retención en las universidades (Centoni & Maruotti, 2021).

Sin embargo, mejorar el rendimiento académico sigue siendo un desafío, particularmente en instituciones de educación superior que enfrentan contextos socioeconómicos y culturales complejos (Núñez-Canal et al., 2022). Las causas del bajo rendimiento académico son múltiples y pueden incluir factores como el nivel de preparación previa, las condiciones económicas del estudiante, el apoyo familiar, y el acceso a recursos educativos adecuados (Rahman et al., 2023). La identificación temprana de estudiantes en riesgo de fracaso académico es crucial para poder intervenir de manera efectiva, evitando así la deserción y promoviendo un mejor desempeño en las aulas (Bañeres et al., 2023).

En los últimos años, la evolución de la tecnología y el acceso a grandes volúmenes de datos han abierto nuevas oportunidades para comprender y abordar el rendimiento académico (Haleem et al., 2022). El análisis de datos en el contexto educativo ha permitido a las instituciones recopilar información sobre las características de los estudiantes, sus hábitos de estudio, su entorno familiar y su participación en actividades académicas (Johar et al., 2023). Este enfoque ha revelado patrones que pueden ser utilizados para diseñar estrategias de intervención más precisas y personalizadas, con el fin de mejorar los resultados de aprendizaje y promover una educación más inclusiva.

Uno de los avances más destacados en este campo ha sido el uso de algoritmos de aprendizaje supervisado. Estas técnicas, que forman parte del aprendizaje automático o machine learning, han demostrado su potencial para predecir el rendimiento académico de manera eficaz (Issah et al., 2023). Al analizar grandes cantidades de datos históricos, los algoritmos pueden identificar relaciones entre factores como el promedio semestral, el nivel socioeconómico y el bienestar psicológico de los estudiantes, permitiendo hacer predicciones sobre su desempeño futuro. Esta capacidad predictiva ofrece una solución tecnológica valiosa para los desafíos actuales de la educación superior (Matz et al., 2023).

Los modelos predictivos basados en aprendizaje supervisado permiten no solo prever el rendimiento académico, sino también ofrecer herramientas que apoyen la toma de decisiones en áreas académicas y administrativas (Matzavela & Alepis, 2021). Al identificar a los estudiantes que podrían requerir apoyo adicional, los docentes y gestores universitarios pueden diseñar programas de tutoría, becas, apoyo psicológico, y otras intervenciones ajustadas a las necesidades de cada grupo estudiantil. Esta aproximación tecnológica busca no solo prevenir el fracaso académico, sino también mejorar el rendimiento general de los estudiantes en diversos contextos educativos (Sghir et al., 2023).

A pesar de la creciente aplicación de estas técnicas, aún existe un vacío en la literatura sobre su implementación en entornos específicos, como las universidades situadas en regiones con características socioeconómicas particulares (Nsanzumuhire & Groot, 2020). Las instituciones en estas áreas enfrentan desafíos únicos, y la aplicación de modelos predictivos adaptados a sus contextos podría marcar una diferencia significativa en los resultados académicos (Kamalov et al., 2023). Es crucial continuar investigando cómo los algoritmos de aprendizaje supervisado pueden contribuir al éxito educativo en estos escenarios.

El objetivo de esta investigación es predecir el rendimiento académico de los estudiantes universitarios mediante el uso de algoritmos de aprendizaje supervisado, identificando los

factores sociales, económicos y académicos que influyen en su desempeño. Con ello, se busca proporcionar a las instituciones herramientas tecnológicas que faciliten la toma de decisiones y mejoren la gestión académica.

2. Materiales y métodos

2.1. Diseño de investigación

Este estudio empleó un enfoque cuantitativo con un diseño no experimental, de tipo correlacional y transversal. El objetivo fue determinar la relación entre diversas variables socioeconómicas, académicas y personales con el rendimiento académico de los estudiantes. No se manipuló ninguna variable de manera intencional, sino que se analizaron datos históricos ya disponibles, recolectados entre los semestres académicos 2010-I y 2020-II.

