Clasificación del desempeño académico en la prueba Saber 11 mediante variables socioeconómicas: estudio comparativo entre redes neuronales, convolucionales y sistemas expertos

Elkin Vladimir Acosta Velásquez, César Augusto Sanabria Casanova

Trabajo de grado para optar el título de Magíster en Análisis de Datos y Sistemas Inteligentes

Director

César Hernando Valencia Niño

Ph.D. Ingeniería Eléctrica

Mg. Kelly Johanna Niño Sandoval

Universidad Santo Tomás, Bucaramanga

División de Ingenierías y Arquitectura

Maestría en Análisis de Datos y Sistemas Inteligentes

2026

Dedicatoria

Dedicamos este trabajo de grado, fruto de un proceso de formación, investigación y crecimiento profesional, a nuestras familias, quienes han sido el principal soporte durante este recorrido académico.

A nuestros padres, por su ejemplo de vida, sus enseñanzas, valores y sacrificios, que constituyen el fundamento de nuestra formación personal y profesional. Su apoyo incondicional y confianza permanente han sido una fuente constante de inspiración para alcanzar cada una de nuestras metas.

A nuestras esposas, por su comprensión, paciencia y acompañamiento durante las largas jornadas de estudio, análisis e investigación que demandó este proyecto. Su respaldo ha sido esencial para mantener la motivación y la perseverancia necesarias para culminar esta importante etapa académica.

A nuestros hijos, quienes representan la mayor motivación para continuar aprendiendo, creciendo y contribuyendo desde el conocimiento a la construcción de una mejor sociedad. En ellos encontramos el propósito que impulsa nuestros esfuerzos y el compromiso permanente con la excelencia.

Extendemos también esta dedicatoria a nuestros familiares, amigos y seres queridos, quienes de una u otra manera contribuyeron con palabras de aliento, apoyo y confianza durante el desarrollo de esta investigación.

Finalmente, dedicamos este logro a todas las personas que creen en el poder transformador de la educación, la ciencia, los datos y la innovación como herramientas para generar conocimiento, promover el desarrollo y aportar soluciones a los desafíos de nuestra sociedad.

Agradecimientos

Expresamos nuestro sincero agradecimiento a Dios por brindarnos la fortaleza, la sabiduría y la perseverancia necesarias para culminar este importante logro académico.

A la Universidad Santo Tomás y a la Maestría en Análisis de Datos y Sistemas Inteligentes, por los conocimientos, espacios de formación y oportunidades de crecimiento profesional que hicieron posible el desarrollo de esta investigación.

Al doctor Cesar Hernando Valencia Niño, director de este trabajo, y a la profesora Kelly Johanna Niño Sandoval, por su orientación, acompañamiento académico y valiosos aportes, los cuales contribuyeron significativamente al fortalecimiento de este proyecto.

Al ICFES, por la disponibilidad de la información utilizada en esta investigación, y a la comunidad académica y científica cuyos aportes sirvieron de base para el desarrollo de este estudio.

Finalmente, agradecemos de manera especial a nuestras familias, especialmente a nuestros padres, esposas e hijos, por su apoyo incondicional, comprensión y motivación permanente durante este proceso de formación y crecimiento profesional.

Contenido

Lista de tablas

Lista de figuras

Resumen

El trabajo de investigación se desarrolló bajo un estudio comparativo de los diferentes modelos de inteligencia artificial “MLP – CNN 1D - LightGBM” para la clasificación del desempeño académico de los resultados de la prueba de Estado Saber 11° en Colombia, a través de las variables socioeconómicas, familiares e institucionales, abordando la problemática en las limitaciones de los enfoques estadísticos tradicionales para capturar las estructuras complejas, no lineales y de alta dimensionalidad de las variables que permiten explicar las brechas educativas en Colombia. Para ello se formuló modelos predictivos multiclases que lograran utilizar un conjunto de datos históricos de aproximadamente de 7.2 millones de registros anonimizados proporcionados por el ICFES correspondiente a los periodos del 2014-2 al 2024-2. Se hizo uso de la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) en el que se implementó un pipeline de procesamiento hibrido y comparativos de las tres clases de modelos de inteligencia artificial. Los resultados experimentales permitieron demostrar que las arquitecturas de redes neuronales profundas “MLP” alcanzaron el mejor desempeño predictivo y mayor estabilidad estadística, superando a la red neuronal convolucional unidimensional “CNN 1D” y el algoritmo de ensamble (LightGBM) al capturar las dependencias no lineales del dataset de cada individuo. En conclusión, las variables socioeconómicas poseen una capacidad predictiva significativa sobre la clasificación en el rendimiento académico, y los modelos aprendizaje profundo son una herramienta tecnológica y metodológica para validar el diseño de un sistema de alerta temprana y apoyo para la toma de decisiones basada en evidencia en el sector educativo colombiano

Palabras clave: deep learning, desempeño académico, educación, saber 11, minería datos, sistemas inteligentes, socioeconómico.

Abstract

The research work was developed as a comparative study of different artificial intelligence models ("MLP – CNN 1D - LightGBM") for the classification of academic performance based on the results of the Saber 11° State examination in Colombia, through socioeconomic, family, and institutional variables. It addresses the problem of the limitations in traditional statistical approaches when capturing the complex, non-linear, and high-dimensional structures of the variables that explain educational gaps in Colombia. To this end, multiclass predictive models were formulated to utilize a historical dataset of approximately 7.2 million anonymized records provided by the ICFES, corresponding to the periods from 2014-2 to 2024-2. The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was employed, implementing a hybrid processing pipeline and comparative analyses of the three classes of artificial intelligence models. The experimental results demonstrated that deep neural network architectures ("MLP") achieved the best predictive performance and highest statistical stability, outperforming the one-dimensional convolutional neural network ("1D CNN") and the ensemble algorithm (LightGBM) by capturing the non-linear dependencies of the dataset for each individual. In conclusion, socioeconomic variables possess a significant predictive capacity regarding academic performance classification, and deep learning models serve as a technological and methodological tool to validate the design of an early warning system and support evidence-based decision-making in the Colombian educational sector.

Keywords: deep learning, academic performance, education, Saber 11, data mining, intelligent systems, socioeconomic.

1 Planteamiento del problema

1.1 Contexto del problema

En el mundo globalizado, el crecimiento económico de los países es fundamental; para ello se requieren inversión extranjera, investigación, innovación, desarrollo tecnológico y generación de empleos. Por ello, es necesario impulsar la educación de calidad que permita la inmersión laboral de la población y así generar mejores sociedades [1]. Para las Naciones Unidas (ONU), los países están llamados a adoptar el cuarto Objetivo de Desarrollo Sostenible (ODS) que hace referencia a "*Garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos*" [2]).

La Organización de las Naciones Unidas (ONU) cuenta con diecisiete objetivos para la transformación del mundo propuesto en el 2015 hasta el 2030, dentro de los objetivos se encuentra la "Educación de Calidad", eslabón fundamental para alcanzar los diferentes objetivos propuestos de desarrollo sostenibles (ODS) [3], siendo unos de los principales eslabones para el crecimiento y desarrollo económico de un país, para ello el Estado colombiano estipula en su Constitución Política de Colombia (CP) los fines esenciales definidos en el artículo 2:

> *… servir a la comunidad, promover la prosperidad general y garantizar la efectividad de los principios, derechos y deberes consagrados en la Constitución; facilitar la participación de todos en las decisiones que los afectan y en la vida económica, política, administrativa y cultural de la Nación…* [4].

El artículo 67 menciona "… *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…*" [4], evaluando la importancia del desarrollo de la persona mediante la transformación del conocimiento a través de la educación,

como herramienta fundamental para generar un estado productivo y social [5].

De acuerdo con el informe desarrollado por la Organización para la Cooperación y el Desarrollo Económicos OECD del 2024, indican que Colombia es uno de los países más desiguales en el mundo, dentro de los aspectos más desiguales se encuentra el acceso a la educación de calidad, lo que ha generado el abandono de la escuela por parte de los estudiantes con habilidades deficientes, sobre todo de los estudiantes con los factores socioeconómicos más desfavorecidos, afectando la productividad y economía del Estado colombiano.

Colombia cuenta con un instrumento de evaluación que permite medir la calidad de educación en el nivel de educación media denominado Examen de Estado de la Educación Media, Saber 11°, el cual está compuesto por cinco pruebas "*Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés*" [6]); Según el informe nacional de resultados para Colombia 2022 permite realizar una comparación entre los diferentes lapsos del examen Saber 11, en el año 2015 obtuvieron un promedio en el Saber de matemáticas de 390 puntos, el 2018 alcanzaron un promedio de 391 y en el 2022 disminuyó con un promedio de 7.5 puntos, obteniendo un promedio de 383 puntos [7], siendo una de las competencias más importantes y fundamentales en la investigación, desarrollo e innovación (I+D+I).

En Estados Unidos se han desarrollado múltiples estudios para identificar cuáles son los factores que pueden mejorar el rendimiento académico los cuales hacen usos de los diferentes tipos de bases de datos para realizar análisis comparativos, patrones y tendencias para comprender el desempeño académico, dentro de los estudios realizados se viene implementando el uso de modelos de Machine Learning como Árboles de Decisión (CART) , Random Forest (RF), XGBoost y LightGBM, los cuales permiten descubrir relaciones complejas entre las diferentes variables como características demográficas, contexto familiar, entornos académicos y los resultados de las pruebas PISA en comparación métodos tradicionales estadísticos, generando un menor error

estadístico y mayor eficiencia en la identificación de factores predictores para la implementación de planes educativos institucional [8].

La Universidad de Educación de Hong Kong, Ting Kok, Hong Kong, China, realizó un estudio para identificar cuáles son los factores que influyen en el rendimiento en los resultados de las pruebas del Programa para la Evaluación Internacional de Estudiantes (PISA) 2015 en la ciencia de los países Singapur y Finlandia mediante uso de modelos de inteligencia artificial; el análisis logró identificar diez factores claves que influyen en los resultados en tres niveles [9].

Factor individual se identificaron cuatro variables, el tiempo de aprendizaje en ciencias, investigación científica, creencia epistemológica y el afecto por el dominio de la ciencia, en el factor familiar identificaron tres variables entre ellas se encuentra la riqueza familiar, el nivel ocupacional de los padres, y los recursos educativos en el hogar, y a nivel escolar se encuentran las variables medios tecnológicos y de comunicaciones disponibles en la escuela y el clima disciplinario en las clases de ciencia, brindando información de valor basado en datos que permita desarrollar una estrategia para fortalecer los saberes en ciencia [9].

De igual forma, Alkan et al. [10] analizó los resultados de las competencia de ciencia, matemáticas y lectura, 25 variables que hacen parte de las pruebas del Programa para la Evaluación Internacional de Estudiantes (PISA) 2018 en 37 países pertenecientes a la Organización para la Cooperación y el Desarrollo Económicos (OCDE) logrando identificar cinco variables claves que influyen positivamente en los resultados de las pruebas PISA 2018: acceso a los recursos TIC en las escuela o en el hogar, la intensidad horaria de clase, conciencia metacognitiva, el índice socioeconómico, el índice ocupacional de los padres. El estudio brindó información objetiva para la implementación de políticas y estrategias académicas orientada a mejorar los resultados en las pruebas PISA.

