Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

https://doi.org/10.23913/ride.v16i32.2841

Artículos científicos

Predicción del rendimiento académico con un modelo de clasificación: una comparación entre estudiantes rurales y urbanos de educación media superior

Prediction Using Two Classification Models: A Comparison between Rural and Urban Upper Secondary Students

Previsão do desempenho acadêmico com um modelo de classificação: uma comparação entre estudantes do ensino médio de áreas rurais e urbanas

Yolanda Moyao Martínez
Benemérita Universidad Autónoma de Puebla, México
yolanda.moyao@correo.buap.mx
https://orcid.org/0000-0002-7259-3525

Carmen Cerón Garnica
Benemérita Universidad Autónoma de Puebla, México
carmen.ceron@correo.buap.mx
https://orcid.org/0000-0001-6480-6810

Resumen

Este trabajo de investigación tuvo como propósito predecir el aprovechamiento académico por debajo de la meta esperada, definida a partir del promedio mínimo aprobatorio establecido por la institución, en estudiantes de nivel medio superior en Puebla a través del uso de modelos predictivos de aprendizaje automático supervisado, basados en técnicas de clasificación y minería de datos. En este estudio, el término *aprovechamiento académico* se emplea como indicador del *rendimiento académico* de los estudiantes. Para ello, se procesaron dos conjuntos de información provenientes de una preparatoria rural y una urbana de la Benemérita Universidad Autónoma de Puebla (BUAP). Se empleó el algoritmo de clasificación de bosques aleatorios *(Random Forest*) propuesto por Breiman (2001), el cual fue entrenado y evaluado de manera independiente en ambos conjuntos de datos, utilizando

como métricas de desempeño la precisión y la sensibilidad (r*ecall*). En ambos conjuntos de datos se obtuvo una precisión del 0.72, lo que indica un desempeño comparable del modelo en los contextos rural y urbano. El modelo permitió identificar elementos clave relacionados con el aprovechamiento académico, tales como, la asistencia a clases y el aprovechamiento anticipado en algunas materias, los cuales mostraron una correspondencia descriptiva con el nivel de escolaridad. Asimismo, los hallazgos revelaron algunas diferencias entre ambos contextos, lo que sugiere la necesidad de implementar propuestas de mejora adaptadas a cada contexto, pero que sean muy particulares a cada entorno. A partir de estos resultados, se derivan implicaciones directas para las organizaciones educativas y el sector público, particularmente en el diseño de estrategias orientadas a la prevención del rezago académico y con ello, la deserción escolar en ambos contextos.

Palabras clave: bosques aleatorios, contexto geográfico, deserción escolar, minería de datos, modelos de clasificación.

Abstract

This study aimed to predict academic performance below the expected target, defined according to the minimum passing grade established by the institution, among upper-secondary education students in Puebla through the use of predictive models based on supervised machine learning, employing classification techniques from data mining. For this purpose, two datasets obtained from a rural and urban upper secondary school of the Benemérita Universidad Autónoma de Puebla (BUAP) were processed. The Random Forest classification algorithm was employed and trained and evaluated independently on both datasets, using accuracy and sensitivity (recall) as performance metrics. An accuracy of 0.72 was obtained for both datasets, indicating comparable model performance in rural and urban contexts. The model made it possible to identify key factors related to academic performance, such as class attendance and prior achievement in specific subjects, which showed a descriptive association with the level of educational attainment. Furthermore, the findings revealed differences between both contexts, suggesting the need for improvement strategies tailored to the specific characteristics of each environment. Based on these results, relevant implications emerge for educational institutions and the public sector, particularly for guiding the design of strategies aimed at preventing academic underachievement and, consequently, reducing school dropout in both rural and urban contexts.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Keywords: random forests, geographical context, school dropout, data mining, classification models.

Resumo

Esta pesquisa teve como objetivo predizer o desempenho acadêmico abaixo da meta esperada, definida pela média mínima de aprovação estabelecida pela instituição, em alunos do ensino médio de Puebla, utilizando modelos preditivos de aprendizado de máquina supervisionado baseados em técnicas de classificação e mineração de dados. Neste estudo, o termo "desempenho acadêmico" é utilizado como indicador do desempenho acadêmico dos alunos. Dois conjuntos de dados foram processados, um de uma escola de ensino médio rural e outro de uma escola de ensino médio urbana da Benemérita Universidad Autónoma de Puebla (BUAP). O algoritmo de classificação Random Forest, proposto por Breiman (2001), foi utilizado, sendo treinado e avaliado independentemente em ambos os conjuntos de dados, utilizando acurácia e recall como métricas de desempenho. Uma acurácia de 0,72 foi obtida em ambos os conjuntos de dados, indicando desempenho comparável do modelo nos contextos rural e urbano. O modelo permitiu a identificação de elementos-chave relacionados ao desempenho acadêmico, como frequência às aulas e aproveitamento precoce em algumas disciplinas, que apresentaram correlação descritiva com o nível de escolaridade. Da mesma forma, os resultados revelaram algumas diferenças entre os dois contextos, sugerindo a necessidade de implementar propostas de melhoria adaptadas a cada contexto, mas altamente específicas para cada ambiente. A partir desses resultados, surgem implicações diretas para organizações educacionais e o setor público, particularmente no desenvolvimento de estratégias voltadas para a prevenção do baixo rendimento acadêmico e, consequentemente, da evasão escolar em ambos os contextos.

Palavras-chave: florestas aleatórias, contexto geográfico, evasão escolar, mineração de dados, modelos de classificação.

Fecha Recepción: Agosto 2025 **Fecha Aceptación:** Febrero 2026

Introducción

En las dependencias educativas del estado de Puebla, constantemente se aborda el tema del aprovechamiento académico por debajo del promedio estatal o de los estándares mínimos establecidos por la SEP y, como una posible consecuencia, se presenta la deserción escolar. Particularmente, este estudio se aplica en el nivel medio superior; sin embargo, esta situación sigue marcando una problemática relevante tanto para las instituciones educativas de nivel medio superior como para las autoridades educativas estatales.

En la práctica es común que, a pesar de que las instituciones educativas cuenten con conjuntos de datos escolares, no se tienen las herramientas tecnológicas suficientes que permitan procesar y analizar dicha información, con el objetivo de desarrollar estrategias educativas que contribuyan a la prevención y mitigación de esta problemática. Por ello, se analizaron dos conjuntos de datos diferenciados por zona geográfica, uno con información de estudiantes urbanos y otro con estudiantes rurales, ambos de nivel medio superior en Puebla (ver sección Metodología para detalles sobre el tamaño de muestra y recolección de datos).

A pesar de que existen algunos trabajos de investigación de nivel internacional que abordan esta problemática a través del uso de estrategias de minería de datos, como los estudios de (Cortez & Silva, 2008; Romero & Ventura, 2020) quienes aplican diferentes modelos de minería de datos a conjuntos de datos académicos de estudiantes de *secondary school* (nivel equivalente al medio superior en el contexto mexicano). El objetivo principal de estos trabajos es identificar factores clave que impacten en el aprovechamiento académico y anticipar, con cierto grado de exactitud, si los estudiantes alcanzarán el éxito o enfrentarán dificultades durante su desarrollo escolar. A partir de estos antecedentes, el presente estudio retoma el enfoque de minería de datos para analizar su aplicabilidad en el contexto educativo del nivel medio superior en Puebla.

En cuanto a las medidas de rendimiento estos estudios no utilizan métricas como la precisión ni la sensibilidad; en su lugar, utilizan la exactitud global (PCC) para valorar las tareas de clasificación y el error cuadrático medio (RMSE) para las tareas de regresión. Según los autores, los mejores resultados de predicción son alcanzados cuando se incluyen las variables (G1 y G2) que representan las calificaciones de los dos primeros periodos escolares (Cortez & Silva, 2008). Por ello, concluyen que los modelos predictivos pueden utilizarse como herramientas valiosas para anticipar el aprovechamiento académico con un alto nivel de exactitud (Cortez & Silva, 2008).

Aunque los estudios internacionales han demostrado la eficacia de los modelos predictivos en educación, su aplicación ha sido en contextos donde las condiciones socioeconómicas y educativas son más igualitarias para los estudiantes. En contraste, el estado de Puebla presenta brechas significativas entre zonas rurales y urbanas que influyen considerablemente en la continuidad y el aprovechamiento académico, y como consecuencia, aumenta el riesgo de abandono escolar. Algunos estudios recientes han evidenciado que los estudiantes de contextos rurales enfrentan en mayor medida desventajas como infraestructura educativa, acceso a recursos tecnológicos y condiciones socioeconómicas (INEGI, 2022; CONEVAL, 2023). Por ello, es fundamental extender estos modelos al contexto local.

