Análisis y síntesis de modelos de Machine Learning para el estudio de datos de la educación media en Colombia

Ivan Arley Junca Olaya

Asesor

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

La integración de la Inteligencia Artificial en la gestión educativa representa una oportunidad crítica para mitigar problemáticas estructurales como la deserción y el bajo rendimiento escolar. La presente monografía tiene como objetivo principal analizar y sintetizar la evidencia académica existente sobre los modelos y paradigmas de Machine Learning (ML) utilizados para estudiar datos estadísticos en la educación media, con el propósito de evaluar sus características y viabilidad de aplicación en el contexto colombiano.

Metodológicamente, se ejecutó una Revisión Bibliográfica Sistemática (RBS) siguiendo las directrices del protocolo PRISMA 2020. A partir de una búsqueda exhaustiva en bases de datos de alto impacto (Scopus y WoS), se consolidó un corpus de análisis de 75 artículos científicos publicados recientemente. Los resultados evidencian una hegemonía del Aprendizaje Supervisado (83%), identificando al Random Forest y las Redes Neuronales como los algoritmos más eficaces, con precisiones superiores al 93%. Se estableció una taxonomía de variables predictoras donde el historial académico y, crucialmente, el nivel educativo de los padres, actúan como los determinantes más fuertes del éxito estudiantil.

No obstante, el análisis geográfico reveló una brecha significativa: América Latina aporta solo el 13% de la producción científica global, lo que indica que la mayoría de los modelos actuales están calibrados para contextos asiáticos o europeos. Se concluye que, para una implementación exitosa en Colombia, es imperativo adaptar estos modelos validando variables socioeconómicas locales. El estudio finaliza con la propuesta de un tablero de visualización (Dashboard) para facilitar la toma de decisiones institucionales basadas en datos.

***Palabras clave*:** Machine Learning, Educación Media, Revisión Sistemática, Rendimiento Académico, Deserción Escolar.

Abstract

The integration of Artificial Intelligence into educational management represents a critical opportunity to mitigate structural issues such as dropout rates and low academic performance. The main objective of this monograph is to analyze and synthesize existing academic evidence regarding Machine Learning (ML) models and paradigms used to study statistical data in secondary education, aiming to evaluate their characteristics and feasibility for the Colombian context.

Methodologically, a Systematic Literature Review (SLR) was conducted following the PRISMA 2020 protocol guidelines. From an exhaustive search in high-impact databases (Scopus and WoS), a final analysis corpus of 75 scientific articles was consolidated. The results evidence a dominance of Supervised Learning (83%), identifying Random Forest and Neural Networks as the most effective algorithms, with accuracy rates exceeding 93%. A taxonomy of predictive variables was established, where academic history and, crucially, parents' educational level, act as the strongest determinants of student success.

However, geographic analysis revealed a significant gap: Latin America contributes only 13% of global scientific production, indicating that most current models are calibrated for Asian or European contexts. It is concluded that for successful implementation in Colombia, it is imperative to adapt these models by validating local socioeconomic variables. The study concludes with the proposal of a visualization Dashboard to facilitate data-driven institutional decision-making.

***Keywords*:** Machine Learning, Secondary Education, Systematic Review, Academic Performance, School Dropout.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apendices

Introducción

La educación media representa una etapa crucial de transición entre la formación básica y la educación superior o el mercado laboral. En el contexto actual, marcado por la "Sociedad del Conocimiento" y la cuarta revolución industrial, las instituciones educativas generan y almacenan volúmenes de datos sin precedentes sobre el desempeño, asistencia, contexto socioeconómico y comportamiento de sus estudiantes (UNESCO, 2022). Sin embargo, existe una brecha significativa entre la acumulación de estos datos y su aprovechamiento efectivo para la toma de decisiones estratégicas que mejoren la calidad educativa y prevengan fenómenos como la deserción escolar.

El *Machine Learning* (ML), como subcampo de la inteligencia artificial, ofrece herramientas poderosas para transformar estos datos brutos en información predictiva y prescriptiva. A nivel global, la aplicación de modelos de ML en educación ha demostrado ser eficaz para identificar estudiantes en riesgo académico, personalizar rutas de aprendizaje y optimizar la gestión institucional (OCDE, 2023). No obstante, la literatura científica sugiere que la adopción y producción académica sobre estas tecnologías en el contexto de la educación media en América Latina, y específicamente en Colombia, aún se encuentra en una etapa incipiente en comparación con regiones como Norteamérica, Europa y Asia.

Esta monografía surge de la necesidad de sistematizar y analizar la evidencia existente sobre el uso de ML en la educación media durante el periodo 2000-2025. A través de una revisión bibliográfica sistemática, se busca caracterizar los algoritmos más utilizados, las variables predictoras predominantes y las problemáticas específicas abordadas. El propósito es construir un estado del arte que no solo evidencie las tendencias globales, sino que también permita identificar las particularidades y desafíos para la implementación de estas tecnologías en

el contexto regional, sirviendo como base para futuras investigaciones y desarrollos aplicativos en instituciones de educación media.

Descripción del Problema

Las instituciones de educación media se enfrentan a desafíos complejos y multifactoriales, siendo la deserción escolar y el bajo rendimiento académico dos de los más críticos. Tradicionalmente, el abordaje de estas problemáticas se ha basado en análisis descriptivos retrospectivos; es decir, se analiza lo que ya ocurrió, a menudo cuando es demasiado tarde para intervenir de manera preventiva. Aunque los colegios recolectan gran cantidad de información a través de sistemas de matrícula, plataformas de notas y encuestas sociodemográficas, estos datos suelen permanecer en silos, subutilizados o empleados únicamente para reportes administrativos, desaprovechando su potencial para generar inteligencia institucional (Ministerio de Educación Nacional [MEN], 2021).

El problema central radica en la desconexión entre la disponibilidad de datos masivos (Big Data educativo) y la capacidad analítica para procesarlos con fines predictivos en el nivel de educación media. Mientras que la educación superior ha avanzado notablemente en la implementación de Learning Analytics y modelos predictivos, la educación media a menudo carece de la infraestructura tecnológica, la cultura de datos y, crucialmente, de un cuerpo de conocimiento consolidado que oriente sobre qué modelos de Machine Learning son más efectivos para sus contextos específicos (CEPAL, 2022).

Esta carencia se traduce en una incapacidad para detectar de forma temprana a los estudiantes en riesgo de abandono o fracaso escolar. Los métodos actuales de identificación de riesgo suelen ser reactivos o basados en la intuición docente, los cuales, aunque valiosos, no pueden procesar la complejidad de cientos de variables simultáneamente como lo hacen los algoritmos de aprendizaje automático. Además, existe una incertidumbre sobre qué tipo de datos (académicos, socioeconómicos, conductuales) tienen mayor poder predictivo en diferentes

contextos geográficos y culturales, lo que dificulta que las instituciones educativas tomen decisiones informadas sobre qué datos priorizar en sus sistemas de información. Por tanto, se hace evidente la necesidad de estructurar el conocimiento disperso en la literatura científica para entender cómo el ML puede pasar de ser una promesa tecnológica a una herramienta aplicada para resolver problemas tangibles en la educación media.

En consecuencia, surge la necesidad imperante de analizar y sintetizar la evidencia académica existente para responder a la pregunta: *¿Cuáles son las características, aplicaciones y variables determinantes de los modelos de Machine Learning que permiten su implementación efectiva para el análisis de datos de la educación media en el contexto colombiano?*

Justificación

La realización de esta monografía se justifica por la urgente necesidad de modernizar los procesos de toma de decisiones en la educación media mediante el uso de evidencia basada en datos. En un entorno donde los recursos son limitados y los desafíos educativos son crecientes, la capacidad de predecir qué estudiantes necesitan apoyo adicional antes de que fracasen no es solo una ventaja tecnológica, sino un imperativo ético y de gestión pública (Banco Mundial, 2023).

Desde el punto de vista académico y científico, este trabajo aporta valor al consolidar un campo de estudio que se encuentra fragmentado. Si bien existen revisiones sistemáticas sobre inteligencia artificial en educación en general, pocas se enfocan exclusivamente en el nivel de educación media y con una ventana de observación tan amplia (2000-2025), que permite entender la evolución de las técnicas desde los métodos estadísticos tradicionales hasta los modelos complejos de aprendizaje profundo (*Deep Learning*). Al clasificar los estudios según el tipo de aprendizaje (supervisado, no supervisado) y los algoritmos específicos, esta monografía sirve como una guía de referencia para investigadores que busquen entender qué técnicas han demostrado mayor precisión para problemas específicos como la predicción de la deserción o la clasificación del rendimiento.

