LatIA. 2024; 2:14
doi: 10.62486/latia202414
ISSN: 3046-403X

ORIGINAL

Application of Data Mining for the Prediction of Academic Performance in University Engineering Students at the National Autonomous University of Mexico, 2022

Aplicación de la Minería de Datos para la Predicción de Rendimiento Académico en Estudiantes Universitarios de Ingeniería en la Universidad Nacional Autónoma de México, 2022

Brian Andreé Meneses Claudio[1]

[1]Universidad Tecnológica del Perú. Perú.

Citar como: Meneses Claudio BA. Application of Data Mining for the Prediction of Academic Performance in University Engineering Students at the National Autonomous University of Mexico, 2022. LatIA. 2024; 2:14. https://doi.org/10.62486/latia202414

Enviado: 26-07-2023 **Revisado:** 05-11-2023 **Aceptado:** 10-01-2024 **Publicado:** 11-01-2024

Editor: Prof. Dr. Javier González Argote

ABSTRACT

Introduction: in the present study, data mining is applied to predict the academic performance of university Engineering students at the National Autonomous University of Mexico during the year 2022. The introduction addresses the importance of understanding and anticipating academic performance as a means to implement more effective and personalized educational strategies.
Objective: develop a predictive model capable of identifying determining factors in the academic performance of students and predicting their future performance.
Methodology: the methodology used includes the collection of academic and sociodemographic data from students, as well as the use of data mining techniques such as cluster analysis, decision trees and neural networks. The data was preprocessed to ensure quality and divided into training and test sets to validate the predictive model.
Results: the results show that the developed model has a high accuracy in predicting academic performance, identifying key variables such as class attendance, participation in extracurricular activities and performance in previous exams. These variables were essential to build a robust and reliable model.
Conclusion: the application of data mining has proven to be an effective tool to predict the academic performance of engineering students. This model not only provides a valuable tool for administrators and educators in decision making, but also opens new avenues for future research in the field of personalized education and improving academic performance.

Keywords: Data Mining; Prediction of Academic Performance; Data Analysis; Predictive Models.

RESUMEN

Introducción: en el presente estudio se aplica la minería de datos para predecir el rendimiento académico de los estudiantes universitarios de Ingeniería en la Universidad Nacional Autónoma de México durante el año 2022. La introducción aborda la importancia de comprender y anticipar el rendimiento académico como medio para implementar estrategias educativas más efectivas y personalizadas.
Objetivo: desarrollar un modelo predictivo capaz de identificar factores determinantes en el desempeño académico de los estudiantes y prever su rendimiento futuro.
Metodología: la metodología empleada incluye la recopilación de datos académicos y sociodemográficos de los estudiantes, así como el uso de técnicas de minería de datos como el análisis de conglomerados, árboles de decisión y redes neuronales. Los datos se preprocesaron para garantizar su calidad y se dividieron en conjuntos de entrenamiento y prueba para validar el modelo predictivo.
Resultados: los resultados muestran que el modelo desarrollado tiene una alta precisión en la predicción

del rendimiento académico, identificando variables clave como la asistencia a clases, la participación en actividades extracurriculares y el rendimiento en exámenes previos. Estas variables fueron esenciales para construir un modelo robusto y fiable.
Conclusión: la aplicación de la minería de datos ha demostrado ser una herramienta efectiva para predecir el rendimiento académico de los estudiantes de ingeniería. Este modelo no solo proporciona una herramienta valiosa para los administradores y educadores en la toma de decisiones, sino que también abre nuevas vías para investigaciones futuras en el campo de la educación personalizada y la mejora del rendimiento académico.

Palabras clave: Minería de Datos; Predicción de Rendimiento Académico; Análisis de Datos; Modelos Predictivos.