En México, y en particular en el contexto poblano, existen pocas investigaciones que permitan comparar resultados y evaluar la eficiencia de estas herramientas y modelos en contextos locales, lo cual limita a las instituciones educativas para tomar decisiones que se sustenten en evidencia.

En el estado de Puebla, el abandono escolar en el nivel medio superior continúa siendo un desafío crucial y requiere de atención inmediata, ya que Puebla muestra una brecha urbano-rural más evidente del país en términos de abandono escolar. De acuerdo con la Secretaría de Educación Pública (SEP, 2023), las comunidades rurales son las que muestran mayores tasas de inasistencia y niveles de rezago. A nivel nacional, la deserción escolar en media superior llegó al 13.5% en zonas rurales, frente al 8.2% en zonas urbanas (CONEVAL, 2023), lo que pone en evidencia que las condiciones de origen impactan significativamente en el rendimiento escolar.

Por otro lado, en Puebla la tasa de abandono escolar alcanzó 11.3 % durante el ciclo escolar 2021–2022. Para los ciclos 2022-2023, 2023-2024 descendió a 9.7% y 9.5%, respectivamente. Aun así, esta sigue siendo la más alta entre los diferentes niveles educativos del estado SEP (2023).

Del mismo modo, diferentes estudios han evidenciado que factores como la distancia a la escuela y la disponibilidad del transporte público marcan una diferencia en el contexto rural. Según la Secretaría de Educación Pública (Dirección General de Planeación, Programación y Estadística Educativa, 2024), los estudiantes tienen que recorrer de 40 a 60 minutos diarios para llegar a la escuela, lo cual impacta la asistencia y como consecuencia, el aprovechamiento escolar. Por el contrario, en las zonas urbanas las variables de mayor impacto son las faltas escolares, la elección por cercanía y algunas actividades extracurriculares, lo que refleja tendencias que el modelo de este estudio busca analizar.

Aunque se observa un leve progreso, el tema sigue siendo un reto para las instituciones educativas. La evidencia empírica permite observar que la diferencia entre los estudiantes de contextos rurales y urbanos no solo se debe a factores individuales, sino también a diferencias estructurales propias de cada contexto. Por ello, se evidencia la necesidad de desarrollar herramientas de análisis comparativo y predictivo del aprovechamiento académico, con el fin de detectar oportunamente factores que contribuyen a la deserción escolar y así planificar estrategias de prevención diferenciadas según el contexto. Esto no solo permite una aplicación práctica en la revisión y mejora de las políticas educativas, sino que también contribuye a entender como las condiciones estructurales y contextuales pueden afectar el aprovechamiento académico en el estado de Puebla.

En este sentido surge la siguiente pregunta de investigación: ¿Qué tan eficaces son los modelos predictivos basados en minería de datos para prevenir el bajo aprovechamiento académico de estudiantes de nivel medio superior en Puebla? El objetivo del presente estudio es diseñar y valorar el uso de técnicas de minería de datos con el objetivo de comprender y anticipar mejor el aprovechamiento académico de estudiantes de nivel medio superior en Puebla. Es por ello que se analizaron dos conjuntos de datos diferenciados por la zona geográfica, uno de tipo rural y otro de tipo urbano, los cuales presentan diferencias educativas asociadas a factores estructurales del territorio Pérez Pérez et al. (2025) y también se utilizaron métricas estadísticas para evaluar el rendimiento del modelo.

El presente artículo se organiza de la siguiente manera: primero, se expone el marco teórico con base en las teorías educativas, psicológicas, sociológicas y computacionales, junto con la revisión del estado del arte relacionado con los modelos de predicción; a continuación, se describe la metodología empleada en el desarrollo del modelo predictivo; luego, se presentan y analizan los resultados obtenidos; y finalmente, se exponen las conclusiones y recomendaciones dirigidas a entidades del sector educativo y público.

Marco teórico

Para tener una comprensión más clara respecto al proceso del aprovechamiento académico y los factores asociados a este, basta con repasar el soporte conceptual proporcionado por algunas de las teorías educativas, psicológicas, sociales y computacionales, así como su integración en modelos de predicción.

En la minería de datos se pueden desarrollar modelos predictivos de aprovechamiento y de riesgo académico, que faciliten la detección temprana de factores que pueden influir en el aprovechamiento académico. Gracias a estos algoritmos se pueden identificar factores de tipo sociodemográficos, psicológicos, económicos, entre otros, con el propósito de implementar estrategias educativas de apoyo a los estudiantes en riesgo.

Es por ello que, para diseñar modelos predictivos efectivos, es importante revisar algunas de las teorías que expliquen las diversas dimensiones que influyen en el aprendizaje, tales como la teoría del aprendizaje significativo, la teoría de la motivación y la teoría del capital cultural (Bourdieu & Passeron, 1990), con el propósito de comprender cómo estas interactúan con el aprovechamiento académico, sentando las bases que permitirán identificar los factores clave en los algoritmos predictivos.

Según Ausubel (1968), la teoría del aprendizaje significativo destaca la importancia de los conocimientos previos como base inicial para generar aprendizaje efectivo, permitiendo anticipar el aprovechamiento de materias específicas. Piaget (1972) resalta el papel del desarrollo cognitivo en la construcción del conocimiento, indicando que la madurez intelectual influye en la capacidad de aprendizaje, lo cual fundamenta la predicción del aprovechamiento académico.

Por su parte, Vygotsky (1978) subraya que el contexto social y la zona de desarrollo próximo (ZPD) influyen significativamente en el desarrollo cognitivo, proporcionando variables clave para la predicción del aprovechamiento académico. Asimismo, estos conceptos se reflejan en la gestión académica diferenciada y necesaria en zonas rurales y zonas urbanas marginales (Pérez Pérez et al., 2025).

En el contexto psicológico, Bandura (2001, pp. 7-10) destaca que quienes tienen alta autoeficacia actúan y piensan sobre la creencia de sus propias capacidades y habilidades, de tal forma que este actuar se traduce en la motivación, confianza e iniciativa de los estudiantes, permitiéndoles alcanzar un mejor aprovechamiento académico.

De igual manera, la teoría de la motivación de Ryan & Deci (2000, pp. 1-3), aplicada en diferentes ámbitos, incluida la educación, vista como un modelo amplio que explica por qué

las personas se sienten motivadas al realizar determinadas actividades y cómo este factor influye positivamente en el aprovechamiento académico.

La Teoría de la Autodeterminación identifica tres necesidades psicológicas básicas:

i. Autonomía: percepción de control personal sobre sus acciones y decisiones;
ii. Competencia: habilidad para ejecutar tareas asignadas y obtener resultados satisfactorios; y
iii. Relación: inclusión en grupos de trabajo con el fin de recibir apoyo social.

Satisfacer estas tres necesidades permite a los estudiantes desarrollar habilidades y conocimientos de manera significativa y duradera, favoreciendo el aprendizaje significativo y el éxito escolar.

Por otro lado, Bourdieu & Passeron (1990, pp. 31-33) sugiere que el aprovechamiento académico puede depender en gran medida del capital cultural[1] transmitido de generación en generación, además del aprovechamiento académico individual y los recursos disponibles. Se puede inferir que estos elementos hacen posible que los estudiantes con mayor capital cultural tengan más oportunidades de alcanzar el éxito académico, a diferencia de aquellos que cuenten con un menor capital cultural, ya que esta desigualdad actúa como una barrera que limita el desarrollo de sus habilidades escolares.

Buschini (2023, pp. 448-452) retoma la teoría de Luhmann para destacar que la estructura organizativa del sistema educativo es un factor significativo que configura el proceso educativo y condiciona los procesos de enseñanza-aprendizaje. La escuela, como sistema social y organizacional, no solo se basa en una estructura jerárquica que integra diferentes funciones, sino que también posee un grado de autonomía que determina cómo se organiza y desarrolla el aprendizaje dentro del sistema, impactando de manera significativa en este proceso.

Finalmente, desde el punto de vista computacional, todos estos elementos pueden integrarse en algoritmos con la funcionalidad de pronosticar el aprovechamiento académico con base en grandes conjuntos de datos educativos. Para ello, se adopta el modelo de proceso de minería de datos *Cross Industry Standard Process for Data Mining* (CRISP-DM; Chapman et al., 2000), reconocido como un método sistemático y estándar para el desarrollo de proyectos de minería de datos y modelos predictivos. Este modelo consiste en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado,

[1] Bourdieu & Passeron (1990) refiere que habitus incluye al capital cultural, los hábitos perdurables y que se compone de los conocimientos, habilidades, capacidades, hábitos y contenidos culturales.

evaluación y despliegue y permite organizar de forma ordenada la construcción y validación de los modelos aplicados en este estudio.