1.2 Descripción clara y delimitada de la problemática.

La evidencia empírica reciente ha demostrado que la brecha educativa asociada a factores socioeconómicos no responde a relaciones lineales simples, sino a interacciones complejas, no lineales y de alta dimensionalidad entre variables individuales, familiares e institucionales [11], [12]. En este contexto, los enfoques tradicionales basados en modelos lineales presentan limitaciones para captar dichas interacciones, lo que limita la capacidad explicativa y predictiva de los modelos [13].

Investigaciones recientes han incorporado modelos de inteligencia artificial y aprendizaje automático para capturar relaciones no lineales y efectos complejos entre dichas variables, logrando mejoras sustanciales en la capacidad predictiva frente a enfoques estadísticos tradicionales [14].

En contraste, los modelos de aprendizaje automático como Random Forest, Gradient Boosting (XGBoost) y redes neuronales artificiales permiten modelar relaciones no lineales, capturar interacciones de alto orden y manejar datos con alta dimensionalidad y ruido, características propias de los datos educativos masivos [15], [16].

Específicamente, los modelos basados en árboles (Random Forest y XGBoost) son altamente robustos frente a datos tabulares heterogéneos y permiten interpretar la importancia de variables, mientras que las redes neuronales densas permiten aproximar funciones complejas mediante aprendizaje jerárquico de representaciones. Por su parte, las redes neuronales convolucionales (CNN), aunque tradicionalmente utilizadas en datos estructurados espaciales como imágenes, han sido adaptadas recientemente para datos tabulares mediante representaciones matriciales o embeddings, mostrando capacidad para identificar patrones locales y relaciones estructurales complejas entre variables ([17], [18], [19]).

En este sentido, la selección y comparación de estas arquitecturas responde a la necesidad de

evaluar cuál de ellas logra una mejor capacidad predictiva del desempeño académico en función de variables socioeconómicas, contribuyendo así al diseño de modelos robustos para la toma de decisiones basadas en evidencia en el contexto educativo colombiano, para lo cual se plantea la siguiente pregunta orientadora:

¿Cómo diseñar e implementar un modelo de clasificación del desempeño académico en la prueba Saber 11 que, mediante redes neuronales, convolucionales y sistemas expertos, aprenda patrones y dependencias implícitas a partir de variables socioeconómicas en Colombia?

1.3 Justificación del proyecto

La literatura científica ha señalado que una de las principales limitaciones en los estudios educativos basados en aprendizaje automático es la ausencia de validaciones rigurosas y comparaciones sistemáticas entre modelos, lo que reduce su utilidad para la toma de decisiones en política pública [20].

En contraste, el contexto colombiano presenta una producción científica limitada en revistas de alto impacto que integre modelos de inteligencia artificial supervisados con análisis profundo de factores socioeconómicos aplicados al examen Saber 11°. Este vacío metodológico restringe el uso efectivo de analítica avanzada en la formulación de políticas educativas basadas en evidencia, lo que justifica la necesidad de desarrollar modelos predictivos robustos que contribuyan al cumplimiento del Objetivo de Desarrollo Sostenible 4 [3].

La complejidad del fenómeno educativo en Colombia, marcado por brechas estructurales de desigualdad [21], exige superar los métodos estadísticos tradicionales. Si bien modelos de aprendizaje supervisado como Random Forest (RF) y XGBoost son el estándar de oro para datos tabulares debido a su capacidad para manejar valores faltantes y capturar interacciones no lineales

mediante ensambles de árboles [8], esta investigación propone una arquitectura de Redes Neuronales Convolucionales (CNN).

La elección de una CNN, tradicionalmente usada en la visión artificial, se justifica aquí por su capacidad de extraer características latentes de estructuras de datos reorganizadas como tensores, lo que permite identificar patrones jerárquicos espaciales entre variables socioeconómicas que las redes densas (MLP) o los modelos de boosting podrían omitir al tratar las variables de forma aislada [22]. La comparación multi modelo permitirá determinar si la ingeniería de características de las CNN ofrece una ventaja competitiva en la clasificación de niveles de desempeño (Muy Bajo, Bajo Medio, Alto, Muy Alto) frente a la robustez de los modelos basados en árboles.

Además, los resultados del análisis bibliométrico que se pueden observar en el numeral 2.3 de este documento evidencian que, aunque existe un interés creciente en la evaluación del desempeño académico, la mayoría de los estudios se fundamentan en enfoques estadísticos tradicionales, con escasa utilización de técnicas de aprendizaje automático. Esta situación limita la capacidad de capturar relaciones complejas entre variables socioeconómicas y resultados académicos, lo que justifica la implementación de modelos de machine learning como Random Forest, XGBoost y redes neuronales en el presente estudio.

1.4 Objetivo general

- Desarrollar un modelo de clasificación del desempeño académico en la prueba Saber 11 en Colombia utilizando redes neuronales, convolucionales, sistemas expertos y variables socioeconómicas.

1.5 Objetivos específicos

- Seleccionar las variables socioeconómicas relevantes del conjunto de datos correspondientes al periodo 2014-2024, para su representación como entradas del modelo de clasificación.
- Implementar un modelo de clasificación del desempeño académico empleando redes neuronales, convolucionales y sistemas expertos que capture dependencias no lineales entre variables socioeconómicas.
- Validar estadísticamente el desempeño del modelo de clasificación del rendimiento académico, mediante el análisis de la estabilidad de las predicciones y la comparación con modelos de referencia.

1.6 Alcance y delimitaciones del proyecto

La investigación se desarrollará en el contexto del sistema educativo colombiano, utilizando datos históricos de los exámenes Saber 11 correspondientes al periodo 2014–2024, disponibles a través del Instituto Colombiano para la Evaluación de la Educación (ICFES).

El alcance del estudio comprende la estructuración y depuración de un conjunto de datos analítico basado en variables socioeconómicas, así como la adopción, parametrización e implementación de un modelo de clasificación del rendimiento académico basado en redes neuronales convolucionales. Dicho modelo será entrenado y validado para la clasificación del desempeño académico por categorías, conforme a los principios del aprendizaje supervisado y la analítica predictiva.

El estudio no tiene como objetivo el desarrollo de una nueva arquitectura de inteligencia artificial, sino la aplicación y validación de un modelo existente en un contexto educativo específico, garantizando su reproducibilidad y utilidad para investigaciones futuras y tareas

similares. Este enfoque es consistente con trabajos recientes en analítica predictiva educativa, donde modelos de aprendizaje automático existentes son adaptados y evaluados en contextos particulares para el análisis del desempeño académico [12].

Como productos del proyecto se obtendrán: (i) una base de datos estructurada y documentada para uso académico, (ii) un modelo funcional y validado que podrá ser reutilizado como referencia en estudios posteriores, y (iii) un artículo científico como resultado del proceso de divulgación de los hallazgos.

El trabajo se limita al análisis predictivo del rendimiento académico a partir de variables socioeconómicas, sin realizar inferencia causal ni evaluación directa de impactos en políticas educativas, de acuerdo con los lineamientos metodológicos de la analítica predictiva y el aprendizaje supervisado propuestos por [23].

1.7 Delimitación Técnica del Problema

En términos operativos, la variable dependiente del estudio corresponde al nivel de desempeño académico en la prueba Saber 11, el cual será modelado como un problema de clasificación supervisada. Esta variable será categorizada en niveles (muy bajo, bajo, medio, alto, muy alto), definidos a partir de percentiles del puntaje global, conforme a criterios establecidos por el ICFES.

La población de estudio está conformada por los estudiantes que presentaron la prueba Saber 11 en Colombia durante el periodo 2014–2024, y la unidad de análisis corresponde a cada estudiante individual con su respectivo conjunto de variables socioeconómicas.

Las variables independientes incluyen factores socioeconómicos, familiares e institucionales tales como nivel educativo de los padres, acceso a recursos tecnológicos, estrato socioeconómico, tipo de institución educativa, entre otros.

- *Variable Dependiente*: Desempeño académico categorizado en cinco niveles (Muy Bajo, Bajo, Medio, Alto, Muy Alto) según el puntaje global de la prueba Saber 11.
- *Variables Independientes*:
  - *Socioeconómicas*: estrato, acceso a internet, computador, número de personas en el hogar.
  - *Escolares:* tipo de jornada, ubicación, bilingüismo, calendario, naturaleza, carácter.
  - *Familiares*: nivel educativo y ocupación del padre y la madre
  - *Población:* Estudiantes de último grado de educación media en Colombia que presentaron la prueba entre 2014 y 2024.
- *Unidad de Análisis*: Registro individual de microdatos del ICFES (estudiante), anonimizado.

2 Marco teórico y estado del arte

2.1 Conceptos teóricos relevantes

Unas de las propuestas desarrolladas por [24] fue establecer si las máquinas pueden ser inteligentes, proponiendo que las máquinas pueden aprender a través de la enseñanza de los niños, mediante diferentes tipos de procesos y métodos educativos, así mismo con la inducción científica que permite que sean producidos por la misma máquina; El aprendizaje automático se encuentra basado en matemáticas aplicadas y computacional convirtiéndose en unos de las subáreas de la inteligencia artificial, así como los modelos de procesamiento como el lenguaje natural, reconocimiento de voz, visión de computadora y otros [24].

Es de gran importancia entender qué es la inteligencia artificial, [25] define la inteligencia artificial como un conjunto de componentes tecnológicos que emulan las funciones del cerebro del ser humano, De igual forma [28] define de la inteligencia artificial en el cual destaca dos

características claves “razonamiento - comportamiento” que pueden desarrollar los modelos de sistemas inteligencia a través de los algoritmos basados en datos para aprender y tomar dediciones como los seres humanos; de igual forma las máquinas han avanzado mediante el aprendizaje automático “Machine Learning” permitiendo extraer y analizar información para logra obtener conocimientos de los datos [26], para [27] el machine learning es la ciencia y el arte de programar computadoras facilitando el aprendizaje en base de datos.

La Inteligencia artificial basada en datos desarrolla el análisis de gran cantidad de información a través de algoritmos estadísticos como el Deep learning, que logra obtener patrones para resolver problemas y pronosticar resultados en las diferentes áreas como la medicina, redes sociales, juegos, cambio climático, investigaciones, la agricultura, la educación, entre otros [25].

Unos de los fundamentos teóricos de la inteligencia artificial fue desarrollado por [29] proponen una serie de teoremas que permite describir el modelo de una red neuronal desde el concepto de la lógica computacional en equivalencia a los modelos biológicos de los sistemas nerviosos del ser humano, estableciendo el funcionamiento del perceptrón a través de estímulos, información histórica, lógica en red por ciclos, memoria codificada dinámica, aportando las bases fundamentales de los sistemas inteligentes actuales.

Para entender cómo funciona una red neuronal profunda podemos tomar la explicación descrita por [28] el cual indica la red neuronal es una serie de capas apiladas que se conectan unas con las otras de forma directa, compuesta de la siguiente forma, capa de datos de entrada, la capa de salida y capa oculta, las cuales son transformadas de forma no lineal, capa por capa mediante la propagación hacia adelante hasta encontrar el valor más preciso y adecuado de los pesos para cada capa (Entrenamiento), así mismo hace uso del backpropagation el cual permite tomar los errores de predicción y realizar los ajustes en los pesos de las neuronas, permitiendo mejorar las predicción de la red neuronal.