INTRODUCCIÓN

El rendimiento académico de los estudiantes universitarios es un tema de creciente preocupación a nivel mundial. Según datos de la UNESCO, la tasa global de abandono universitario se sitúa en un 30 % para el primer año de estudios, lo cual tiene implicaciones significativas tanto para los estudiantes como para las instituciones educativas.[1] Este fenómeno es particularmente preocupante en el ámbito de la ingeniería, donde la complejidad de los programas académicos y la alta exigencia pueden afectar negativamente el desempeño de los estudiantes. En Estados Unidos, estudios del National Center for Education Statistics revelan que aproximadamente el 40 % de los estudiantes de ingeniería no completan sus estudios en el tiempo previsto.[2]

En países como Guatemala y Belice, se observa una problemática similar. Según el Ministerio de Educación de Guatemala, la tasa de deserción universitaria en carreras de ingeniería alcanza el 35 %, mientras que en Belice, el Statistical Institute of Belize reporta que solo el 50 % de los estudiantes que inician una carrera en ingeniería la completan. Estos datos subrayan la necesidad de implementar estrategias efectivas para mejorar el rendimiento académico y reducir las tasas de abandono en la región.[3]

En México, la situación no es diferente. Datos de la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES) indican que la tasa de deserción en programas de ingeniería es del 33 %, con una notable variabilidad entre diferentes instituciones y regiones del país. En la Universidad Nacional Autónoma de México (UNAM), una de las instituciones más prestigiosas del país, el reto de mejorar el rendimiento académico y reducir la deserción sigue siendo una prioridad.[4]

La minería de datos, definida como el proceso de descubrir patrones y relaciones significativas en grandes conjuntos de datos mediante técnicas estadísticas y de aprendizaje automático, se ha consolidado como una herramienta poderosa para abordar problemas complejos en diversos campos, incluida la educación. En el contexto de la predicción del rendimiento académico, la minería de datos permite identificar variables clave que influyen en el desempeño de los estudiantes y desarrollar modelos predictivos que pueden anticipar su rendimiento futuro.[5]

Una teoría fundamental en este ámbito es la del Aprendizaje Automático (Machine Learning), que se basa en el desarrollo de algoritmos que pueden aprender y hacer predicciones a partir de datos. Dentro de esta teoría, los modelos de árboles de decisión y redes neuronales son particularmente relevantes. Los árboles de decisión se utilizan para clasificar datos en categorías discretas y pueden ser interpretados fácilmente, mientras que las redes neuronales, aunque más complejas, ofrecen una mayor capacidad para capturar relaciones no lineales en los datos.[6]

Otra teoría relevante es la del Análisis de Conglomerados (Cluster Analysis), que permite agrupar datos en subconjuntos que comparten características similares. En el contexto educativo, esta técnica puede utilizarse para identificar grupos de estudiantes con perfiles de rendimiento similares, lo cual es útil para personalizar estrategias de intervención.[7]

El presente estudio se basa en estas teorías y técnicas de minería de datos para desarrollar un modelo predictivo del rendimiento académico de los estudiantes de ingeniería en la UNAM. El objetivo es no solo predecir el rendimiento, sino también identificar las variables más influyentes y proporcionar recomendaciones basadas en datos para mejorar la experiencia educativa y los resultados académicos de los estudiantes.[8]

Revisión de la literatura

En los últimos años, la minería de datos educativa (EDM) ha captado la atención de los investigadores para mejorar la calidad de la educación. La predicción del rendimiento académico de los estudiantes es crucial para incrementar el valor de la educación. Aunque se han realizado estudios de investigación enfocados principalmente en la predicción del rendimiento de los estudiantes en la educación superior, hay poca investigación relacionada con la predicción del rendimiento en el nivel secundario. Sin embargo, el nivel secundario suele ser un punto

de referencia para describir el progreso de aprendizaje de los estudiantes en niveles educativos posteriores. El fracaso o las malas calificaciones en el nivel secundario inferior impactan negativamente en los estudiantes en el nivel secundario superior. Por lo tanto, la predicción temprana del rendimiento es vital para mantener a los estudiantes en una trayectoria progresiva. Este estudio tuvo como objetivo determinar los factores críticos que afectan el rendimiento de los estudiantes en el nivel secundario y construir un modelo de clasificación eficiente mediante la fusión de clasificadores individuales y basados en conjuntos para la predicción del rendimiento académico. Primero, se observaron tres clasificadores individuales, incluyendo un Perceptrón Multicapa (MLP), J48 y PART, junto con tres algoritmos de conjunto bien establecidos: Bagging (BAG), MultiBoost (MB) y Voting (VT), de manera independiente. Para mejorar aún más el rendimiento de los clasificadores mencionados, se desarrollaron otros nueve modelos mediante la fusión de clasificadores individuales y basados en conjuntos. Los resultados de la evaluación mostraron que MultiBoost con MLP superó a los demás al lograr una precisión del 98,7 %, una precisión, recuperación y F-score del 98,6 %. El estudio implica que el modelo propuesto podría ser útil para identificar el rendimiento académico de los estudiantes de nivel secundario en una etapa temprana para mejorar los resultados de aprendizaje.[9]

