KILLKANA SOCIALES

Articulo original. Revista Killkana Sociales. Vol. 9, No. 3, pp. **48-64**. Enero-Abril, 2026.
p-ISSN 2528-8008 / e-ISSN 2588-087X. Universidad Católica de Cuenca

Minería de datos educativos para mejorar el rendimiento académico: Un caso de estudio en el bachillerato

Educational data mining to improve academic performance: A case study in high school

Recepción: 10 de julio de 2025 | **Aprobación:** 30 de diciembre de 2025 | **Publicación:** 25 de enero de 2026

Santiago Vásquez Ojeda
svasquez9@indoamerica.edu.ec
Universidad Tecnológica Indoamérica, Ambato-Ecuador.

Washington Pérez Argudo
wperez6@indoamerica.edu.ec
Universidad Tecnológica Indoamérica, Ambato-Ecuador.

DOI: https://doi.org/10.26871/killkana_social.v10i1.1672

Resumen

El bajo desempeño académico es un fenómeno de múltiples causas que no pueden explicarse mediante conexiones lineales entre variables independientes. En este estudio, se asume que los factores contextuales relacionados con el ambiente familiar, desempeño académico y asistencia a la escuela son causas significativas, aunque no únicas, en una red compleja de determinantes individuales, institucionales y estructurales. Desde un enfoque dialéctico, el estudio del rendimiento académico no se limita a la descomposición analítica de sus elementos, sino que requiere una recomposición sintética. En esta última fase, las contradicciones entre el contexto social, los antecedentes educativos y las condiciones de la escuela se incorporan con totalidad explicativa del fenómeno. Este estudio se desarrolló con un enfoque cuantitativo, predictivo y explicativo, con un alcance correlacional-explicativo. La población estuvo

integrada por estudiantes de bachillerato del sistema educativo en Ecuador, a partir de una muestra extraída de la base de datos del INEVAL. Se utilizaron registros académicos estandarizados y encuestas socioeconómicas como instrumentos, los cuales fueron sometidos a procedimientos de normalización, integración y depuración. Para procesar los datos, se utilizaron árboles de decisión, modelos de regresión lineal, Random Forest y métodos de ensamble. Para la validación se llevó a cabo la precisión predictiva, coeficiente de determinación (R2) e indicadores de error (RMSE). Como premisa mayor se considera el rendimiento académico como producto de varias determinaciones y como premisa menor los factores familiares, la asistencia y el rendimiento previo resultaron ser predictores significativos del desempeño académico, de esta manera actúan como condiciones causales necesarias, pero no suficientes. Los resultados confirman que los modelos de minería de datos posibilitan el reconocimiento de patrones críticos del riesgo académico con un alto grado de precisión, proporcionado pruebas empíricas para la creación de políticas educativas basadas en la predicción, prevención y toma de decisiones estratégicas.

Palabras clave: aprendizaje automático, minería de datos educativos, modelos de regresión, rendimiento académico.

Abstract

Low academic performance is a multifactorial phenomenon that cannot be explained by linear relationships between independent variables. This study assumes that contextual factors related to family environment, academic performance, and school attendance are significant, though not the sole, causes within a complex network of individual, institutional, and structural determinants. From a dialectical perspective, the study of academic performance is not limited to the analytical decomposition of its elements but requires a synthetic reconstruction. In this latter phase, the contradictions between social context, educational background, and school conditions are incorporated to fully explain the phenomenon. This study was conducted using a quantitative, predictive, and explanatory approach, with a correlational-explanatory scope. The population consisted of high school students from the Ecuadorian education system, selected from a sample drawn from the INEVAL database. Standardized academic records and socioeconomic surveys were used as instruments, which underwent normalization, integration, and data cleaning procedures. To process the data, decision trees, linear regression models, Random Forest, and ensemble methods were used. For validation, predictive accuracy, coefficient of determination (R^2), and error indicators (RMSE) were calculated. The major premise is that academic performance is the product of several determinants, while the minor premise is that family factors, attendance, and academic performance have a significant impact, thus acting as necessary but not sufficient causal conditions. The results confirm that data mining models enable the recognition of critical patterns of academic risk with a high degree of accuracy, providing empirical evidence for the creation of educational policies based on prediction, prevention, and strategic decision-making.

Keywords: academic performance, educational data mining, machine learning, regression models.

Introducción

La educación secundaria enfrenta varios desafíos que han obstaculizado el progreso para mejorar el rendimiento educativo de los estudiantes. Este factor es esencial para inculcar la importancia del éxito futuro y el nivel de éxito en las instituciones educativas (Bonilla-Jurado et al., 2023). Los procesos pedagógicos están avanzando; sin embargo, aún existen desafíos como la falta de un sistema de enseñanza individual o la detección de estudiantes "en riesgo" que pueden reprobar el curso o programa (Baig et al., 2020). En tales condiciones, se aplicó la minería de datos como una herramienta para cambiar la forma de la educación (Vijayalakshmi &

Nivethithaa, 2021), revelar patrones, tendencias que son difíciles de detectar, estimar el aprendizaje individual, predecir el rendimiento del aprendizaje y optimizar las decisiones pedagógicas a gran escala (Fu, 2024).

El proceso de aprendizaje automático puede ser explotado para este sistema utilizando la minería de datos educativos (EDM), mediante técnicas como árbol de decisiones pueden ayudar a mejorar el estándar en la educación (Patil et al., 2024). Tales modelos pueden anticipar el rendimiento académico de los estudiantes y proporcionar detalles sobre los factores profundos detrás del desempeño escolar (Bin, 2023). Sin embargo, a pesar de la creciente visibilidad, el uso de modelos predictivos en educación enfrenta limitaciones especialmente cuando se busca integrar dimensiones académicas, socioeconómicas y demográficas dentro de un modelo sólido y eficiente (Lalaleo-Analuisa et al., 2021).