2.2. Población y muestra

La población de estudio estuvo conformada por los estudiantes ingresantes al primer semestre de la carrera de Ingeniería de Sistemas e Informática de una universidad en la selva peruana, lo que sumó un total de 861 registros de estudiantes. Se utilizó un muestreo censal, donde la totalidad de los registros fueron analizados sin excluir a ningún sujeto. La muestra incluyó datos relevantes de los estudiantes como edad, género, estado civil, modalidad de ingreso, situación laboral, bienestar psicológico, y situación socioeconómica, además de su promedio ponderado semestral.

2.3. Fuente de datos

Los datos fueron obtenidos de la base proporcionada por la Dirección de Asuntos Académicos de la universidad, previa solicitud formal. Estos datos incluyeron registros de los estudiantes, los cuales fueron organizados en atributos académicos, sociales y económicos. Se aseguraron las consideraciones éticas durante la recopilación de los datos, garantizando la confidencialidad y anonimato de los sujetos estudiados. La base de datos fue limpiada para eliminar duplicados y registros incompletos.

2.4. Algoritmos de aprendizaje supervisado

Se emplearon tres algoritmos de aprendizaje supervisado para la predicción del rendimiento académico:

K-Vecinos más cercanos (K-NN): Este algoritmo clasifica las instancias nuevas basándose en la cercanía a las instancias previamente clasificadas en el conjunto de entrenamiento. Se utilizó el valor óptimo de K para mejorar la precisión del modelo, determinado a través de validación cruzada.

Naive Bayes: Basado en el teorema de Bayes, este clasificador probabilístico asume que las variables predictoras son independientes entre sí. A pesar de esta suposición, el modelo es eficaz en muchos contextos educativos.

Árbol de decisión: Este modelo se construyó a partir de la segmentación iterativa de los datos en ramas, donde cada nodo representa una decisión basada en un atributo, lo que facilita la interpretación de los factores que más influyen en el rendimiento académico.

2.5. Preparación de los datos

En la fase de preparación de los datos, se realizó la limpieza y transformación de estos. Se eliminaron registros duplicados, y los valores faltantes se trataron mediante imputación estadística. Atributos como la dependencia estudiantil y la situación socioeconómica fueron transformados en variables categóricas, mientras que las calificaciones se mantuvieron como

valores continuos. Los datos fueron normalizados para mejorar el rendimiento de los algoritmos de clasificación.

2.6. División del conjunto de datos

El conjunto de datos fue dividido en dos subconjuntos: el 75% de los registros fue utilizado para el entrenamiento de los algoritmos y el 25% restante para la validación y prueba. Esta división permitió evaluar el desempeño de los modelos y evitar sobreajuste. Se empleó la técnica de validación cruzada de K-pliegues para garantizar la robustez de los resultados y minimizar el sesgo en la evaluación de los modelos.

2.7. Métricas de evaluación

La evaluación de los modelos predictivos se basó en varias métricas de desempeño:

Exactitud (Accuracy): Representa la proporción de instancias correctamente clasificadas sobre el total.

Precisión (Precision): Mide la cantidad de verdaderos positivos entre las instancias clasificadas como positivas.

Sensibilidad (Recall): Indica la proporción de verdaderos positivos identificados correctamente por el modelo.

Puntaje ROC-AUC: Evalúa la capacidad del modelo para diferenciar entre clases, siendo 1 el valor ideal.

Cada una de estas métricas fue calculada para los tres modelos y los resultados fueron comparados para seleccionar el algoritmo más adecuado para la predicción del rendimiento académico.

2.8. Herramientas y entornos de desarrollo

El procesamiento y análisis de los datos fueron realizados utilizando el lenguaje de programación Python, junto con las bibliotecas especializadas en aprendizaje automático como Scikit-learn para la implementación de los algoritmos. Las simulaciones y pruebas de los modelos se ejecutaron en Google Colab, aprovechando su capacidad de procesamiento y su integración con bibliotecas de machine learning. Para la visualización de resultados, se emplearon las librerías Matplotlib y Seaborn.