Se puede concluir que sus teorías sientan las bases para entender cómo a través de la integración de diferentes factores y tomando en cuenta el enfoque educativo y psicológico, se puede inferir que el aprendizaje relevante, la motivación o la autoeficacia impactan en el aprovechamiento académico. Por último, se puede entender que las teorías computacionales generan las condiciones para integrar estos elementos en los modelos predictivos.

Estos enfoques teóricos han sido ampliamente abordados en la literatura. Diversos estudios (Peña-Ayala, 2014; Romero & Ventura, 2020; Taylor *et al*., 2016) señalan que las herramientas para el análisis de datos educativos han demostrado su efectividad y son útiles para descubrir tendencias significativas que impactan en el aprovechamiento académico. Además, facilitan la toma de decisiones basadas en evidencia con el objetivo de mejorar el proceso de enseñanza-aprendizaje a través de la planificación de estrategias adecuadas para reducir las brechas educativas.

De manera similar, los estudios de Contreras-Bravo *et al*. (2021), Gil-Vera & Quintero (2021), Castrillón *et al*. (2020) y Díaz-Martínez *et al*. (2021) señalan que la minería de datos educativa emplea diversos algoritmos, tales como los bosques aleatorios, redes neuronales artificiales (RNA) y regresión logística. Estos algoritmos se utilizan para analizar grandes volúmenes de datos académicos, detectar patrones relevantes y predecir el riesgo de bajo aprovechamiento académico, con el fin de identificar estudiantes en situación de riesgo, diseñar intervenciones adecuadas y optimizar los procesos de enseñanza-aprendizaje.

Además, la incorporación de estrategias de minería de datos educativa (*Educational Data Mining*, EDM) ha demostrado ser un recurso útil para identificar elementos clave en el aprovechamiento académico, con el objetivo de prevenir riesgos de abandono escolar y diseñar acciones de mejora. Estudios recientes han revelado cómo el análisis de datos facilita el desarrollo de modelos de predicción que ayudan a la toma de decisiones bien informadas en ámbitos educativos, contribuyendo a mejorar la eficacia organizacional (Romero & Ventura, 2020; Merceron & Tato, 2023; Ordoñez-Avila *et al*., 2023).

Principales algoritmos usados en modelos de predicción educativa

Los modelos de predicción han evolucionado hasta convertirse en herramientas útiles para identificar elementos que impactan en el aprovechamiento académico de los estudiantes, apoyando la toma de decisiones basadas en evidencia por parte de docentes y responsables de la gestión educativa. Los algoritmos más utilizados incluyen: bosques aleatorios, RNA, regresión logística y algoritmos de agrupamiento (*clustering*).

Según Aguilar-Reyes (2025), el análisis del aprovechamiento académico a través de modelos de aprendizaje automático se abordó a partir de dos enfoques predictivos: la regresión logística multinomial y los árboles de clasificación, los cuales permitieron identificar elementos significativos asociados al rendimiento estudiantil. En sus resultados, el autor reporta que el modelo de regresión logística alcanzó una precisión del 100% en la clasificación, mientras que el modelo basado en árboles de clasificación obtuvo una precisión del 70.83%, bajo las condiciones experimentales y el esquema de validación descritos en su estudio.

En otros estudios de investigación, el análisis se ha realizado con estudiantes de nivel superior mediante la aplicación de un modelo predictivo basado en RNA. A partir de sus resultados, el autor concluye que variables como la relación familiar positiva, el número de faltas y las calificaciones previas influyen de forma significativa en el aprovechamiento académico, y reporta que el modelo alcanzó una efectividad del 87%[2] en la predicción (Choque-Aguilar, 2024).

De igual forma, Gil-Vera & Quintero (2021) presentan el desarrollo de un modelo predictivo que se basa en RNA con el objetivo de identificar el riesgo de bajo aprovechamiento académico en estudiantes de educación media superior. En su estudio se consideraron variables como tiempo dedicado al estudio, las faltas escolares y el uso de redes sociales, entre otras variables relevantes. El modelo alcanzó una precisión del 73%[3] en la clasificación de los estudiantes y las variables más significativas en la predicción fueron el tiempo de estudio, ausencias a clases y el tiempo dedicado al uso de redes sociales.

Específicamente en el caso de México, Rico-Páez (2022) realizó una investigación en la que utilizó tres algoritmos: Naïve Bayes, k-NN (k vecinos más cercanos) y árbol de decisión C4.5, para predecir el aprovechamiento académico de los estudiantes de una universidad

[2] Choque-Aguilar (2024). Destacan variables clave como la relación positiva familiar, número de faltas y calificaciones previas.
[3] Gil-Vera & Quintero (2021). Destacan como variables más significativas el tiempo dedicado al estudio, las ausencias a clases y el uso de redes sociales.

pública. Los resultados muestran que, a partir de las tres primeras actividades evaluables del curso, es posible alcanzar una precisión del 70% en la predicción de si el estudiante aprobará la asignatura, lo que evidencia la factibilidad de realizar una evaluación temprana y de orientar oportunamente a los estudiantes.

Por otro lado, Acosta-Gonzaga y Ramírez-Arellano (2020), investigadores del Instituto Politécnico Nacional, realizaron un estudio de predicción del aprovechamiento académico y mostraron que tanto las técnicas de minería de datos, como las máquinas de vectores de soporte (*Suppport Vector Machine*, SVM) como los métodos de estadística tradicional, en particular regresión lineal resultan efectivas. Sin embargo, las SVM destacan por su mayor desempeño predictivo, mientras que la regresión lineal identificar y analizar los factores más significativos asociados al aprovechamiento académico.

Metodología

En esta sección se describió la estrategia metodológica del estudio, incluyendo el diseño de la investigación, la forma en que se abordó el problema planteado, el tipo de datos requeridos, los instrumentos de recolección y los métodos de análisis que se emplearán para responder la pregunta de investigación.

Enfoque metodológico

El enfoque metodológico que se utilizó es el cuantitativo, ya que permitió analizar conjuntos de datos escolares a través del uso de herramientas estadísticas y computacionales propias de la minería de datos, tales como los algoritmos de clasificación y las métricas de avaluación del aprovechamiento predictivo. Este enfoque fue fundamental para anticipar el aprovechamiento académico de los estudiantes a partir de la identificación de patrones en los datos. Para sustentar la elección de un diseño no experimental y el enfoque cuantitativo, se retomaron las aportaciones de la metodología general propuestas por Hernández *et al*. (2014).

Además, la evaluación de estos modelos predictivos se realizó mediante el método *hold-out*, particionando el conjunto de datos en dos subconjuntos: un 80% asignado al entrenamiento del modelo de bosques aleatorios y un 20% reservado para su evaluación. Las métricas estadísticas, como la precisión y la sensibilidad, así como las reportadas en las tablas 1 y 2, fueron calculadas sobre el conjunto de prueba, con el fin de estimar la capacidad predictiva del modelo en datos no vistos. Esto explica que el soporte mostrado en dichas

tablas sea menor que el tamaño total de la muestra y permite evitar sesgos por sobreajuste, garantizando una evaluación neutral de su desempeño.

Asimismo, se llevó a cabo un análisis comparativo de los resultados obtenidos, con el propósito de plantear estrategias que contribuyan a la mejora de los procesos de toma de decisiones en organizaciones educativas. Todo ello se realizó mediante un enfoque cuantitativo, basado en muestras de tamaño estadísticamente suficiente para garantizar la confiabilidad de los resultados. Finalmente, este enfoque permitió asegurar la consistencia metodológica al comparar el desempeño de los modelos en ambos contextos de análisis.

Este estudio se abordó mediante un diseño de investigación no experimental, de tipo descriptivo-correlacional. Asimismo, el estudio es de tipo transversal, ya que el análisis se realizó a partir de datos recolectados en un único momento temporal. Este enfoque es adecuado debido a la ausencia de manipulación de variables y el carácter observacional del análisis, ya que se trabajó con conjuntos de datos existentes de estudiantes a nivel medio superior. En particular, se analizaron los datos correspondientes a estudiantes que viven en Tehuacán y ciudad de Puebla en el estado de Puebla, con el objetivo de identificar patrones, relaciones y elementos fundamentales asociados al aprovechamiento académico.

Este enfoque descriptivo caracteriza al grupo de estudio a partir de las variables sociodemográficas y escolares, mientras que el enfoque correlacional la relación entre dichas variables y el aprovechamiento académico.