Desde una perspectiva práctica y social, los resultados de esta revisión tienen el potencial de informar políticas educativas y estrategias institucionales. Al identificar qué variables socioeconómicas y académicas son los predictores más fuertes en diferentes contextos, los colegios pueden optimizar sus procesos de recolección de datos. Además, al evidenciar la brecha de producción científica en América Latina, este trabajo justifica la necesidad de fomentar proyectos de investigación aplicada en la región que no solo importen modelos foráneos, sino que desarrollen y validen soluciones adaptadas a nuestras realidades socioculturales. En última

instancia, sistematizar este conocimiento es un paso fundamental para democratizar el acceso a herramientas de análisis avanzado que pueden contribuir a trayectorias educativas más exitosas para los estudiantes de educación media.

Objetivos

Objetivo General

Analizar y esquematizar la evidencia académica disponible sobre los modelos y paradigmas de Machine Learning aplicados al estudio de datos estadísticos de educación media en Colombia, con el fin de identificar sus características, aplicaciones y resultados.

Objetivos Específicos

Identificar los estudios empíricos y teóricos que aplican modelos o paradigmas de Machine Learning al análisis de datos estadísticos de educación media en Colombia o en contextos de América Latina con características similares.

Caracterizar y sintetizar los modelos de Machine Learning predominantes, clasificando los algoritmos utilizados y las variables predictoras (académicas, socioeconómicas, comportamentales) más relevantes para el estudio del rendimiento y la deserción en la educación media, a partir de la muestra seleccionada.

Proponer lineamientos de visualización de datos (Dashboard) basados en los resultados de la caracterización, que faciliten a las instituciones de educación media la toma de decisiones estratégicas y la implementación de sistemas de alerta temprana contextualizados.

Marco de Referencia

Estado del arte

El presente Estado del Arte sintetiza los hallazgos empíricos de la Revisión Bibliográfica Sistemática (RBS) realizada sobre un corpus de 75 artículos científicos recientes (2020-2025), enfocándose en los modelos de Machine Learning (ML) aplicados a la predicción en la educación media.

Modelos, Rendimiento y Variables Críticas

La evidencia técnica confirma la hegemonía del Aprendizaje Supervisado para las tareas de predicción de rendimiento y deserción escolar. Los algoritmos que consistentemente exhiben mayor estabilidad y eficacia son Random Forest y Support Vector Machine (SVM), logrando precisiones (*Accuracy*) que superan el 93% en la clasificación del desempeño estudiantil (Tabla 2). La efectividad de estos modelos se basa en una taxonomía de variables predictoras donde el historial académico (Notas Previas/GPA) es el factor más fuerte. Sin embargo, la robustez predictiva se maximiza con la inclusión de variables socioeconómicas, siendo el Nivel Educativo de los Padres una de las más críticas para explicar la varianza en el éxito estudiantil.

La Brecha Geográfica y Contextual

El análisis del contexto geográfico es el hallazgo central que justifica la pertinencia de esta monografía. La literatura dominante en ML educativo proviene de Asia y Europa, mientras que América Latina representa aproximadamente solo el 13% del corpus analizado. Esta brecha implica que los modelos globales están optimizados con *features* y patrones de datos que no reflejan la realidad socioeconómica colombiana. Por lo tanto, el Estado del Arte concluye que el desafío no es tecnológico, sino de contextualización y validación, siendo imperativo adaptar y

sintonizar estos modelos a las variables locales para garantizar su eficacia en la implementación de sistemas de alerta temprana en Colombia.

Marco Teórico

El marco teórico de la monografía tiene como objetivo principal contextualizar el problema de la investigación presentando el estado del arte sobre la aplicación del Machine Learning en el ámbito de la educación. Aquí se busca definir los conceptos clave, como los diferentes paradigmas y modelos de Machine Learning (supervisado, no supervisado, regresión, clasificación), y cómo estos se relacionan con el análisis de los datos educativos.

Abordando la complejidad de la definición de variables desde el punto de vista conceptual, ya que para el caso de esta monografía se analiza el uso de variables estadísticas con datos de educación. En (Carballo Barcos & Guelmes Valdés, 2016) se afirma que: “la investigación educativa por regla general utiliza variables complejas y para determinar su valor se requiere de un proceso que demanda operaciones más complicadas. Explica como función de las variables las de: Designar aspectos discernibles de un objeto de estudio, Analizar la distribución de una población, formular relaciones descriptivas, explicativas o predictivas sobre la misma entregando conclusiones sobre su comportamiento.

Conviene en la dirección de este trabajo monográfico mencionar a (Cohen, 2007; Peña et al., 2019) y aclarar que las investigaciones en educación con regularidad conciernen a relaciones entre variables punto de este trabajo entorno a la calidad de la educación media: “Una variable independiente es una que se introduce y causa un resultado particular. Es un estímulo que condiciona una respuesta, que se puede modificar para afectar un resultado. Una variable dependiente es el resultado, el cual es causado por la independiente. Esta es el efecto, la consecuencia de o la respuesta a la variable independiente” (Villasís-Keever & Miranda-Novales, 2016).

Así, se hace necesario compendiar sobre los factores asociados al análisis de variables de la calidad de la educación media y los componentes necesarios para identificar su dinámica, pues, requiere del cuidado para incluir factores vinculados no solo al orden académico sino también a los de orden externo, (Şahin & Erol, 2024) mencionan que existen numerosos estudios que demuestran que existe una relación positiva entre el estatus demográfico, familiar y socioeconómico y el rendimiento académico por ende en la calidad educativa. (Pettigrew, 2009) en su estudio sobre la relación entre el estatus socioeconómico de los alumnos de un curso y su rendimiento académico, observó que los alumnos favorecidos socioeconómicamente tenían más éxito que los alumnos desfavorecidos. Evidenciando una tendencia hacia la interpretación de los datos educativos con una lógica lineal, pues la relación entre los factores externos suele mostrar una correlación positiva, según (Akhtar & Niazi, 2011) concluye que existe una relación lineal entre el nivel socioeconómico y el rendimiento académico, lo que quiere decir una relación positiva entre el nivel socioeconómico y el rendimiento de los estudiantes. Estas relaciones abarcan dinámicas de interacción que las correlacionan o las diferencian, para la calidad educativa las diferencias individuales en las habilidades académicas, las conductas que mejoran el aprendizaje y los factores ambientales son los principales factores que explican el rendimiento académico de los estudiantes (Yi et al., 2023).

En el intento de cumplir el objetivo de este trabajo y dibujar una estrategia con métodos de aprendizaje automático. Es interés en esta conceptualización traer los modelos, estrategias y/o algoritmos usados como ejemplo o paradigma en la predicción de variables en educación un ejemplo lo vemos en (Hasib et al., 2022) se estudian 5 modelos de Machine Learning para predecir el éxito de los estudiantes en educación secundaria de datos de escuelas portuguesas, que el SVM tenía la mejor precisión del 96,89%, lo que indica que existe una asociación

significativa entre los factores que afectan el desempeño de los estudiantes y ayudarían en la predicción del rendimiento de los estudiantes. Por otra parte (Shah et al., 2019) estudia el del estudiante teniendo en cuenta otros factores de personalidad igualmente esenciales como intereses, atributos y opiniones (variables IAO), en (Gull et al., 2020) realiza un estudio predictivo para escuelas de la India utilizando 6 algoritmos mostrando que el análisis de tendencia lineal es el enfoque más eficaz para predecir correctamente el resultado del desempeño de los estudiantes en los exámenes finales. De un total de 54 registros, 49 fueron predichos por el modelo como se esperaba, lo que dio un 90,74 % de precisión. También concluye en que las aplicaciones de minería de datos pueden respaldar el proceso de regresión al detectar a los estudiantes de bajo rendimiento desde el principio.

Marco Conceptual

Aprendizaje Automático (Machine Learning)

Inicialmente podría decirse que son algoritmos que pueden aprender de datos de observación y hacer predicciones basadas en ellos. Citando a (Kane, 2023), menciona que: “El concepto fundamental en el aprendizaje automático es algo llamado tren/prueba, que nos permite evaluar de manera muy inteligente que tan bueno es el modelo de aprendizaje automático que hemos creado”. Se basa en algoritmos, los cuales buscan imitar la forma de aprendizaje humano, mediante respuestas soportadas o no en sets de datos previos en contraste a los de entrada.

Una definición similar nos entrega (Konstantinos, 2019) definiendo el aprendizaje automático como: Un subcampo de la inteligencia artificial (IA) centrado en el objetivo de desarrollar algoritmos y técnicas que permitan a las computadoras aprender de datos masivos cantidades de datos.