2.9. Ética y confidencialidad

Se aseguraron los más altos estándares éticos durante el manejo de los datos, cumpliendo con las normativas institucionales para la investigación académica. Los datos fueron anonimizados para proteger la privacidad de los estudiantes, y el acceso a los mismos estuvo restringido al equipo de investigación. La investigación fue monitoreada y validada por especialistas en el manejo de datos educativos.

3. Resultados y discusión

El presente estudio se desarrolló aplicando la metodología CRISP-DM, que permitió llevar a cabo un análisis detallado mediante seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En este caso, la comprensión del negocio estuvo centrada en la mejora del rendimiento académico de los estudiantes universitarios, un aspecto clave para la toma de decisiones en las áreas académicas y administrativas. La predicción del rendimiento académico a través de algoritmos de aprendizaje supervisado puede ser un recurso esencial para identificar a tiempo a los estudiantes en riesgo de bajo rendimiento y optimizar el uso de los recursos educativos.

3.1. Comprensión de los datos

En la fase de comprensión de los datos, se trabajó con un conjunto de datos proporcionado por la Dirección de Asuntos Académicos de la Universidad Nacional Amazónica de Madre de Dios. Estos datos están disponibles públicamente en el sitio web de Kaggle bajo el enlace https://www.kaggle.com/dsv/4462348. El conjunto de datos incluye información de 861 estudiantes matriculados en el primer semestre de la carrera de Ingeniería de Sistemas e Informática, recopilados entre los semestres 2010-I y 2020-II.

Los atributos considerados para este estudio fueron los siguientes: tipo de admisión, dependencia del estudiante, género, edad, estado civil, preparación universitaria, bienestar psicológico, situación laboral del estudiante, situación socioeconómica y el promedio ponderado semestral. Cada uno de estos atributos se seleccionó por su potencial relevancia en el rendimiento académico de los estudiantes.

En cuanto a los resultados descriptivos, se puede observar que, de los 861 estudiantes ingresantes, 48 estudiantes (5.6%) no contaban con información suficiente en algunos atributos, lo que los excluyó de ciertos análisis, dejando un total de 813 registros completos. La mayoría de los estudiantes, 550 (67.7%), ingresaron por admisión regular, mientras que otros lo hicieron a través de modalidades especiales como traslados internos y externos. En términos de dependencia estudiantil, el 60.3% de los estudiantes proviene de contextos dependientes (familias con mayor intervención en el soporte educativo), lo que sugiere la influencia del entorno familiar en el rendimiento académico.

La distribución por género muestra que el 63.2% de los estudiantes son hombres (514 estudiantes), mientras que el 36.8% son mujeres (299 estudiantes). La edad de los estudiantes varía considerablemente, con un promedio de 18.81 años. La mayoría de los estudiantes (el 50% del total) tiene entre 17 y 20 años, lo que es consistente con el rango de edad esperado para estudiantes de primer semestre. En cuanto al estado civil, el 96.4% de los estudiantes son solteros (784 estudiantes), mientras que un 3.6% son casados (29 estudiantes).

En relación con la preparación académica previa, se encontró que el 70% de los estudiantes tienen una formación previa en educación técnica o preuniversitaria. Este dato resalta la importancia de los programas de preparación académica en el rendimiento universitario, ya que aquellos con mayor preparación tienden a tener mejores promedios semestrales. El bienestar psicológico de los estudiantes reveló que el 85% de los estudiantes no presentaron indicadores de estrés severo, mientras que el 15% reportó algún tipo de afectación psicológica.

En cuanto a la situación laboral, el 40% de los estudiantes trabaja a tiempo parcial o completo, lo que podría influir en su capacidad para dedicar tiempo al estudio. Los datos sobre la situación socioeconómica muestran que el 55% de los estudiantes proviene de hogares con ingresos bajos o muy bajos, lo cual puede estar relacionado con dificultades para cubrir los costos asociados con su educación.

