UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTAD DE INGENIERÍAS

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

TEMA:

MINERÍA DE DATOS EDUCATIVOS PARA MEJORAR EL RENDIMIENTO ACADÉMICO: UN CASO DE ESTUDIO EN EL BACHILLERATO

Trabajo de Titulación previo a la obtención del título de Magíster en BIG DATA Y CIENCIA DE DATOS.

Autor(a)
Ing. Santiago Javier Vásquez Ojeda
Tutor(a)
Ing. Washington O. Pérez A., Mg.

AMBATO – ECUADOR
2025

AUTORIZACIÓN POR PARTE DEL AUTOR PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE TITULACIÓN

Yo, Santiago Javier Vásquez Ojeda, declaro ser autor del Trabajo de Titulación con el nombre "MINERÍA DE DATOS EDUCATIVOS PARA MEJORAR EL RENDIMIENTO ACADÉMICO: UN CASO DE ESTUDIO EN EL BACHILLERATO", como requisito para optar al grado de Magíster en BIG DATA Y CIENCIA DE DATOS y autorizo al Sistema de Bibliotecas de la Universidad Indoamérica, para que con fines netamente académicos divulgue esta obra a través del Repositorio Digital Institucional (RDI-UTI).

Los usuarios del RDI-UTI podrán consultar el contenido de este trabajo en las redes de información del país y del exterior, con las cuales la Universidad tenga convenios. La Universidad Indoamérica no se hace responsable por el plagio o copia del contenido parcial o total de este trabajo.

Del mismo modo, acepto que los Derechos de Autor, Morales y Patrimoniales, sobre esta obra, serán compartidos entre mi persona y la Universidad Indoamérica, y que no tramitaré la publicación de esta obra en ningún otro medio, sin autorización expresa de la misma. En caso de que exista el potencial de generación de beneficios económicos o patentes, producto de este trabajo, acepto que se deberán firmar convenios específicos adicionales, donde se acuerden los términos de adjudicación de dichos beneficios.
Para constancia de esta autorización, en la ciudad de Ambato a los treinta y un días del mes de agosto de 2025, firmo conforme:

Autor: Santiago J. Vásquez O.
Firma:

Firmado electrónicamente por:
SANTIAGO JAVIER VASQUEZ OJEDA
Validar únicamente con FirmaEC
Número de Cédula: 180420800-5
Dirección: Tungurahua, Ambato, La Península, Catiglata.
Correo Electrónico: santiago_vasquez180@hotmail.com
Teléfono: 0984972790

APROBACIÓN DEL DIRECTOR

En mi calidad de Director del Trabajo de Titulación "MINERÍA DE DATOS EDUCATIVOS PARA MEJORAR EL RENDIMIENTO ACADÉMICO: UN CASO DE ESTUDIO EN EL BACHILLERATO" presentado por Santiago Javier Vásquez Ojeda, para optar por el Título de Magíster en BIG DATA Y CIENCIA DE DATOS.

CERTIFICO

Que dicho Trabajo de Titulación ha sido revisado en todas sus partes y considero que reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte los Examinadores que se designe.

Ambato, 28 de agosto del 2025

……

Ing. Washington O. Pérez A., Mg.

DIRECTOR

DECLARACIÓN DE AUTENTICIDAD

Quien suscribe, declaro que los contenidos y los resultados obtenidos en el presente Trabajo de Titulación, como requerimiento previo para la obtención del Título de Magíster en BIG DATA Y CIENCIA DE DATOS, son absolutamente originales, auténticos y personales y de exclusiva responsabilidad legal y académica del autor

Ambato, 31 de agosto del 2025

.....

Santiago Javier Vásquez Ojeda

180420800-5

APROBACIÓN DE EXAMINADORES

El Trabajo de Titulación ha sido revisado, aprobado y autorizada su impresión y empastado, sobre el Tema: MINERÍA DE DATOS EDUCATIVOS PARA MEJORAR EL RENDIMIENTO ACADÉMICO: UN CASO DE ESTUDIO EN EL BACHILLERATO, previo a la obtención del Título de Magíster en BIG DATA Y CIENCIA DE DATOS, reúne los requisitos de fondo y forma para que el estudiante pueda presentarse a la sustentación del Trabajo de Titulación.

Ambato, 31 de agosto del 2025

……..

Ing. QUEZADA SARMIENTO PABLO ALEJANDRO, PhD.

EXAMINADOR

……..

Ing. MIRANDA VILLACIS ALBA DE LOS CIELOS, PhD.