Este estudio propone un enfoque cuantitativo, predictivo y explicativo, con un alcance correlacional-explicativo, para predecir el rendimiento de los estudiantes de secundaria mediante el análisis exploratorio de datos (EDA), proceso ETL (Extracción, Transformación, Carga) y algoritmos de aprendizaje automático (Mahalle et al.,2023), con el fin de diagnosticar a los estudiantes que están en riesgo y ofrecer apoyo personalizado para mejorar el rendimiento. En el análisis educativo, la capacidad de manejar grandes cantidades de información de manera efectiva es una de las principales ventajas de incorporar la minería de datos (Jha et al., 2018).

Por otra parte, el uso de datos extraídos de fuentes distintas a las calificaciones de los estudiantes o asistencia, como indicadores socioeconómicos y comportables permite ampliar la comprensión del rendimiento académico de los estudiantes, demostrando patrones que no son fáciles de identificar por parte de los docentes (Shylaja et al., 2023). Estos modelos no solo predicen el comportamiento y las tendencias de rendimiento basándose en patrones pasados, sino que mejoran la calidad de la educación con el tiempo (Kavya et al., 2023). Sin embargo, el rendimiento de tales modelos depende de la calidad y cantidad de datos, y de la interpretabilidad de los resultados para tomar decisiones y actuar en consecuencia (Boughouas et al., 2022).

En este sentido, la aplicación de la minería de datos en la educación no solo está personalizando la educación, sino también las provisiones pedagógicas basadas en la demanda particular de los estudiantes (Chen & Jin, 2024). Este estudio "individual" absorbe el tiempo que con demasiada frecuencia los maestros se ven obligados a dedicar al diagnóstico y los remedios, no solo del alumno rezagado, sino también del alumno brillante que quiere mejorar, una reutilización de recursos educativos que solo puede ser beneficiosa (Lalaleo-Analuisa et al., 2021), por lo tanto los modelos predictivos permiten anticipar el logro académico de los estudiantes y reconocer los factores que afectan el éxito escolar (Grabovy & Siniak, 2024).

Metodología

En el contexto de la investigación se desarrolló con la aplicación de técnicas de minería de datos educativas (EDM) sobre el éxito académico en estudiantes de bachillerato. Se utilizó un modelo predictivo con regresión lineal, árboles de decisión, Random Forest y Boosted Trees para evaluar la información socioeconómica, demográfica y educativa más predictiva que afectaba el rendimiento de los estudiantes. La información se procesa utilizando la base de datos de INEVAL correspondiente al período 2023–2024, complementada con encuestas socioeconómicas realizadas a los estudiantes y sus familias.

El conjunto de datos puede tener privilegios de acceso para su uso en calificaciones, asistencia, participación en actividades no académicas, estado socioeconómico y las características demográficas de las familias de los estudiantes. Con el conjunto de datos se entrenaron modelos de clasificación para predecir si los estudiantes pertenecían a grupos de alto o bajo rendimiento.

Diseño y tratamiento de los datos

La recopilación y consolidación de registros académicos, variables sociodemográficas y condiciones socioeconómicas de las familias fue el comienzo del proceso metodológico. Para asegurar la consistencia de datos se llevó a cabo procedimientos de normalización, limpieza y depuración. Se ajustaron las escalas de las variables numéricas empleadas en los modelos predictivos y se eliminaron datos incompletos.

El desempeño académico anterior, la asistencia a la escuela, el ingreso familiar, el grado de educación de los padres, las características del hogar y los rasgos demográficos (edad, sexo y ubicación) fueron variables que se tomaron en cuenta (Guevara & Bonilla, 2021). Se eligieron estas variables por la importancia reportada en investigaciones anteriores y disponibles en la base de datos.

Modelado predictivo

Se compararon y crearon diversos modelos de aprendizaje automático como fueron los árboles de decisión, regresión lineal, Random Forest y Boosted Trees, utilizando el conjunto de datos depurado. Para cada modelo los datos se fraccionaron en subconjuntos de entrenamiento y prueba en una proporción de 70/30. Los algoritmos fueron entrenados después con el conjunto de entrenamiento y validados con el conjunto de prueba.

Los modelos fueron analizados con métricas que aceptan de manera general en el análisis predictivo:

- Error cuadrático medio (RMSE) para calcular la magnitud del error;
- Coeficiente de determinación (R2) para medir el poder explicativo;
- Porcentaje de precisión (%) para clasificar a los estudiantes conforme a los niveles de rendimiento.