En términos de rendimiento académico, el promedio ponderado semestral de los estudiantes es de 11.68, con una desviación estándar de 3.92, lo que indica una amplia variabilidad en el rendimiento académico. Aproximadamente el 40% de los estudiantes tiene un promedio entre 10 y 14, lo que indica un rendimiento aceptable. Sin embargo, un 15% de los estudiantes presenta promedios por debajo de 10, lo que sugiere un riesgo de fracaso académico, mientras que solo el 10% tiene un promedio superior a 14.

3.2. Preparación de Datos

En la fase de preparación de datos, se llevaron a cabo varios pasos clave para garantizar que los datos estuvieran listos para ser procesados por los modelos predictivos. El primer paso fue la limpieza de datos, donde se eliminaron duplicados y registros incompletos, lo que redujo el conjunto inicial de 861 registros a 813 completos. Se utilizó la imputación estadística para tratar los valores faltantes, y se realizó una normalización de los datos numéricos para mejorar la

eficacia de los algoritmos de clasificación. Además, las variables categóricas, como la situación socioeconómica y el estado civil, fueron transformadas en variables numéricas mediante codificación binaria y polinominal.

Una parte crucial de la preparación de datos fue el análisis de correlación entre los atributos del conjunto de datos y el rendimiento académico, medido por el promedio ponderado semestral. A continuación, se presenta la tabla completa con las correlaciones calculadas entre los atributos y el promedio semestral.

Tabla 1. Correlación de atributos con el rendimiento académico (promedio semestral)

Atributo	Correlación con el Promedio Semestral
Dependencia del estudiante	153.482
Situación Socioeconómica	136.245
Sexo	58.816
Bienestar Psicológico	-1.672
Estado Civil	-22.913
Condición Laboral	-92.197
Edad	-100.916
Tipo de Admisión	-137.798
Preparación Universitaria	-154.236

El análisis de correlación revela que la dependencia del estudiante y la situación socioeconómica son los factores con mayor influencia positiva en el rendimiento académico, con valores de 153.482 y 136.245, respectivamente. Esto sugiere que los estudiantes que cuentan con un mayor apoyo familiar y mejores recursos económicos tienden a obtener mejores resultados. Estas variables reflejan la importancia del entorno familiar y la estabilidad económica para facilitar el éxito académico, ya que los estudiantes en condiciones más favorables tienen mayor capacidad para centrarse en sus estudios sin las distracciones o limitaciones que generan la falta de recursos.

Por otro lado, variables como la edad, la condición laboral, el tipo de admisión y la preparación universitaria mostraron correlaciones negativas significativas, lo que indica que los estudiantes mayores (-100.916) o aquellos que trabajan (-92.197) enfrentan más dificultades académicas, probablemente debido a la falta de tiempo y el desgaste por cumplir con responsabilidades adicionales. Asimismo, el tipo de admisión (-137.798) y la preparación universitaria (-154.236) también reflejan desafíos para aquellos que ingresaron mediante programas especiales o con una formación previa insuficiente, lo que les dificulta mantener un rendimiento académico óptimo. Aunque el bienestar psicológico y el estado civil mostraron correlaciones menos significativas, es posible que estos factores influyan en casos particulares donde el estrés o las responsabilidades familiares afecten el desempeño académico.

3.3. Evaluación de los modelos

En la fase de evaluación, se probaron tres algoritmos de aprendizaje supervisado: K-Vecinos más cercanos (KNN), Naive Bayes (NB) y Árbol de Decisión (AD). A continuación, se presenta el análisis de las matrices de confusión y las métricas de evaluación de cada modelo, expresadas con precisión numérica.