EXAMINADOR

DEDICATORIA

Dedico este trabajo con profundo agradecimiento y amor a mi familia, cuya presencia y apoyo han sido esenciales a lo largo de este proceso académico. A mi esposa Alexandra, por su comprensión, paciencia y constante aliento, que me han impulsado a alcanzar esta meta. A mis hijos Samara y Benjamín, quienes con su ternura y alegría me motivan cada día a ser mejor. Este logro también les pertenece.

Gracias por ser mi fuerza, mi inspiración y mi razón para seguir adelante.

AGRADECIMIENTO

Agradezco a Dios por brindarme la fortaleza y sabiduría necesarias para culminar esta etapa académica. A mi familia, por su amor, comprensión y apoyo incondicional. A mis hijos, fuente de inspiración y motivación constante. A la Universidad Tecnológica Indoamérica y a sus docentes, por su excelencia académica y valiosa orientación durante todo el proceso formativo.

A todos quienes, de manera directa o indirecta, contribuyeron a la realización de este trabajo, mi más sincero agradecimiento.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA
FACULTAD DE INGENIERÍAS
MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

TEMA: MINERÍA DE DATOS EDUCATIVOS PARA MEJORAR EL RENDIMIENTO ACADÉMICO: UN CASO DE ESTUDIO EN EL BACHILLERATO

AUTOR(A): Ing. Santiago Javier Vásquez Ojeda
TUTOR (A): Ing. Washington O. Pérez A., Mg.

RESUMEN EJECUTIVO

El objetivo de este estudio es analizar el rendimiento académico de los estudiantes de bachillerato utilizando Minería de datos y modelos predictivos para identificar las variables clave que influyen en su desempeño. La metodología empleada se basa en el uso de técnicas estadísticas y de aprendizaje automático, como la regresión lineal, árboles de decisión, Random Forest y Boosted Trees, aplicadas a datos académicos y socioeconómicos obtenidos de la base de datos del Ineval se analizaron datos de los años 2023–2024. Los resultados indican que el modelo de Random Forest es el más preciso, alcanzando una precisión del 85%, seguido por Boosted Trees con un 83%. Las variables socioeconómicas, como el ingreso familiar y el nivel educativo de los padres, junto con el rendimiento académico previo y la asistencia a clases, fueron identificadas como los factores más influyentes en el rendimiento de los estudiantes. En conclusión, este estudio subraya la importancia de integrar Data Mining en la educación, ya que permite personalizar las estrategias pedagógicas y tomar decisiones informadas para mejorar el rendimiento académico de los estudiantes, considerando un enfoque multidimensional que abarque tanto los aspectos académicos como los socioeconómicos.

DESCRIPTORES: Aprendizaje automático, Educación secundaria, Minería de datos, Modelos predictivos, Rendimiento académico

ABSTRACT

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTY OF ENGINEERING

MASTER'S IN BIG DATA AND DATA SCIENCE

AUTHOR: VASQUEZ OJEDA SANTIAGO JAVIER

TUTOR: MSc. PEREZ ARGUDO WASHINGTON

ABSTRACT

Educational Data Mining to Improve Academic Performance: A Case Study in High School.

The objective of this research is to analyze the academic performance of high school students using data mining and predictive models to identify the key variables that influence their performance. The methodology employed is based on the use of statistical and machine learning techniques, such as linear regression, decision trees, Random Forest, and Boosted Trees, applied to academic and socioeconomic data obtained from the Ineval database. Data from the years 2023–2024 were analyzed. The results indicate that the Random Forest model is the most accurate, achieving 85% accuracy, followed by Boosted Trees with 83%. Socioeconomic variables, such as family income and parents' educational level, along with previous academic performance and class attendance, were identified as the most influential factors in student performance. In conclusion, this study highlights the importance of integrating data mining into education, as it allows for the personalization of teaching strategies and informed decision-making to improve student academic performance, considering a multidimensional approach that encompasses both academic and socioeconomic aspects.

KEYWORDS: Keywords: academic achievement, academic performance, data mining, predictive models.

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

Resumen

El objetivo de este estudio es analizar el rendimiento académico de los estudiantes de bachillerato utilizando Minería de datos y modelos predictivos para identificar las variables clave que influyen en su desempeño. La metodología empleada se basa en el uso de técnicas estadísticas y de aprendizaje automático, como la regresión lineal, árboles de decisión, Random Forest y Boosted Trees, aplicadas a datos académicos y socioeconómicos obtenidos de la base de datos del Ineval se analizaron datos de los años 2023–2024. Los resultados indican que el modelo de Random Forest es el más preciso, alcanzando una precisión del 85%, seguido por Boosted Trees con un 83%. Las variables socioeconómicas, como el ingreso familiar y el nivel educativo de los padres, junto con el rendimiento académico previo y la asistencia a clases, fueron identificadas como los factores más influyentes en el rendimiento de los estudiantes. En conclusión, este estudio subraya la importancia de integrar Data Mining en la educación, ya que permite personalizar las estrategias pedagógicas y tomar decisiones informadas para mejorar el rendimiento académico de los estudiantes, considerando un enfoque multidimensional que abarque tanto los aspectos académicos como los socioeconómicos.