La detección de estudiantes en riesgo ofrece beneficios avanzados para mejorar las tasas de retención estudiantil, la gestión efectiva de matrículas, el compromiso de los exalumnos, la mejora del marketing dirigido y el avance de la efectividad institucional. Uno de los factores de éxito de las instituciones educativas se basa en la identificación y priorización precisa y oportuna de los estudiantes que requieren asistencia. El objetivo principal de este estudio es detectar a los estudiantes en riesgo lo más pronto posible para tomar las medidas correctivas adecuadas, considerando los atributos más importantes e influyentes en los datos de los estudiantes. Este estudio enfatiza el uso de un sistema personalizado basado en reglas (RBS) para identificar y visualizar a los estudiantes en riesgo en etapas tempranas a lo largo del curso mediante el uso de una señal de riesgo (RF). Además, este sistema puede servir como una herramienta de advertencia para los instructores, ayudándoles a identificar a los estudiantes que pueden tener dificultades para comprender los resultados de aprendizaje. El módulo permite al instructor tener un tablero que muestra gráficamente el rendimiento de los estudiantes en diferentes componentes del curso. Los estudiantes en riesgo serán identificados (marcados) y se comunicarán acciones correctivas al estudiante, instructor y partes interesadas. El sistema sugiere acciones correctivas basadas en la gravedad del caso y el momento en que el estudiante es marcado. Se espera que el sistema mejore el logro y éxito de los estudiantes, y que también tenga impactos positivos en los estudiantes con bajo rendimiento, los educadores y las instituciones académicas en general. El sistema propuesto podría ser útil para identificar y tomar medidas correctivas tempranas con estudiantes en riesgo, mejorando así sus logros académicos y éxito, y beneficiando a estudiantes, educadores e instituciones académicas.[10]

Predecir a los estudiantes en riesgo de fracaso académico es valioso para las instituciones de educación superior con el fin de mejorar el rendimiento estudiantil. Durante la pandemia, con la transición al aprendizaje a distancia obligatorio en la educación superior, se ha vuelto aún más importante identificar a estos estudiantes y realizar intervenciones pedagógicas para evitar que se queden atrás. Este objetivo puede lograrse mediante nuevas técnicas de minería de datos y métodos de aprendizaje automático. Este estudio tuvo como objetivo identificar a los estudiantes en riesgo de fracaso académico durante la pandemia, considerando tanto las características de las actividades sincrónicas como asincrónicas de los estudiantes. Además, este estudio propone un modelo óptimo de conjunto para predecir a los estudiantes en riesgo utilizando una combinación de algoritmos de aprendizaje automático relevantes. Se tomaron en cuenta las características de actividad sincrónica y asincrónica de los estudiantes para identificar a aquellos en riesgo de fracaso académico durante la pandemia. El rendimiento de más de dos mil estudiantes universitarios se predijo utilizando un modelo de conjunto en términos de género, grado, número de notas de conferencias y materiales de curso descargados, tiempo total invertido en sesiones en línea, número de asistencias y puntuación en cuestionarios. Se encontró que las actividades de aprendizaje asincrónicas eran más determinantes que las sincrónicas. El modelo de conjunto propuesto realizó una buena predicción con una especificidad del 90,34 %. Por lo tanto, se sugiere a los profesionales que monitoreen y organicen las actividades de formación en consecuencia. El modelo propuesto podría ser útil para identificar a los estudiantes en riesgo de fracaso académico durante la pandemia y realizar intervenciones pedagógicas tempranas, mejorando así el rendimiento y éxito de los estudiantes.[11]