Análisis de matrices de confusión

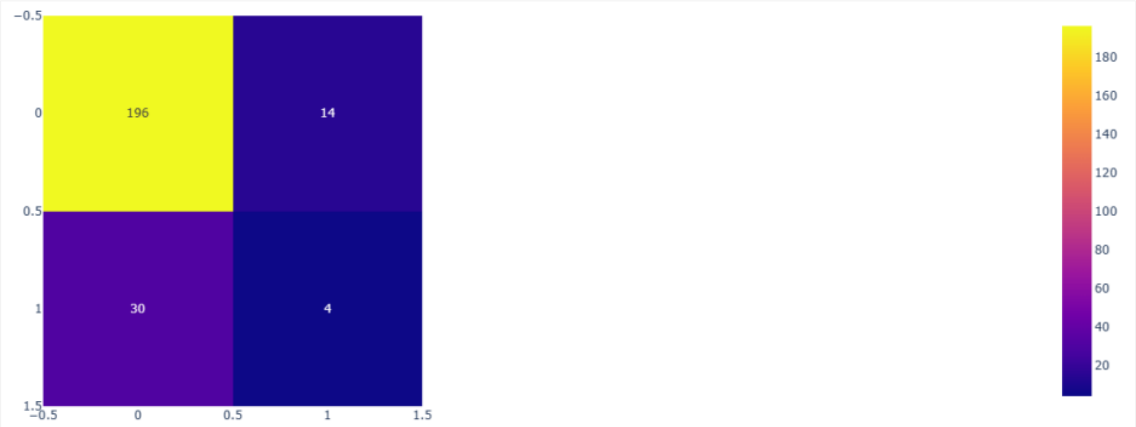

Figura 1. Matriz de confusión del modelo K-Vécinos más cercanos

El KNN identificó correctamente a 196 estudiantes de bajo rendimiento (TN), pero clasificó erróneamente a 14 estudiantes de bajo rendimiento como si tuvieran un alto rendimiento (FP). En cuanto a los estudiantes con alto rendimiento, el modelo logró identificar correctamente solo a 4 estudiantes (TP), mientras que dejó de clasificar correctamente a 30 estudiantes con alto rendimiento (FN). Este análisis indica que el modelo KNN tuvo un buen rendimiento en la clasificación de estudiantes con bajo rendimiento, pero tuvo dificultades significativas en identificar correctamente a los estudiantes con alto rendimiento.

Naive Bayes (NB):