Características de la muestra

El tamaño de la muestra es de 198 participantes para el contexto urbano y de 384 para el contexto rural, lo que permite contar con un número de observaciones suficiente para sustentar la fiabilidad de los resultados, considerando un nivel de confianza del 95% y un margen de error aceptable en estudios educativos. Estos tamaños muestrales favorecen la representatividad de los datos y la confiabilidad de los resultados, al proporcionar una base apropiada para el análisis cuantitativo y la validación de los modelos predictivos.

Los criterios de inclusión considerados fueron:

a) estudiantes inscritos en el nivel medio superior de la BUAP;
b) contar con información completa en las variables académicas y sociodemográficas utilizadas en el estudio;
c) pertenecer a alguno de los contextos definidos (urbano o rural).

Se excluyeron los registros que presentaban datos incompletos, inconsistencias en la información académica.

Se empleó un muestreo no probabilístico por conveniencia, considerando los conjuntos de datos disponibles en os contextos rural y urbano.

El instrumento de recolección de datos fue un cuestionario estructurado, como se muestra en el Apéndice A, el cual fue diseñado para recopilar información clave para el estudio. En particular, permitió obtener datos sobre factores: a) académicos; b) personales y demográficos; c) familiares y sociales; y d) contextuales, con el propósito de contar con una visión integral de los elementos que influyen en el aprovechamiento académico de los estudiantes de nivel medio superior en Puebla.

Asimismo, el cuestionario fue diseñado con base en instrumentos utilizados en estudios previos y actualmente se encuentra en proceso de evaluación de su confiabilidad mediante el coeficiente alfa de Cronbach, con el fin de garantizar la consistencia interna de los datos considerados.

Variables de estudio y operacionalización

La variable del estudio fue el aprovechamiento académico, categorizado en dos niveles: “Alto” y “Medio”, con base en el promedio de calificaciones de los estudiantes y siguiendo los criterios institucionales de evaluación. Se consideró como “Alto” el rango [8.5, 10.0] y como “medio” el rango [7.0, 8.49]. Esta categorización permitió abordar el problema como una tarea de clasificación supervisada para el entrenamiento del modelo predictivo.

Resultados

Análisis del modelo bosques aleatorios

Para el conjunto de datos rural se utilizó el criterio de impureza de Gini (*Gini impurity*) y para el conjunto de datos urbano el criterio de entropía (*entropy)* en la construcción de los árboles de decisión del modelo de bosques aleatorios. La profundidad máxima de los árboles se fijó en 10 niveles como estrategia de regularización. Se entrenaron 200 árboles para mejorar la estabilidad de la predicción. Además, se configuraron los parámetros con un tamaño mínimo de hoja de una muestra y un mínimo de dos muestras para la división de nodos, lo que permitió un particionamiento de los datos.

Las tablas 1 y 2 muestran los resultados obtenidos en cuanto al desempeño general del modelo para los enfoques rural y urbano, respectivamente, lo que permite observar lo siguiente:

- La precisión del modelo fue ligeramente mayor en el enfoque urbano que en el rural, lo que sugiere una mejor capacidad de clasificación global en este contexto.
- La precisión macro obtenida fue de 0.72 para el conjunto de datos rural y de 0.73 para el conjunto urbano, lo que refleja un desempeño similar del modelo en los dos entornos en términos de clasificación correcta para las dos clases.
- En ambos contextos, la sensibilidad macro alcanzó un valor de 0.71, lo que indica una capacidad equivalente del modelo para identificar correctamente los casos reales de cada clase tanto en el ámbito rural como en el urbano.
- El *F1-score* macro fue de 0.71 en ambos conjuntos de datos, lo que evidencia un equilibrio adecuado entre precisión y sensibilidad del modelo en ambos entornos analizados.

Reporte de clasificación

Las Tablas 1 y 2 muestran los resultados de las métricas de evaluación alcanzadas por el modelo en el conjunto de datos de prueba para cada una de las clases de aprovechamiento académico “Alto” y “Medio” en los contextos rural y urbano, respectivamente.

Tabla 1. Indicadores de desempeño en el contexto rural

Clase	Precisión	Sensibilidad	*F1-score*	Soporte *(Support)*
Alto	0.70	0.76	0.73	21
Medio	0.74	0.67	0.70	21
Exactitud *(Accuracy)*			0.71	42
Promedio macro *(Macro avg)*	0.72	0.71	0.71	42
Promedio ponderado *(weighted avg)*	0.72	0.71	0.71	42

Fuente: Elaboración propia a partir de análisis con Python

- Clase “Alto”:
    - En el conjunto de datos rural se registró una sensibilidad de 0.76, mientras que en el contexto urbano fue de 0.85, lo que indica una capacidad adecuada de identificación de los estudiantes con un aprovechamiento académico alto.
    - La precisión fue moderada, con un valor de 0.70 en el contexto rural y de 0.69 en el contexto urbano, lo que evidencia que una parte de las instancias clasificadas como “Alto” en realidad correspondieron a la clase “Medio”.
    - El *F1-score* fue de 0.73 en el conjunto de datos rural y de 0.76 en el urbano, lo que sugiere un buen equilibrio entre precisión y sensibilidad, por tanto, un desempeño global adecuado para esta categoría.

Tabla 2. Indicadores de desempeño en el contexto urbano

Clase	Precisión	Sensibilidad	*F1-score*	Soporte
Alto	0.69	0.85	0.76	13
Medio	0.78	0.58	0.67	12
Exactitud			0.72	25
Promedio macro	0.73	0.71	0.71	25
Promedio ponderado	0.73	0.72	0.71	25

Fuente: Elaboración propia a partir de análisis con Python

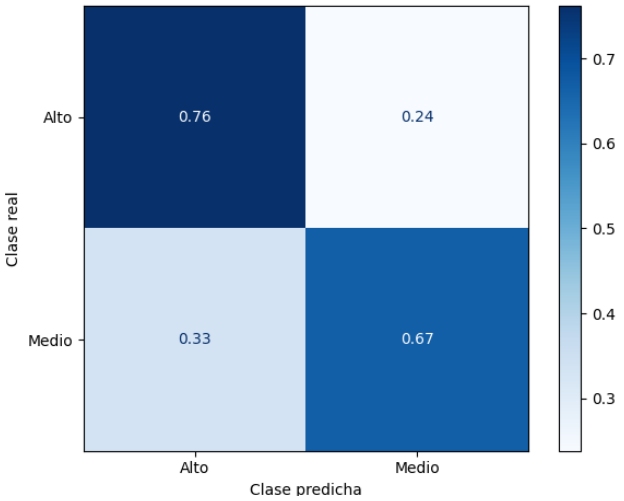
- Clase “Medio”:
    - La precisión fue de 0.74 en el conjunto de datos rural y de 0.78 en el urbano, lo que indica que la mayoría de las predicciones realizadas para esta categoría fueron correctas.
    - En cuanto a la sensibilidad, se obtuvo un valor de 0.67 para el contexto rural y de 0.58 para el urbano, lo que indica que una proporción de los estudiantes pertenecientes a esta clase fue clasificada como “Alto”.
    - El *F1-score* fue de 0.70 en el conjunto de datos rural y de 0.67 en el urbano, lo que refleja un desempeño aceptable al considerar de manera conjunta la precisión y la sensibilidad para esta categoría.

Matriz de confusión normalizada

Para los conjuntos de datos rural y urbano, las figuras 1 y 2 muestran la matriz de confusión, cuyas entradas representan la proporción de casos normalizados para cada combinación de clases reales y predichas. Se observa que, cuando la clase real es “Alto”, el modelo clasifica de forma correcta 0.76 de los casos en el contexto rural y 0.85 en el urbano, mientras que comete errores en 0.24 y 0.15 de los casos, respectivamente, al asignarlos a la clase “Medio”.

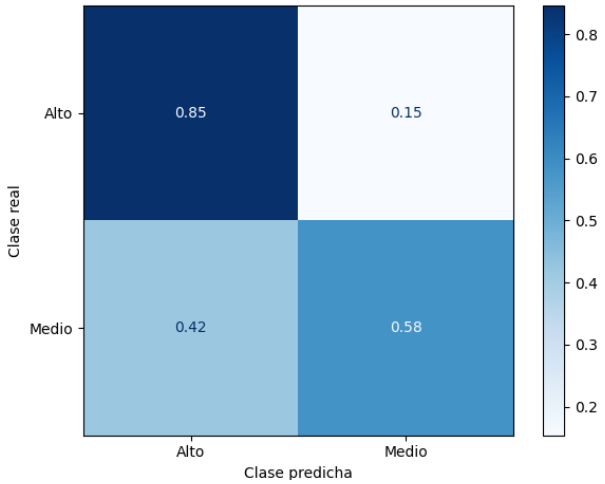
Asimismo, cuando la clase real es “Medio”, el modelo clasifica correctamente 0.67 de los casos en el contexto rural y 0.58 en el urbano, mientras que clasifica de manera incorrecta 0.33 y 0.42 de los casos, respectivamente, al asignarlos a la clase “Alto”, lo que sugiere una tendencia a sobreclasificar la categoría “Alto”.