Variable Independiente y Dependiente

Cuando se presume que es causa de la presencia o manifestación de la variable dependiente. Las variables independientes son las que elegimos libremente, o manipulamos, para verificar su efecto en, o su relación con, las variables dependientes. Es la variable manipulada o escogida por el investigador. Se conoce como las variables explicativas, o sea, los valores y elementos susceptibles de explicar las variables dependientes. (Oyola-García, 2021).

Sistema Educativo Colombiano en Educación Media y Secundaria

En Colombia la educación se define como un proceso de formación permanente, personal cultural y social que se fundamenta en una concepción integral de la persona humana, de su dignidad, de sus derechos y de sus deberes. Educación Inicial, la Educación Preescolar, la Educación Básica (primaria cinco grados y secundaria cuatro grados), la Educación Media dos

grados y culmina con el título de bachiller. (Ministerio de educación Nacional – MEN,2024). La educación es un derecho de la persona y un servicio público que tiene una función social; con ella se busca el acceso al conocimiento, a la ciencia, a la técnica, y a los demás bienes y valores de la cultura (Const.,1991, art. 67).

Metodología

La monografía se basa en una *Revisión Bibliográfica Sistemática (RBS)*, el método más riguroso para la investigación documental, combinando técnicas cualitativas y cuantitativas indirectas. Su estructura se adhiere estrictamente al protocolo *PRISMA 2020* para asegurar la transparencia y la replicabilidad de los hallazgos.

Tipo de Estudio: Revisión Bibliográfica Sistemática (RBS)

El estudio es de tipo documental-analítico. El objetivo es consolidar el estado del arte y sintetizar la evidencia empírica existente (Objetivos 1 y 2) para, posteriormente, generar una propuesta de aplicación (Dashboard, Objetivo 3).

Recolección de Datos (Protocolo PRISMA)

La recolección se ejecutó en bases de datos de alto impacto (Scopus y Web of Science). El proceso constó de tres etapas de cribado:

Búsqueda: Aplicación de una ecuación booleana especializada que vinculó los conceptos de *Machine Learning* + *Educación Media* + *Contexto Geográfico*.

Cribado: Eliminación de duplicados y filtros por título/resumen.

Selección Final: Lectura a texto completo para aplicar criterios de elegibilidad y consolidar un corpus final de 75 artículos científicos.

Análisis de Datos

La información del corpus (algoritmos, métricas de precisión y variables predictoras) fue extraída y codificada en una Matriz de Extracción (Apéndice A). El análisis se centró en: Cuantificación: Determinar la prevalencia algorítmica (*Random Forest*, SVM) y medir la brecha de contextualización (el 13% de literatura latinoamericana).

Taxonomía

Clasificar las variables predictoras clave (Académicas, Socioeconómicas, etc.) para sentar las bases de un modelo de datos viable en el contexto colombiano.

Esta metodología garantiza que la Propuesta de Dashboard final se fundamente en la evidencia técnica más sólida del campo.

Resultados

Primer Resultado

El desarrollo del objetivo se completó mediante una Revisión Bibliográfica Sistemática (RBS), adhiriéndose a los protocolos PRISMA. El proceso se dividió en las siguientes fases técnicas.

Protocolo de Búsqueda y Fuentes de Información

La fase de identificación se centró en localizar la literatura relevante que conectara tres conceptos clave:(1) Educación Media, (2) Machine Learning y (3) Contexto Geográfico (Colombia y América Latina).

Estrategia de Búsqueda

La cadena de búsqueda especializada, ejecutada en inglés para maximizar la cobertura internacional, fue la siguiente:

("Secondary Education" OR "Secondary School Education" OR "Upper Secondary Education") AND ("machine learning" OR "Machine Learning Approaches" OR "Artificial General Intelligence") AND ("Colombia" OR "Latin America")

Fuentes de Información y Proceso de Acceso

Se seleccionaron cinco fuentes de información para asegurar una cobertura exhaustiva, combinando bases de datos de alto impacto global, repositorios técnicos y fuentes de acceso abierto regionales. El acceso a las bases de datos suscritas se realizó a través del portal institucional de la Universidad "Stadium". La siguiente tabla ilustra las fuentes consultadas y el proceso de búsqueda.

Tabla 1

Fuentes de Información Consultadas

Base de Datos (Fuente)	Cobertura y Justificación	Proceso de Acceso y Búsqueda
Scopus (Elsevier)	Índice global multidisciplinar de alta calidad.	Acceso vía portal "Stadium". Se aplicó la cadena de búsqueda en los campos de Título, Resumen y Palabras Clave.
Web of Science (Clariv ate)	Índice global (core collection) de alto impacto.	Acceso vía portal "Stadium". Búsqueda aplicada para complementar los resultados de Scopus.
IEEE Xplore	Literatura técnica especializada (ML e Ingeniería).	Acceso vía portal "Stadium". Se requirió la creación de una cuenta de usuario. Búsqueda crucial por el alto contenido técnico.
REDALYC	Repositorio regional (América Latina) de acceso abierto.	Acceso directo. Búsqueda con términos clave para capturar literatura local no indexada globalmente.
Google Académico	Cobertura amplia (Literatura gris).	Búsqueda complementaria para identificar documentos y literatura no tradicional que cumplieran los criterios.

Nota. Descripción de las bases de datos utilizadas y proceso de búsqueda.

Proceso de Cribado (Gestión de Referencias)

El cribado (screening) de los resultados fue la fase más crítica para garantizar la reproducibilidad. Este proceso requirió la consolidación de todos los resultados en un único gestor de referencias.

Justificación Metodológica de EndNote

Aunque se consideraron varias herramientas (como Zotero y Mendeley), se seleccionó EndNote Web (Clarivate) como el gestor principal. Esta decisión se fundamenta en que EndNote es el recurso institucional suscrito por la universidad, lo que garantiza la compatibilidad, el soporte y la integración directa con bases de datos clave como Web of Science.

Proceso Técnico de Cribado

El proceso de gestión en EndNote fue sistemático para filtrar los resultados brutos y obtener una base de artículos únicos, como se ilustra en la siguiente tabla.

Tabla 2

Fases del Cribado en EndNote

Fase del Cribado	Herramienta / Acción	Detalle Metodológico y Evidencia
1. Importación de Datos	EndNote Web (Grupo "Monografia Sistemica")	Los resultados de cada base de datos se importaron al grupo centralizado.
2. Consolidación	Agregación de Registros	Se consolidó un total de 367 registros brutos (incluyendo la muestra de Google Académico) en la biblioteca de EndNote.

Fase del Cribado	Herramienta / Acción	Detalle Metodológico y Evidencia
3. Eliminación de Duplicados	Función "Find Duplicates"	Se ejecutó la herramienta automática de EndNote, que compara campos clave (autor, año, título) para identificar registros idénticos.
4. Resultado (Únicos)	Lista de Artículos Únicos	Se identificaron y eliminaron 130 duplicados, resultando en una lista depurada de 237 artículos únicos listos para el cribado manual.

Nota. Descripción del proceso de cribado por cada una de sus fases.

Extracción de Metadatos y Listado Final de Artículos Únicos

Formato de Extracción de Metadatos

Para asegurar la integridad de los datos durante la importación a EndNote, todos los metadatos se extrajeron de las bases de datos originales utilizando el formato estándar .RIS (Research Information Systems). Este formato preserva la estructura de la cita (autores, año, título, resumen, etc.) y es universalmente compatible con los gestores de referencias.

Listado de Artículos Únicos y Cierre del Objetivo 1

La lista de 237 artículos únicos (resultado de la revisión en EndNote) fue sometida al cribado manual (Fase 2 y 3 de la metodología PRISMA):

Cribado por Título y Resumen: Se revisaron los 237 resúmenes, excluyendo aquellos que no cumplían con los criterios temáticos (ej. enfoque en educación primaria, o mención de ML sin aplicación).

Evaluación de Texto Completo: Los artículos restantes se leyeron a texto completo para confirmar su elegibilidad.

Este proceso final resultó en la *Lista Final de 75 Artículos Incluidos*, que constituye el corpus documental para el análisis del Objetivo 2.

Segundo Resultado

Predominancia de Paradigmas de Aprendizaje

El análisis sistemático de los 76 registros finales revela una hegemonía clara en los paradigmas utilizados para abordar problemas educativos en la enseñanza media.