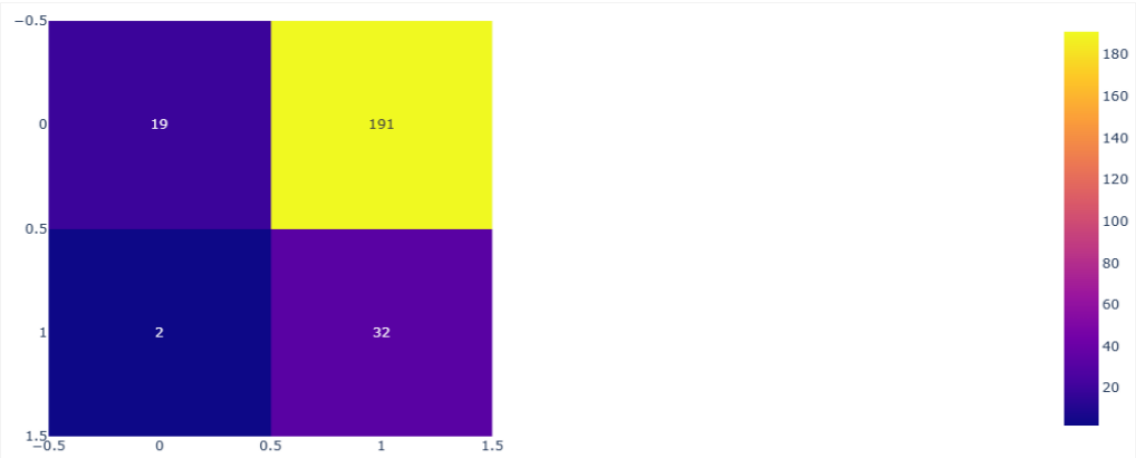

Figura 2. Matriz de confusión del modelo Naive Bayes

El NB identificó correctamente a 32 estudiantes con alto rendimiento (TP), mientras que solo 2 estudiantes con alto rendimiento fueron clasificados incorrectamente como de bajo rendimiento (FN). Sin embargo, el modelo clasificó incorrectamente a 191 estudiantes con bajo rendimiento como si tuvieran un buen desempeño (FP), y solo identificó correctamente a 19 estudiantes de bajo rendimiento (TN). Este análisis sugiere que, aunque el NB tiene un buen desempeño en la clasificación de estudiantes con alto rendimiento, tiene una alta tasa de falsos positivos, lo que significa que asigna incorrectamente a muchos estudiantes con bajo rendimiento como si tuvieran un buen rendimiento académico.

Árbol de Decisión (AD):

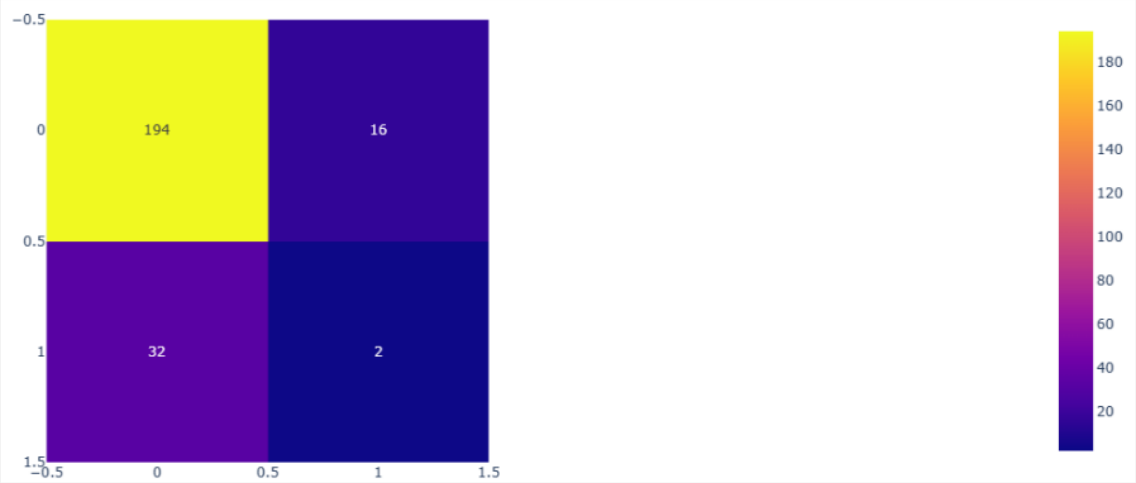

Figura 3. Matriz de confusión del modelo Árbol de Decisión

El modelo AD identificó correctamente a 2 estudiantes con alto rendimiento (TP), mientras que clasificó incorrectamente a 32 estudiantes con alto rendimiento como si tuvieran bajo rendimiento (FN). En cuanto a los estudiantes con bajo rendimiento, el modelo logró identificar correctamente a 194 de ellos (TN), pero clasificó erróneamente a 16 estudiantes de bajo rendimiento como si tuvieran un buen desempeño (FP). Este análisis muestra que el modelo AD es relativamente bueno para identificar a los estudiantes con bajo rendimiento, pero tiene dificultades importantes para identificar a los estudiantes con alto rendimiento, como lo demuestra el bajo número de verdaderos positivos (TP) y el alto número de falsos negativos (FN).