Figura 1. Matriz de confusión normalizada por filas para el contexto rural



Fuente: Modelo predictivo implementado en Python

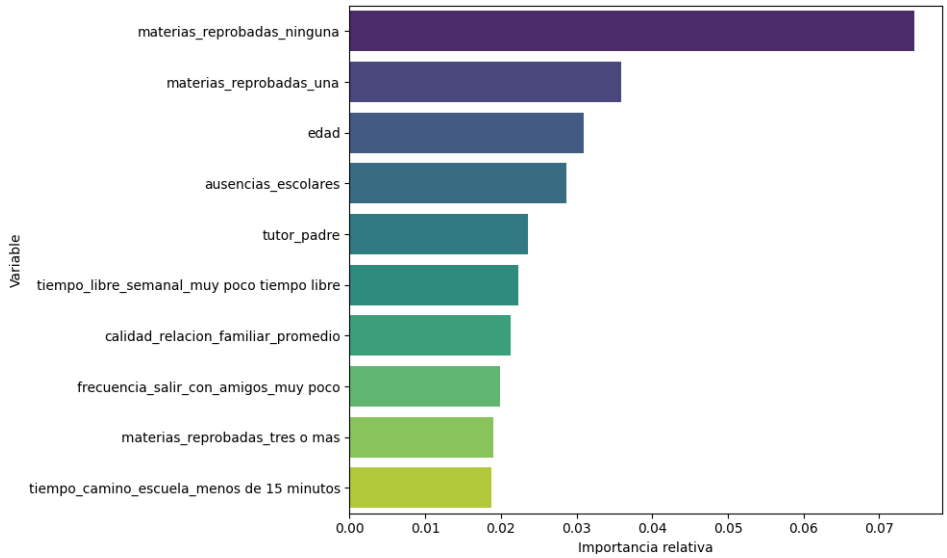
Figura 2. Matriz de confusión normalizada por filas para el contexto urbano



Fuente: Modelo predictivo implementado en Python

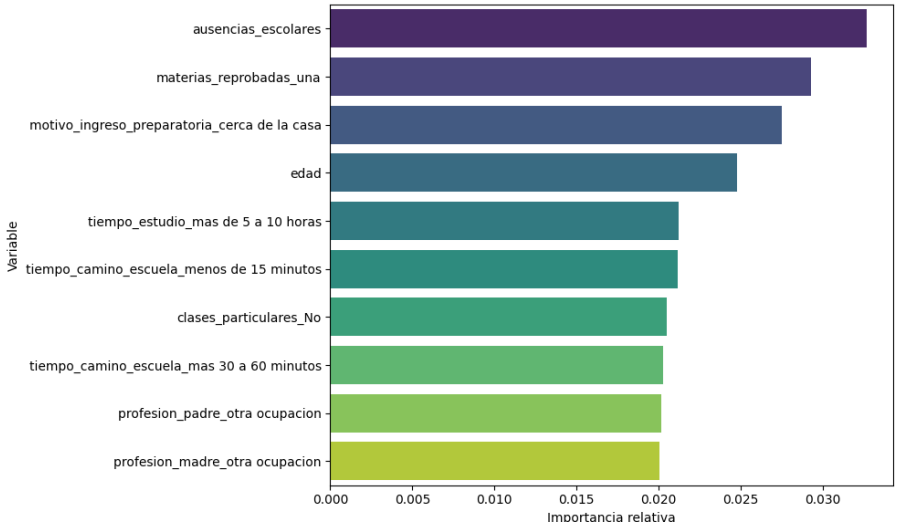
Para los conjuntos de datos rural y urbano, los gráficos de barras de las figuras 3 y 4 respectivamente, muestran las variables clave utilizadas por el modelo para predecir el aprovechamiento académico de los estudiantes. El nivel de importancia de cada variable se evalúo con base en su contribución proporcional a la disminución de la impureza en los árboles del modelo, medida mediante el criterio de Gini (*mean decrease in impurity*).

Figura 3. Diez variables más significativas para el contexto rural.



Fuente: Modelo predictivo implementado en Python

Figura 4. Diez variables más significativas para el contexto urbana.



Fuente: Modelo predictivo implementado en Python

Análisis inferencial por contexto rural y urbano

Con el fin de complementar los resultados del modelo predictivo y analizar a detalle las diferencias estructurales entre ambos contextos, se realizó un análisis inferencial comparativo entre los estudiantes de zonas rurales y urbanas. Este análisis incluyó variables categóricas y numéricas, con el propósito de determinar si las diferencias observadas entre ambos grupos son estadísticamente significativas.

Variables categóricas

Los resultados obtenidos muestran que diversas variables de los estudiantes presentan diferencias estadísticamente significativas entre los contextos rural y urbano. Entre ellas se encuentra la variable tipo de preparatoria, la cual mostró una diferencia estadísticamente significativa ($\chi^2(1) = 575.55$, $p < 0.001$).

Asimismo, se identificaron diferencias importantes en variables asociadas al entorno familiar, tales como la convivencia con los padres ($\chi^2(1) = 4.79$, $p = 0.029$), el nivel educativo de la madre ($\chi^2(3) = 22.45$, $p < 0.001$) y el nivel educativo del padre ($\chi^2(3) = 34.96$, $p < 0.001$).

También fueron identificadas diferencias significativas en las ocupaciones de los padres profesión de la madre ($\chi^2(4) = 22.86$, $p < 0.001$), profesión del padre ($\chi^2(4) = 27.97$, $p < 0.001$), así como en variables asociadas a tiempos de traslado y hábitos académicos, como el tiempo para llegar a la escuela ($\chi^2(3) = 10.09$, $p = 0.018$) y el tiempo dedicado al estudio ($\chi^2(3) = 10.36$, $p = 0.016$).

Asimismo, variables como la frecuencia con la que los estudiantes salen con amigos ($\chi^2(4) = 14.19$, $p = 0.007$) y el consumo de alcohol entre semana ($\chi^2(4) = 23.70$, $p < 0.001$) y durante el fin de semana ($\chi^2(5) = 19.37$, $p = 0.002$) mostraron diferencias evidentes entre contextos.

En contraste, variables como sexo, apoyo escolar, apoyo familiar, clases particulares, intención de estudiar una licenciatura, salud y actividades extraescolares no mostraron diferencias de importancia entre ambos grupos.

Variables numéricas

Para el caso de las variables numéricas, los resultados muestran diferencias estadísticamente significativas únicamente en la edad ($U = 17{,}884.00$, $p < 0.001$, $r = 0.296$), con valores ligeramente mayores en los estudiantes del contexto rural. El tamaño del efecto corresponde al coeficiente de Rosenthal, lo que indica una magnitud de efecto moderada.

Por otra parte, ni el número de ausencias escolares ($U = 16{,}086.50$, $p = 0.923$) ni la calificación del primer semestre ($U = 21{,}696.00$, $p = 0.262$) mostraron diferencias estadísticamente significativas entre ambos contextos.

Interpretación general

En conjunto, el análisis inferencial evidencia diferencias estadísticamente significativas entre los contextos rural y urbano en variables familiares, educativas y conductuales. Sin embargo, las variables directamente asociadas al aprovechamiento académico inicial, como las ausencias y la calificación semestral, no mostraron diferencias estadísticamente significativas

Discusión

A continuación, se presenta un análisis comparativo con estudios previos que abordan la predicción del aprovechamiento académico. Asimismo, se destaca la contribución específica de este estudio en el contexto educativo del estado de Puebla.

Los resultados obtenidos a través del modelo de bosques aleatorios muestran una precisión aproximada de 0.72 a 0.73 tanto en el contexto rural como en el urbano, lo cual es consistente con los resultados reportados en investigaciones previas (Cortez & Silva, 2008; Gil-Vera & Quintero, 2021), estos hallazgos sugieren la viabilidad del uso de este tipo de modelos como herramientas de apoyo en la predicción del aprovechamiento académico en contextos educativos.

No obstante, este estudio aporta un análisis que permite comparar a estudiantes de contextos rurales y urbanos en Puebla e identificar variables que presentan una mayor influencia en cada contexto, En el contexto rural, la asistencia, el apoyo familiar y la cantidad de materias reprobadas muestran una relevancia más significativa, mientras que en el urbano las faltas escolares, la cantidad de materias reprobadas y el motivo de ingreso a la preparatoria aparecen como factores con mayor peso, de acuerdo con los resultados presentados en las figuras 3 y 4.