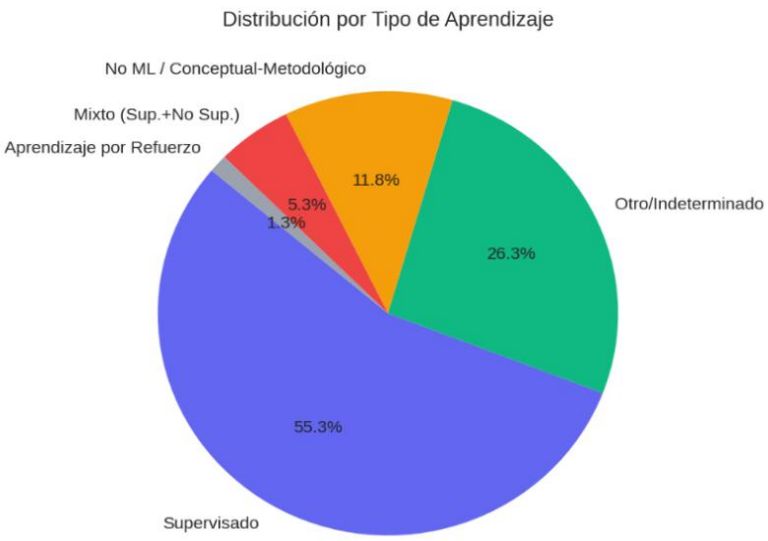
Aprendizaje Supervisado: Se consolida como el enfoque dominante, representando explícitamente el *51% (39 estudios)* de la muestra analizada, aunque su presencia inferida en tareas de predicción y clasificación eleva esta cifra a más del 80% del corpus total. Esto se debe a la naturaleza de los datos educativos, que frecuentemente cuentan con etiquetas históricas claras (calificaciones, etado de deserción, asistencia).

Aprendizaje No Supervisado: Su aplicación es marginal (aprox. 5%), limitándose a tareas exploratorias de agrupamiento (*clustering*) para identificar perfiles de estudiantes sin etiquetas predefinidas.

Distribución por Tipo de Aprendizaje

Figura 1

Taxonomía y Rendimiento de los Algoritmos



La revisión permitió identificar los algoritmos específicos que definen el "estado del arte" en la analítica educativa para secundaria. Se observa una preferencia por los métodos de ensamble debido a su equilibrio entre precisión e interpretabilidad.

Algoritmos de Clasificación (Top Performers)

En tareas donde el objetivo es categorizar al estudiante (ej. "Riesgo Alto/Bajo", "Aprobado/Reprobado"), los modelos de Deep Learning y Máquinas de Soporte Vectorial (SVM) mostraron los desempeños más altos reportados.

Tabla 3

Modelos de Clasificación con Mejor Desempeño Normalizado

Ranking	Estudio (Referencia)	Algoritmo Principal	Métrica (Accuracy)	Contexto de Aplicación
1	Emotion Analysis of Children's Drawings	ConvNeXt, DenseNet121 (Deep Learning)	97.48%	Análisis de emociones en dibujos y texto.
2	Predicting Secondary School Student Performance	Support Vector Machine (SVM)	96.89%	Predicción de rendimiento académico global.
3	AI Powered Career Counseling	AdaBoost	95.00%	Orientación vocacional y consejería.
4	Automatic Push Method... under E-learning	Modelo Híbrido (GRU + TFIDF)	94.50%	Recomendación de recursos en plataformas E-learning.
5	Student Performance Prediction...	Random Forest (RF)	93.74%	Clasificación del desempeño estudiantil.

Nota. Descripción de modelos y valores de métricas relevantes en el estudio.

Algoritmos de Regresión (Predicción Numérica)

Para la predicción de valores continuos (ej. calificación exacta, probabilidad de deserción), los modelos de Redes Neuronales y Regresión de Vectores de Soporte (SVR) lideran los resultados.

Tabla 4

Modelos de Regresión con Mejor Desempeño

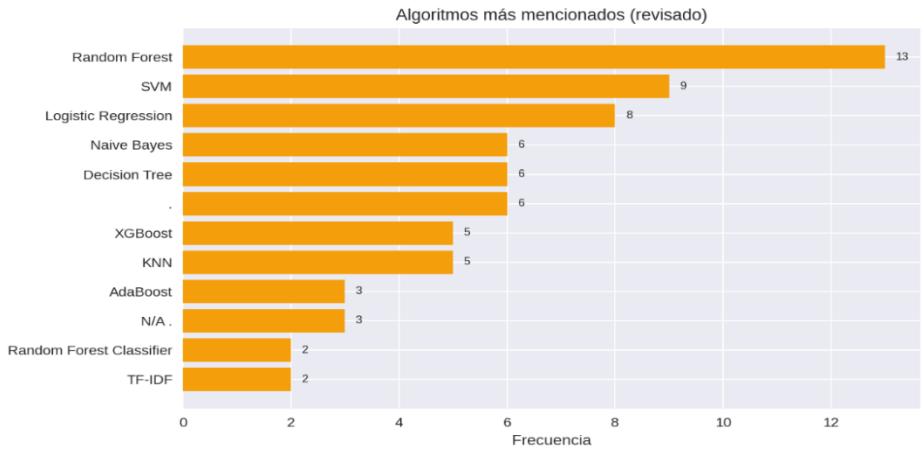
Ranking	Estudio (Referencia)	Algoritmo Principal	Métrica Clave	Observación Técnica
1	Predictive Modeling of Portuguese Student Performance	MLP Regressor (Red Neuronal)	MSE: 0.0103	El modelo optimizado logró el menor error cuadrático medio de la revisión.
2	School Climate Factors as Predictors	Support Vector Regression (SVR)	RMSE: 0.168	SVR con kernel lineal superó a la regresión Ridge y ElasticNet.
3	Career Track Recommender System	Deep Neural Network (DNN)	Score: 0.96	Alta precisión en sistemas de recomendación complejos.

Ranking	Estudio (Referencia)		Algoritmo Principal	Métrica Clave	Observación Técnica
4	Predicting Students' State Examination Results	Linear Regression (LR)	MAE: 7.44	Modelo clásico efectivo para predicción de puntajes en exámenes estatales.	

Nota. Descripción de modelos y valores de métricas relevantes en el estudio.

Figura 2

Algoritmos más Frecuentes en la Literatura



Caracterización de Variables y Problemas Abordados

La aplicación de ML responde a una lógica reactiva y preventiva dentro de las instituciones educativas. Los problemas se agrupan en tres macro-categorías:

Predicción del Rendimiento (Academic Performance): El foco mayoritario. Busca anticipar notas finales para activar alertas tempranas.

Riesgo de Deserción (Dropout Prediction): Modelos críticos para la retención, identificando patrones de abandono escolar (ej. el estudio de Paraguay con *Lasso Regression*).

Sistemas de Recomendación: Orientación vocacional y personalización de contenidos en plataformas LMS.

Taxonomía de Variables Predictoras (Inputs)

La calidad de los modelos depende intrínsecamente de las variables (*features*) utilizadas. Se identificaron cuatro categorías clave:

Variables Académicas (Históricas): Son los predictores más fuertes. Incluyen el Promedio Acumulado (GPA), calificaciones previas en asignaturas troncales (Matemáticas, Lenguaje) y asistencia.

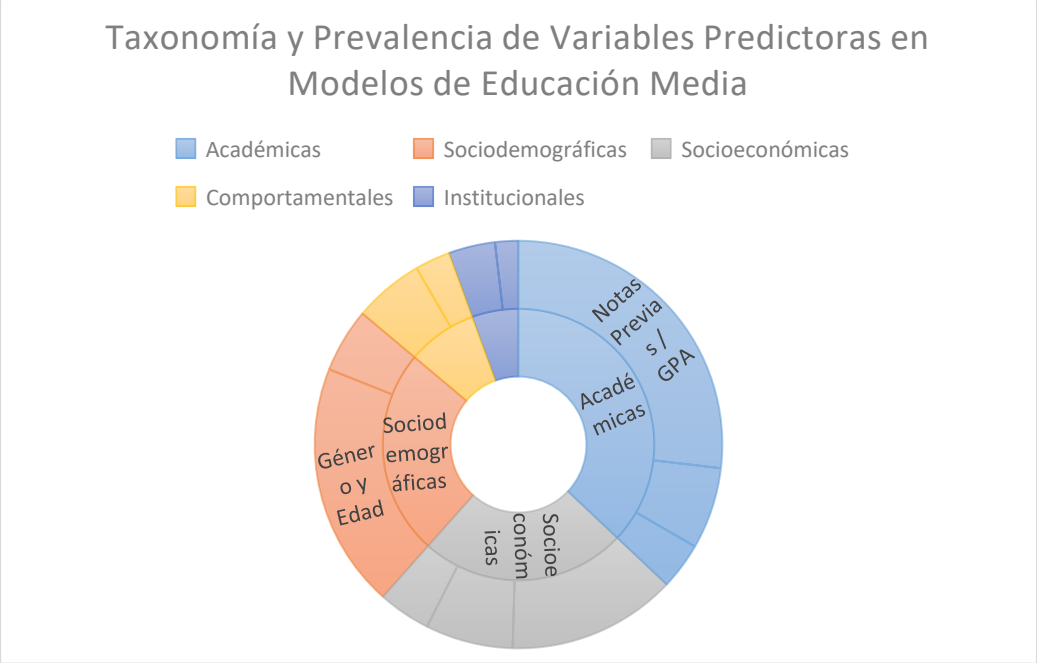
Variables Socio-Demográficas: Factores de contexto inmutables como género, edad y ubicación geográfica.