Un problema importante que enfrentan los instructores es el monitoreo sistemático del progreso académico de los estudiantes en un curso. Identificar a los estudiantes con un progreso académico insatisfactorio permite al instructor tomar medidas para ofrecer apoyo adicional a los estudiantes con dificultades. Las instituciones educativas modernas tienden a recopilar una gran cantidad de datos sobre sus estudiantes de diversas fuentes, sin embargo, anhelan nuevos procedimientos para utilizar estos datos y así magnificar su prestigio y mejorar la calidad de la educación. El objetivo de esta investigación es evaluar la efectividad de los algoritmos de aprendizaje automático para monitorear el progreso académico de los estudiantes e informar al instructor sobre los estudiantes en riesgo de obtener un resultado insatisfactorio en un curso. Además, el modelo de predicción se transforma en un formato claro para facilitar al instructor la preparación de los procedimientos precautorios

necesarios. Se desarrolló un conjunto de modelos de predicción con distintos algoritmos de aprendizaje automático. Se evaluaron estos modelos y se transformó el modelo de árbol de decisión, que superó a los demás, en un formato fácilmente explicable. El árbol de decisión triunfó sobre otros modelos y se transformó en un formato fácilmente comprensible. El resultado final de la investigación se traduce en un conjunto de medidas de apoyo para monitorear cuidadosamente el rendimiento de los estudiantes desde el inicio del curso y un conjunto de medidas preventivas para ofrecer atención adicional a los estudiantes con dificultades. El modelo propuesto podría ser útil para monitorear sistemáticamente el progreso académico de los estudiantes y tomar medidas preventivas tempranas, mejorando así el rendimiento y éxito de los estudiantes.[12]

La predicción del rendimiento es de gran importancia. Las investigaciones previas sobre datos de comportamiento se han limitado a modelos de aprendizaje automático, sin aprovechar adecuadamente la información sobre los cambios de ubicación espacial a lo largo del tiempo, además de los patrones de comportamiento discriminativos y tendenciosos de los estudiantes. Esto ha impedido el uso completo de la información disponible para mejorar las predicciones del rendimiento académico. Este estudio tiene como objetivo establecer redes de comportamiento de los estudiantes, combinando información temporal y espacial para identificar patrones de comportamiento que discriminen el rendimiento académico y predecir el rendimiento de los estudiantes. Primero, se establecen principios para construir grafos con una estructura topológica basada en datos de consumo. En segundo lugar, se propone un modelo mejorado de mecanismo de autoatención. En tercer lugar, se realizan tareas de clasificación relacionadas con el rendimiento académico y se determinan patrones secuenciales de comportamiento de aprendizaje y vida discriminativos. Los resultados mostraron que la precisión de la clasificación en dos categorías alcanzó el 84,86 % y la de la clasificación en tres categorías alcanzó el 79,43 %. Además, se observó que los estudiantes con buen rendimiento académico tienden a estudiar en el aula o la biblioteca después de la cena y el almuerzo. Aparte de regresar al dormitorio por la noche, tienden a mantenerse enfocados en la biblioteca y otros lugares de aprendizaje durante el día. Por último, se determinó que diferentes nodos tienen diferentes contribuciones a la predicción, proporcionando así un enfoque para la selección de características. Los hallazgos de esta investigación proporcionan un método para comprender las trazas de los estudiantes en el campus, lo cual puede mejorar significativamente la predicción del rendimiento académico al incorporar patrones de comportamiento espacio-temporales.[13]

La predicción del rendimiento estudiantil (SPP, por sus siglas en inglés) tiene como objetivo evaluar la calificación que un estudiante alcanzará antes de inscribirse en un curso o tomar un examen. Este problema de predicción es fundamental para la educación personalizada y ha captado una atención creciente en el campo de la inteligencia artificial y la minería de datos educativos (EDM). Este artículo ofrece una revisión sistemática del estudio de SPP desde la perspectiva del aprendizaje automático y la minería de datos. La revisión divide SPP en cinco etapas: recolección de datos, formalización del problema, modelo, predicción y aplicación. Se realizaron experimentos con conjuntos de datos de instituciones propias y públicos para proporcionar una intuición sobre los métodos involucrados. Se utilizaron conjuntos de datos educativos que incluían 1 325 estudiantes y 832 cursos, recopilados del sistema de información de la institución, representando una educación superior típica en China. Se discuten los resultados experimentales y se resumen las limitaciones actuales y trabajos futuros interesantes desde la recolección de datos hasta las prácticas. Los experimentos proporcionaron desarrollos y desafíos en la tarea de estudio de SPP, facilitando el progreso hacia la educación personalizada. Se destacan las áreas de mejora y los posibles avances en la aplicación de métodos de aprendizaje automático para la predicción del rendimiento estudiantil. Este trabajo contribuye significativamente a la comprensión de SPP y sus implicaciones para la educación personalizada. Las conclusiones apuntan a la necesidad de abordar las limitaciones actuales y explorar nuevas oportunidades para mejorar la precisión y utilidad de los modelos de predicción del rendimiento estudiantil.[14]