Abstract

The objective of this study is to analyze the academic performance of high school students using data mining and predictive models to identify key variables that influence their performance. The methodology employed is based on the use of statistical and machine learning techniques, such as linear regression, decision trees, Random Forest, and Boosted Trees, applied to academic and socioeconomic data obtained from the Ineval database. Data from the years 2023–2024 were analyzed. The results indicate that the Random Forest model is the most accurate, reaching an accuracy of 85%, followed by Boosted Trees with 83%. Socioeconomic variables, such as family income and parents' educational level, along with previous academic performance and class attendance, were identified as the most influential factors in student performance. In conclusion, this study highlights the importance of integrating data mining into education, as it allows for personalized pedagogical strategies and informed decision-making to improve students' academic performance, considering a multidimensional approach that encompasses both academic and socioeconomic aspects.

Keywords: Data Mining, Academic Performance, Academic Achievement, Predictive Models.

Introducción

La educación secundaria se enfrenta varios impedimentos que han obstaculizado el progreso de la educación secundaria 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 el data mining como una herramienta para cambiar la forma de la educación (Vijayalakshmi & Nivethithaa, 2021).La minería de datos puede revelar patrones y tendencias que son difíciles de detectar y tiene el poder de 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) y nuevamente se dice que un árbol de decisiones también puede ayudar a mejorar el estándar en la educación (Patil et al., 2024). Tales modelos pueden anticipar aún más 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 de Data Mining no es nada obvio, especialmente cuando se busca integrar varias dimensiones académicas, socioeconómicas y demográficas dentro de un modelo sólido y eficiente (Lalaleo-Analuisa et al., 2021).

Este estudio introduce un método cuantitativo y predictivo para predecir el rendimiento de los estudiantes de secundaria según el análisis exploratorio de datos (EDA) y el proceso ETL (Extracción, Transformación, Carga) y el modelado predictivo 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 ámbito del análisis educativo, la capacidad de manejar grandes cantidades de información de manera efectiva es una de las principales ventajas de incorporar el Data Mining (Jha et al., 2018).

Los datos extraídos de fuentes distintas a las calificaciones de los estudiantes, asistencia, etc., por ejemplo, utilizando datos de aprendizaje automático, podemos calcular otros factores que afectan el rendimiento académico de los estudiantes sobre la base de los estudiantes y predecir estos factores que los maestros no pueden predecir (Shylaja et al., 2023). Estos modelos no solo predicen el comportamiento y las tendencias de rendimiento basándose en patrones pasados, sino que también intervienen de manera 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).

Los modelos predictivos pueden predecir el logro académico de los estudiantes y estos factores que afectan el éxito escolar (Grabovy & Siniak, 2024). En este sentido, la aplicación del Data Mining en la educación no solo está personalizando la educación, sino también personalizando 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).

Metodología

Los fundamentos metodológicos se desarrollaron en el contexto de la investigación de data mining sobre el éxito académico. Utilizamos 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 de secundaria. La información se procesa utilizando la base de datos de Ineval, y la escuela ecuatoriana, el año de análisis de los datos es 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 y el estado socioeconómico y las características demográficas de las familias de los estudiantes. Es a lo largo de estos conjuntos de datos que se entrenaron modelos de clasificación para predecir si los estudiantes pertenecían a grupos de alto o bajo rendimiento.

Variables y Fuentes de Datos

Las fuentes de información utilizadas en este estudio son los registros académicos y una encuesta de condiciones socioeconómicas. Los siguientes parámetros del modelo fueron dominantes:

El logro académico de los estudiantes se evalúa promediando las calificaciones finales de los módulos. Este logro está influenciado por varios rasgos socioeconómicos, como el ingreso familiar, el nivel educativo de los padres y la vivienda, que afectan el acceso a la educación y las experiencias de aprendizaje (Guevara & Bonilla, 2021). Por el contrario, los factores relacionados con la escuela, la participación en clases y el éxito académico parecen ser predictores importantes del logro continuo de los estudiantes. Las características demográficas (edad y sexo) del niño influirán en el logro académico junto con el lugar de residencia (rural/urbano), que el contexto social puede variar y las instalaciones educativas pueden divergir considerablemente.