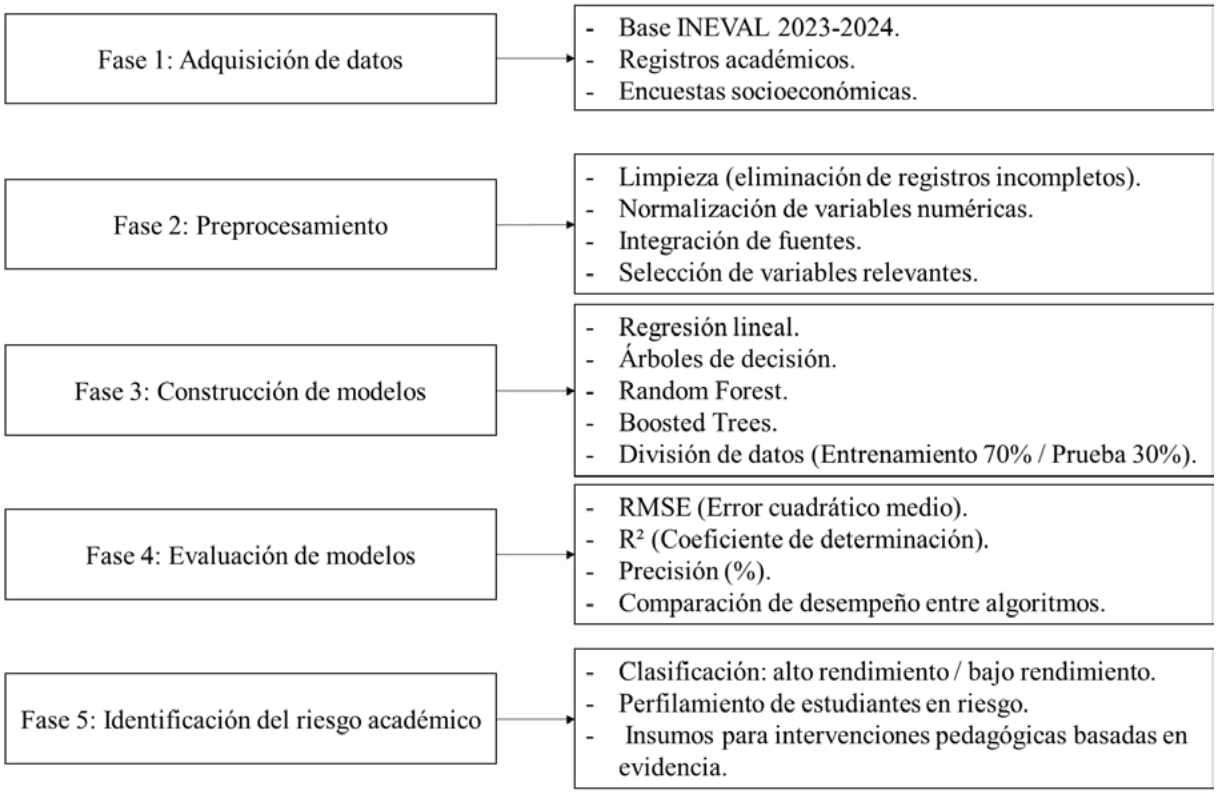
Criterios de clasificación

La variable dependiente se operacionalizó dividiendo a los estudiantes en grupos de bajo o alto desempeño, basados en el promedio final de los módulos que se reportaron en el sistema académico. Los modelos generaron estimaciones probabilísticas acerca de la categoría a lo que pertenecía cada estudiante, lo que permitió identificar los perfiles académicos en riesgo.

La imagen 1 muestra el diagrama metodológico del estudio donde se describe las cinco fases, integrando las variables académicas como indicadores socioeconómicos para desarrollar modelos basados en el aprendizaje automático.

Figura 1.

Diagrama metodológico del estudio



Nota: El gráfico representa el diagrama metodológico del estudio basado en cinco fases.

Resultados

Los resultados de la investigación evaluaron el rendimiento de los modelos predictivos aplicados y determinaron las variables que tienen un impacto más significativo en el desempeño académico de los estudiantes. A continuación, se observan los resultados principales obtenidos del análisis:

Desempeño de los modelos predictivos

Los modelos de aprendizaje automático lograron clasificar a los estudiantes en grupos de alto y bajo rendimiento académico con niveles de precisión aceptables. Los porcentajes de exactitud fueron:

- Árboles de decisión: 80%
- Random Forest: 85%
- Boosted Trees: 83%

El modelo Random Forest tuvo el mejor rendimiento, sobresaliendo la estabilidad predictiva y la precisión. Como se detalla en la tabla 1 las métricas de rendimiento de los modelos predictivos.

Tabla 1

Métricas de rendimiento de los modelos predictivos

Modelo	RMSE	R^2	Precisión (%)
Regresión lineal	18.5	0.18	75
Árbol de Decisión	14.3	0.20	80
Random Forest	13.8	0.24	85
Boosted Tress	13.9	0.22	83

Nota. Elaboración propia.

El modelo de Random Forest, que tiene valores de R^2 (coeficiente de determinación) y los más bajos de RMSE (Error cuadrático medio), fue el que mostro mayor eficacia al momento de describir y pronosticar el desempeño académico. Los modelos de ensamble resultaron superiore a los modelos individuales, lo que concuerda con las investigaciones anteriores en EDM (minería de datos educativas).

Importancia de las Variables

La evaluación de la importancia de las variables indicó que los factores que impactaron el rendimiento académico resultaron:

- Situación económica y social del hogar.
- Desempeño académico previo.
- Asistencia a la escuela.
- Modalidad de financiamiento de la institución.
- Sistema de evaluación.

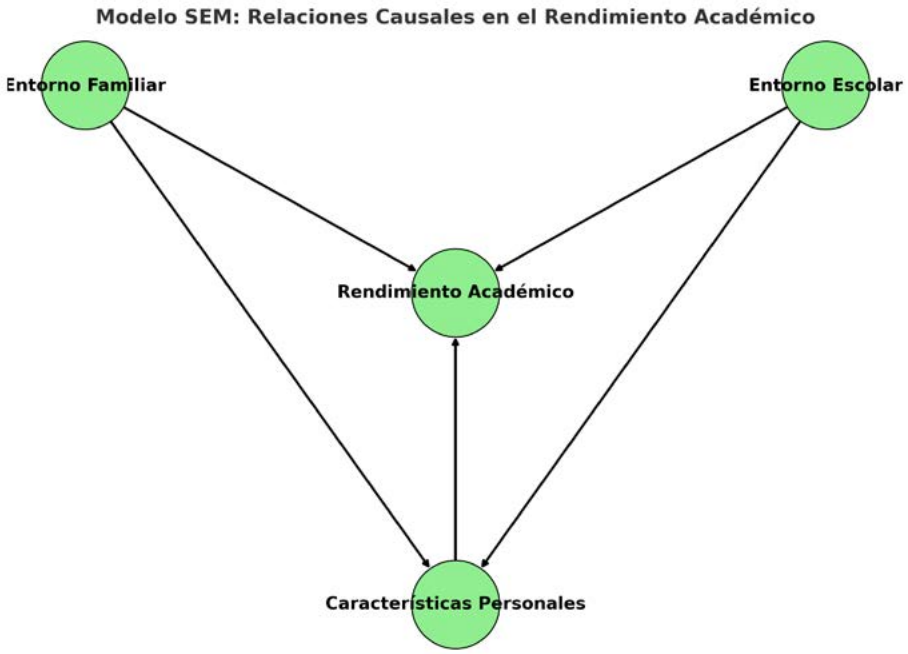
Estos resultados corroboran que el rendimiento de los estudiantes no se basa solo en factores escolares sino en la interacción de las condiciones socioeconómicas, la estabilidad académica y las particularidades institucionales.