Además, el desempeño del modelo no se evaluó únicamente mediante la exactitud, sino que incorpora métricas complementarias como la sensibilidad y *el F1-score*, con el fin de analizar el equilibrio entre la capacidad de identificación de los estudiantes en riesgo y la precisión de las calificaciones realizadas. En particular, el valor obtenido del *F1-score* macro (≈ 0.71 para ambos contextos) indica un balance adecuado entre sensibilidad y precisión, lo que favorece la identificación oportuna de estudiantes con posible riesgo académico.

Aunque se identifican dificultades en la clasificación entre niveles “Medio” y “Alto”, principalmente en el contexto urbano, donde se observa una menor sensibilidad para la clase “Medio”, estos resultados permiten reconocer áreas de mejora en el modelo. En particular, la confusión entre ambas categorías sugiere la necesidad de refinar los criterios de separación entre niveles de rendimiento y proporciona una base para el diseño de intervenciones individualizadas orientadas a reducir el abandono escolar y la desigualdad educativa.

En el contexto de las preparatorias rurales, los resultados muestran que los estudiantes sin materias reprobadas presentan un mejor aprovechamiento académico. Asimismo, las inasistencias a clases parecen tener un efecto negativo en su aprovechamiento académico. Otro factor relevante es el tiempo disponible para estudiar y el trayecto de la casa a la escuela; cuando los estudiantes disponen de poco tiempo o deben recorrer distancias largas, su aprovechamiento académico tiende a ser menor. Estas condiciones podrían influir en su nivel de energía, puntualidad y capacidad para realizar tareas escolares.

Por otra parte, se observa que el aspecto familiar tiene una gran influencia en los estudiantes, especialmente en relación con el tutor principal. Cuando el padre ejerce como tutor, se observa una mayor estabilidad y apoyo familiar en la educación del estudiante. Dada la importancia de la relación familiar, esta convivencia podría relacionarse con un mejor estado emocional de los estudiantes, lo que favorecería su aprovechamiento académico. Además, otro factor relevante es la convivencia con amigos; los resultados sugieren que una

menor interacción social o carencia de habilidades interpersonales se asocia con un bajo aprovechamiento académico.

Por otro lado, en el contexto urbano, se observan nuevamente las necesidades de los estudiantes con respecto a su tiempo, ya que muchos reportan una ubicación cercana a la preparatoria y un tiempo de traslado corto. Esto indica que el tiempo es un recurso muy valioso para los estudiantes y probablemente, para sus familias. Cuando el tiempo se invierte en estudiar, se observa una influencia positiva en el aprovechamiento académico. Asimismo, la edad de los estudiantes también se relaciona con su aprovechamiento; se plantea como hipótesis que aquellos con edades fuera del promedio podrían experimentar dificultades de integración al grupo; lo que podría afectar su aprovechamiento académico.

Además, se observa nuevamente que reprobar materias y las ausencias escolares influyen en el aprovechamiento académico. Sin embargo, a diferencia del contexto rural, la convivencia social no parece ser un factor relevante, mientras que la ocupación de la madre y del padre juega un rol significativo. Por último, las clases particulares se presentan como un aspecto importante; este fenómeno podría explicarse por el hecho de que los salones urbanos suelen tener un mayor número de estudiantes, lo que podría limitar la atención individualizada del profesor. En contraste, en las preparatorias rurales, la inversión de tutores privados tiende a ser menos frecuente.

Como se muestra en las figuras 3 y 4, estas variables más relevantes por contexto fueron significativamente asociadas con el aprovechamiento académico $(X^2, p < 0.005)$, destacando su relevancia en zonas rurales como urbanas.

Análisis del contexto local de Puebla

En los resultados obtenidos en este estudio se pueden identificar algunos puntos clave en el ámbito socioeducativo de Puebla, los cuales difieren de los encontrados en estudios internacionales previos. En particular, se detectaron variables como la asistencia a clase, el apoyo familiar, la cantidad de materias reprobadas y la edad de los estudiantes, las cuales podrían estar influenciadas por aspectos estructurales y culturales propios de esta región.

En el contexto rural, con base en la gran importancia de la asistencia y del apoyo familiar, se infiere que este comportamiento depende de varios aspectos de la zona. Muchas comunidades pequeñas de Puebla enfrentan carencias relacionadas con la infraestructura y el acceso a recursos educativos. Además, las dificultades en el transporte influyen de manera significativa en la puntualidad y regularidad de la asistencia a clases.

Por otro lado, con base en los resultados arrojados por el modelo, los cuales muestran que el acompañamiento del padre y un entorno familiar de apoyo juegan un papel importante para los estudiantes del contexto rural, se sugiere que estas condiciones pueden ayudar a mitigar las desigualdades estructurales. En consecuencia, se requieren estrategias que integren programas de apoyo social y apoyo comunitario específico para zonas rurales.

En cambio, en el contexto urbano, se puede observar que las variables que juegan un papel importante son las siguientes: ausencias escolares; elección de la preparatoria por su cercanía; horas dedicadas al estudio; y edad, reflejando una tendencia más vinculada a hábitos y decisiones personales y estructurales.

Se sugiere que la edad podía relacionarse con situaciones de repetición de cursos o rezago escolar, debido a condiciones económicas y sociales presentes en contextos urbanos de Puebla, las cuales propician que los estudiantes interrumpan su desarrollo académico.

De igual forma, la ausencia de clases particulares y un tiempo de traslado largo podrían estar relacionados con variables urbanas, como la distribución geográfica y la limitada disponibilidad de materiales de apoyo adicionales. Estas condiciones requieren atención especializada, tanto para favorecer la igualdad educativa como, de manera potencial, para mejorar el aprovechamiento escolar.

Los resultados obtenidos muestran que, a diferencia de lo que suelen considerar algunos estudios internacionales, en Puebla es necesario y fundamental tener en cuenta la diversidad presente en las diferentes regiones rurales y urbanas, con el objetivo de evaluar el aprovechamiento académico de manera más precisa.

La influencia de las variables sociofamiliares en entornos rurales, en comparación con las personales en el contexto urbano, sugiere que los planes educativos y las estrategias de intervención deben enfocarse a cada contexto, con el propósito de atender la situación socioeconómica y cultural específica del estado de Puebla. Esto se basa en resultados donde se muestran que, en zonas rurales, factores como la asistencia a clases y el apoyo familiar son determinantes del aprovechamiento académico, mientras que en contextos urbanos predominan variables relacionadas con los hábitos, tiempo de estudio y edad de los estudiantes.

En conjunto, estos hallazgos sugieren que este análisis a nivel local no solo funciona como una herramienta complementaria, sino que también evidencia la validez e importancia de los modelos predictivos para tomar decisiones educativas bien fundamentadas y respaldadas.

Recomendaciones

Con base en los resultados presentados, se recomienda implementar acciones que promuevan la permanencia escolar e impulsen el apoyo docente en contextos rurales y urbanos mediante capacitación especializada y recursos adecuados (Hernández et al., 2014; Cedillo-Arce et al., 2024). Asimismo, se sugiere entrenar al personal educativo en el uso de modelos predictivos que le permitan identificar no solo a estudiantes en situación de riesgo, sino también a estudiantes con alto rendimiento, con el objetivo de apoyarlos de manera oportuna. Para ello, el monitoreo constante de las variables más significativas de cada contexto resulta vital para la detección y la implementación de estrategias educativas efectivas.

Específicamente, para las preparatorias rurales se recomienda implementar acciones comunitarias que contribuyan a acortar los tiempos de traslado a las escuelas, por ejemplo, mediante rutas escolares o transporte compartido, con el fin de disminuir el esfuerzo de traslado y mejorar la asistencia escolar. Asimismo, se sugiere promover la participación familiar a través de talleres, programas de orientación para padres y tutores, con el objetivo de fortalecer el apoyo familiar y favorecer el aprovechamiento académico de los estudiantes.

Para la educación media superior en lugares urbanos, se recomienda implementar asesorías individualizadas para estudiantes en riesgo, especialmente aquellos que no cuenten con recursos para tutorías privadas, con el propósito de realizar un seguimiento del progreso académico y favorecer su aprovechamiento. De manera complementaria, se sugiere promover hábitos de estudio autónomos mediante el uso de plataformas virtuales y aplicaciones educativas, lo que permitirá a los estudiantes reforzar sus conocimientos en horarios disponibles, incrementando así la efectividad del aprendizaje y la retención de contenidos temáticos.