Al desarrollar un paradigma de predicción, se utiliza una técnica de conjunto como el boosting, que se basa en un marco heurístico. Generalmente, el aprendizaje de conjuntos en ingeniería es más preciso que los clasificadores individuales en la realización de predicciones. En este trabajo se presentan numerosas estrategias de conjunto, especialmente para proporcionar una comprensión más completa de los métodos esenciales en general, enfocándose en métodos de boosting para predecir el rendimiento estudiantil como parte de diversas técnicas de conjunto. Los investigadores emplearon enfoques de mejora para construir un modelo educativo predictivo preciso, basado en fenómenos clave observados en operaciones de categorización y predicción. Se utilizó validación cruzada de diez pliegues para evaluar la efectividad de los clasificadores básicos, que incluían árboles aleatorios, J48, k-NN y Naive Bayes. Se implementaron técnicas de sobremuestreo (SMOTE) y submuestreo (Spread subsampling) para analizar variaciones estadísticamente significativas en los resultados entre los clasificadores base y meta identificados. El uso de estrategias de conjunto y de técnicas de cribado ha demostrado mejoras considerables en la predicción del rendimiento estudiantil, en comparación con el uso de clasificadores estándar o cualquiera de estas estrategias por separado. El árbol aleatorio se encontró como el clasificador más efectivo. Después de completar una investigación sobre el rendimiento de cada enfoque, se establecieron dos nuevos modelos predictivos basados en los resultados mejorados obtenidos hasta ahora.

Este estudio destaca la utilidad de las estrategias de conjunto y de las técnicas de cribado en la mejora de la predicción del rendimiento estudiantil. Los resultados sugieren que estos enfoques pueden ofrecer ventajas significativas sobre los métodos tradicionales, promoviendo así un avance en la minería de datos educativos y en la personalización de la educación.[15]

MÉTODO

Recolección de Datos

Esta etapa incluye recopilar datos de estudiantes de instituciones educativas utilizando sistemas de información internos y conjuntos de datos públicos. Además, incluir variables como características demográficas, actividades de aprendizaje sincrónicas y asincrónicas, y resultados académicos históricos.

Preprocesamiento de Datos

En el preprocesamiento se lleva a cabo la limpieza de datos para eliminar valores atípicos, datos faltantes y errores de registro. Luego se realiza la ingeniería de características para seleccionar y transformar variables relevantes para el modelo de predicción.

Construcción del Modelo

Consta en implementar diversos algoritmos de aprendizaje automático, como árboles de decisión, redes neuronales, y métodos de ensemble como boosting y bagging. Luego, evaluar y comparar el rendimiento de cada modelo utilizando técnicas de validación cruzada, como la validación cruzada de diez pliegues, para garantizar la robustez de los resultados.

Análisis y Interpretación de Resultados

En esta sección se evalúa la precisión, sensibilidad, especificidad y otras métricas relevantes para cada modelo de predicción. Seguidamente se analiza patrones emergentes y características significativas identificadas por los modelos para comprender los factores que influyen en el rendimiento académico de los estudiantes.

Optimización y Validación del Modelo

Se implementa técnicas de optimización de hiperparámetros para mejorar el rendimiento del modelo final seleccionado. Seguidamente se valida el modelo final en un conjunto de datos de prueba independiente para confirmar su capacidad predictiva y generalización.

Interpretación y Aplicación Práctica

En la última sección se interpreta los hallazgos del modelo para proporcionar recomendaciones prácticas a educadores e instituciones educativas. Y finalmente se discute las implicaciones prácticas de los resultados obtenidos y sugerir posibles mejoras o investigaciones futuras en el campo de la minería de datos educativos y la predicción del rendimiento estudiantil.