La educación ha venido transformándose a través de la dataficación mediante la recolección, procesamientos y análisis de la información mejorando la calidad de la educación, unos de los ejemplos de la dataficación es cuando un grupo de estudiantes realizan un examen y se crean una ingesta de datos que permite interpretar y generar predicciones probabilistas de los resultados de los exámenes, este proceso solo es posible mediante el uso de la inteligencia artificial. Generando para la educación mejoras en los procesos académicos, políticas públicas, experiencia de enseñanza personalizadas y fortalecimiento en la investigación científica académica, innovación continua en la educación [30].

Es de gran importancia tener en cuenta que para realizar los diferentes dataficación de los colegios es necesario tener en cuenta varios factores de los que menciona recursos de capital físico como la tecnología, equipo, ubicación geográfica, materias primas, Recurso capital Humano se encuentra las capacitaciones, experiencia, inteligencia, trabajadores y los recursos de capital organizacional entre ellos se encuentra la estructura organizacional, sistema de planificación y control, procesos internos y el entorno de la organización, así mismo es de gran importancia la ventaja historia de la organización su trayectoria en el tiempo y espacio lo que permite obtener experiencia y reputación organizacional; de igual forma unas de las ventajas competitiva de una organización son los fenómenos sociales compuestas por la relación y cultura entre sus stakeholder y Partners [29], [31].

2.2 Modelos, técnicas o enfoques existentes

En los diferentes trabajos de investigación o proyectos de ciencias de datos e inteligencia artificial hacen uso de la metodología Cross-Industry Standard Process for Data Mining “CRISP-DM” con el propósito de evaluar los resultados de clasificación y predicción de los diferentes

modelos de inteligencia artificial en el uso de los resultados de los exámenes de evaluación de los estudiantes tanto a nivel nacional e internacional como las Programa para la Evaluación Internacional de Estudiantes "PISA" y Saber 11° en Colombia, [32] utiliza el modelo CRIPS-MD como base de extracción del conocimiento con el fin de predecir el nivel de desempeño de los estudiantes tomando los resultados de los años 2017 al 2019 del examen saber 11° del instituto Colombiano para el Fomento de la educación superior "ICFES" y aplicando cuatro modelos de machine learning supervisado como J48 (C4.5), LMT, PART y Perceptrón Multicapa (MLP), evaluando cual brinda mejor capacidad predictiva.

Así mismo lo realiza [33] en su trabajo denominado "*A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations*", quien hace uso de la metodología CRIPS-MD para identificar las regiones con índices de pobreza económica y cuáles son las variables que influyen en los resultados del saber 11° en Colombia de los año 2016 al 2018 aplicando nueve modelos de algoritmo de clasificación de los cuales se encuentra los más destacados Extreme Gradient Boosting Machine (XGBM), Light Gradient Boosting Machine (LGBM) y Perceptrón Multicapa (MLP): en Aplicación de Técnicas de Minería de Datos para el Análisis de los factores socioculturales que determinan el puntaje de las pruebas saber 11° del año 2018, 2019 y 2020.

El trabajo realizado por [34] nombrado "*Machine Learning Approaches for Predicting U.S. Students' Scientific Literacy: An Analysis of Key Factors Across Performance Levels and Socioeconomic Statuses*", se encuentra a lineado las fases implementada bajo la metodología CRISP-DM aunque no se menciona de forma específica la implementación, este estudio hizo uso los modelos de Machine Learnin Regresión tradicional (modelo estadístico base), Random Forest, XGBoost (eXtreme Gradient Boosting), lo que permitió identificar factores claves de desempeño en el rendimiento académico del programa para la evaluación internacional de estudiantes "PISA"

del año 2006 al 2015 de los Estados Unidos.

2.3 Revisión de trabajos previos relacionados

Para fundamentar la viabilidad técnica y científica de la presente investigación, se realizó una revisión sistemática de la literatura reciente (2020-2025) en bases de datos de alto impacto como Scopus y Web of Science. Esta búsqueda se centró en identificar estudios que emplean modelos de aprendizaje automático y aprendizaje profundo para la clasificación del desempeño académico. A continuación, en la Tabla 1, se presenta una síntesis comparativa de los trabajos más relevantes, destacando las arquitecturas utilizadas, el origen de los datos y los hallazgos principales que sirven de base para la comparación multi modelo propuesta en este estudio.

Tabla 1. *Síntesis de investigaciones previas sobre predicción del rendimiento académico mediante Machine Learning.*

Autor (Año)	Modelo(s) Evaluado(s)	Datos / Población	Resultado Principal	DOI
Zhu et al. (2025)	XGBoost, Random Forest, LightGBM	PISA 2022 (EE. UU.)	XGBoost demostró la mayor capacidad predictiva en matemáticas, superando a modelos lineales tradicionales.	10.1016/j.ijer.2025.102537
Alvarez-Garcia et al. (2024)	Clúster Analysis & XAI (SHAP)	PISA 2022	Logró identificar perfiles de riesgo académico basados en el sentido de pertenencia y ansiedad matemática.	10.1016/j.compedu.2024.105166
Gomez-Talal et al. (2025)	Interpretable ML (IML)	PISA Global	Validó que los modelos de "caja negra" pueden ser explicados para toma de decisiones en políticas públicas.	10.1016/j.eswa.2024.125100
Xi & Panoutsos (2018/2020)	**CNN**, RBF, Fuzzy Logic	Datos Tabulares Complejos	Demostró que las **CNN** extraen características automáticas en datos no-imagen superando a redes densas.	10.1109/IS.2018.8710470
Bernardo et al. (2021)	ANN (Redes Neuronales), RF	PISA Filipinas	Las variables socioeconómicas del hogar superaron en peso predictivo a las variables escolares.	10.1016/j.heliyon.2021.e06387
Demir & Karaboğa (2021)	Deep Learning (MLP)	PISA 2018	El uso de redes neuronales profundas permitió clasificar niveles de logro con una precisión superior al 85%.	10.1007/s12528-021-09273-y

Autor (Año)	Modelo(s) Evaluado(s)	Datos / Población	Resultado Principal	DOI
Castrillón et al. (2020)	RF, SVM, Naive Bayes	Saber 11 (Colombia)	El modelo **Random Forest** fue el más robusto para predecir el desempeño en contextos universitarios iniciales.	10.17081/invinno.8.1.3544
Löhr et al. (2020)	ML Assisted Solutions	PISA 2018	Implementación de técnicas de regularización para evitar el sobreajuste en modelos de predicción educativa.	10.1186/s40594-020-00227-z
Suaza-Medina et al. (2024)	ML Predictivo	Educación Media	Establece marcos de referencia para la intervención temprana basada en clasificaciones de riesgo social.	10.1016/j.procs.2024.01.001
Jeganathan et al. (2024)	OptCatB (CatBoost Opt.)	PISA Immigrant Data	El ajuste de hiperparámetros en modelos de boosting es clave para reducir el error en minorías.	10.1038/s41598-024-55671-w

Nota: la tabla fue adaptada a partir de la revisión de literatura (2026).

La revisión sistemática de la literatura evidencia una tendencia creciente hacia el uso de modelos de ensamble como XGBoost [13] y Random Forest por su eficiencia en datos estructurados. Sin embargo, surge una oportunidad de investigación en la aplicación de Redes Neuronales, y en especial de Redes Neuronales Convolucionales (CNN) para el contexto colombiano. Siguiendo la premisa de [22] esta tesis propone que la arquitectura CNN puede identificar jerarquías de variables socioeconómicas que los modelos tradicionales omiten, aportando una nueva dimensión técnica al análisis de las pruebas Saber 11.

Asimismo, con el propósito de contextualizar el uso de modelos de aprendizaje automático en la predicción del desempeño académico, la Tabla 2 presenta una síntesis comparativa de los 5 estudios relevantes reportados en la literatura científica reciente. Estos trabajos evidencian la evolución desde enfoques tradicionales hacia modelos avanzados capaces de capturar relaciones no lineales y patrones complejos en datos educativos.

Tabla 2. *Comparación de modelos de aprendizaje automático aplicados a la predicción del desempeño académico y problemas de clasificación en datos educativos.*

Autor	Modelo	Datos	Resultado
Zhu et al. (2025)	XGBoost	PISA	Alta precisión predictiva
Alkan et al. (2025)	ML Ensemble	PISA	Identificación de variables clave
Kotsiantis (2007)	ML revisión	Educación	Mejora sobre modelos clásicos
Fernández-Delgado (2014)	Comparación ML	179 datasets	RF y boosting dominan

Los estudios revisados relacionados en la tabla anterior coinciden en que los modelos de ensemble como Random Forest y XGBoost presentan un desempeño superior en datos tabulares, debido a su capacidad para capturar relaciones no lineales y manejar variables heterogéneas. Sin embargo, la aplicación de redes neuronales profundas, particularmente CNN, en este tipo de datos sigue siendo limitada y requiere validación empírica rigurosa, lo que representa una oportunidad de investigación en el contexto colombiano.

2.4 Identificación de vacíos o limitaciones en la literatura

A pesar del crecimiento de la analítica educativa, estudios recientes evidencian que la aplicación de modelos avanzados de aprendizaje automático en contextos latinoamericanos, particularmente en Colombia, sigue siendo limitada en comparación con países desarrollados ([12], [35]). Esta brecha se refleja tanto en la baja producción científica indexada como en la escasa aplicación de modelos comparativos robustos sobre datos educativos a gran escala.

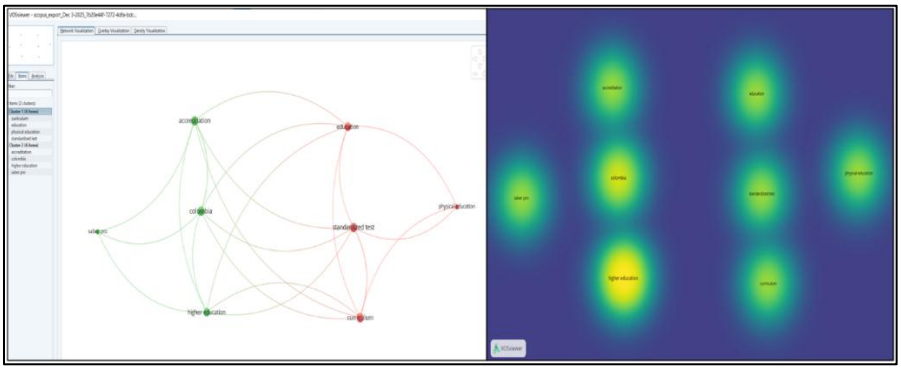
Por otra parte, el rendimiento académico ha sido ampliamente estudiado desde enfoques multidimensionales, destacándose la influencia de factores socioeconómicos, familiares y escolares [37], en un metaanálisis seminal, concluye que el nivel socioeconómico es uno de los predictores más consistentes del desempeño académico, explicando una proporción significativa de su variabilidad. Este enfoque se fundamenta en la teoría del capital humano [36] y los recursos de la

empresa y ventaja competitiva sostenida de [31], la cual establece que la educación es un mecanismo central para la acumulación de habilidades productivas y el desarrollo económico, y en su teoría del capital humano, plantea que la educación constituye una inversión que incrementa la productividad y los ingresos futuros de los individuos.