Límites específicos del estudio en el contexto de Puebla

Pese a que este estudio aporta hallazgos significativos sobre la predicción del aprovechamiento académico en estudiantes de nivel medio superior en Puebla, es importante señalar algunas limitaciones específicas que pueden afectar la interpretación de los resultados y restringir su exploración a la totalidad de la población estudiantil del estado.

En primer lugar, debido a la limitada disponibilidad de recursos administrativos, el estudio se limitó a dos zonas geográficas de la entidad: la ciudad de Puebla para el contexto urbano y la región de Tehuacán para el contexto rural, bajo un muestreo de tipo no probabilístico por

conveniencia. Esta delimitación implica que los resultados no son directamente generalizables a otras regiones del estado que presentan características socioeconómicas y culturales diferentes a las consideradas en este análisis.

De igual forma, al haberse obtenido la información mediante un cuestionario autoadministrado, no se descarta la posibilidad de que algunos estudiantes hayan contestado de manera socialmente deseable, particularmente en temas relacionados con el apoyo familiar, la convivencia y los hábitos de estudio, lo que podría generar sesgos en las respuestas. Asimismo, puede existir un sesgo de autoselección, ya que es probable que los estudiantes con mayor interés académico hayan mostrado una mayor disposición a participar en el estudio.

Por lo tanto, estas limitaciones deben considerarse al interpretar los resultados, ya que las respuestas socialmente deseables y el sesgo de autoselección podrían influir en el desempeño del modelo, posiblemente inflando métricas como la precisión.

Otro aspecto significativo es la desproporción entre los tamaños de las muestras rural y urbana, lo que podría impactar en el rendimiento del modelo y en la comparación entre ambos contextos. Asimismo, algunas variables, como el nivel de aprovechamiento académico, están determinadas por criterios institucionales que pueden variar entre escuelas, lo que afecta la compatibilidad de los resultados y limita su validez externa.

Por otro lado, cabe mencionar que ciertos factores, como recursos tecnológicos a disposición de los hogares y las escuelas; la calidad de la enseñanza; y la información detallada sobre las condiciones socioeconómicas, no fueron consideradas dentro del conjunto de datos analizado, lo cual podría ser fundamental para fortalecer el modelo de predicción del aprovechamiento académico, particularmente en zonas rurales, donde estos elementos suelen desempeñar un papel determinante en el acceso a oportunidades educativas y en la reducción de desigualdades.

Finalmente, a pesar de que el modelo bosques aleatorios alcanzó un rendimiento estable, la menor precisión observada en la clasificación de las clases “Medio” y “Alto”, particularmente por la confusión entre ambas categorías, sugiere que el modelo puede mejorar mediante ajustes metodológicos. Estos podrían incluir la incorporación de variables no consideradas en este estudio, el ajuste de hiperparámetros, el balanceo de clases, o la implementación de un enfoque de análisis longitudinal, con el propósito de incrementar la efectividad en la identificación de estudiantes en situación de riesgo académico.

En conclusión, estas delimitaciones se pueden interpretar como líneas de investigación futura que consideren todas las zonas geográficas de la entidad, que integren múltiples y variadas variables, e incorporen métodos de seguimiento longitudinal, así como enfoques cualitativos complementarios, para analizar con mayor profundidad la evolución del aprovechamiento académico.

Conclusiones

Los resultados sugieren que el modelo de bosques aleatorios alcanza un aprovechamiento académico equilibrado y consistente en ambos contextos cuando se analiza la sensibilidad por clase. En el contexto rural, el modelo identifica correctamente a los estudiantes de la clase “Alto” en un 0.76 mientras que en el contexto urbano esta tasa alcanza un 0.85. En contraste, la identificación correcta de la clase “Medio” es menor, con valores de 0.67 en el contexto rural y de 0.58 en el urbano.

Estos resultados indican que el modelo es más efectivo cuando reconoce a los estudiantes con aprovechamiento académico alto, aunque presenta mayores dificultades para diferenciar aquellos pertenecientes a la clase “Medio” en ambos contextos, lo que sugiere la necesidad de fortalecer la discriminación entre estas categorías intermedias de rendimiento.

El balance entre las métricas sugiere que el modelo es conveniente en contextos educativos donde la identificación temprana de ambas clases de aprovechamiento es prioritaria.

En el contexto rural, según la matriz de confusión normalizada, se observa una tendencia a confundir la clase “Medio” con la clase “Alto”, con una proporción de 0.33 de los casos, y en menor medida a confundir la clase “Alto” con la clase “Medio”, con un valor de 0.24.

En el contexto urbano, esta confusión es más pronunciada para la clase “Medio”, que es clasificada como “Alto” en 0.42 de los casos, mientras que la clase “Alto” es confundida como “Medio” en una proporción de 0.15.

En general, el modelo bosques aleatorios presenta un aprovechamiento académico adecuado, aunque puede ser mejorado, especialmente en el reconocimiento de la clase “Medio” en ambos contextos, donde la sensibilidad alcanza valores de 0.67 en el contexto rural y de 0.58 ene el urbano. Para el contexto rural, el modelo identifica algunas variables clave para la predicción del aprovechamiento académico, tales como el historial de reprobación, la edad, las ausencias y el entorno familiar. Asimismo, variables relacionadas

con el tiempo de estudio, la interacción social y el trayecto a la escuela también juegan un papel importante.

Esta información constituye una herramienta útil para diseñar y ejecutar programas de prevención educativos ajustados para responder a las diferencias culturales y socioeconómicas propias de los contextos rural y urbano en Puebla (Pérez Pérez et al., (2025), con el propósito de mejorar la atención a sus demandas específicas y promover el aprovechamiento académico de los estudiantes.

Para los estudiantes de zonas rurales, se sugiere el desarrollo de estrategias de apoyo enfocadas en cubrir necesidades de transporte y de acceso a herramientas tecnológicas y materiales educativos. En cambio, en las zonas urbanas adquiere mayor relevancia el acompañamiento individual de los estudiantes clasificados en la categoría de riesgo “Medio”, según la clasificación generada por el modelo predictivo.

De igual forma, se reconoce la importancia de integrar, en las dependencias educativas locales, modelos predictivos basados en minería de datos como una herramienta tecnológica que apoye los procesos de planificación y seguimiento. Asimismo, estos modelos contribuyen a la educación eficaz de los planes educativos y de las estrategias de prevención, fortaleciendo la toma de decisiones basada en evidencia.

Con estas herramientas de predicción se pretende apoyar la implementación de acciones orientadas a reducir las altas tasas de abandono escolar, particularmente en las zonas rurales, donde existen diferencias socioeconómicas marcadas. Este modelo de predicción ayuda a evidenciar las brechas educativas y a justificar la ejecución de acciones tanto generales como diferenciadas, ya que los estudiantes rurales presentan escaso o nulo acceso a recursos tecnológicos, transporte escolar, actividades fuera del horario escolar y sesiones de apoyo para reforzar las carencias en sus aprendizajes.

Los organismos educativos y de gobierno pueden beneficiarse mediante el uso de la minería de datos como apoyo para elaborar iniciativas de acompañamiento focalizadas, canalizar recursos de manera equitativa y adoptar medidas para la distribución de los recursos educativos entre regiones.

Este trabajo de investigación contribuye a la identificación temprana de factores asociados al bajo aprovechamiento académico en los primeros semestres, los cuales constituyen una etapa clave para la continuidad educativa. Asimismo, proporciona una base objetiva para orientar acciones de tutoría, acompañamiento académico y asignación de recursos mediante el uso de estrategias de minería de datos.

Futuras líneas futuras de investigación

Con los resultados obtenidos en este estudio surge la necesidad de realizar un análisis más exhaustivo en aspectos como la selección de variables, los algoritmos de clasificación y los procesos de validación del modelo, con el objetivo de optimizar los modelos de predicción y diseñar acciones más efectivas para prevenir el abandono escolar. En trabajos futuros, para profundizar en los factores de mayor impacto, será importante revisar técnicas de selección de características (*feature selection*) que permitan optimizar su identificación y medir su identificación y medir de manera más precisa su influencia en el aprovechamiento académico.

Más adelante, se propone poner en práctica modelos de datos longitudinales para monitorear periódicamente la evolución del aprovechamiento académico y evaluar de manera anticipada el riesgo de abandono escolar.

Asimismo, se plantea medir el impacto de las estrategias propuestas, dirigidas a las variables que el modelo identifica como más relevantes, con el fin de valorar su efectividad en la mejora del aprovechamiento académico.

Para ello, se sugiere el desarrollo de una plataforma digital que brinde a las dependencias educativas una herramienta para la implementación operativa de este modelo, con el objetivo de localizar oportunamente a los estudiantes en alto riesgo de deserción escolar.