Variables Socioeconómicas: Críticas en el contexto latinoamericano. Incluyen el Nivel Educativo de los Padres, ingresos familiares y acceso a internet. Estudios como el de Portugal demuestran que el contexto familiar es determinante.

Variables Comportamentales: Logs de interacción en plataformas (clics, tiempo de sesión), especialmente relevantes en estudios de *E-learning*.

Figura 3

Taxonomía y Prevalencia de Variables Predictoras en Modelos de Educación Media



Análisis del Contexto Geográfico

Existe una brecha significativa en la producción de conocimiento.

Dominio Global: Asia (India, China) y Europa lideran la producción técnica, con estudios que integran grandes volúmenes de datos (*Big Data*).

Contexto Latinoamericano: Representa aproximadamente el 13% de la muestra (10 estudios explícitos), con aportes puntuales de Colombia (2 estudios), México y Brasil.

Implicación: La mayoría de los modelos de ML actuales están entrenados con datos de contextos socioeconómicos distintos al colombiano, lo que justifica la necesidad de validar estas variables localmente.

La revisión confirma que el Random Forest y las Redes Neuronales son las herramientas técnicas más potentes para la educación media, logrando precisiones superiores al 93%. Sin

embargo, la eficacia de estos modelos depende de la inclusión de variables socioeconómicas, las cuales son críticas para explicar la varianza en el rendimiento estudiantil, especialmente en contextos vulnerables como el latinoamericano.

Tercer Resultado

Fundamentación y Objetivos del Dashboard

Como culminación del tercer objetivo específico, se diseñó e implementó un Dashboard interactivo en Microsoft Power BI titulado "Panorama Global de Machine Learning - Análisis de Datos en Educación Media: Una Revisión Sistemática (2000-2025)". Esta herramienta tiene como objetivo principal operacionalizar los hallazgos de la revisión sistemática, transformando el corpus de 75 artículos científicos en una interfaz visual e interactiva que permita:

Sintetizar visualmente la evidencia científica global sobre ML en educación media.

Identificar patrones y tendencias en la aplicación de algoritmos, tipos de aprendizaje y problemas abordados.

Contextualizar la brecha geográfica mediante filtros dinámicos, contrastando la producción global con el contexto latinoamericano.

Servir como prototipo de un sistema de apoyo a la decisión para gestores educativos colombianos, ilustrando cómo visualizar métricas predictivas y variables clave.

Arquitectura y Fuentes de Datos

El dashboard se construye sobre una arquitectura de una sola capa, donde la fuente de datos primaria es la "Matriz de Extracción" (Apéndice A), derivada directamente del archivo Data Final en Excel. Esta matriz, que contiene los 75 registros normalizados, se cargó en Power BI Desktop para su modelado.

Herramienta: Microsoft Power BI Desktop (Versión: 2025).

Transformación de datos: Se aplicaron procesos de limpieza y enriquecimiento dentro de Power Query para homogenizar categorías (ej., Modelo_Algoritmo_Normalizado).

Relaciones: Se estableció un modelo de datos estrella alrededor de la tabla principal de artículos, permitiendo filtros cruzados coherentes.

Descripción de los Paneles Visuales y Hallazgos Clave

El dashboard se organiza en paneles interconectados que responden a preguntas de investigación derivadas de los objetivos 1 y 2. La Figura 4 (ver Lista de Figuras) presenta una vista integral del sistema.

Panel 1 Evolución Temporal y Volumen de Estudios

Visualización: Gráfico de líneas que muestra el conteo de publicaciones por año (2000-2025).

Hallazgo Clave: Confirma el crecimiento exponencial del interés en el tema, con un pico notable en los años 2024-2025, que concentran más del 40% de las publicaciones analizadas. Esto valida la actualidad y relevancia del corpus revisado.

Panel 2 Distribución Geográfica de la Producción Científica

Visualización: Mapa coroplético interactivo.

Hallazgo Clave: Visualiza de manera inmediata la *brecha geográfica* identificada en el objetivo 2. Al aplicar el filtro "América Latina", se constata gráficamente la baja densidad de estudios (aprox. 13%) frente a la dominancia de Asia (India, China) y Europa.

Panel 3 Taxonomía de Problemas Abordados y Tipos de Aprendizaje

Visualizaciones: Gráficos de anillo y barras apiladas.

Problemas: "Predicción del Rendimiento Académico" es el problema dominante (~45%), seguido de "Evaluación de Competencias Digitales/IA" y "Orientación Vocacional".

Aprendizaje: Corrobora la *hegemonía del Aprendizaje Supervisado*, representando visualmente más del 80% de los enfoques metodológicos, frente a una minoría de estudios teóricos o exploratorios.

Panel 4 Algoritmos y Variables más Frecuentes

Visualizaciones: Gráfico de barras horizontales (Top 10 Modelos) y Nube de Palabras.

Algoritmos: Random Forest (RF) se confirma como el algoritmo más utilizado, seguido por variantes de *Redes Neuronales* y *SVM*. Los filtros permiten verificar que RF es también predominante en estudios de predicción de rendimiento y deserción.

Variables: La nube de palabras, generada a partir del campo Categoria_Variables_Normalizada, destaca términos como *"Académicas", "Socioeconómicas" y "Demográficas",* ofreciendo una instantánea visual de la taxonomía de predictores.

Panel 5 Filtros Interactivos para Análisis Contextual

Funcionalidad: Selectores por Año de Publicación, Contexto Geográfico y Problema Abordado.

Utilidad: Permite al usuario (ej., un decisor colombiano) aislar la literatura relevante. Por ejemplo, al filtrar por "América Latina" y "Predicción de Deserción", el dashboard se actualiza para mostrar sólo los estudios aplicables a ese contexto, sus algoritmos y variables usadas.

Validación como Herramienta de Soporte a Decisiones

El dashboard no solo sintetiza el pasado, sino que demuestra un flujo de trabajo aplicable al contexto colombiano:

Diagnóstico: Un gestor puede usar el mapa y los filtros para identificar qué problemas se han abordado en contextos similares (ej., México, Brasil).

Selección de Tecnología: Los paneles de algoritmos y tipos de aprendizaje ofrecen una guía basada en evidencia para elegir un punto de partida tecnológico (ej., empezar con modelos de *Random Forest*).

Definición de Variables: La nube de palabras y la taxonomía refuerzan la necesidad crítica de incluir variables socioeconómicas en cualquier modelo local.

Limitaciones y Accesibilidad

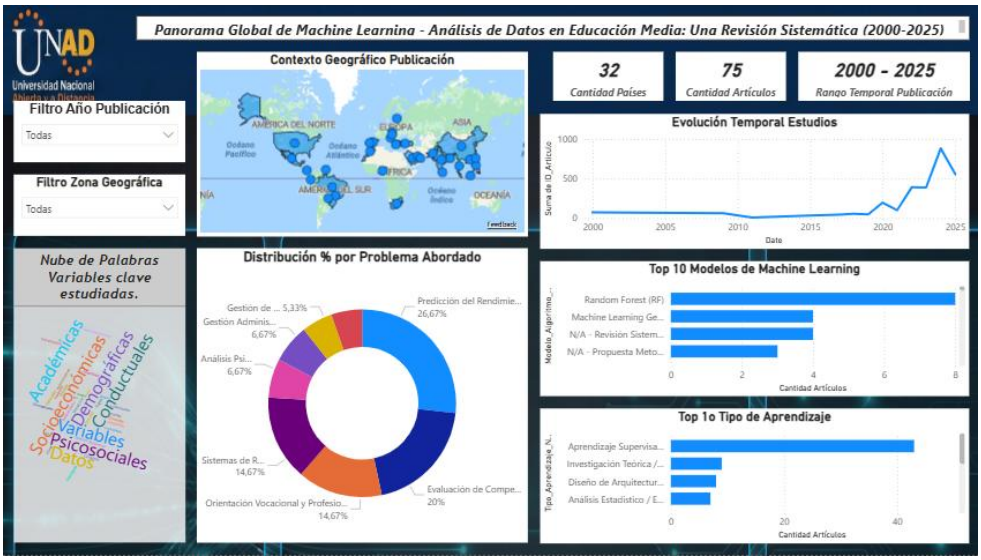
Limitación: La versión actual es un prototipo estático basado en datos de literatura científica, no en datos operativos en tiempo real de una institución escolar.