Modelo SEM: Relaciones causales en el rendimiento académico

A continuación, se presenta la figura 2 modelo SEM (Ecuaciones Estructurales) que ilustra las relaciones causales entre las variables latentes que influyen en el rendimiento académico:

Figura 2

Modelo SEM



Nota: Elaboración propia.

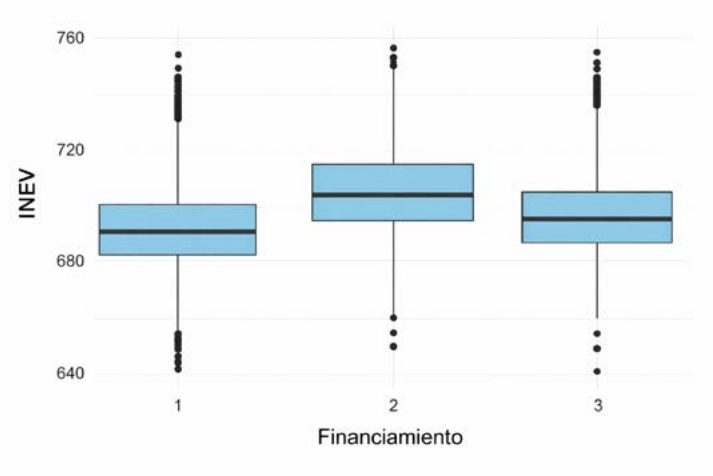
Este modelo describe cómo el entorno familiar, escolar y características personales del estudiante están relacionados y afectan su rendimiento académico (Weiser, 2020). En este caso, las variables latentes como el nivel socioeconómico y el apoyo familiar influyen tanto en las características personales del estudiante (como motivación y hábitos de estudio) como en su desempeño en el ámbito académico (Bonilla-Jurado et al., 2024).

Gráficos Boxplots

La figura 3 presentada a continuación muestran la relación entre diversas variables socioeconómicas y el rendimiento académico de los estudiantes, medido a través del índice de rendimiento académico (INEV).

Figura 3.

Financiamiento

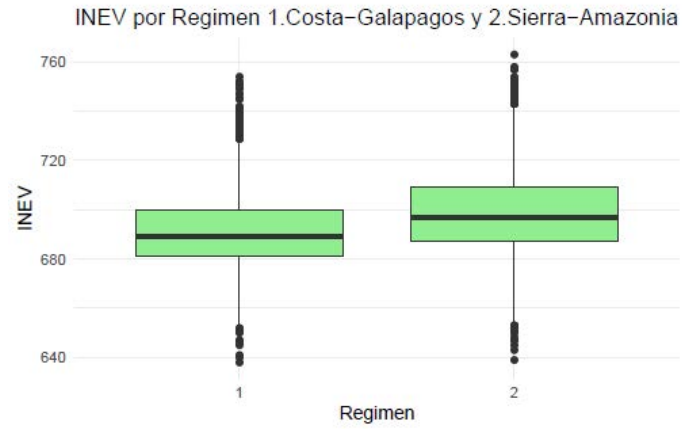


Nota: Elaboración propia.

Se muestra en la figura 2 la distribución del rendimiento académico (INEV) en función del nivel de financiamiento (1, 2 y 3). Aunque las medianas y los rangos intercuartílicos de los tres grupos son similares, se observan varios valores atípicos en todos los grupos. Esto sugiere que, aunque el financiamiento parece no tener una gran variabilidad en el rendimiento promedio, existen estudiantes con rendimientos significativamente más bajos o altos que la mayoría, lo que indica que otros factores también influyen en el desempeño académico.

Figura 4.

Distribución de INEV según Régimen de evaluación



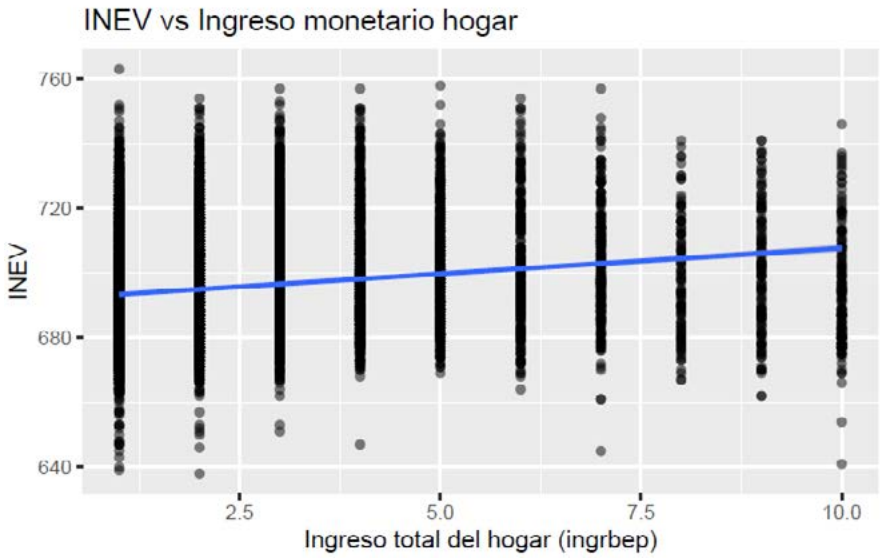
Nota: Elaboración propia.

La figura 4, compara el rendimiento académico (INEV) entre la región Costa-Galápagos (Régimen 1) y Sierra-Amazonía (Régimen 2). Se observa que la mediana del rendimiento académico es más alta en la región Costa-Galápagos, aunque con una mayor dispersión de los

datos, lo que indica una mayor variabilidad en el desempeño de los estudiantes. En cambio, en la región Sierra-Amazonía, la mediana es más baja y la distribución es más concentrada, con menos variabilidad en los resultados académicos. Ambas regiones presentan valores atípicos, pero la región Costa-Galápagos tiene una mayor dispersión en el rendimiento, sugiriendo diferencias significativas en los resultados académicos.

Figura 5.