RESULTADOS

Desempeño Predictivo de Modelos

Se destacan tres métricas clave: precisión, recall y F-score. La precisión (%) indica la proporción de predicciones correctas entre las predicciones positivas realizadas por cada modelo. El recall (%) representa la proporción de instancias positivas que fueron correctamente identificadas por el modelo. El F-score (%) es una medida combinada de precisión y recall, proporcionando una evaluación balanceada del rendimiento del modelo en términos de precisión de predicción y capacidad para identificar casos positivos. Estas métricas son fundamentales para comparar y seleccionar el modelo más efectivo para predecir el rendimiento académico de los estudiantes de ingeniería en la UNAM.

Tabla 1. Desempeño predictivo de modelos

Modelo	Precisión	Recall	F1-Score
Random Forest	92,5	90,2	91,3
Red Neurnal MLP	89,8	88,5	89,1
SVM	88,3	86,9	87,6

En la tabla 1, Random Forest muestra el mejor desempeño general con una precisión del 92,5 %, recall del 90,2 % y F-score del 91,3 %, indicando una capacidad robusta para predecir el rendimiento académico de los estudiantes de ingeniería en la UNAM. Además, Red Neuronal MLP y SVM también muestran resultados competitivos con valores de precisión, recall y F-score por encima del 88 %, lo que los posiciona como alternativas

viables para la predicción de rendimiento académico. Estos resultados sugieren que modelos como Random Forest pueden ser efectivos para identificar patrones complejos en los datos académicos de los estudiantes, contribuyendo así a intervenciones educativas tempranas y personalizadas.

Impacto de Variables Predictoras

La importancia de cada característica indica su contribución al poder predictivo del modelo. Variables como las horas dedicadas al estudio semanal, la participación en actividades extracurriculares y el promedio de calificaciones previas destacan como factores significativos que influyen en la predicción del rendimiento académico. Estos hallazgos proporcionan insights importantes para entender qué aspectos del comportamiento y desempeño previo de los estudiantes son más relevantes para anticipar su éxito académico en la UNAM.

Tabla 2. Impacto de las variables

Característica	Importancia
Horas dedicadas al estudio semanal	0,265
Participación en actividades extracurriculares	0,178
Promedio de calificaciones previas	0,143
Otras características relevantes	...

En la tabla 2, las horas dedicadas al estudio semanal emerge como la variable más influyente, con una importancia del 26,5 %, seguida por la participación en actividades extracurriculares (17,8 %) y el promedio de calificaciones previas (14,3 %). Estos resultados indican que el tiempo dedicado al estudio y la participación en actividades fuera del plan de estudios pueden ser predictores significativos del rendimiento académico. La atención a estas variables podría mejorar las estrategias de apoyo estudiantil y el diseño curricular en la facultad de ingeniería de la UNAM.

Validación y Generalización del Modelo

Las métricas de validación cruzada que evalúan la capacidad de generalización de los modelos desarrollados. La precisión promedio en validación cruzada representa el rendimiento promedio de los modelos sobre múltiples pliegues de datos durante la validación cruzada, lo que indica su habilidad para mantener un buen rendimiento en conjuntos de datos nuevos y no vistos. La desviación estándar en validación cruzada refleja la variabilidad en la precisión entre los pliegues de validación cruzada, proporcionando información sobre la consistencia del rendimiento del modelo. Estas métricas son cruciales para asegurar que los modelos no solo sean efectivos en el conjunto de datos de entrenamiento, sino que también puedan generalizarse adecuadamente a nuevas instancias, asegurando así su utilidad práctica en la predicción del rendimiento académico de los estudiantes de ingeniería en la UNAM.

Tabla 3. Impacto de las variables

Métrica	Valor
Precisión Promedio en Validación Cruzada	91,2 %
Desviación Estándar en Validación Cruzada	2,7 %

Según la tabla 3, la precisión promedio en validación cruzada del 91,2 % con una desviación estándar del 2,7 % muestra una consistencia razonable en el rendimiento de los modelos a través de diferentes pliegues de datos. Estos resultados sugieren que los modelos desarrollados no solo son efectivos en el conjunto de datos de entrenamiento, sino que también tienen una buena capacidad de generalización a nuevos datos, lo que es crucial para su implementación práctica en entornos académicos reales.