Ahora bien, a partir del análisis bibliométrico realizado sobre literatura indexada en Scopus, se evidencia que la producción científica relacionada con evaluación educativa en Colombia, particularmente en torno a pruebas estandarizadas como Saber Pro y Saber 11, es reciente y en crecimiento, concentrándose principalmente en los años 2024 y 2025. Asimismo, los resultados muestran una alta concentración temática en enfoques tradicionales de evaluación, tales como análisis de calidad educativa, acreditación y comparaciones curriculares.

Sin embargo, se identifica una limitada incorporación de metodologías avanzadas de analítica de datos, particularmente modelos predictivos basados en aprendizaje automático, así como un bajo énfasis en el análisis sistemático de variables socioeconómicas como determinantes del rendimiento académico, como se evidencia en la ilustración 1. Esta brecha metodológica evidencia una oportunidad de investigación relevante, orientada al desarrollo de modelos predictivos robustos que contribuyan al fortalecimiento de la toma de decisiones basada en evidencia en el sistema educativo colombiano.

Figura 1. *Análisis de Co-ocurrencia y Análisis de Densidad y Temporalidad.*



Nota: Figura a la izquierda muestra la. red de co-ocurrencia de términos en la literatura sobre evaluación educativa (Scopus),

mientras que la figura a la derecha corresponde al mapa de densidad temática de términos clave en la literatura analizada (Scopus). El tamaño de los nodos representa la frecuencia de aparición y la proximidad indica la relación entre términos.

Nota: La figura fue adaptada a partir de datos indexados en Scopus, procesados mediante el software VOSviewer.

El análisis bibliométrico realizado permite identificar patrones relevantes en la literatura existente. En primer lugar, se evidencia una fuerte concentración temática en términos como “higher education”, “accreditation”, “educational quality” y “Colombia”, lo cual sugiere un enfoque predominantemente institucional y evaluativo. En segundo lugar, la presencia de términos como “curriculum”, “physical education” y “standardized test” indica un interés en la coherencia curricular y la medición del desempeño académico.

No obstante, la baja frecuencia de términos asociados a técnicas avanzadas como “data mining”, “machine learning” o “predictive modeling” evidencia una limitada adopción de enfoques analíticos avanzados en este campo. Este hallazgo refuerza la necesidad de incorporar modelos de aprendizaje automático en el análisis del rendimiento académico, constituyendo un vacío claro en la literatura.

- Analítica educativa y Educational Data Mining
- Inteligencia artificial aplicada a educación
- Machine Learning supervisado
- Redes neuronales profundas
- CNN en datos tabulares
- Variables socioeconómicas y desempeño académico
- Vacíos de investigación

3 Metodología.

3.1 Tipo y enfoque de la investigación

Esta investigación aplicada se enmarca bajo el paradigma positivista con un enfoque cuantitativo y con un alcance correlacional teniendo en cuenta que este estudio busca establecer la correlacional entre los factores socioeconómicos del estudiante, y el examen saber 11 mediante la implementación de modelos de clasificación multiclases (redes neuronales y machine learning), permitiendo analizar el nivel de precisión y exactitud de los modelos al momento de clasificar la variable objetivo (Target), contribuyendo al sistema de alerta temprana y estrategias de mejora institucional.

La presente investigación se desarrolla bajo un enfoque cuantitativo, aplicado y experimental, orientado a la construcción, entrenamiento y comparación de modelos supervisados clasificación para predecir el desempeño académico en la prueba Saber 11 a partir de variables socioeconómicas, familiares, escolares e institucionales. El estudio se enmarca en el campo de Educational Data Mining y Learning Analytics, disciplinas que emplean técnicas de minería de datos, aprendizaje automático y aprendizaje profundo para descubrir patrones relevantes en datos educativos y apoyar procesos de toma de decisiones basados en evidencia [20]. La pertinencia de este enfoque se sustenta en investigaciones recientes que han demostrado la utilidad de modelos de aprendizaje automático para predecir el rendimiento académico y analizar variables asociadas al desempeño estudiantil en pruebas estandarizadas [13], [33].

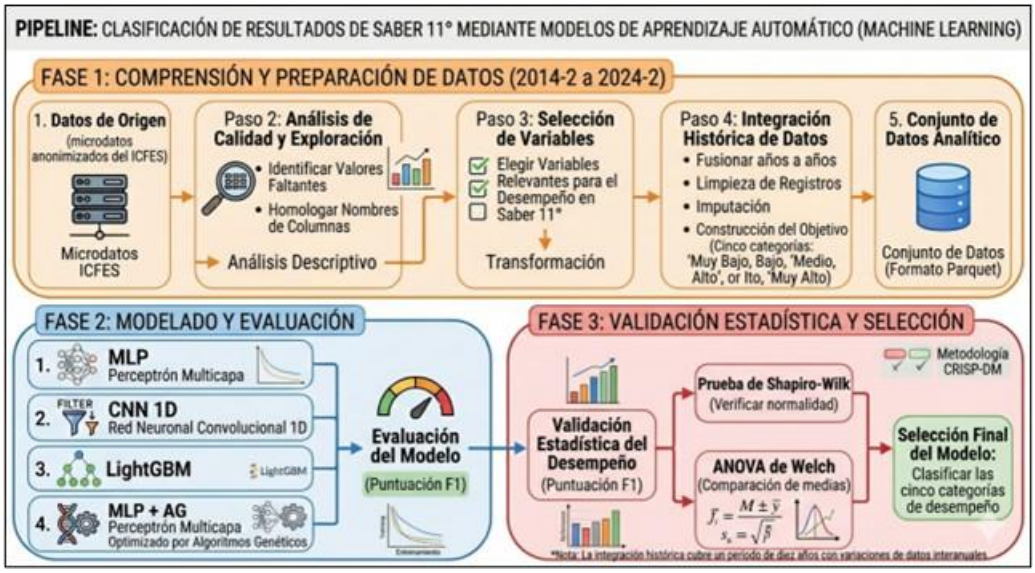
3.2 Descripción de las fases del proyecto

El desarrollo del trabajo de grado de investigación aplicada se realizó bajo la orientación de la

metodología de minería de datos CRISP-DM (Cross-Industry Standard Process for Data Mining) el cual permitirá implementar un modelo de clasificación del desempeño académico en la prueba Saber 11 usando redes neuronales convolucionales y variables socioeconómicas, las fase a usar de la metodología serán las fases de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación de los modelos de Inteligencia artificial, lo que permitirá transformar datos en conocimiento [38].

El desarrollo metodológico del proyecto fue estructurado siguiendo las fases de la metodología CRISP-DM, permitiendo organizar el proceso analítico desde la comprensión del problema hasta la validación estadística de los modelos implementados, como se muestra en la imagen.

Figura 2. *Pipeline de metodología CRISP_MD para la clasificación Resultados Saber 11°.*



Nota: La figura fue elaborado mediante herramienta Gemini de Google IA basados metodología CRISP-MD

Fase 1. Comprensión del problema y definición analítica: La primera fase consistió en identificar el problema de investigación y establecer los objetivos analíticos del proyecto. En esta etapa se realizó una revisión sistemática de literatura científica indexada en Scopus, Web of Science y ScienceDirect con el propósito de identificar modelos de aprendizaje automático aplicados a predicción del rendimiento académico.

La revisión bibliográfica permitió identificar que modelos como Random Forest, XGBoost y redes neuronales profundas han mostrado resultados relevantes en pruebas educativas internacionales como PISA y en sistemas educativos nacionales.

Asimismo, se identificó una limitada producción científica en Colombia relacionada con modelos avanzados de aprendizaje profundo aplicados al examen Saber 11, especialmente utilizando CNN adaptadas a datos tabulares.

A partir de esta revisión se formuló la pregunta orientadora del proyecto y se definieron los objetivos específicos asociados a:

- Selección de variables socioeconómicas;
- Implementación de modelos de clasificación;
- Validación estadística comparativa.

Fase 2. Comprensión y adquisición de datos: En esta fase se descargaron y analizaron los microdatos oficiales anonimizados publicados por el ICFES correspondientes al periodo 2014-2 a 2024-2.

Los archivos originales se encontraban distribuidos por periodo en formato TXT, con estructuras parcialmente heterogéneas dependiendo del año de aplicación del examen, durante esta etapa se realizó:

- Exploración inicial de variables.
- Análisis de calidad de datos.
- Identificación de valores faltantes.
- Homologación de nombres de columnas.
- Validación de consistencia temporal.
- Análisis descriptivo preliminar.

La documentación oficial de DataIcfes permitió identificar las variables relevantes para el proyecto, incluyendo información:

- Personal.
- Académica.
- Socioeconómica.
- Familiar.
- Institucional.

La exploración inicial evidenció alta heterogeneidad entre periodos, diferencias en cardinalidad de variables y presencia de categorías inconsistentes, lo cual hizo necesaria una etapa intensiva de preparación y normalización.

Fase 3. Preparación y transformación de datos: La preparación de los datos constituyó una de las etapas más críticas del proyecto debido a la magnitud del conjunto de información y a la heterogeneidad estructural presente entre periodos históricos del examen Saber 11 [39], [40], en esta fase se realizó:

- Integración histórica de datasets;
- Limpieza de registros;
- Imputación de valores faltantes;
- Transformación de variables;

- Construcción de variable objetivo.
- Generación del dataset analítico final.

La integración de los datos permitió consolidar aproximadamente 7.2 millones de registros estudiantiles en un único archivo estructurado en formato Parquet, optimizando el rendimiento computacional y el procesamiento analítico, posteriormente se eliminaron:

- Variables identificadoras.
- Columnas redundantes.
- Registros inconsistentes.
- Campos sin relevancia predictiva.

Las variables categóricas de baja cardinalidad fueron transformadas mediante OneHotEncoder, mientras que variables de alta cardinalidad fueron codificadas mediante Target Encoding, estrategia recomendada para problemas tabulares de alta dimensionalidad y variables categóricas complejas [41], [42], para las variables numéricas se implementó RobustScaler debido a su menor sensibilidad frente a valores atípicos.

La variable objetivo fue construida a partir del puntaje global del examen Saber 11, transformándolo en cinco categorías ordinales:

- Muy bajo (0 - 100) puntos.
- Bajo (100 - 200) puntos.
- Medio (200 - 300) puntos.
- Alto (200 - 300) puntos.
- Muy alto (200 - 300) puntos.

La categorización se realizó utilizando percentiles del puntaje global, permitiendo estructurar un problema de clasificación multiclase.

Finalmente, el conjunto de datos fue dividido en subconjuntos de entrenamiento, validación y

prueba utilizando estratificación por clase para preservar la distribución original de las categorías.

Fase 4. Diseño e implementación de modelos: En esta etapa se implementaron diferentes arquitecturas de aprendizaje automático y aprendizaje profundo, los modelos desarrollados incluyeron:

- MLP con función de activación LeakyReLU.
- MLP con función de activación Softplus.
- CNN 1D.
- LightGBM.
- MLP con hiperparámetros optimizados mediante algoritmos genéticos.

Cada arquitectura fue desarrollada mediante notebooks independientes en Python 3.12.13 utilizando TensorFlow 2.20.0, Keras 3.13.2 Scikit-learn 1.6.1, NumPy 2.0.2, Matplotlib 3.10.0, SciPy 1.16.3 y LightGBM 4.6.0, las arquitecturas MLP fueron configuradas utilizando:

- Capas densas decrecientes.
- Batch Normalization.
- Dropout;
- Optimizador Adam;
- Funciones de activación no lineales como Relu, Leaky Relu y Sofplus.