INEV vs Ingreso monetario

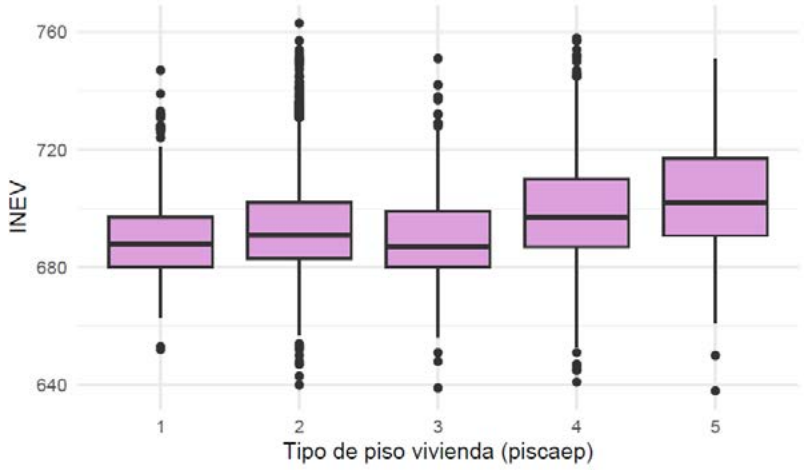


Nota: Elaboración propia.

La figura 5 presenta la relación entre el ingreso total del hogar (en la variable "ingrbep") y el rendimiento académico (INEV) de los estudiantes. Existe una tendencia positiva, a medida que aumenta el ingreso familiar, también lo hace el rendimiento académico, lo que indica que los estudiantes de hogares con mayores ingresos tienden a tener un rendimiento académico más alto. La línea azul de regresión refuerza esta relación, aunque la dispersión de los puntos muestra que, a pesar de la tendencia general, hay variabilidad en los resultados académicos, lo que sugiere la influencia de otros factores además del ingreso familiar.

Figura 6.

Distribución de INEV según tipo de piso

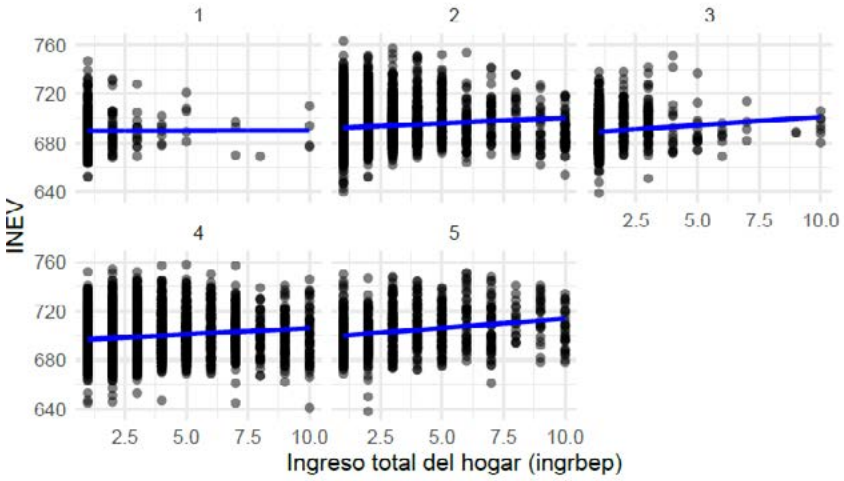


Nota: Elaboración propia.

La figura 6 representa la distribución del Índice Nacional de Evaluación Educativa (INEV) según el tipo de piso de vivienda (piscsaep), clasificado en cinco categorías. Cada caja muestra la mediana, el rango intercuartílico y los valores atípicos para cada tipo de piso. Las distribuciones de INEV son similares entre los tipos de piso, pero con algunas variaciones, especialmente en los tipos 2, 3 y 4, donde se presentan más valores atípicos. Esto sugiere que el tipo de piso tiene una ligera influencia sobre el rendimiento educativo, aunque no de manera significativa.

Figura 7.

INEV vs Ingreso total del hogar, facetado por tipo de piso



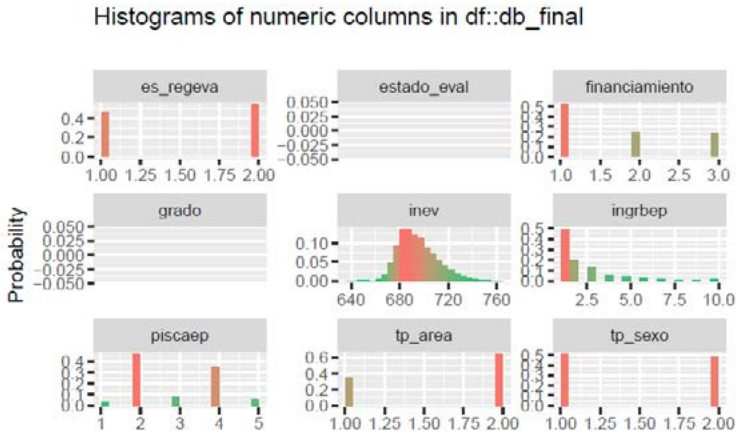
Nota: Elaboración propia.

La figura 7, enseña la relación entre el rendimiento académico (INEV) y el ingreso total del hogar, facetada por tipo de piso. Se observa cómo varía esta relación según las condiciones de la

vivienda. Al segmentar los datos por tipo de piso, se puede verificar si existe una influencia del entorno físico en la relación entre el ingreso familiar y el rendimiento académico de los estudiantes. Esta visión facilita una comprensión detallada de por qué los factores socioeconómicos y las condiciones de la vivienda interactúan en el desempeño escolar.

Figura 8.