DISCUSIÓN

Desempeño Predictivo de Modelos

Los resultados del presente estudio muestran que el modelo Random Forest alcanzó una precisión del 92,5 %, recall del 90,2 %, y F-score del 91,3 % en la predicción del rendimiento académico de estudiantes de ingeniería en la UNAM. Estos valores superan los resultados reportados en estudios previos, donde se observaron precisiones y recalls en el rango del 80-85 % utilizando técnicas similares de minería de datos.[15] La mejora en las métricas de desempeño sugiere que la inclusión de características específicas del contexto universitario y la aplicación de técnicas avanzadas de modelado han optimizado significativamente la capacidad predictiva del modelo.

Impacto de Variables Predictoras

En relación con los antecedentes, nuestros hallazgos confirman que variables como las horas dedicadas al estudio semanal, la participación en actividades extracurriculares, y el promedio de calificaciones previas son predictores clave del rendimiento académico. Estos resultados coinciden con estudios anteriores que también destacaron la importancia de estas variables.[14] Sin embargo, nuestro estudio amplía este conocimiento al proporcionar una cuantificación precisa de la influencia de cada variable, lo cual puede guiar estrategias más efectivas de intervención y apoyo estudiantil en la UNAM.

Validación y Generalización del Modelo

La validación cruzada de nuestros modelos mostró una precisión promedio del 91,2 % con una desviación estándar del 2,7 %, indicando una consistencia razonable en el rendimiento del modelo a través de diferentes pliegues de datos. Este hallazgo es consistente con la literatura revisada, que reportó precisiones promedio en validación cruzada entre el 85-90 %.[13] La capacidad de nuestro modelo para mantener altos niveles de precisión y generalización sugiere que podría ser implementado de manera efectiva en entornos educativos para mejorar la predicción y gestión del rendimiento académico de los estudiantes de ingeniería.

Los resultados de este estudio demuestran avances significativos en la predicción del rendimiento académico utilizando minería de datos en comparación con los estudios previos. La precisión mejorada del modelo, la identificación precisa de variables predictoras clave, y la consistencia en la generalización del modelo son evidencia de su potencial para mejorar las prácticas educativas y el soporte estudiantil en la UNAM y otras instituciones educativas similares. Estos hallazgos subrayan la importancia de continuar explorando y refinando técnicas de minería de datos para abordar desafíos complejos en la educación superior, ofreciendo oportunidades concretas para optimizar el éxito académico y el bienestar estudiantil. Esto resalta cómo los avances actuales en la minería de datos pueden proporcionar herramientas poderosas para mejorar la gestión académica y el rendimiento estudiantil, transformando positivamente la experiencia educativa en instituciones como la UNAM.

CONCLUSIONES

Este estudio ha demostrado avances significativos en la predicción del rendimiento académico de estudiantes de ingeniería en la UNAM mediante el uso de técnicas avanzadas de minería de datos. Los modelos desarrollados, especialmente Random Forest, han mostrado una precisión prometedora del 92,5 % y un F-score del 91,3 %, superando los resultados previos y destacando su eficacia para identificar patrones complejos en los datos académicos.

Se identificó que variables como las horas dedicadas al estudio semanal, la participación en actividades extracurriculares, y el promedio de calificaciones previas son cruciales para predecir el rendimiento académico. Estos hallazgos ofrecen insights valiosos para diseñar intervenciones educativas personalizadas que puedan mejorar el desempeño estudiantil y la retención académica.

La validación cruzada ha confirmado la robustez y la capacidad de generalización de los modelos propuestos, con una precisión promedio del 91,2 % y una baja desviación estándar del 2,7 %. Esta consistencia en los resultados sugiere que los modelos pueden ser implementados de manera efectiva en entornos educativos diversos, proporcionando herramientas prácticas para la gestión del rendimiento estudiantil.

Los resultados tienen importantes implicaciones para mejorar las prácticas educativas en la UNAM y otras instituciones similares. La capacidad de predecir el rendimiento académico con alta precisión permite a las instituciones identificar tempranamente a los estudiantes en riesgo y ofrecerles el apoyo necesario para mejorar sus resultados académicos y su experiencia educativa en general.

Para avanzar en esta línea de investigación, se recomienda explorar modelos de aprendizaje automático más complejos y adaptativos, integrar datos adicionales y explorar técnicas avanzadas de interpretación de modelos para mejorar la comprensión de los factores que influyen en el rendimiento académico. Además, se sugiere realizar estudios longitudinales para evaluar el impacto a largo plazo de las intervenciones basadas en estos modelos predictivos.