El procedimiento comenzó con la recopilación de datos de los estudiantes y fue seguido por la manipulación y normalización de los datos. Se eliminaron los registros con datos incompletos y para el análisis se estandarizaron los parámetros socioeconómicos y de distancia interpupilar. En segundo lugar, se construyeron modelos con los datos de los estudiantes empleando algoritmos de predicción (regresión lineal, árboles de decisión, Random Forest y Boosted Trees). Los datos se dividieron en conjuntos de entrenamiento y prueba. El entrenamiento y la prueba entrenaron el modelo con el conjunto de entrenamiento y el conjunto de prueba, respectivamente. El rendimiento de los modelos

se evaluó en términos de RMSE y porcentaje de estudiantes clasificados correctamente en bandas de rendimiento.

Resultados

Los modelos predictivos alcanzaron las siguientes precisiones en la clasificación de estudiantes según su rendimiento académico:

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

Las variables más influyentes en los modelos fueron el nivel socioeconómico de la familia, el rendimiento académico previo de los estudiantes y su asistencia a clases. Los estudiantes provenientes de familias con mayores ingresos y un nivel educativo más alto de los padres mostraron un mejor rendimiento académico. Asimismo, el rendimiento académico previo fue uno de los mejores predictores del desempeño futuro, lo que subraya la importancia de la continuidad educativa y la retroalimentación constante sobre el rendimiento de los estudiantes.

Importancia de las Variables

En la siguiente tabla se presentan las métricas de rendimiento obtenidas con los diferentes modelos predictivos:

Tabla 1: Métricas de rendimiento de los modelos predictivos.

Modelo	**RMSE (Error cuadrático medio)**	**R^2(Coeficiente determinación)**	**Precisión (%)**
Regresión Línea	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

Fuente: Elaboración propia

Los resultados obtenidos muestran que el **Random Forest** tuvo un rendimiento superior en términos de precisión (85%) y coeficiente de determinación (R^2), mientras que los modelos, ofreció una precisión de 80%, lo que aún indica un rendimiento adecuado para la predicción del rendimiento académico.

Modelo SEM: Relaciones Causales en el Rendimiento Académico

A continuación, se presenta un **Modelo SEM (Ecuaciones Estructurales)** que ilustra las relaciones causales entre las variables latentes que influyen en el rendimiento académico:

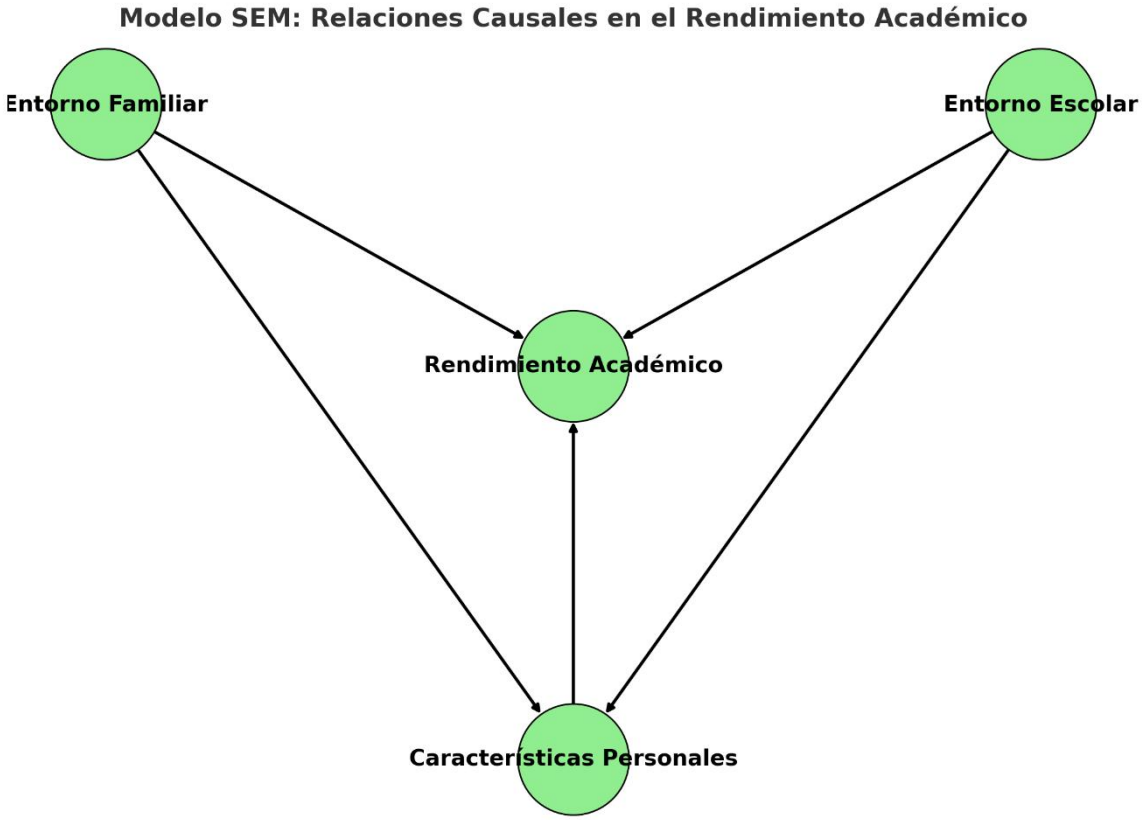


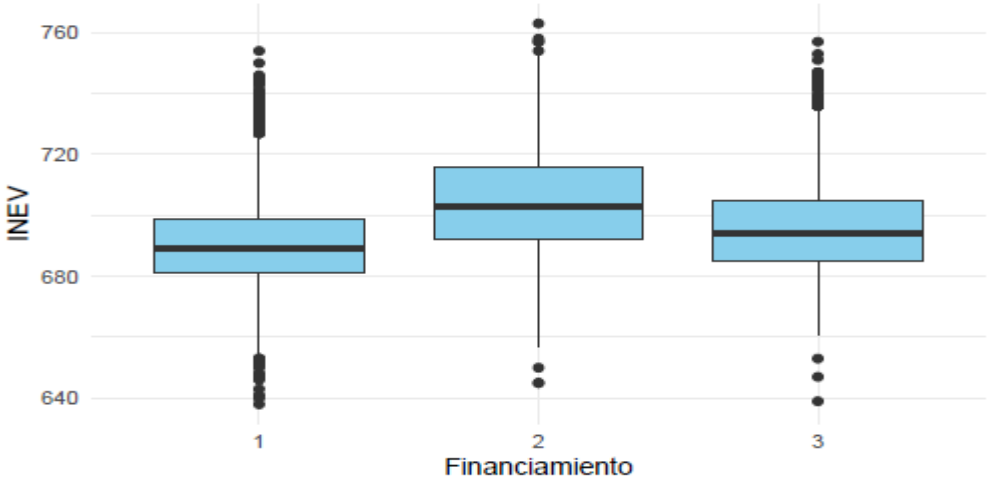
Fig 1. Modelo SEM

Este modelo describe cómo el entorno familiar, el entorno escolar y las 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

Las gráficas presentadas 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).

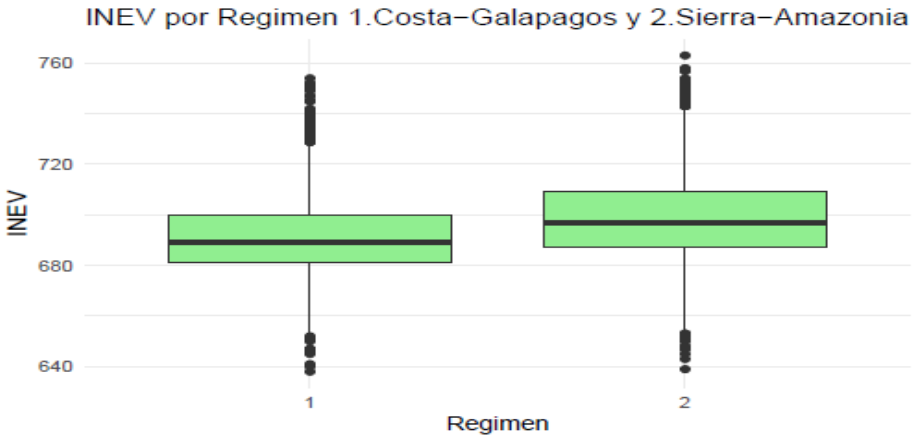
Gráfico 1. Financiamiento



Fuente: Elaboración propia

En La gráfica 1. muestra 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.