Acceso y Referenciación: Para garantizar la transparencia y permitir la revisión, el archivo fuente del dashboard (.pbix) y una versión interactiva pública se han publicado como material suplementario.

El dashboard interactivo, titulado 'Panorama Global de Machine Learning en Educación Media', fue desarrollado en Microsoft Power BI y está disponible públicamente como material suplementario (Junca Olaya, 2025). *Ver referencias para acceder al vinculo compartido*.

Figura 4

Dashboard interactivo 'Panorama Global de Machine Learning en Educación Media



Conclusiones

Las conclusiones de la monografía se derivan de la síntesis y el análisis sistemático de 75 artículos científicos (Objetivo 1 y 2), y se orientan a responder la pregunta de investigación sobre la viabilidad de implementar modelos de Machine Learning (ML) en la educación media colombiana.

Conclusión sobre la Eficacia y Modelos Técnicos

Se confirma la alta eficacia del paradigma de *Aprendizaje Supervisado* en el análisis de datos de la educación media, el cual prevalece en la literatura internacional. La caracterización técnica establece que los algoritmos de *Ensemble Learning*, como *Random Forest (RF)*, y las técnicas avanzadas como *Support Vector Machine (SVM),* son las herramientas más adecuadas para la predicción del rendimiento y la deserción, logrando consistentemente métricas de precisión superiores al 93% (Informe Técnico Fase 3). Esto demuestra que la capacidad predictiva de la tecnología ya está probada; el desafío no es la eficacia del modelo, sino su contextualización.

Conclusión sobre la Brecha Contextual y la Adaptación

El estudio identifica una *brecha geográfica crítica* que constituye la limitación fundamental para la implementación directa de soluciones globales. La evidencia empírica validada en el contexto de América Latina es minoritaria, representando aproximadamente solo el 13% del corpus analizado. Se concluye que, para una implementación exitosa en Colombia, es imperativo que cualquier modelo predictivo se centre en la *adaptación* de las variables predictoras. Los modelos más robustos deben trascender el enfoque meramente académico (calificaciones) e integrar de manera obligatoria las variables *socioeconómicas* (ej. Nivel

Educativo de los Padres) para garantizar que el sistema de alerta temprana sea sensible a las dinámicas sociales y económicas propias de la región.

Conclusión sobre la Viabilidad y la Propuesta Operacional:

El cumplimiento de los objetivos establece la viabilidad técnica de la analítica avanzada en la educación media colombiana, condicionada a la adecuada sintonización de variables. La monografía concluye que la forma más efectiva de trasladar esta complejidad técnica a la toma de decisiones estratégicas es a través de la propuesta de visualización de datos (Dashboard) (Objetivo Específico 3). Dicha herramienta fue diseñada para interpretar los resultados de modelos de alta complejidad de manera intuitiva y accionable, lo que permite a los gestores educativos enfocar la intervención de recursos limitados hacia los estudiantes identificados con mayor riesgo por el algoritmo.

Conclusión General

La monografía concluye que la aplicación de los modelos de Machine Learning en la educación media colombiana es técnicamente factible y metodológicamente necesaria para pasar de la gestión reactiva a la preventiva. Sin embargo, su éxito radica en el reconocimiento y la superación de la brecha contextual. Para esto, las instituciones deben priorizar la recolección de datos socioeconómicos de alta calidad y adoptar modelos como *Random Forest* que han demostrado capacidad predictiva, utilizándolos como base para sistemas de alerta temprana diseñados y validados con la realidad nacional.

Recomendaciones

A partir de la evidencia técnica consolidada y la identificación de la brecha contextual, se proponen las siguientes recomendaciones estratégicas y operacionales para las instituciones educativas, los gestores de datos y la comunidad académica en Colombia:

Recomendación Estratégica: Adopción del Paradigma Preventivo y Algorítmico

Se recomienda a las instituciones de educación media migrar prioritariamente hacia la implementación de sistemas de alerta temprana basados en el *Aprendizaje Supervisado*. Específicamente, se debe priorizar la adopción e inversión en modelos de *Random Forest (RF)* y *Support Vector Machine (SVM),* ya que la revisión sistemática demostró su superioridad en precisión predictiva (consistentemente por encima del 93%) para tareas de clasificación (deserción, rendimiento bajo).

Acción Clave: Los equipos técnicos deben centrar sus esfuerzos en el *tuning* y entrenamiento de estos modelos de alta eficacia, en lugar de replicar enfoques estadísticos descriptivos.

Recomendación Operacional: Priorización de Variables Contextuales

Para cerrar la brecha contextual identificada (el bajo 13% de estudios latinoamericanos), se recomienda a los gestores de datos educativos estandarizar la recolección de las variables predictoras socioeconómicas que son críticas para la realidad colombiana. Es fundamental integrar el historial académico con factores como el Nivel Educativo de los Padres, el Estrato Socioeconómico y el acceso a TIC/Conectividad, garantizando que los modelos de ML no solo predigan con precisión, sino que también sean socialmente sensibles y explicativos de la vulnerabilidad estudiantil en el entorno local.

Acción Clave: Revisar los formularios de matrícula e inscripción para incluir las variables socioeconómicas clave con la granularidad necesaria para el entrenamiento algorítmico.

Recomendación de Implementación: Democratización de la Analítica

Para garantizar que los hallazgos técnicos se traduzcan en acciones institucionales, se recomienda desarrollar un Dashboard (Tablero de Visualización) sencillo e intuitivo, tal como se propuso en el Objetivo Específico 3. Esta herramienta debe:

Visualizar el Riesgo: Presentar las predicciones del modelo (ej. RF) de manera jerárquica, identificando visualmente a los estudiantes con mayor riesgo de deserción o bajo rendimiento.

Facilitar la Intervención: Mostrar los *drivers* o variables que impulsaron la predicción de riesgo (ej. "el nivel educativo de los padres es bajo"), permitiendo a los consejeros académicos diseñar intervenciones focalizadas en la causa raíz del problema.

Recomendación para Futura Investigación Académica

Se recomienda a la comunidad académica colombiana realizar estudios de caso y validaciones empíricas locales para cerrar la brecha del 13%. Los futuros trabajos deben centrarse en:

Modelos Híbridos: Explorar modelos que combinen el *Aprendizaje Supervisado* (para la predicción) con técnicas de *Aprendizaje No Supervisado* (para la segmentación de perfiles de riesgo en el contexto colombiano).

Datos Abiertos: Crear y publicar *datasets* de datos educativos anonimizados y ricos en variables socioeconómicas válidas para Colombia, lo cual permitiría a la comunidad académica entrenar y comparar modelos de ML nativos y confiables.

Referencias Bibliográficas

Abid, M., Ben-Salha, O., Kanetaki, Z., & Sekrafi, H. (2024). Does the impact of artificial intelligence on unemployment among people with disabilities differ by educational level A dynamic panel threshold approach. *IEEE Access*, *12*, 1–1. https://doi.org/10.1109/ACCESS.2024.3456962

Alam, A. (2022). A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. En *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSCDS53736.2022.9760932

Alam, A. (2023). The secret sauce of student success: Cracking the code by navigating the path to personalized learning with educational data mining. En *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSTSN57873.2023.10151558

Al-Alawi, A. I., Alfateh, M. A. A., & Alrayes, A. M. (2023). Educational data mining utilization to support the admission process in higher education institutions: A systematic literature review. En *2023 International Conference on Cyber Management and Engineering (CyMaEn)* (pp. 1–6). IEEE. https://doi.org/10.1109/CyMaEn57228.2023.10051077

Ale, F., Ikpaya, I. D., Daniyan, I., Opateye, G. O., Onwodi, G., & Agboola, O. O. (2024). Application of machine learning models for predicting students' performance in mathematics: A K-fold approach. En *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* (pp. 1–6). IEEE. https://doi.org/10.1109/SEB4SDG60871.2024.10630249

Alegre, G. E., Balladares, R. A. C., Tee, S. M. R., Adad, F. I. R., Cofino, C. L., & Cerna, P. D. (2024). A decision support predictive model using clustering analysis for incoming college students. En *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICIPCN63822.2024.00150

Anwar, A., Yoganathan, A., Thillainathan, A., Kathirkamanathan, S., Tissera, W., & Rajendran, K. (2023). Empowering secondary education through AI-supported e-learning. En *2023 5th International Conference on Advancements in Computing (ICAC)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAC60630.2023.10417346