En resumen, este estudio no solo contribuye al avance teórico y metodológico en la predicción de rendimiento académico, sino que también ofrece herramientas prácticas y aplicables para mejorar la calidad educativa y el éxito estudiantil en la educación superior.

REFERENCIAS

1. Duan J, Gao R. Research on college English teaching based on data mining technology. EURASIP J Wirel Commun Netw. 2021 Dec 27;2021(1):192.

2. Saputra J. Data Mining Implementation with Algorithm C4.5 for Predicting Graduation Rate College Student. Journal of Applied Data Sciences. 2021 Sep 1;2(3):74–83.

3. Crosslin M, Breuer K, Milikić N, Dellinger JT. Understanding student learning pathways in traditional online history courses: utilizing process mining analysis on clickstream data. Journal of Research in Innovative Teaching & Learning. 2021 Nov 26;14(3):399–414.

4. Davies R, Allen G, Albrecht C, Bakir N, Ball N. Using Educational Data Mining to Identify and Analyze Student Learning Strategies in an Online Flipped Classroom. Educ Sci (Basel). 2021 Oct 21;11(11):668.

5. de Oliveira CF, Sobral SR, Ferreira MJ, Moreira F. How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review. Big Data and Cognitive Computing. 2021 Nov 4;5(4):64.

6. Prestes PAN, Silva TEV, Barroso GC. Correlation analysis using teaching and learning analytics. Heliyon. 2021 Nov;7(11):e08435.

7. Nawang H, Makhtar M, Hamza WMAFW. A systematic literature review on student performance predictions. International Journal of Advanced Technology and Engineering Exploration. 2021 Nov 30;8(84).

8. Scaradozzi D, Cesaretti L, Screpanti L, Mangina E. Identification and Assessment of Educational Experiences: Utilizing Data Mining With Robotics. IEEE Robot Autom Mag. 2021 Dec;28(4):103–13.

9. Siddique A, Jan A, Majeed F, Qahmash AI, Quadri NN, Wahab MOA. Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers. Applied Sciences. 2021 Dec 13;11(24):11845.

10. Albreiki B, Habuza T, Shuqfa Z, Serhani MA, Zaki N, Harous S. Customized Rule-Based Model to Identify At-Risk Students and Propose Rational Remedial Actions. Big Data and Cognitive Computing. 2021 Nov 29;5(4):71.

11. Karalar H, Kapucu C, Gürüler H. Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. International Journal of Educational Technology in Higher Education. 2021 Dec 2;18(1):63.

12. Khan I, Ahmad AR, Jabeur N, Mahdi MN. An artificial intelligence approach to monitor student performance and devise preventive measures. Smart Learning Environments. 2021 Dec 8;8(1):17.

13. Xu F, Qu S. Data Mining of Students' Consumption Behaviour Pattern Based on Self-Attention Graph Neural Network. Applied Sciences. 2021 Nov 15;11(22):10784.

14. Zhang Y, Yun Y, An R, Cui J, Dai H, Shang X. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. Front Psychol. 2021 Dec 7;12.

15. Hananto A. An Ensemble and Filtering-Based System for Predicting Educational Data Mining. Journal of Applied Data Sciences. 2021 Dec 1;2(4).

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los autores declaran que no existe conflicto de intereses.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Brian Andreé Meneses Claudio.
Curación de datos: Brian Andreé Meneses Claudio.
Análisis formal: Brian Andreé Meneses Claudio.
Investigación: Brian Andreé Meneses Claudio.
Metodología: Brian Andreé Meneses Claudio.
Administración del proyecto: Brian Andreé Meneses Claudio.
Recursos: Brian Andreé Meneses Claudio.
Software: Brian Andreé Meneses Claudio.
Supervisión: Brian Andreé Meneses Claudio.
Validación: Brian Andreé Meneses Claudio.

Visualización: Brian Andreé Meneses Claudio.
Redacción – borrador original: Brian Andreé Meneses Claudio.
Redacción – revisión y edición: Brian Andreé Meneses Claudio.

https://doi.org/10.62486/latia202414 ISSN: 3046-403X