Gráfico 2. Distribución de INEV según Regimen de evaluación

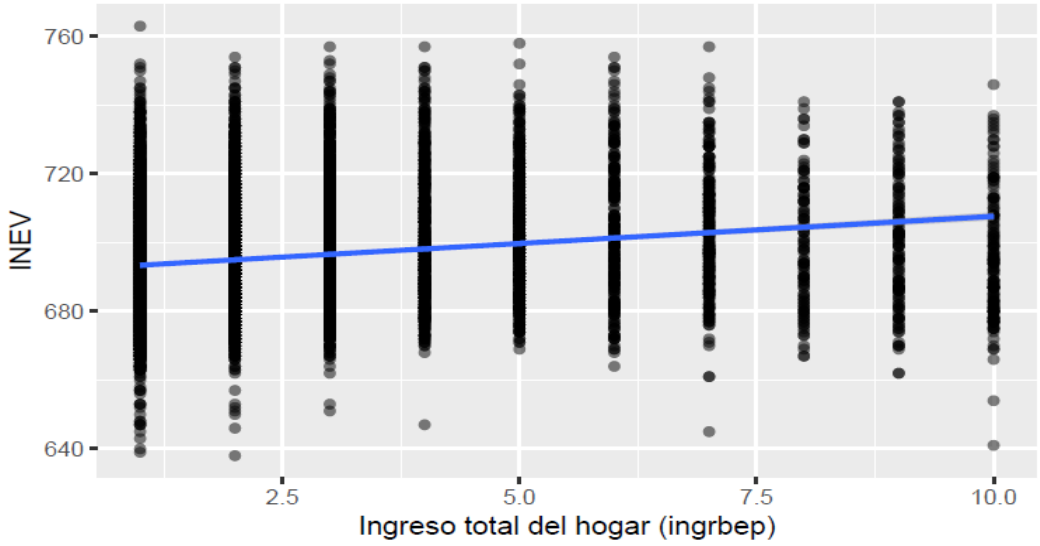


Fuente: Elaboración propia

La gráfica 2. compara el rendimiento académico (INEV) entre dos regiones de Ecuador: Costa-Galápagos (Regimen 1) y Sierra-Amazonía (Regimen 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.

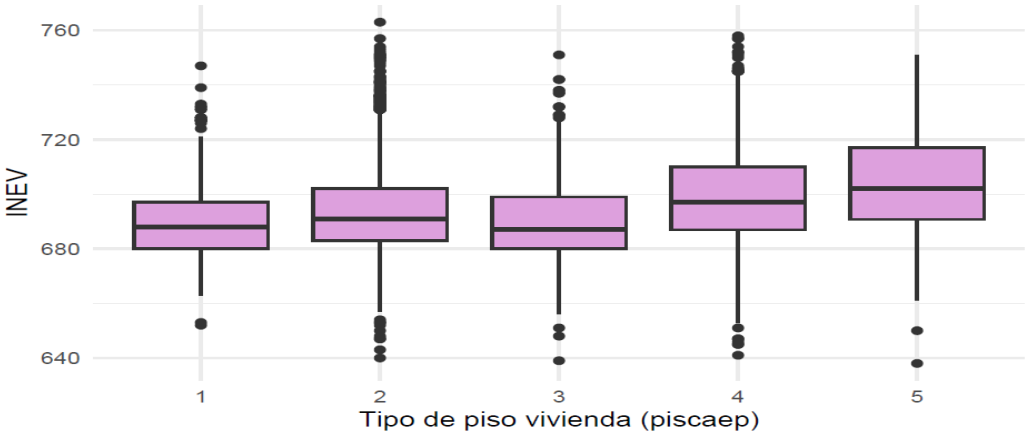
Gráfico 3. Inev vs Ingreso monetario



Fuente: Elaboración propia

La gráfica 3. presenta la relación entre el ingreso total del hogar (en la variable "ingrbep") y el rendimiento académico (INEV) de los estudiantes. Se observa una tendencia positiva, ya que, 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.

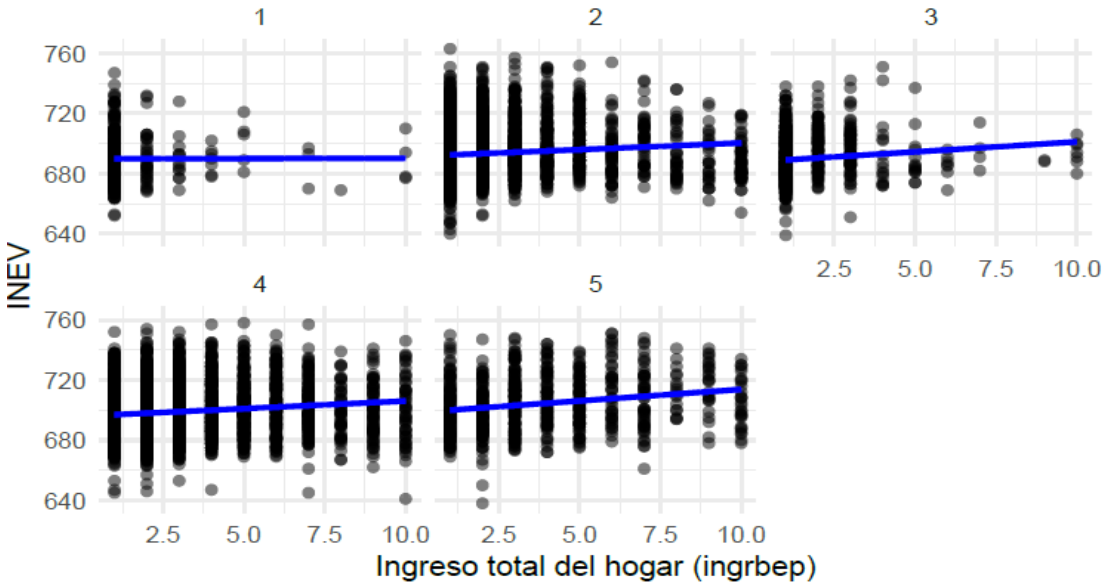
Gráfico 4. Distribución de INEV según tipo de piso