Finalmente, se busca expandir el estudio a más zonas geográficas de México con el propósito de evaluar la validez externa del modelo y determinar si los comportamientos observados en los estudiantes de nivel media superior en Puebla se manifiestan de manera consistente en otros contextos, o si existen particularidades asociadas a la zona geográfica que limiten su generalización.

Referencias

Acosta-Gonzaga, Elizabeth, & Ramirez-Arellano, Aldo. (2020). Estudio comparativo de técnicas de analítica del aprendizaje para predecir el rendimiento académico de los estudiantes de educación superior. *CienciaUAT*, 15(1), 63-74. Epub 22 de diciembre de 2020. https://doi.org/10.29059/cienciauat.v15i1.1392

Aguilar-Reyes, J. E., Mejía-Peñafiel, E. F., Morocho-Barrionuevo, T. P., & Velasco Castelo, G.-M. (2025). Estudio del rendimiento académico mediante la comparación de modelos de regresión y árboles de clasificación. *Telos*: *Revista de estudios Interdisciplinarios en ciencias sociales, 27*(1), 94-115. https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1317-05702025000100094

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart and Winston. https://archive.org/details/in.ernet.dli.2015.112045/page/n3/mode/2up?utm_source=chatgpt.com

Bandura, A. (2001). *Social cognitive theory: An agentic perspective*. *Annual Review of Psychology*, *52*, 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society and culture*. SAGE Publications. https://archive.org/details/reproductionined0000bour

Buschini, J. D. (2023). Niklas Luhmann y la teoría general de los sistemas sociales. En A. A. M. Camou (Coord.), *Cuestiones de teoría social contemporánea* (pp. 443–471). La Plata: Universidad Nacional de La Plata; EDULP. https://www.memoria.fahce.unlp.edu.ar/libros/pm.5846/pm.5846.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación Universitaria, 13*(1), 93–102. https://doi.org/10.4067/S0718-50062020000100093

Cedillo-Arce, J. M., Beltrán-Abreo, H. M., Saltos-Arce, M. I., & Soriano-Barzola, F. R. (2024). Explorando la minería de datos en la gestión educativa superior: desafíos y

oportunidades en la era digital. *Reincisol, 3*(5), 1368–1385. https://doi.org/10.59282/reincisol.V3(5)1367-1385

Choque-Aguilar, M. R. (2024). Red neuronal para predecir el rendimiento académico. *Revista Simón Rodríguez, 4*(8), 22–35. https://doi.org/10.62319/simonrodriguez.v.4i8.31

CONEVAL. (2023). Informe de pobreza multidimensional 2022: Resultados nacionales y por entidad federativa. Consejo Nacional de Evaluación de la Política de Desarrollo Social. https://www.coneval.org.mx

Contreras-Bravo, L. E., Fuentes-López, H. J., & Rivas-Trujillo, E. (2021). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. *Boletín Redipe*, *10*(13) 171–190. https://doi.org/10.36260/rbr.v10i13.1737

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. En A. Brito & J.Teixeira (Eds.), *Proceedings of the 5th Annual Future Business Technology Conference* (pp. 5-12). EUROSIS. https://doi.org/10.24432/C5TG7T

Díaz-Martínez, M. A., Ahumada-Cervantes, M. de los Ángeles, & Melo-Morín, J. P. (2021). Árboles de decisión como metodología para determinar el rendimiento académico en educación superior. *Revista Lasallista de Investigación, 18*(2), 94–104. https://revistas.unilasallista.edu.co/index.php/rldi/article/view/2724

Gil-Vera, V. D., & Quintero-López, C. (2021). Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *Información Tecnológica, 32*(6), 221–228. https://doi.org/10.4067/s0718-07642021000600221

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (6.ª ed.). McGraw-Hill. https://uniclanet.unicla.edu.mx/assets/contenidos/254857_DOC_2023-03-01_18:46:18.pdf

INEGI. (2022). Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH 2022). Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/programas/dutih/

Merceron, A., & Tato, A. (2023). Introduction to neural networks and uses in educational data mining. En M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th*

International Conference on Educational Data Mining (pp. 578–581). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115774

Ordoñez-Avila, R., Salgado Reyes, N., Meza, J., & Ventura, S. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *Heliyon, 9*(3), e13939. https://doi.org/10.1016/j.heliyon.2023.e13939

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Pérez Pérez, A. M., Custodio Valenzuela, M., Cerón Garnica, C., Mila Avendaño, V. M., & Moyao Martínez, Y. (2025). La gestión académica en centros educativos urbanos marginales y zonas rurales orientada a la preparación del docente para la alfabetización inicial. *Edutec*. *Revista Electrónica de Tecnología Educativa, 33*(1), 1–12. https://doi.org/10.58299/edutec.v33i1.335

Piaget, J. (1972). *Psychology and epistemology: Towards a theory of knowledge.* Penguin. https://books.google.com.mx/books/about/Psychology_and_Epistemology.html?id=7DkdAQAAMAAJ&redir_esc=y

Rico-Páez, Andrés. (2022). Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(24), e044. https://doi.org/10.23913/ride.v12i24.1196

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Secretaría de Educación Pública. (2023). Estadística e indicadores educativos: Puebla, ciclo escolar 2022–2023. Gobierno de México. https://planeacion.sep.gob.mx

Secretaría de Educación Pública. Dirección General de Planeación, Programación y Estadística Educativa. (2024). Estadística educativa. Puebla. Ciclo escolar 2023–2024 [Informe]. https://planeacion.sep.gob.mx/Doc/estadistica_e_indicadores/EstIndEntFed2023/21_PUE.pdf

Taylor, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. En A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, methodologies, and applications* (pp. 379–396). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch16

Macías-Ureta, K. T., & Ordóñez-Valencia, E. V. (2025). Metodologías activas para el desarrollo de habilidades matemáticas: Un análisis bibliográfico. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, 6*(2), 3431–3450. https://doi.org/10.56712/latam.v6i2.3917

Vygotsky, L. S. (1978). Mind in society: *The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Eds. & Trans.). Harvard University Press. https://autismusberatung.info/wp-content/uploads/2023/09/Vygotsky-Mind-in-society.pdf

Rol de Contribución	Autor (es)
Conceptualización	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Metodología	Dra. Yolanda Moya Martínez, rol principal Dra. Carmen Cerón Garnica, rol que apoya
Software	Programación, desarrollo de software; Diseño de programas informáticos; Implementación del código informático y algoritmos de soporte; Pruebas de componentes de código existentes. Dra. Yolanda Moya Martínez, rol principal Dra. Carmen Cerón Garnica, rol que apoya
Validación	Verificación, ya sea como parte de la actividad o por separado, de la replicación / reproducibilidad total de los resultados / experimentos y otros productos de la investigación. Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Análisis Formal	Dra. Yolanda Moya Martínez, rol principal Dra. Carmen Cerón Garnica, rol igual
Investigación	Dra. Yolanda Moya Martínez, rol principal Dra. Carmen Cerón Garnica, rol que apoya
Recursos	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Curación de datos	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Escritura - Preparación del borrador original	Dra. Yolanda Moya Martínez, rol principal Dra. Carmen Cerón Garnica, rol que apoya
Escritura - Revisión y edición	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Visualización	Dra. Yolanda Moya Martínez, rol principal Dra. Carmen Cerón Garnica, rol que apoya
Supervisión	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Administración de Proyectos	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual
Adquisición de fondos	Dra. Yolanda Moya Martínez, rol igual Dra. Carmen Cerón Garnica, rol igual

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Apéndice

Cuestionario aplicado a estudiantes rurales y urbanos

1. Elige alguna preparatoria
2. Sexo
3. Edad
4. Dirección (Rural o Urbana)
5. Cantidad de miembros en tú familia
6. Estado de convivencia de los padres
7. Nivel educativo de la madre
8. Nivel educativo del padre
9. Profesión de la madre
10. Profesión del padre
11. Por qué elegiste esta preparatoria
12. Quién es el tutor
13. Tiempo que te lleva llegar a la escuela
14. Tiempo semanal dedicado al estudio
15. Número de materias reprobadas
16. Recibes algún apoyo escolar
17. Recibes algún apoyo familiar
18. Recibes clases particulares
19. Participas en actividades extraescolares
20. Asististe a la guardería
21. Tienes la intención de estudiar la licenciatura
22. Tienes acceso a internet en casa
23. Estás en una relación romántica
24. Calidad de las relaciones familiares
25. Cantidad de tiempo libre semanal
26. Frecuencia con la que sales con amigos
27. Consumo de alcohol durante la semana
28. Consumo de alcohol durante el fin de semana
29. Estado de salud
30. Cantidad de ausencias escolares
31. Calificación del primer semestre

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467