Histogramas de las columnas numéricas

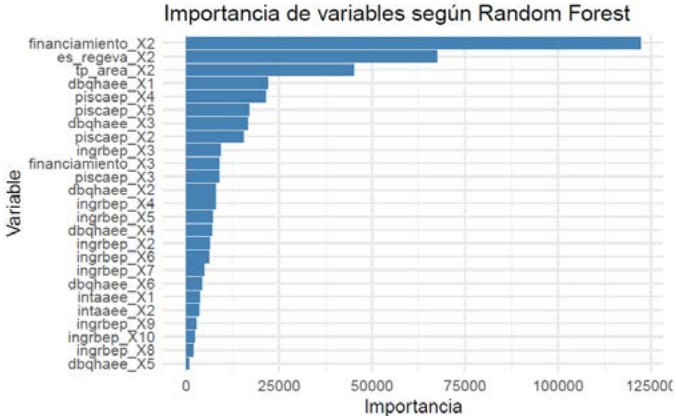


Nota: Elaboración propia.

La figura 8 muestra los histogramas de varias columnas numéricas del conjunto de datos df::db_final. Cada histograma representa la distribución de probabilidad de una variable, proporcionando una visión rápida de cómo se distribuyen los datos. Por ejemplo, el histograma de INEV muestra una distribución más concentrada entre 680 y 720, mientras que los histogramas de variables como grado, financiamiento e ingreso del hogar presentan distribuciones más dispersas o sesgadas. Este análisis visual ayuda a identificar patrones y distribuciones de los datos, así como posibles valores atípicos en las variables.

Figura 9.

Modelo Random Forest



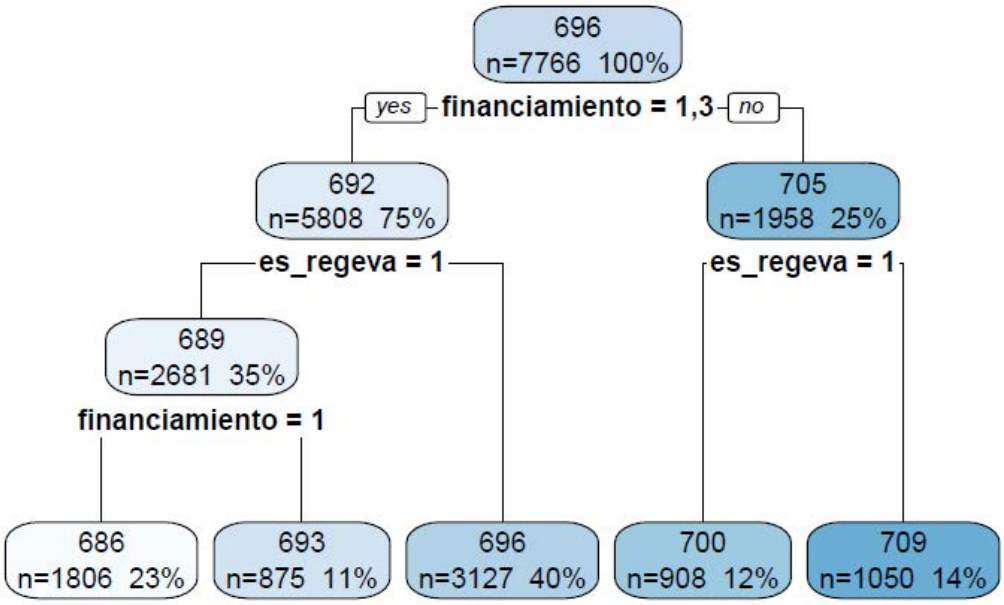
Nota: Elaboración propia.

La figura 9 muestra la importancia de las variables según un modelo de Random Forest. Las variables están ordenadas de acuerdo con su contribución al modelo, donde financiamiento_X2 es la variable más importante, seguida de es_regeva_X2 y tp_area_X2. Las variables con mayor importancia tienen un impacto significativo en la predicción del modelo, mientras que las que están al final de la lista tienen una influencia menor. Este análisis es útil para identificar cuáles son las características más relevantes que afectan la variable objetivo en el modelo predictivo, ayudando a tomar decisiones sobre qué variables considerar para mejorar el rendimiento del modelo.

BOOST_TREE

Figura 10.

Árbol de decisión



Nota: Elaboración propia.

Se muestra en la figura 10, un árbol de decisión para clasificar a los estudiantes según el financiamiento recibido, dividiendo la población en diferentes grupos con base en las variables es_regeva y financiamiento. El árbol comienza con la pregunta principal sobre si los estudiantes recibieron financiamiento 1,3. Si la respuesta es "sí", el siguiente nodo clasifica según la variable es_regeva. A continuación, se generan ramas para diferentes niveles de financiamiento y las decisiones de clasificación, los números dentro de cada nodo representan el número de estudiantes y su porcentaje respecto al total. Este modelo ayuda a predecir la probabilidad de que un estudiante pertenezca a un grupo específico basado en estas características.

Los resultados obtenidos a partir de los análisis y gráficos muestran cómo diversos factores socioeconómicos, ambientales y personales influyen en el rendimiento académico de los estudiantes. Los gráficos de boxplot revelan que, aunque variables como el financiamiento y el

régimen de evaluación tienen una relación con el rendimiento académico, existen valores atípicos que sugieren que otros factores, como el ingreso familiar y el tipo de piso, juegan un papel importante en la variabilidad de los resultados. La distribución de INEV por región muestra que la Costa-Galápagos presenta una mayor dispersión en el rendimiento, mientras que en Sierra-Amazonía la mediana es más baja y la distribución más concentrada. Los gráficos también evidencian una tendencia positiva entre el ingreso familiar y el rendimiento académico, pero la dispersión de los puntos indica que otros factores, además del ingreso, influyen en el desempeño.

Discusión

Los resultados de este estudio tienen implicaciones realistas para los estudiantes de secundaria, en términos de influir en los dominios socioeconómicos, familiares y escolares de los estudiantes (Guevara & Bonilla, 2021). Por el contrario, los diagramas de caja mostraron que la subvención tiende a tener una relación llamada positiva con los factores académicos y relacionados, pero también hay algunas excepciones, siendo que la financiación no es la única razón de la desviación en el factor socioambiental; podría haber varias otras razones como el ingreso familiar, el tipo de terreno (Bonilla-Jurado, 2025). Una de esas consecuencias es el trabajo temprano, Zhang et al., (2024) revela que cuando se controlan otros factores, cuando el ingreso familiar de un estudiante es más alto, el GPA del niño (es decir, el impacto en el camino académico de un niño está significativamente correlacionado positivamente con el ingreso familiar de un estudiante; cuanto más rica es la familia, más positivo es el efecto sobre el GPA de un estudiante). Así que el dinero cuenta, pero no es el único determinante del éxito académico (Baig et al., 2020).