Fuente: Elaboración propia

El gráfico 4. representa la distribución del Índice Nacional de Evaluación Educativa (INEV) según el tipo de piso de vivienda, clasificado en cinco categorías (piscsaep). Cada caja muestra la mediana, el rango intercuartílico y los valores atípicos para cada tipo de piso. Se observa que, en general, 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.

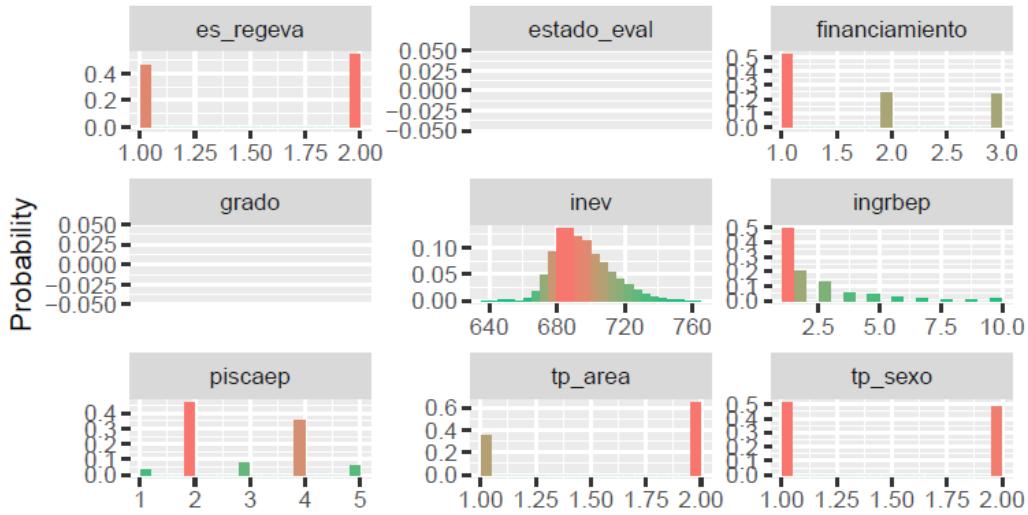
Gráfico 5. INEV vs Ingreso total del hogar, facetado por tipo de piso



Fuente: Elaboración propia

La gráfica 5. Enseña la relación entre el rendimiento académico (INEV) y el ingreso total del hogar, facetada por tipo de piso, permite observar 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.

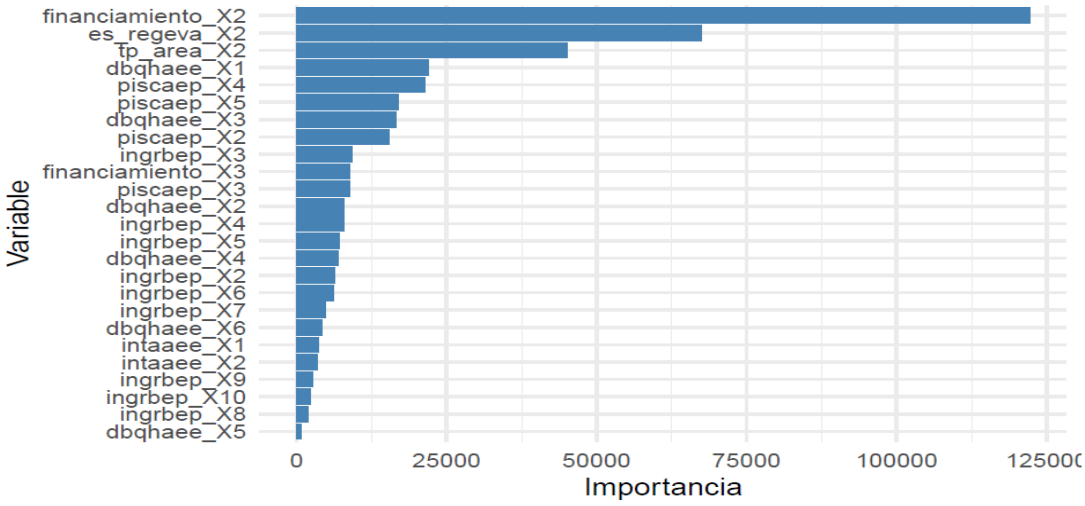
Gráfico 6. Histogramas de las columnas numéricas



Fuente: Elaboración propia

La gráfica 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.