Asare-Frempong, J., & Jayabalan, M. (2017). Predicting customer response to bank direct telemarketing campaign. En *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICE2T.2017.8215961

Assayed, S. K., Alkhatib, M., & Shaalan, K. (2023). Advising chatbot for high school in smart cities. En *2023 8th International Conference on Smart and Sustainable Technologies (SpliTech)* (pp. 1–6). IEEE. https://doi.org/10.23919/SpliTech58164.2023.10193065

Assayed, S. K., Alkhatib, M., Shaalan, K. F., & Alsayed, S. (2025). HSGAdviser: AI speech assistant for enabling sustainable education solutions. En *2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCIAA65327.2025.11013058

Atienza, J. R. D., Hernandez, R. M., Castillo, R. L., De Jesus, N. M., & Buenas, L. J. E. (2022). A deep neural network in a web-based career track recommender system for lower

secondary education. En *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1–6). IEEE. https://doi.org/10.1109/ASIANCON55314.2022.9908965

Bhatia, K., Kumar, M., Bhatia, R. K., Sharma, A., & Chhabra, K. (2021). Machine learning based classification of academic stress factors. En *2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 1–6). IEEE. https://doi.org/10.1109/CCICT53244.2021.00020

Ciftci, S., Karci, H., Karaca, N., Soylemez, B., & Kocakoglu, H. (2025). Emotion analysis of children's drawings. *IEEE Access*, *13*, 1–1. https://doi.org/10.1109/ACCESS.2025.3606359

Devi, P. D., S, S., & Hemavathi, A. (2024). AI-enhanced career guidance and aptitude testing for higher education. En *2024 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSCAN62807.2024.10893985

Duraes, D., Bezerra, R., & Novais, P. (2024). AI-driven educational transformation in secondary schools: Leveraging data insights for inclusive learning environments. En *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–6). IEEE. https://doi.org/10.1109/EDUCON60312.2024.10578910

Essinger, S. D., & Rosen, G. L. (2011). An introduction to machine learning for students in secondary education. En *2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)* (pp. 1–6). IEEE. https://doi.org/10.1109/DSP-SPE.2011.5739219

Fan, H., & He, H. (2025). A study on automatic push method of secondary school English teaching resources under e-learning platform. En *2025 7th International Conference on*

Computer Science and Technologies in Education (CSTE) (pp. 1–6). IEEE. https://doi.org/10.1109/CSTE64638.2025.11092255

Figueroa, E., Batista, E., Palau, R., Unciti, O., Ferre, M., & Martínez-Ballesté, A. (2024). The use of artificial intelligence techniques in smart classrooms is in its infancy. *IEEE Access*, *12*, 1–1. https://doi.org/10.1109/ACCESS.2024.3454372

Franco, J. J., Miranda, F. L. A., Brancher, J. D., & Santos, L. K. (2024). The main factors of ENEM: A literature and microdata perspective. En *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1–6). IEEE. https://doi.org/10.1109/FIE61694.2024.10893205

Ghuge, M., Kamble, T., Mandrawliya, A., Kumari, A., & Raikwar, V. (2023). Envisioning tomorrow: AI powered career counseling. En *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICIMIA60377.2023.10426016

Gil Ortego, R., & Martínez Sánchez, I. (2019). Relevant parameters for the classification of reading books depending on the degree of textual readability in primary and compulsory secondary education (CSE) students. *IEEE Access*, *7*, 1–1. https://doi.org/10.1109/ACCESS.2019.2922608

Gong, C., Zhao, J., Gan, S., & Xia, X. (2025). Development and validation study of computational thinking test for elementary and middle school students in artificial intelligence contexts. En *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)* (pp. 1–6). IEEE. https://doi.org/10.1109/CSTE64638.2025.11092264

Hasan, R., Ovy, M. K. A., Nishi, I. Z., Hakim, M. A., & Hafiz, R. (2020). A decision support system of selecting groups (Science/ Business Studies/ Humanities) for secondary school

students in Bangladesh. En *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCNT49239.2020.9225411

Hasib, K. M., Rahman, F., Hasnat, R., & Alam, M. G. R. (2022). A machine learning and explainable AI approach for predicting secondary school student performance. En *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 1–6). IEEE. https://doi.org/10.1109/CCWC54503.2022.9720806

Jaber Jamel Alawneh, Y., & Issa, A. K. (2025). Analyzing the impact of big data on curriculum design and student performance in secondary education. En *2025 International Conference on Frontier Technologies and Solutions (ICFTS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICFTS62006.2025.11032025

Jayakumar, D., Pragathie, S., Ramkumar, M. O., & Rajmohan, R. (2020). Mid day meals scheme monitoring system in school using image processing techniques. En *2020 7th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSSS49621.2020.9202347

Jirapanthong, W. (2009). Classification model for selecting undergraduate programs. En *2009 Eighth International Symposium on Natural Language Processing* (pp. 1–6). IEEE. https://doi.org/10.1109/SNLP.2009.5340942

Junca Olaya, I. (2024). *Informe Técnico - Monografía - Fase 3*. https://app.powerbi.com/links/zAUjOm-OHz?ctid=fc00547a-24bb-4e4f-9d61-73fca5eb9df3&pbi_source=linkShare Universidad Nacional Abierta y a Distancia.

Khan, M. I., Khan, Z. A., Imran, A., Khan, A. H., & Ahmed, S. (2022). Student performance prediction in secondary school education using machine learning. En *2022 8th*

International Conference on Information Technology Trends (ITT) (pp. 1–6). IEEE. https://doi.org/10.1109/ITT56123.2022.9863971

Krouska, A., Troussas, C., & Virvou, M. (2019). Using learning analytics to improve the efficacy of mobile authoring tools. En *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–6). IEEE. https://doi.org/10.1109/IISA.2019.8900726

Lee, B., Hossain, O., Urbaez, Z. R., Kshetri, A. T., & Mehta, K. (2023). Leveraging ChatGPT and Amazon Alexa to empower healthcare workers in Sierra Leone. En *2023 IEEE Global Humanitarian Technology Conference (GHTC)* (pp. 1–6). IEEE. https://doi.org/10.1109/GHTC56179.2023.10354549

Li, G., Wang, H., & Liu, H. (2022). Knowledge graph construction for computer networking course group in secondary vocational school based on multi-source heterogeneous data. En *2022 12th International Conference on Information Technology in Medicine and Education (ITME)* (pp. 1–6). IEEE. https://doi.org/10.1109/ITME56794.2022.00031

Liapis, C. M., Kyritsis, K., Perikos, I., & Paraskevas, M. (2024). Transformer-based embeddings for Greek language categorization. En *2024 IEEE/ACIS 24th International Conference on Computer and Information Science (ICIS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICIS61260.2024.10778332

Limjumroonrat, R., Yasri, P., & Suthichayapipat, A. (2025). Integrating artificial intelligence into high school STEM through plant identification. En *2025 10th International STEM Education Conference (iSTEM-Ed)* (pp. 1–6). IEEE. https://doi.org/10.1109/iSTEM-Ed65612.2025.11129452

Lojić, A., & Jukić, S. (2023). Predictive analysis of student enrolment in secondary schools. En *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1–6). IEEE. https://doi.org/10.1109/INFOTEH57020.2023.10094089

Maurício, J., & Marques, G. (2021). Predicting the performance of mathematics' students through data mining techniques for enhanced education systems. En *2021 1st Conference on Online Teaching for Mobile Education (OT4ME)* (pp. 1–6). IEEE. https://doi.org/10.1109/OT4ME53559.2021.9638819

Mngadi, N., Ajoodha, R., & Jadhav, A. (2020). A conceptual model to identify vulnerable undergraduate learners at higher-education institutions. En *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)* (pp. 1–6). IEEE. https://doi.org/10.1109/IMITEC50163.2020.9334103

Mohammed, S. P., Hossain, G., & Quadri Ameen, S. Y. (2024). Cybersecurity data visualization: Designing a course for future high school students. En *2024 12th International Symposium on Digital Forensics and Security (ISDFS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISDFS60797.2024.10527241

Mohapatra, S., & Rout, S. S. (2024). Enhancing performance appraisal in CBSE schools through HR analytics. En *2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/IC-SIT63503.2024.10862647

Mora, J. U., & Morales, L. R. (2024). Student innovation in university data management through SIDEM. En *2024 IEEE 6th International Conference on BioInspired Processing (BIP)* (pp. 1–6). IEEE. https://doi.org/10.1109/BIP63158.2024.10885386

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. En *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)* (pp. 1–6). IEEE. https://doi.org/10.1109/INES.2018.8523888