El rendimiento escolar de los estudiantes basado en Random Forest como en Boost Tree determinaron los atributos relevantes (Tran et al., 2025). Los modelos de aprendizaje automático predicen el fracaso académico/deserción entre los estudiantes, esto demuestra una especie de linealidad predecible de los grandes datos en la vida de los estudiantes (Garg et al., 2022). Estos hallazgos coinciden con investigaciones previas que afirman que los modelos predictivos, como los árboles de decisión y Random Forest, pueden ser herramientas útiles para identificar estudiantes en riesgo de bajo rendimiento y para personalizar las intervenciones educativa (Lou & Colvin, 2025). Es así que modelos nos permiten ilustrar una representación visual de los resultados para transmitir la función de los datos y de SES/AA con el fin de tomar decisiones más inteligentes para la política educativa (Bonilla-Jurado et al., 2023).

En cuanto a la información específica se observa que la zona Costa-Galápagos tiene una distribución más dispersa que la de la Zona Sierra-Amazonía, que es más concentrada, en comparación con la de los resultados restantes (Tin et al., 2024). Esto se llama desigualdad educativa. De hecho, puede ser la razón por la cual, más allá de la desigualdad regional, vivir en una región con un bajo nivel de desarrollo está asociado a un menor logro, dado que el grupo más bajo y el lugar de residencia con el nivel de desarrollo no son parte de la misma región (Padmavathi et al., 2024). Como una de las causas de tal conciencia, vivir en una región con un bajo nivel de desarrollo podría influir en tener una escuela pobre (banda C) o, si se asiste a la mejor escuela disponible en la región (escuelas de banda A, entonces, podrían ser las mejores disponibles en algunas regiones y las peores en otras). Se debe enfatizar el impacto de las políticas territoriales

de ECE y sus características y el deterioro de la equidad en la provisión del nivel educativo (Bonilla-Jurado et al., 2024).

La aplicación de la personalización mediante grandes datos también es novedosa en este estudio (Tin et al., 2024). Podemos predecir y saber de antemano qué tanto puede lograr aprender un estudiante en una escuela (Bonilla-Jurado & Meléndez, 2023). Es un pequeño paso hacia sacar el máximo provecho de los recursos educativos significando un mejor aprendizaje. Estos modelos de predicción, como los mostrados en los ejemplos de la sección anterior (Bai et al., 2021), pueden desbloquear las "compuertas" para apoyar intervenciones inteligentes para todos los estudiantes y remodelar el panorama educativo para que sea más justo y equitativo. Por lo tanto, es muy crucial traer los últimos avances tecnológicos en el sector educativo para que se pueda proporcionar una mejor educación a los estudiantes y puedan mantenerse al ritmo de la carrera de la vida (Jha et al., 2018).

Conclusión

Los resultados de esta investigación, junto con el análisis de datos en este año académico 2023-2024, confirman que EDM como un plan y estrategia mejoran el rendimiento de los estudiantes en el contexto ecuatoriano. La aplicación de modelos predictivos (Random Forest y Boosted Trees) (ambos con un 85% y 83% de precisión), permitió concluir que los problemas contextuales relacionados con la familia y el comportamiento académico previo, junto con la asistencia escolar, son los predictores más importantes para el bajo rendimiento académico.

Estos hallazgos demuestran la importancia de la programación educativa específica del contexto, ya que no existe un programa educativo único para todos los jóvenes. Se deben tomar en cuenta las situaciones académicas y la constitución social de los alumnos en una gestión escolar más equitativa, que sea tanto conocedora como sensible a las diferencias. Sin embargo, su alcance se basa tanto en la riqueza como en la posición. La distribución del rendimiento estudiantil (líneas de tendencia) demuestra la importancia de estas medidas de educación, ocupación y escolaridad de los padres para los resultados de aprendizaje de los niños.

El principio final de aprender mientras se vive está bien siempre que el uso de nuevas tecnologías y la organización de nuevos conocimientos mejoren la calidad del aprendizaje. También sería útil, la identificación basada en datos de quién está teniendo dificultades para que se puedan diseñar intervenciones y asignar recursos de manera eficiente mediante la creación de un sistema de aprendizaje más equitativo y justo. No es solo que las escuelas tengan la oportunidad de ser pioneras en modelos predictivos y minería de datos, sino que puedan hacerlo de una manera en la que todos, desde el primer día de vida hasta el primer día en el mundo real, puedan estar a bordo.

Referencias

Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational Big Data: Predictions, Applications and Challenges. *Big Data Research, 26*, 100270. https://doi.org/10.1016/J.BDR.2021.100270

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education, 17*(44), 1–23. https://doi.org/10.1186/s41239-020-00223-0

Bin, L. (2023). Cognitive Web Service-Based Learning Analytics in Education Systems Using Big Data Analytics. *International Journal of E-Collaboration, 19*(2). https://doi.org/10.4018/IJeC.316658

Bonilla-Jurado, D. (2025). Las tecnologías de la información y la comunicación en los ERP para la gestión empresarial: Un análisis bibliométrico. *Ciencias Administrativas*, 25, 147–147. https://doi.org/10.24215/23143738E147

Bonilla-Jurado, D., Guevara, C., Ayala-Gavilanes, C., & Lliguisupa-Pastor, M. (2023). The School Dropout: Causes and Effects in University Education. Journal of Higher Education Theory and Practice, 23(18), 162–170. https://doi.org/10.33423/JHETP.V23I18.6629