Por su parte, la arquitectura CNN 1D fue adaptada para datos tabulares mediante convoluciones unidimensionales orientadas a identificar patrones locales entre variables transformadas.

Los modelos fueron entrenados utilizando múltiples iteraciones experimentales con el propósito de evaluar estabilidad estadística y reproducibilidad.

Fase 5. Evaluación estadística y validación: La fase de evaluación consistió en comparar el desempeño de los modelos mediante métricas estadísticas de clasificación, las principales métricas utilizadas fueron:

- F1-score.
- Precisión.
- Recall.
- Exactitud.
- Matrices de confusión.
- Estabilidad estadística.

Adicionalmente, se evaluó la dispersión estadística de cada modelo mediante desviación estándar, observando mayor estabilidad en las arquitecturas MLP frente a CNN y LightGBM.

Finalmente, se realizaron análisis comparativos para determinar diferencias de desempeño entre modelos y evaluar su capacidad de generalización, el desarrollo metodológico se organizó en seis fases:

OE1 — Selección de variables socioeconómicas relevantes del periodo 2014–2024: Actividad 1.1. Adquisición y consolidación de microdatos. Se descargaron los archivos de microdatos Saber 11 desde las fuentes oficiales del ICFES para el periodo 2014-2 al 2024-2. Posteriormente, se realizó la exploración inicial por periodo, la limpieza, estandarización y unificación de las bases históricas en un único conjunto consolidado, almacenado en formato Parquet para su procesamiento eficiente.

Actividad 1.2. Selección de variables predictoras. Con base en la revisión de la literatura y el análisis de correlación con la variable de salida, se identificaron y seleccionaron las variables socioeconómicas con mayor relevancia teórica y empírica para la explicación del desempeño académico en la prueba.

Actividad 1.3. Construcción de la variable objetivo y partición del conjunto de datos. Se definió la variable objetivo-categórica a partir del puntaje global Saber 11, utilizando niveles de desempeño oficiales. El conjunto resultante se dividió en subconjuntos de entrenamiento, validación y prueba con estratificación por clase.

OE2 — Implementación de modelos CNN 1D, MLP y sistemas expertos: Actividad 2.1. Diseño y configuración de arquitecturas. Se definieron las arquitecturas de los modelos MLP y CNN 1D, especificando capas, funciones de activación, optimizadores y criterios de parada temprana. Paralelamente, se configuró el sistema experto basado en reglas derivadas del conocimiento del dominio y de los patrones identificados en la fase anterior.

Actividad 2.2. Entrenamiento de los modelos. Cada modelo fue entrenado de forma independiente sobre el subconjunto de entrenamiento, con ajuste de hiperparámetros sobre el subconjunto de validación. Se documentaron las configuraciones finales de cada modelo para garantizar la reproducibilidad de los experimentos.

Actividad 2.3. Registro y control de experimentos. Se implementó un esquema de registro sistemático de los experimentos, incluyendo configuraciones, curvas de aprendizaje y métricas de validación por época, con el fin de asegurar la trazabilidad y comparabilidad de los modelos entrenados.

OE3 — Validación estadística del desempeño y comparación con modelos de referencia: Actividad 3.1. Evaluación de desempeño sobre el conjunto de prueba. Se calcularon las métricas de clasificación —exactitud, precisión, F1-score— para cada modelo sobre el subconjunto de prueba. Se analizaron las matrices de confusión para identificar patrones de error diferenciados por nivel de desempeño.

Actividad 3.2. Análisis estadístico de estabilidad. Se aplicó validación cruzada estratificada y pruebas estadísticas de comparación entre modelos (prueba de McNemar o equivalente) para determinar si las diferencias observadas en el desempeño son estadísticamente significativas.

Actividad 3.3. Interpretación y síntesis comparativa. Se analizó la contribución de las variables

socioeconómicas al desempeño de cada modelo mediante técnicas de importancia de características. Los resultados se sintetizaron en una comparación estructurada de los modelos en términos de desempeño predictivo, complejidad computacional y capacidad de generalización.

3.3 Fuentes de datos y criterios de selección

La fuente principal corresponde a los microdatos públicos anonimizados del examen Saber 11 publicados por el ICFES para el periodo 2014-2 a 2024-2. La base consolidada contiene más de siete millones de registros individuales y variables asociadas a características personales, familiares, socioeconómicas, escolares, institucionales y resultados académicos. La variable continua punt_global fue transformada en una variable categórica multiclase, permitiendo representar el desempeño académico en niveles: muy bajo (0 – 100) puntos, bajo (100 – 200) puntos, medio (200 – 300) puntos, alto (300 - 400) puntos y muy alto (400 – 500). puntos

Las bases de datos seleccionadas del repositorio de datos abiertos del Instituto Colombiano para la evaluación de la educación “ICFES” denominadas DataIcfes, las cuales cuentan con información anonimizadas de las evaluaciones estandarizadas del examen saber 11° en Colombia de los periodos desde el 20214.2 hasta 2024.2, lo que equivale 10 años aproximadamente de información las cuales cuentan con información sociodemográfico como información personal, información académica, información socioeconómica [39].

Estas bases de datos cuentan con información entre 86 columnas y 7239505 filas, cada variable depende del periodo en que se realizó el examen del saber 11° durante el periodo 2014.2 – 2024.2, como se muestra en la tabla 1 Variables sociodemográficas por tipo de información [39].

Tabla 3. *Variables Sociodemográficas por tipo de información.*

Grupo de Información	Nombre de la Variable (Campo)	Descripción Breve
Información Personal	estu_tipodocumento	Tipo de documento de identidad del estudiante.
	estu_consecutivo	Identificador único y público del estudiante en el Examen.
	estu_nacionalidad	País de nacionalidad del estudiante evaluado.
	estu_genero	Género (Masculino o Femenino) del estudiante.
	estu_fechanacimiento	Fecha de nacimiento registrada del estudiante.
	estu_pais_reside	País de residencia del estudiante.
	estu_depto_reside	Nombre del departamento de residencia del estudiante.
	estu_cod_reside_depto	Código DANE del departamento de residencia.
	estu_mcpio_reside	Nombre del municipio de residencia del estudiante.
	estu_cod_reside_mcpio	Código DANE del municipio de residencia.
	estu_tieneetnia	Indica si el estudiante se reconoce como parte de una etnia.
	estu_privado_libertad	Indica si el estudiante presenta el examen en condición de reclusión.
	estu_depto_presentacion	Departamento donde el estudiante presentó el examen.
	cod_depto_presentacion	Código DANE del departamento de presentación.
	estu_mcpio_presentacion	Municipio donde el estudiante presentó el examen.
	cod_mcpio_presentacion	Código DANE del municipio de presentación.
Información Académica	periodo	Año y semestre de aplicación de la prueba (ej: 20242).
	estu_estudiante	Define si la inscripción fue individual o por establecimiento.
	estu_grado	Grado escolar que cursa el estudiante al momento del examen.
	estu_repite	Indica si el estudiante ha repetido años escolares previamente.
	estu_estadoinvestigacion	Estado de validez o investigación del resultado obtenido.
	estu_pilopaga	Indica si es potencial beneficiario del programa Ser Pilo Paga.
	estu_generacione	Indica si es beneficiario del programa Generación E.
	punt_lectura_critica	Puntaje numérico en la prueba de Lectura Crítica.
	percentil_lectura_critica	Percentil nacional en la prueba de Lectura Crítica.
	desemp_lectura_critica	Nivel de desempeño cualitativo en Lectura Crítica.
	punt_matematicas	Puntaje numérico en la prueba de Matemáticas.
	percentil_matematicas	Percentil nacional en la prueba de Matemáticas.
	desemp_matematicas	Nivel de desempeño cualitativo en Matemáticas.
	punt_c_naturales	Puntaje numérico en la prueba de Ciencias Naturales.
	percentil_c_naturales	Percentil nacional en la prueba de Ciencias Naturales.
	desemp_c_naturales	Nivel de desempeño cualitativo en Ciencias Naturales.
	punt_sociales_ciudadanas	Puntaje numérico en la prueba de Sociales y Ciudadanas.
	percentil_sociales_ciudadanas	Percentil nacional en la prueba de Sociales y Ciudadanas.
	desemp_sociales_ciudadanas	Nivel de desempeño cualitativo en Sociales y Ciudadanas.
	punt_ingles	Puntaje numérico en la prueba de inglés.
	percentil_ingles	Percentil nacional en la prueba de inglés.
	desemp_ingles	Nivel de desempeño según el Marco Común Europeo para inglés.
	punt_global	Suma ponderada de puntajes de las 5 áreas (escala 0-500).
	percentil_global	Percentil general del estudiante a nivel nacional.

Grupo de Información	Nombre de la Variable (Campo)	Descripción Breve
Información Personal	estu_tipodocumento	Tipo de documento de identidad del estudiante.
	estu_consecutivo	Identificador único y público del estudiante en el Examen.
	estu_nacionalidad	País de nacionalidad del estudiante evaluado.
	estu_genero	Género (Masculino o Femenino) del estudiante.
	estu_fechanacimiento	Fecha de nacimiento registrada del estudiante.
	estu_pais_reside	País de residencia del estudiante.
	estu_depto_reside	Nombre del departamento de residencia del estudiante.
	estu_cod_reside_depto	Código DANE del departamento de residencia.
	estu_mcpio_reside	Nombre del municipio de residencia del estudiante.
	estu_cod_reside_mcpio	Código DANE del municipio de residencia.
	estu_tieneetnia	Indica si el estudiante se reconoce como parte de una etnia.
	estu_privado_libertad	Indica si el estudiante presenta el examen en condición de reclusión.
	estu_depto_presentacion	Departamento donde el estudiante presentó el examen.
	cod_depto_presentacion	Código DANE del departamento de presentación.
	estu_mcpio_presentacion	Municipio donde el estudiante presentó el examen.
	cod_mcpio_presentacion	Código DANE del municipio de presentación.
Planteles (Colegios)	cole_codigo_icfes	Código asignado por el ICFES a la sede educativa.
	cole_cod_dane_establecimiento	Código DANE que identifica a la institución educativa.
	cole_nombre_establecimiento	Nombre oficial de la institución educativa principal.
	cole_genero	Población estudiantil que atiende (Masculino/Femenino/Mixto).
	cole_naturaleza	Tipo de administración del colegio (Oficial o No Oficial).
	cole_calendario	Calendario académico de la institución (A o B).
	cole_caracter	Tipo de formación ofrecida (Académico/Técnico/Otros).
	cole_cod_dane_sede	Código DANE específico de la sede donde estudia el alumno.
	cole_nombre_sede	Nombre de la sede específica de la institución.
	cole_sede_principal	Indica si la sede es la principal de la institución.
	cole_area_ubicacion	Ubicación geográfica de la sede (Urbana o Rural).
	cole_cod_mcpio_ubicacion	Código DANE del municipio donde se ubica el colegio.
	cole_mcpio_ubicacion	Nombre del municipio donde se encuentra el plantel.
	cole_cod_depto_ubicacion	Código DANE del departamento donde se ubica el colegio.
	cole_depto_ubicacion	Nombre del departamento donde se encuentra el plantel.
	cole_jornada	Jornada escolar asignada al estudiante.
Información Socioeconómica	fami_estratovivienda	Estrato de los servicios públicos de la vivienda del hogar.
	fami_personashogar	Número total de personas que conforman el hogar.
	fami_cuartoshogar	Número de habitaciones exclusivas para dormir en el hogar.
	fami_educacionpadre	Nivel educativo más alto alcanzado por el padre.
	fami_educacionmadre	Nivel educativo más alto alcanzado por la madre.
	fami_ocupacionpadre	Ocupación u oficio principal desempeñado por el padre.