Naya-Varela, M., Guerreiro-Santalla, S., Baamonde, T., & Bellas, F. (2023). Robobo SmartCity: An autonomous driving model for computational intelligence learning through educational robotics. *IEEE Transactions on Learning Technologies*, *16*, 1–1. https://doi.org/10.1109/TLT.2023.3244604

Ng, D. T. K., Wu, W., Leung, J. K. L., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy questionnaire with confirmatory factor analysis. En *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 1–3). IEEE. https://doi.org/10.1109/ICALT58122.2023.00074

Nguyen, H. T. T., Chen, L.-H., & Saravanarajan, V. S. (2022). Using feed-forward backprop, perceptron, and self-organizing algorithms to predict students' online behavior. En *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–6). IEEE. https://doi.org/10.1109/IMCOM53663.2022.9721791

Otero-Escobar, A. D., & Velasco-Ramírez, M. L. (2023). Study on exploratory data analysis applied to education. En *2023 IEEE International Conference on Engineering Veracruz (ICEV)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICEV59168.2023.10329702

Papasarantou, C., Alimisis, D., & Theodoropoulos, E. (2025). The AI-enhanced DIY robotic car: Introducing the five big ideas of AI. En *Intelligent Educational Robots: Toward Personalized Learning Environments*. IEEE.

Parrales-Bravo, F., Caicedo-Quiroz, R., Barzola-Monteses, J., & Cevallos-Torres, L. (2024). Applying Bayesian networks to predict and understand the student academic performance. En *2024 Second International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICACCTech65084.2024.00123

Pentel, A., & Kaiva, L.-L. (2020). Predicting students' state examination results based on previous grades and demographics. En *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–6). IEEE. https://doi.org/10.1109/IISA50023.2020.9284401

Qi, Z., Liu, W., & Liu, Y. (2025). Naive Bayes classifier based digital intelligence literacy prediction for secondary school teachers. En *2025 14th International Conference on Educational and Information Technology (ICEIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICEIT64364.2025.10976158

Quammie, M., & Hosein, P. (2024). School climate factors as predictors of school performance: A machine learning approach. En *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICETSIS61505.2024.10459425

Rai, A., S, H. V., Qureshi, M. I., Raghumahti, R., & Kaur, G. (2024). Integrating artificial intelligence in early childhood education: A review of current practices and future Directions. En *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)* (pp. 1–6). IEEE. https://doi.org/10.1109/IDICAIEI61867.2024.10842823

Ranawaka, U. M., & Rajapakse, C. (2020). Predicting examination performance using machine learning approach: A case study of the Grade 5 scholarship examination in Sri Lanka. En *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (pp. 1–6). IEEE. https://doi.org/10.1109/SCSE49731.2020.9313029

Rathi, S., Wamanacharya, S., Shah, S., & Khedkar, V. (2024). Enhanced agricultural decision-making: Merging climatic data with machine learning algorithms. En *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICDICI62993.2024.10810951

Rizk, F. H., Mohamed, M. E., Sameh, B., Zaki, A. M., Eid, M. M., & El-Kenawy, E.-S. M. (2024a). Enhancing student performance prediction with Greylag Goose Optimization algorithm. En *2024 International Telecommunications Conference (ITC-Egypt)* (pp. 1–6). IEEE. https://doi.org/10.1109/ITC-Egypt61547.2024.10620568

Rizk, F. H., Mohamed, M. E., Sameh, B., Zaki, A. M., Eid, M. M., & El-Kenawy, E.-S. M. (2024b). Predictive modeling of Portuguese student performance: Comparative machine learning analysis. En *2024 International Telecommunications Conference (ITC-Egypt)* (pp. 1–6). IEEE. https://doi.org/10.1109/ITC-Egypt61547.2024.10620557

S, W. G. C., S, S. S. G. T., P, B. K. H., & Gamage, A. (2022). AI and machine learning based e-learning system for secondary education. En *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)* (pp. 1–6). IEEE. https://doi.org/10.1109/I2CT54291.2022.9824643

Sahu, S., Kumar, A., Kumar, R., Shukla, A., Chole, V., & Moolchandani, J. (2024). Predicting student academic performance using machine learning: Analyzing socio-economic and personal factors from secondary education in Portugal. En *2024 4th International*

Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICTACS62700.2024.10841127

Samira, R., Ahmed, B., Abdellatif, E. A., Said, B., & Mouad, E. (2024). Smart guidance systems for Moroccan students. En *2024 Mediterranean Smart Cities Conference (MSCC)* (pp. 1–6). IEEE. https://doi.org/10.1109/MSCC62288.2024.10697009

Sancho, M.-R., Cañabate, A., & Sabate, F. (2015). Contextualizing learning analytics for secondary schools at micro level. En *2015 International Conference on Interactive Collaborative and Blended Learning (ICBL)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICBL.2015.7387638

Song, J. (2021). Design and application of business English sand table simulation training course in secondary vocational school based on machine learning. En *2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)* (pp. 1–6). IEEE. https://doi.org/10.1109/TOCS53301.2021.9688789

ST. (2000). Key predictors of school dropout in Paraguay: A big data analysis. *[Publicación Técnica]*.

ST. (2024). AI-driven educational transformation in secondary schools: Leveraging data insights for inclusive learning environments. *[Publicación Técnica]*.

ST. (2025). Career recommendation using machine learning for secondary education. *[Publicación Técnica]*.

Suryawanshi, S., Kolhe, C., More, P., Barhate, J., Deshmukh, R., & Hingmire, A. (2025). Machine learning approach for career guidance using brain mapping technique and psychometric test. En *2025 1st International Conference on AIML-Applications for*

Engineering & Technology (ICAET) (pp. 1–6). IEEE. https://doi.org/10.1109/ICAET63349.2025.10932262

Tkáčová, Z., Šnajder, L., & Guniš, J. (2020). Artificial intelligence - a new topic in Computer Science curriculum at primary and secondary schools: Challenges, opportunities, tools and approaches. En *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 1–6). IEEE. https://doi.org/10.23919/MIPRO48935.2020.9245429

Vadwala, A. (2025). Career recommendation using machine learning for secondary education. En *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (pp. 1–6). IEEE. https://doi.org/10.1109/IATMSI64286.2025.10985503

Wang, J., Sun, J., Lu, R., Chen, Y., Su, C., & Zhao, Z. (2024). A practical study on statistical analysis of secondary school achievement and case teaching based on data mining. En *2024 14th International Conference on Information Technology in Medicine and Education (ITME)* (pp. 1–6). IEEE. https://doi.org/10.1109/ITME63426.2024.00105

Xiang, F., Zhang, X., Cui, J., Carlin, M., & Song, Y. (2022). Algorithmic bias in a student success prediction models: Two case studies. En *2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)* (pp. 1–6). IEEE. https://doi.org/10.1109/TALE54877.2022.00058

Yeguas-Bolívar, E., Alcalde-Llergo, J. M., Aparicio-Martínez, P., Taborri, J., Zingoni, A., & Pinzi, S. (2022). Determining the difficulties of students with dyslexia via virtual reality and artificial intelligence: An exploratory analysis. En *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural*

Engineering (MetroXRAINE) (pp. 1–6). IEEE. https://doi.org/10.1109/MetroXRAINE54828.2022.9967589

Yeh, J. H.-J., Bartholio, C., Shackleton, E., Costello, L., Perera, M., Yeh, K., & Yeh, C. (2020). Environmentally embedded Internet-of-Things for secondary and higher education. En *2020 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICICT50521.2020.00092

Yu, M. (2025). A cross-national study on education and employment outcomes using machine learning-based predictive modeling. En *2025 8th International Conference on Computer Information Science and Application Technology (CISAT)* (pp. 1–6). IEEE. https://doi.org/10.1109/CISAT66811.2025.11181746

Zewde, T. A. (2025). Enhancing pre-college students readiness through interactive summer AI workshops. En *2025 IEEE Integrated STEM Education Conference (ISEC)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISEC64801.2025.11147382

Zheng, Y., Meng, H., & Jia, W. (2022). Application research and challenges of artificial intelligence in primary and secondary education. https://doi.org/10.1109/ITME56794.2022.00043.

Apéndices

Apéndice A

Vídeo Presentación de la Opción de Grado

https://drive.google.com/drive/folders/16ZRv-2W0rKUObteEagK5YVaM0YBYSVhz?usp=drive_link

Apéndice B

Vinculo de Acceso a la Matriz de Xtracción

https://drive.google.com/drive/folders/16ZRv-2W0rKUObteEagK5YVaM0YBYSVhz?usp=drive_link