Tabla 2. Métricas de evaluación de los algoritmos

Algoritmo	Precisión	Recuperación	Exactitud (Accuracy)	AUC-ROC
KNN	0.2222	0.1176	0.8197	0.5255
Naive Bayes	0.1435	0.9412	0.2090	0.5158
Árbol de Decisión	0.1111	0.0588	0.8033	0.4913

De los tres modelos evaluados, KNN se destacó por su exactitud (0.8197) y un AUC-ROC (0.5255), lo que lo convierte en el modelo con mejor equilibrio general. Aunque su recuperación fue baja (0.1176), lo que significa que no detectó correctamente a muchos estudiantes con alto rendimiento, su precisión moderada (0.2222) sugiere que, cuando el modelo clasifica un estudiante como de alto rendimiento, lo hace con mayor certeza en comparación con los otros modelos. Esto lo posiciona como un buen predictor global del rendimiento académico.

Por otro lado, NB mostró una recuperación extremadamente alta (0.9412), indicando que el modelo identificó correctamente a casi todos los estudiantes de alto rendimiento. Sin embargo, su precisión baja (0.1435) y su exactitud (0.2090) limitan su utilidad, ya que el modelo comete muchos errores al clasificar estudiantes de bajo rendimiento como si tuvieran alto rendimiento. Esto sugiere que, aunque NB es bueno detectando casos positivos, no es confiable en términos generales.

Finalmente, AD presentó el desempeño más bajo en todas las métricas, con una precisión de 0.1111 y una recuperación de 0.0588, lo que refleja su incapacidad para clasificar correctamente tanto a estudiantes con alto rendimiento como de bajo rendimiento. Aunque su exactitud (0.8033) fue relativamente alta, no alcanzó el nivel de KNN, lo que lo hace menos confiable en este contexto.

4. Discusión

Los resultados de esta investigación demuestran que los algoritmos de aprendizaje supervisado, en especial K-Vecinos más cercanos (KNN), ofrecen un modelo predictivo eficaz para identificar el rendimiento académico de los estudiantes. La exactitud del 81.97% alcanzada por este modelo indica que es capaz de clasificar correctamente a la mayoría de los estudiantes, lo cual concuerda con estudios previos que señalan el potencial del análisis de datos para mejorar la toma de decisiones en la educación superior (Haleem et al., 2022). Al permitir la identificación temprana de estudiantes con bajo rendimiento, el uso de estos algoritmos facilita la implementación de estrategias de intervención, tales como tutorías o becas, para mejorar los resultados académicos (Sghir et al., 2023). Este hallazgo refuerza la relevancia de aplicar modelos predictivos en contextos educativos para mejorar la retención y la experiencia de los estudiantes, como se indica en la literatura (Centoni & Maruotti, 2021).

A pesar de los avances logrados, el estudio también destaca las limitaciones de los modelos predictivos en ciertos contextos. La baja recuperación de KNN sugiere que, si bien el modelo clasifica correctamente a los estudiantes con alto rendimiento, no es igualmente efectivo para identificar a aquellos que están en riesgo de fracasar. Esto coincide con estudios que subrayan la importancia de ajustar los modelos a las características específicas de los estudiantes, particularmente en entornos con desafíos socioeconómicos y culturales (Núñez-Canal et al., 2022). Las instituciones educativas, especialmente en regiones con características particulares, pueden necesitar adaptar los modelos a sus necesidades, tal como sugieren Nsanzumuhire y Groot (2020), para que los algoritmos logren una mayor sensibilidad y precisión en la identificación de estudiantes en riesgo.

En última instancia, la investigación confirma que el uso de algoritmos de aprendizaje supervisado ofrece soluciones tecnológicas valiosas para los desafíos actuales de la educación superior (Issah et al., 2023). No obstante, también resalta la necesidad de seguir explorando cómo estos modelos pueden ser optimizados y personalizados según el contexto institucional. Al igual que Matz et al. (2023) sugieren, el análisis de factores como el nivel socioeconómico y el bienestar psicológico es crucial para entender el rendimiento académico. Esto indica que las instituciones deben combinar estas tecnologías con políticas de apoyo social y económico, fortaleciendo las capacidades predictivas de los algoritmos para lograr un impacto positivo en la educación superior.