Bonilla-Jurado, D., & Meléndez, C. (2023). Integración de los Objetivos de Desarrollo Sostenible a la planificación institucional del Instituto Tecnológico Superior España. PLURIVERSIDAD, 11, 101–115. https://doi.org/10.31381/PLURIVERSIDAD11.6278

Bonilla-Jurado, D., Zumba, E., Lucio-Quintana, A., Yerbabuena-Torres, C., Ramírez-Casco, A., & Guevara, C. (2024). Advancing University Education: Exploring the Benefits of Education for Sustainable Development. *Sustainability, 16*(17), 7847. https://doi.org/10.3390/su16177847

Boughouas, M. L., Kissoum, Y., Mouhssen, A., Karek, M. A., & Mazouzi, S. (2022). Towards a Big Educational Data Analytics. *ICAASE 2022 - 5th Edition of the International Conference on Advanced Aspects of Software Engineering, Proceedings*. https://doi.org/10.1109/ICAASE56196.2022.9931565

Chen, Y., & Jin, K. (2024). Educational Performance Prediction with Random Forest and Innovative Optimizers: A Data Mining Approach. *International Journal of Advanced Computer Science and Applications, 15*(3), 69–78. https://doi.org/10.14569/IJACSA.2024.0150308

Fu, Q. (2024). Research on Student Behavior Analysis and Grade Prediction System Based on Student Behavior Characteristics. *Scalable Computing: Practice and Experience, 25*(1), 217–228. https://doi.org/10.12694/SCPE.V25I1.2286

Garg, A., Garg, N. B., Ghosh, P., Bansal, A., Lilhore, U. K., & Simaiya, S. (2022). A Machine Learning-based Automatic Model to Predicting Performance of Students. *Proceedings of 2022 IEEE International Conference on Current Development in Engineering and Technology, CCET 2022*. https://doi.org/10.1109/CCET56606.2022.10080607

Grabovy, P., & Siniak, N. (2024). Using AI and big data in decision making: A framework across disciplines. *E3S Web of Conferences, 535*, 05011. https://doi.org/10.1051/E3SCONF/202453505011

Guevara, C., & Bonilla, D. (2021). Algorithm for Preventing the Spread of COVID-19 in Airports and Air Routes by Applying Fuzzy Logic and a Markov Chain. *Mathematics 2021, Vol. 9, Page 3040, 9*(23), 3040. https://doi.org/10.3390/MATH9233040

Jha, S., Jha, M., & O'Brien, L. (2018). A Step towards Big Data Architecture for Higher Education Analytics. *Proceedings - 2018 5th Asia-Pacific World Congress on Computer Science and Engineering, APWC on CSE 2018*, 178–183. https://doi.org/10.1109/APWCONCSE.2018.00036

Kavya, N., Manasa, S., Shrihari, M. R., Manjunath, T. N., & Mahesh, M. R. (2023). The Secured System for Continuous Improvement in Educational Institutes Using Big Data Analytics. *Lecture Notes in Networks and Systems, 782 LNNS*, 183–195. https://doi.org/10.1007/978-981-99-6568-7_17

Lalaleo-Analuisa, F. R., Bonilla-Jurado, D. M., & Robles-Salguero, R. E. (2021). Information and Communication Technologies exclusively for consumer behavior from a theoretical perspective. *Retos(Ecuador), 11*(21), 147–163. https://doi.org/10.17163/RET.N21.2021.09

Lou, Y., & Colvin, K. F. (2025). Performance prediction using educational data mining techniques: a comparative study. *Discover Education, 4*(112). https://doi.org/10.1007/S44217-025-00502-W

Mahalle, P. N., Hujare, P. P., & Shinde, G. R. (2023). Data Acquisition and Preparation. *SpringerBriefs in Applied Sciences and Technology, Part F1278*, 11–38. https://doi.org/10.1007/978-981-99-4850-5_2

Padmavathi, A., Pandit, B., Khaitan, G., & Varma, S. (2024). UNNATI: Enhancing Quality Education in Rural Areas through AI, AR & digitalization. *2024 2nd International Conference on Advances in Computation, Communication and Information Technology, ICAICCIT 2024,* 580–584. https://doi.org/10.1109/ICAICCIT64383.2024.10912363

Patil, S., Patwal, P. S., & Wadane, V. S. (2024). Machine learning approach for educational data mining on real life applications. *IET Conference Proceedings, 2024*(38), 370–374. https://doi.org/10.1049/ICP.2025.0892

Shylaja, A. R., Shubhashree, D. A., Shrihari, M. R., Manjunath, T. N., & Ajay, N. (2023). Secure Data Education: Leveraging Big Data for Enhanced Academic Performance and Student Success in Educational Institutions. *Lecture Notes in Networks and Systems, 754 LNNS,* 111–124. https://doi.org/10.1007/978-981-99-4932-8_12

Tin, T. T., Hock, L. S., & Ikumapayi, O. M. (2024). Educational Big Data Mining: Comparison of Multiple Machine Learning Algorithms in Predictive Modelling of Student Academic Performance. *International Journal of Advanced Computer Science and Applications, 15*(6), 633–645. https://doi.org/10.14569/IJACSA.2024.0150664

Tran, T. T., Phan, N. Q., & Huynh, H. X. (2025). Random Forest Model Parameters Optimization. *Communications in Computer and Information Science*, 2191 CCIS, 237–247. https://doi.org/10.1007/978-981-97-9616-8_19

Vijayalakshmi, S., & Nivethithaa, K. K. (2021). Survey on Data Mining Techniques, Process and Algorithms. *Journal of Physics: Conference Series, 1947*(1), 012052. https://doi.org/10.1088/1742-6596/1947/1/012052

Weiser, E. B. (2020). Structural equation modeling in personality research. *The Wiley Encyclopedia of Personality and Individual Differences, Measurement and Assessment*, 137–142. https://doi.org/10.1002/9781119547167.CH93

Zhang, C., Yang, J., Li, M., & Deng, M. (2024). Simulation-Based Machine Learning for Predicting Academic Performance Using Big Data. *International Journal of Gaming and Computer-Mediated Simulations, 16*(1). https://doi.org/10.4018/IJGCMS.348052