Grafico 7. Modelo random forest

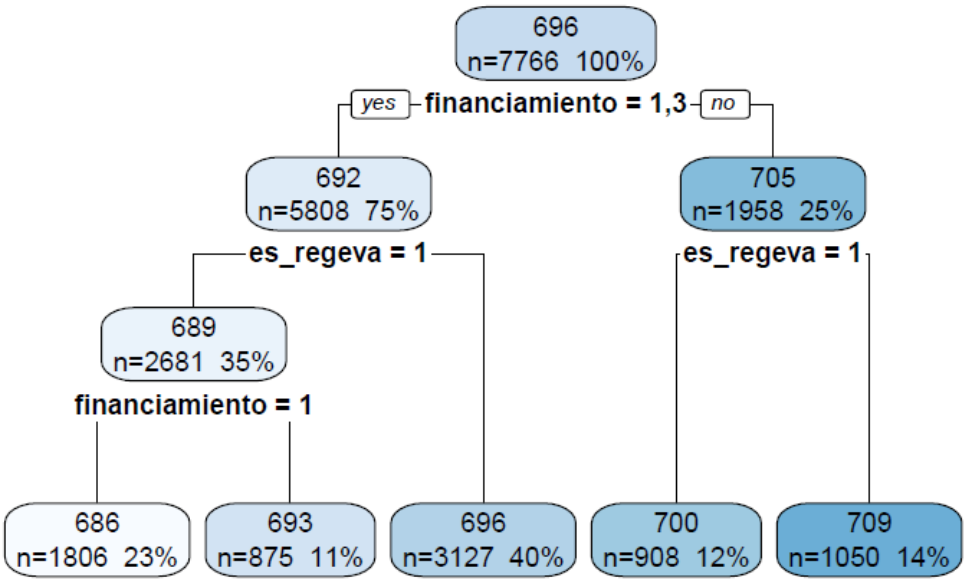


Fuente: Elaboración propia

La gráfica 7. muestra la importancia de las variables según un modelo de Random Forest. En el gráfico, 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



Fuente: Elaboración propia

El gráfico muestra un árbol de decisión utilizado 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, también 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, especialmente en términos de influir en los dominios socioeconómicos, familiares y escolares de los estudiantes (Guevara & Bonilla, 2021). Por el contrario, los diagramas de caja también mostraron que la subvención tiende a tener una relación llamada positiva con los factores académicos y relacionados, pero también hay algunas excepciones, lo que nos lleva al hecho de 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 de Zhang et al., (2024), que revela que cuando se controlan otros factores, cuando el ingreso familiar de un estudiante es más alto, el GPA del niño también lo es (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).

Finalmente, el rendimiento escolar de los estudiantes basado tanto en Random Forest como en Boost Tree se determinaron los atributos relevantes (Tran et al., 2025). Los modelos de aprendizaje automático son buenos para predecir el fracaso académico/deserción entre los estudiantes. Esto demuestra una especie de linealidad predecible de los grandes datos en las vidas de nuestros 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). Mejor aún, los 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 del espacio indicada en la Fig. 4, se observará que la información de la zona costa-Galápagos tiene una distribución más dispersa que la de la zona sierra-Amazonas, que es más concentrada, en comparación con la de los resultados restantes (Tin et al., 2024). Esto se llama desigualdad educativa. De hecho, esa es probablemente la razón por la cual, más allá de la propia casta o desigualdad regional, vivir en una región con un bajo nivel de desarrollo está asociado a un menor logro, dado que el grupo de casta más bajo y el lugar de residencia con el nivel de desarrollo más bajo no son necesariamente 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), el logro académico. 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 tan bien 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 que, con suerte,

significará 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 el EDM como un plan y como una estrategia para mejorar 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) también nos 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. Más bien, hemos abogado por una intervención que, si bien toma en cuenta las situaciones académicas, considera también 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. No importa el "coche usado". Lo que al final subraya la importancia de lo que las escuelas y las familias, y las comunidades locales están haciendo para hacer disponible la mejor escuela posible, no solo para los niños de familias de bajos ingresos, sino para los niños de familias de altos ingresos.

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, lo que añaden, sería la identificación basada en datos de quién está teniendo dificultades para que se puedan diseñar intervenciones para asignar recursos de manera eficiente y crear un sistema de aprendizaje más equitativo y justo. No es solo que las escuelas tengan la oportunidad de ser pioneras en modelos predictivos y data mining, sino que pueden 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*(1), 1–23. https://doi.org/10.1186/S41239-020-00223-0/FIGURES/8

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/S1

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*(1). 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