5. Conclusiones

El estudio permitió determinar que el modelo K-Vecinos más cercanos (KNN) es el más adecuado para predecir el rendimiento académico de los estudiantes universitarios, logrando una exactitud del 81.97%. A pesar de tener una recuperación moderada, su capacidad para clasificar con precisión a los estudiantes con alto rendimiento lo convierte en una herramienta útil para la toma de decisiones en entornos educativos. La implicancia de estos hallazgos sugiere que KNN puede ser utilizado de manera eficaz por las instituciones para identificar a los estudiantes que requieren apoyo académico temprano, mejorando así las estrategias de intervención y personalización del proceso educativo. Esto podría llevar a una mejora en el rendimiento global de los estudiantes y reducir las tasas de deserción.

Financiamiento

Ninguno.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Contribución de autores

Conceptualización: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Curación de datos: Vargas-Quispe, Alex Ali.

Análisis formal: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Investigación: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Metodología: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Software: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Supervisión: Prieto-Luna, Jaime Cesar.

Validación: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Visualización: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Escritura - preparación del borrador original: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Escritura - revisión y edición: Vargas-Quispe, Alex Ali y Prieto-Luna, Jaime Cesar.

Referencias bibliográficas

Abelha, M., Fernandes, S., Mesquita, D., Seabra, F., & Ferreira-Oliveira, A. T. (2020). Graduate Employability and Competence Development in Higher Education—A Systematic Literature Review Using PRISMA. Sustainability, 12(15), 5900. https://doi.org/10.3390/su12155900

Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A.-E., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. International Journal of Educational Technology in Higher Education, 20(1), 3. https://doi.org/10.1186/s41239-022-00371-5

Centoni, M., & Maruotti, A. (2021). Students' evaluation of academic courses: An exploratory analysis to an Italian case study. Studies in Educational Evaluation, 70, 101054. https://doi.org/10.1016/j.stueduc.2021.101054

Goss, H. (2022). Student Learning Outcomes Assessment in Higher Education and in Academic Libraries: A Review of the Literature. The Journal of Academic Librarianship, 48(2), 102485. https://doi.org/10.1016/j.acalib.2021.102485

Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Understanding the role of digital technologies in education: A review. Sustainable Operations and Computers, 3, 275–285. https://doi.org/10.1016/j.susoc.2022.05.004

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Decision Analytics Journal, 7, 100204. https://doi.org/10.1016/j.dajour.2023.100204

Johar, N. A., Kew, S. N., Tasir, Z., & Koh, E. (2023). Learning Analytics on Student Engagement to Enhance Students' Learning Performance: A Systematic Review. Sustainability, 15(10), 7849. https://doi.org/10.3390/su15107849

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. Sustainability, 15(16), 12451. https://doi.org/10.3390/su151612451

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports, 13(1), 5705. https://doi.org/10.1038/s41598-023-32484-w

Matzavela, V., & Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. Computers and Education: Artificial Intelligence, 2, 100035. https://doi.org/10.1016/j.caeai.2021.100035

Nsanzumuhire, S. U., & Groot, W. (2020). Context perspective on University-Industry Collaboration processes: A systematic review of literature. Journal of Cleaner Production, 258, 120861. https://doi.org/10.1016/j.jclepro.2020.120861

Núñez-Canal, M., de Obesso, M. de las M., & Pérez-Rivero, C. A. (2022). New challenges in higher education: A study of the digital competence of educators in Covid times. Technological Forecasting and Social Change, 174, 121270. https://doi.org/10.1016/j.techfore.2021.121270

Rahman, S., Munam, A. M., Hossain, A., Hossain, A. S. M. D., & Bhuiya, R. A. (2023). Socio-economic factors affecting the academic performance of private university students in Bangladesh: a cross-sectional bivariate and multivariate analysis. SN Social Sciences, 3(2), 26. https://doi.org/10.1007/s43545-023-00614-w

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). Education and Information Technologies, 28(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0