Grupo de Información	Nombre de la Variable (Campo)	Descripción Breve
Información Socioeconómica	fami_ocupacionmadre	Ocupación u oficio principal desempeñado por la madre.
	fami_tieneinternet	Indica si el hogar cuenta con servicio de internet.
	fami_tienecomputador	Indica si en el hogar hay disponibilidad de computadora.
	fami_tienelavadora	Indica si el hogar cuenta con lavadora de ropa.
	fami_tienehornomicroogas	Indica si el hogar cuenta con horno microondas o de gas.
	fami_tieneserviciotv	Indica si el hogar cuenta con televisión por cable o satélite.
	fami_tieneautomovil	Indica si en el hogar se dispone de automóvil particular.
	fami_tienemotocicleta	Indica si en el hogar se dispone de motocicleta.
	fami_tieneconsolavideojuegos	Indica si en el hogar hay consola de videojuegos.
	fami_numlibros	Cantidad aproximada de libros disponibles en el hogar.
	fami_comecarnepescadohuevo	Frecuencia de consumo de carnes o huevos en la semana.
	fami_comecerealfrutoslegumbre	Frecuencia de consumo de cereales y frutas en la semana.
	fami_comelechederivados	Frecuencia de consumo de leche y lácteos en la semana.
	estu_dedicacionlecturadiaria	Tiempo que el estudiante dedica diariamente a la lectura.
	estu_dedicacioninternet	Tiempo que el estudiante dedica diariamente al uso de internet.
	estu_horassemanatrabaja	Número de horas que el estudiante labora por semana.
	estu_tiporemuneracion	Tipo de compensación recibida por el trabajo del estudiante.
	estu_inse_individual	Índice de Nivel Socioeconómico calculado individualmente.
	estu_nse_individual	Nivel Socioeconómico del estudiante (escala 1 a 4).
	estu_nse_establecimiento	Nivel Socioeconómico promedio del colegio.

Nota: La tabla fue elaboración basado en data ICFES.

Esta información disponible de 10 años de los periodos del 2014.2 al 2024.2 se realizará la unificación de la información las cuales están en formato .txt, para formar una sola base de datos de tipo paquet, así mismo se tomará la variable “punt_globla” puntaje total obtenido por el estudiante dividas en cinco categorías, “muy bajo (0 – 100) puntos; bajo (100 – 200) puntos; medio (200 – 300) puntos; alto (300 - 400) puntos; muy alto (400 – 500). puntos” como una variable clase e implementar un modelo de clasificación del desempeño académico a través de las 45 categorías

de los resultados del examen del saber 11° usando redes neuronales convolucionales y las variables sociodemográficas**.**

El examen Saber 11 constituye una prueba estandarizada nacional orientada a evaluar competencias genéricas en: lectura crítica; matemáticas; ciencias naturales; sociales y ciudadanas; e inglés.

Además de los puntajes académicos, los microdatos incluyen información asociada a: condiciones socioeconómicas; características familiares; acceso a recursos tecnológicos; información institucional; y contexto geográfico, la selección de esta fuente de información se justifica por múltiples razones, las cuales se seleccionaron los siguiente, así:

Tabla 4. *Selección de Variables socioeconómicas para implementar modelos de IA.*

N°	Nombre Variable	Grupo Información
1	cole_area_ubicacion	Planteles (Colegio)
2	cole_bilingue	
3	cole_calendario	
4	cole_caracter	
5	cole_depto_ubicacion	
6	cole_genero	
7	cole_jornada	
8	cole_mcpio_ubicacion	
9	cole_naturaleza	
10	estu_depto_reside	Información Personal
11	estu_genero	
12	estu_mcpio_reside	
13	estu_nacionalidad	
14	estu_privado_libertad	
15	fami_cuartoshogar	Información Socioeconómica
16	fami_educacionmadre	
17	fami_educacionpadre	
18	fami_estratovivienda	
19	estu_inse_individual	
20	estu_nse_individual	
21	fami_personashogar	
22	fami_tieneautomovil	
23	fami_tienecomputador	
24	fami_tieneinternet	

N°	Nombre Variable	Grupo Información
24	fami_tieneinternet	Información Socioeconómica
25	fami_tienelavadora	
26	fami_tieneserviciotv	
27	fami_numlibros	
28	estu_discapacidad	
29	estu_nse_establecimiento	
30	estu_repite	
31	estu_dedicacioninternet	
32	estu_dedicacionlecturadiaria	
33	estu_horassemanatrabaja	
34	estu_tiporemuneracion	
35	fami_comecarnepescadohuevo	
36	fami_comecerealfrutoslegumbre	
37	fami_comelechederivados	
38	fami_situacioneconomica	
39	fami_tieneconsolavideojuegos	
40	fami_tienehornomicroogas	
41	fami_tienemotocicleta	
42	fami_trabajolabormadre	
43	fami_trabajolaborpadre	
44	target	información Académica
45	periodo	
46	estu_grado	

Nota: La tabla fue elaboración basado en data ICFES.

La elección del periodo 2014-2 a 2024-2 se fundamenta en criterios de comparabilidad estadística y consistencia metodológica establecidos por el ICFES.

La documentación oficial indica que desde 2014-2 el examen Saber 11 presenta una estructura homogénea compatible con análisis longitudinales, como criterios de selección de variables se consideraron:

- Relevancia teórica;
- Disponibilidad histórica;
- Calidad estadística;
- Potencial predictivo.

Se excluyeron variables: identificadoras; redundantes; con alta proporción de valores faltantes; o sin relación conceptual con el desempeño académico como se representa en la tabla4.

Adicionalmente, las variables fueron clasificadas según: tipo de dato; cardinalidad; distribución estadística; y comportamiento temporal.

El almacenamiento final se realizó en formato Parquet debido a sus ventajas en:

- Compresión;
- Velocidad de lectura;
- Optimización de memoria;
- Procesamiento distribuido.

3.4 Herramientas, técnicas y tecnologías empleadas

El desarrollo del proyecto requirió la integración de múltiples herramientas computacionales orientadas al procesamiento masivo de datos, aprendizaje automático, aprendizaje profundo y análisis estadístico, por ello se requiere un lenguaje flexible, con amplia ecosistemas de bibliotecas como Scikit-learn, TensorFlow, PyTorch, Pandas, Numpy, compatibilidad con la inteligencia artificial, y amplia adopción científica, entre las bibliotecas utilizadas se encuentran:

Pandas 2.2.2:

- Manipulación tabular.
- Limpieza.
- Integración.
- Transformación de datos.

NumPy 2.0.2:

- Empleada para operaciones matriciales y procesamiento numérico.

Scikit-learn 1.6.1:

- Preprocesamiento;
- Partición de datos;
- Escalamiento;
- Codificación;
- Métricas estadísticas.

TensorFlow 2.20 - Keras 3.13.2:

- Implementadas para el diseño y entrenamiento de redes neuronales profundas.

LightGBM 4.6.0

- Utilizada para implementar modelos boosting optimizados para datos tabulares.

Matplotlib 3.10.0:

- Utilizadas para visualización estadística y análisis gráfico.

PyArrow 18.1:

Empleada para procesamiento eficiente de archivos Parquet, dentro de las técnicas implementadas se encuentran:

- Imputación de valores faltantes;
- Escalamiento robusto;
- Codificación categórica;
- Regularización;
- Normalización por lotes;
- Validación estratificada;
- Optimización mediante algoritmos genéticos.

El entorno de ejecución principal correspondió a Google Colab, permitiendo acceso a recursos

computacionales escalables y aceleración mediante GPU, los entornos de ejecución G4 (RAM 176 GB – GPU 95.6 – Disco 112.6 GB), Entorno de Ejecución L4 (RAM 53 GB – GPU 22.5 – Disco 112.6 GB),

El preprocesamiento híbrido integró RobustScaler, OneHotEncoder y TargetEncoder.

- El uso de RobustScaler permitió reducir sensibilidad frente a valores atípicos.
- OneHotEncoder fue utilizado para variables nominales de baja cardinalidad.
- TargetEncoder permitió reducir dimensionalidad en variables categóricas complejas.

Las arquitecturas profundas utilizaron los diferentes tipos de características de hiperparametros como la función de activación “ReLu, LeakyReLU, Softplus, Dropout”, Batch Normalization, función de optimización “Adam”; y el método de regularización “Early Stopping”, lo que permite a la red neuronal multicapa aprender los diferentes parámetros no lineales y evitando el sobreajuste “overfiting”.

El preprocesamiento incluyó imputación de valores faltantes, depuración de variables no predictivas o identificadoras, escalamiento robusto de variables numéricas, codificación one-hot para variables categóricas de baja cardinalidad y codificación por objetivo (target encoding) para variables categóricas de alta cardinalidad. El uso de RobustScaler se justifica por su menor sensibilidad a valores extremos, lo que permite mejorar la estabilidad del modelo en presencia de outliers [43], mientras que OneHotEncoder resulta apropiado para variables nominales sin orden intrínseco [41]. Para variables de alta cardinalidad, el target encoding permite reducir la dimensionalidad y conservar información predictiva [42], aunque debe aplicarse exclusivamente sobre el conjunto de entrenamiento para evitar fuga de información (data leakage), lo cual puede generar sobreestimación del desempeño del modelo [44].

En el mundo de la inteligencia artificial existen diferentes arquitecturas como el Machine

Learning “ML”, redes neuronales profundas “DNN”, Inteligencia Cuántica “QAI”, Inteligencia Artificial Generativa “IAGen”, entre otras, las cuales permiten simular circuitos neuronales humanos facilitando generación de contenido, predicciones y clasificación de la información, permitiendo realizar diferentes tareas facilitando la solución de problemas [28].

En la arquitectura de machine learning se destaca tres tipo de aprendizaje automático como el aprendizaje supervisado, aprendizaje no supervisado y el aprendizaje reforzado, los cuales se diferencia según los datos de entrenamiento, para el aprendizaje supervisado se hace uso de los datos etiquetados los cuales permiten realizar las predicciones futuras, en el aprendizaje supervisado se encuentra la subcategoría de predicción por Etiquetas con datos discretos las cuales se centra en la observaciones pasadas y las predicciones con resultados continuos se encuentran denominada de análisis de regresión, con ello permita clasificar o predecir los nuevos datos de entrada [45].

Por parte del aprendizaje no supervisado cuenta con los subgrupos de agrupamiento o por clúster, los cuales permiten definir grupos que comparten semejanzas, pero difieren de otros grupos, el subgrupo de reducción de dimensionalidad es utilizando para eliminar ruidos de los datos o comprimir datos a un espacio dimensional facilitando la visualización en dimensiones de una a tres dimensiones [45].

Los modelos de redes neuronales artificiales realiza su proceso de abstracción de patrones de datos y funciones complejas de forma automática, las redes neuronales se encuentra compuestas por múltiples capas de procesamientos y múltiples niveles de abstracción lineal y no lineales, las capas de las redes neuronales están compuestas por “Imput Layer, Hidder Layer, Output Layer”, de las cuales pueden existir múltiples capas ocultas “Hidder Layer” con una variedades de neuronas interconectadas, cada neurona cuenta con un valor de sesgo “b” el cual tiene el valor de uno (1), de igual forma cada neurona cuenta con una función de activación no lineal “Tanh, ReLu, Leaky

Relu, SoftPlus,", que permite modelar patrones complejos, y las funciones de la capas de salida cuentan con funciones de activación como (Softmax, Sigmoid) según caso de uso del problema de clasificación binaria o multiclase [46].

De igual forma las redes neuronales cuentan con el optimizador iterativo "Descenso del Gradiente", el cual permite graduar los parámetros del modelo para minimizar la función del costo o función perdida hasta llegar al parámetro con el mínimo global, así mismo el Learning rate es el rango escalar de aprendizaje del gradiente, estableciendo la velocidad del aprendizaje, así mismo el learnin rate decay el cual realiza el rango de aprendizaje de forma progresiva medida que va ejecutando las epochs; así mismo para seleccionar la función de pérdida adecuada, hay que tener en cuenta el formato de salida, para el caso de formatos categóricos múltiple se utilizara "categorical_crossentropy", cuando la clasificación es binaria se usará "binary_crossentropy", y cuando es de tipo de problema de regresión se utiliza "Mean Squared Error - mse" [46], con estos hiperparametros permitirá estructurar una red neuronales profundas "DNN" para el procesamiento y predicción de los datos.

En las arquitecturas de las redes neuronales se puede encontrar las redes neuronales convolucionales "CNN", las cuales son especializadas en manejar datos espaciales a través de las matemáticas, tratando los pixeles de las imágenes para extraer e interpretar la correlación de la información de estructura compleja como las imágenes obteniendo datos relevantes considerando el contexto del entorno inmediato; las redes neuronales convolucionales realizan el análisis y modificación de los datos recorriendo las imágenes a través matriz del filtro de núcleo compuesto por cero menos la posición central, con ello selecciona un fragmento de la imagen para realizar una operación de producto punto, este proceso se repite desplazándose por toda la imagen hasta lograr la convolucional final [47].

En las redes neuronales convolucionales cuenta con tres tipos de filtro, núcleo de detección de

borde, núcleo de detención de relieve y el núcleo de resplandor suave, las cuales permite destacar y resaltar las diferentes características de la imagen, después de cada proceso convolucional se realiza una operación de agrupación o de reducción llamada Pooling, esta capa permite extraer los valores más representativos de cada región manteniendo los detalles principales sin modificación, lo que conlleva que una estructura de red CNN debe ser lógica y jerárquica, donde las primera capa permitiendo identificar elementos básicos como bordes y texturas, las capas posteriores "Intermedias" interpretan patrones complejos, logrando identificar los objetos u elementos que están representados en la imagen y las capas finales de la CNN son capas densas, que permite integrar, esquematizar las características de las imágenes y facilitar la clasificación de los datos a través del proceso de entrenamiento del backpropagation lo que permite ajustar sus núcleo de forma gradual, y automática, optimizando la predicción de forma precisa y efectiva [47].

Por eso es de gran importancia realizar pruebas y comparaciones entre los diferentes modelos de inteligencia artificial, ya que cada modelo tiene sus sesgos, por lo que se requiere realizar pruebas con las diferentes arquitecturas y seleccionar el mejor rendimiento y precisión en el entrenamiento y validación de la clasificación o predicción de los datos [45].

Además, para garantizar la validez experimental, el ajuste de codificadores supervisados como TargetEncoder debe realizarse únicamente con los datos de entrenamiento. Aplicarlo antes de la partición entrenamiento-validación-prueba puede introducir una fuga de información (data leakage), dado que el codificador utiliza información de la variable objetivo. Por tanto, la versión final del pipeline debe estructurarse como: división inicial de datos, ajuste del preprocesador con $X_{train} e y_{train}$e, transformación de los conjuntos de validación y prueba, y posterior entrenamiento del modelo [44], [49]. Este principio es fundamental en el diseño de pipelines de aprendizaje automático, ya que la presencia de fuga de información puede generar estimaciones optimistas del desempeño del modelo y comprometer su capacidad de generalización en datos no observados [44].

3.5 Métodos de análisis y criterios de evaluación

La evaluación metodológica de los modelos implementados fue diseñada bajo un enfoque comparativo experimental orientado a garantizar consistencia estadística, reproducibilidad y capacidad de generalización, siguiendo recomendaciones metodológicas de validación experimental en aprendizaje automático y minería de datos educativa [15], [49]. En consecuencia, el proceso de validación se estructuró considerando principios ampliamente utilizados en aprendizaje automático supervisado y minería de datos educativa.

El problema abordado en la investigación corresponde a un problema de clasificación multiclase, donde la variable objetivo representa categorías ordinales del desempeño académico derivadas del puntaje global de la prueba Saber 11. Debido a la naturaleza categórica y parcialmente desbalanceada del problema, se definió un conjunto de métricas orientadas a evaluar tanto el desempeño global como el comportamiento específico de cada modelo frente a las diferentes clases.

La evaluación experimental se estructuró bajo un esquema de separación estratificada de datos en subconjuntos de entrenamiento, validación y prueba, preservando la distribución proporcional de las categorías objetivo. Este enfoque permite reducir sesgos asociados al desbalance de clases y mejorar la representatividad estadística de cada subconjunto.

Con el propósito de garantizar la reproducibilidad experimental, cada arquitectura fue entrenada mediante múltiples iteraciones independientes, permitiendo evaluar la estabilidad de las predicciones y la consistencia estadística del comportamiento de los modelos frente a diferentes inicializaciones y configuraciones.

Las métricas definidas para evaluar el desempeño de los modelos corresponden a:

Exactitud (Accuracy): La exactitud representa la proporción total de predicciones correctas respecto al número total de observaciones. Esta métrica permite obtener una visión general del desempeño del modelo; sin embargo, puede verse afectada en escenarios de desbalance de clases.

Precisión (Precision): La precisión mide la proporción de observaciones clasificadas correctamente dentro del total de predicciones positivas realizadas por el modelo. Esta métrica resulta particularmente útil para evaluar la confiabilidad de las predicciones generadas.

Exhaustividad (Recall): El recall evalúa la capacidad del modelo para identificar correctamente las observaciones pertenecientes a cada categoría real. Esta métrica es relevante en problemas educativos debido a que permite identificar qué tan eficiente es el modelo detectando estudiantes pertenecientes a determinados niveles de desempeño.

F1-score: El F1-score corresponde a la media armónica entre precisión y recall, constituyéndose en una métrica robusta para problemas de clasificación multiclase y conjuntos parcialmente desbalanceados. Debido a que combina simultáneamente precisión y exhaustividad, el F1-score fue definido como la principal métrica comparativa del proyecto.

La literatura reciente relacionada con clasificación educativa basada en aprendizaje automático recomienda el uso del F1-score como indicador principal en contextos donde existe heterogeneidad entre clases y diferencias en distribución poblacional, especialmente en escenarios educativos con distribución desigual entre categorías de desempeño [13], [15], [33].

Matriz de confusión: Las matrices de confusión permiten analizar el comportamiento específico de cada modelo respecto a las clases objetivo, identificando patrones de error, confusiones entre categorías y comportamiento diferencial entre niveles de desempeño académico.

Diseño metodológico de validación experimental: El diseño experimental contempló la comparación sistemática entre diferentes arquitecturas de aprendizaje automático y aprendizaje profundo utilizando un protocolo homogéneo de entrenamiento y evaluación.

Cada modelo fue implementado bajo condiciones experimentales equivalentes considerando:

- Conjunto de entrenamiento;
- Variables predictoras;
- Variable objetivo;
- Subconjuntos de validación;
- Conjunto de prueba.

Esta estrategia metodológica permite garantizar comparabilidad experimental y minimizar sesgos derivados de diferencias en la preparación de los datos.

Asimismo, se implementaron técnicas orientadas a reducir sobreajuste y mejorar capacidad de generalización, incluyendo:

- Batch Normalization;
- Dropout;
- Early Stopping;
- Regularización implícita;
- Ajuste iterativo de hiperparámetros.

En el caso de las redes neuronales profundas, el proceso de entrenamiento contempló múltiples épocas de aprendizaje y monitoreo continuo de métricas de validación para identificar convergencia y estabilidad.

Por otra parte, el proyecto incorporó mecanismos de control frente a fuga de información (*data leakage*), particularmente durante el uso de Target Encoding. Para evitar contaminación entre

subconjuntos, el ajuste de codificadores supervisados fue realizado exclusivamente sobre los datos de entrenamiento antes de transformar los conjuntos de validación y prueba.

La literatura especializada establece que la fuga de información constituye uno de los principales riesgos metodológicos en proyectos de aprendizaje automático, debido a que puede generar sobreestimaciones artificiales del desempeño predictivo [44].

Validación estadística: La evaluación estadística de los modelos fue diseñada para analizar.

- Estabilidad experimental;
- Variabilidad de resultados;
- Consistencia de entrenamiento;
- Capacidad de generalización.

Para ello se contemplaron:

- Análisis descriptivos;
- Medidas de tendencia central;
- Medidas de dispersión;
- Validación iterativa;
- Pruebas estadísticas de normalidad.

Adicionalmente, se consideró la aplicación de pruebas estadísticas comparativas orientadas a determinar si las diferencias observadas entre modelos poseen significancia estadística.

Entre las pruebas contempladas se encuentran:

- Prueba de Shapiro-Wilk para evaluación de normalidad;
- Pruebas paramétricas “ANOVA de Welch” para comparación de distribuciones;
- Análisis comparativos entre iteraciones experimentales.

La incorporación de validación estadística permite fortalecer el rigor científico del proyecto y

garantizar que las diferencias observadas entre modelos no correspondan exclusivamente a variaciones aleatorias derivadas del entrenamiento.

Metodología de modelos implementados: Con el propósito de estructurar comparativamente las arquitecturas evaluadas en la investigación, la Tabla 3 presenta las principales metodologías y configuraciones conceptuales implementadas.

La siguiente tabla presenta las principales metodologías y configuraciones conceptuales implementadas.

Tabla 5. *Modelos y metodologías de Arquitectura de Inteligencia artificial.*

Modelo	Tipo de arquitectura	Técnica principal	Función de activación	Objetivo metodológico
MLP LeakyReLU V1	Red neuronal profunda	Capas densas	LeakyReLU	Establecer baseline inicial
MLP LeakyReLU V2	Red neuronal profunda	Capas densas optimizadas	LeakyReLU	Ajuste arquitectónico y estabilidad
MLP Softplus	Red neuronal profunda	Capas densas	Softplus	Evaluar suavidad de convergencia
MLP Softplus + Decay	Red neuronal profunda	Ajuste dinámico de aprendizaje	Softplus	Reducir oscilaciones de entrenamiento
CNN 1D	Red neuronal convolucional	Convolución unidimensional	ReLU	Capturar patrones locales en datos tabulares
LightGBM	Ensemble Boosting	Árboles de decisión graduales	No aplica	Modelo comparativo robusto para datos tabulares
Algoritmos Genéticos	Optimización evolutiva	Búsqueda heurística	Variable	Optimización de hiperparámetros

Nota: La elaboración de la tabla fue basada estructura de las arquitecturas de IA.

La comparación metodológica de estas arquitecturas permite analizar diferentes enfoques de aprendizaje supervisado aplicados al contexto educativo colombiano, considerando tanto modelos basados en árboles como arquitecturas profundas y métodos evolutivos.

4 Desarrollo de la solución.

4.1 Arquitectura general de la solución.

La solución desarrollada corresponde a una arquitectura analítica modular implementada en Python sobre Google Colab. El flujo general inicia con la carga de microdatos Saber 11 por periodo, continúa con la limpieza y homologación de variables, consolida la información en un archivo Parquet, ejecuta un pipeline de preprocesamiento híbrido y finaliza con el entrenamiento comparativo de modelos supervisados. Este tipo de arquitectura es consistente con los enfoques modernos de analítica de datos para grandes volúmenes de información estructurada, donde se integran procesos de ingestión, transformación y modelado en pipelines reproducibles [18], [50]. La arquitectura se diseñó para soportar datos tabulares masivos, alta cardinalidad en variables territoriales y heterogeneidad entre variables numéricas y categóricas.

El diseño arquitectónico se fundamentó en principios de escalabilidad, trazabilidad, modularidad y reproducibilidad científica, aspectos ampliamente recomendados en proyectos de Educational Data Mining (EDM) y Learning Analytics orientados al procesamiento de grandes volúmenes de datos educativos [49]. La estructura general del sistema fue concebida como un pipeline analítico compuesto por etapas secuenciales de adquisición de datos, transformación, entrenamiento supervisado, validación estadística y comparación experimental.

La arquitectura implementada se diseñó considerando las particularidades del conjunto de datos utilizado, el cual contiene aproximadamente 7.2 millones, con 45 variables socioeconómicos de los registros estudiantiles correspondientes al periodo 2014-2 a 2024-2. La magnitud del dataset obligó a utilizar estrategias de optimización de almacenamiento y procesamiento compatibles con escenarios de Big Data educativo. En consecuencia, el sistema incorporó almacenamiento columnar en formato Parquet, lo que permitió reducir consumo de memoria, acelerar lectura de

archivos y mejorar el rendimiento general del pipeline analítico.

La primera capa funcional de la arquitectura corresponde al subsistema de adquisición y consolidación de datos. Esta etapa fue desarrollada mediante notebooks especializados orientados a la lectura y homologación de archivos históricos del ICFES. Debido a que los microdatos oficiales presentan diferencias estructurales entre periodos, fue necesario implementar procedimientos de estandarización de variables y validación semántica para garantizar consistencia longitudinal. La documentación oficial DataIcfes permitió identificar los campos relevantes asociados a información socioeconómica, académica e institucional, garantizando coherencia metodológica entre los diferentes periodos históricos analizados.

Posteriormente, la arquitectura incorpora una segunda capa asociada al pipeline híbrido de preparación y transformación de datos. Esta fase constituye uno de los componentes más importantes del sistema debido a que los modelos supervisados utilizados requieren representaciones numéricas estructuradas y estadísticamente consistentes. La literatura especializada en aprendizaje automático aplicado a datos tabulares establece que la calidad del preprocesamiento posee un impacto directo sobre el desempeño predictivo de los modelos [41].

En este contexto, la arquitectura implementó una estrategia híbrida de transformación compuesta por RobustScaler, OneHotEncoder y TargetEncoder. Las variables numéricas fueron transformadas mediante RobustScaler debido a la presencia de distribuciones asimétricas y valores extremos frecuentes en variables socioeconómicas. Pedregosa et al. en el 2011 señaló que este tipo de escalamiento resulta particularmente adecuado cuando los datos presentan sensibilidad frente a outliers, ya que utiliza medidas robustas basadas en mediana y rango intercuartílico.

Por otra parte, las variables categóricas nominales de baja cardinalidad fueron procesadas mediante OneHotEncoder, permitiendo representar categorías discretas sin introducir relaciones ordinales artificiales. Las variables categóricas de alta cardinalidad, particularmente aquellas

relacionadas con municipios, instituciones educativas y características territoriales, fueron transformadas mediante Target Encoding. Esta técnica permitió reducir dimensionalidad preservando capacidad predictiva, estrategia ampliamente utilizada en problemas tabulares complejos de gran escala [42].

Un aspecto metodológico fundamental dentro de la arquitectura corresponde al control de fuga de información (*data leakage*). Kaufman et al. en el 2012) estableció que la fuga de información constituye uno de los principales riesgos metodológicos en proyectos de minería de datos, dado que puede producir sobreestimaciones artificiales del desempeño predictivo. En consecuencia, la arquitectura implementada ajustó los codificadores supervisados exclusivamente sobre los datos de entrenamiento antes de transformar los subconjuntos de validación y prueba, garantizando independencia estadística entre conjuntos.

La tercera capa arquitectónica corresponde al sistema de modelado supervisado. Esta capa fue diseñada como un entorno experimental comparativo orientado a evaluar diferentes familias de algoritmos de inteligencia artificial aplicados a clasificación multiclase. Las arquitecturas implementadas pertenecen a dos grandes grupos: modelos de aprendizaje profundo y modelos boosting basados en árboles.

Dentro de las arquitecturas profundas se desarrollaron modelos MLP (Multilayer Perceptron) y CNN 1D (Convolutional Neural Networks), mientras que la familia boosting fue representada mediante LightGBM. La decisión de implementar múltiples arquitecturas responde a la necesidad metodológica de comparar diferentes paradigmas de aprendizaje supervisado sobre un mismo conjunto de datos, evitando asumir previamente la superioridad de un modelo específico. Esta aproximación experimental se encuentra alineada con investigaciones recientes de minería de datos educativa y clasificación tabular avanzada [15], [18].

Las arquitecturas MLP fueron diseñadas para capturar relaciones no lineales complejas entre

variables socioeconómicas y desempeño académico. Las redes neuronales profundas poseen capacidad para modelar interacciones de alta dimensionalidad mediante transformaciones jerárquicas sucesivas, característica particularmente relevante en contextos educativos donde múltiples variables interactúan simultáneamente [51].

La arquitectura CNN 1D fue adaptada al contexto de datos tabulares con el propósito de evaluar la capacidad de las convoluciones unidimensionales para capturar patrones locales y dependencias implícitas entre variables estructuradas. Aunque las redes convolucionales fueron inicialmente desarrolladas para visión por computador, investigaciones recientes han demostrado que las convoluciones unidimensionales pueden generar representaciones relevantes en datos tabulares mediante identificación de patrones secuenciales entre características [18].

La arquitectura general incorpora además mecanismos de regularización y estabilización del entrenamiento profundo. Entre ellos destacan Batch Normalization, Dropout, y EarlyStopping. Batch Normalization fue incorporado para estabilizar la distribución interna de activaciones y acelerar la convergencia del entrenamiento, reduciendo sensibilidad frente a inicializaciones y mejorando estabilidad numérica [52]. Por su parte, Dropout fue implementado como mecanismo de regularización orientado a reducir sobreajuste mediante desactivación aleatoria de neuronas durante el entrenamiento [53].

Finalmente, la arquitectura incorpora una capa de validación experimental y análisis estadístico a través de la prueba de Shapiro-Wilk [54]. Esta etapa permite almacenar resultados experimentales, métricas de clasificación y comportamiento iterativo de los modelos mediante archivos CSV y notebooks especializados de análisis estadístico. La separación modular de componentes facilita reproducibilidad científica, auditoría metodológica y extensión futura del sistema hacia arquitecturas adicionales.

4.2 Diseño del modelo, sistema o metodología

En el diseño del modelo y sistemas inteligentes se implementaron y compararon tres familias de modelos. La primera corresponde a redes neuronales densas tipo MLP, evaluadas con funciones de activación Relu, LeakyReLU y Softplus, normalización por lotes, Dropout y optimización Adam. Este tipo de arquitectura ha demostrado ser eficaz en la modelación de relaciones no lineales en datos tabulares [55], [56].

La segunda corresponde a una CNN 1D adaptada a datos tabulares, con el propósito de evaluar si las convoluciones unidimensionales pueden capturar patrones locales entre variables transformadas, una línea de investigación reciente en Deep learning aplicado a datos estructurados [18].

La última familia corresponde al modelo de machine learning "LightGBM", seleccionado como modelo de referencia por su alto desempeño reportado en datos tabulares y problemas de clasificación supervisada, especialmente en contextos de alta dimensionalidad y heterogeneidad de variables [13], [50].

El diseño metodológico de la solución fue estructurado como un entorno experimental comparativo orientado a evaluar diferentes arquitecturas de inteligencia artificial sobre un problema de clasificación multiclase relacionado con el desempeño académico en la prueba Saber 11. La variable objetivo fue construida a partir del puntaje global oficial del examen, transformándolo en categorías ordinales de desempeño mediante técnicas de discretización basadas en percentiles.

Esta transformación permitió convertir el problema original en una tarea de clasificación supervisada multiclase, facilitando la evaluación comparativa de arquitecturas profundas y modelos boosting. La categorización del puntaje global responde además a necesidades prácticas asociadas a interpretación educativa y generación de alertas tempranas en contextos institucionales.

El diseño experimental del sistema se fundamentó en investigaciones recientes relacionadas con minería de datos educativa, aprendizaje profundo y clasificación supervisada de datos tabulares. Estudios desarrollados por [59] y [66] evidencian que las variables socioeconómicas y familiares poseen capacidad predictiva significativa sobre el desempeño académico, particularmente cuando son procesadas mediante modelos no lineales de aprendizaje automático.

Diseño de arquitecturas MLP: Las arquitecturas MLP implementadas en la investigación fueron diseñadas utilizando capas densas profundas organizadas jerárquicamente. Cada arquitectura incluyó múltiples capas Dense con reducción progresiva de dimensionalidad, permitiendo aprendizaje de representaciones abstractas de las variables de entrada.

Las redes neuronales profundas implementadas utilizaron funciones de activación LeakyReLU y Softplus. La función LeakyReLU fue seleccionada debido a su capacidad para reducir el problema de neuronas muertas asociado a ReLU tradicional [56], demostraron que LeakyReLU permite mantener flujo de gradiente para valores negativos, favoreciendo estabilidad durante el entrenamiento profundo.

Softplus fue incorporada como función de activación alternativa debido a su suavidad matemática y continuidad diferencial la cual se calcula . Esta función permite reducir discontinuidades abruptas en el gradiente, favoreciendo procesos de optimización más estables y suaves durante el aprendizaje iterativo.

Las arquitecturas MLP implementadas integraron además Batch Normalization y Dropout. Batch Normalization estabiliza las distribuciones internas de activaciones y reduce sensibilidad frente a variaciones durante entrenamiento profundo [52]. Simultáneamente, Dropout reduce riesgo de sobreajuste mediante desactivación aleatoria de neuronas, mejorando capacidad de generalización [52].

El optimizador principal utilizado fue Adam, debido a su eficiencia computacional y capacidad

adaptativa frente a problemas de alta dimensionalidad. [55] demostraron que Adam combina ventajas de Momentum y RMSProp, permitiendo convergencia eficiente en redes neuronales profundas.

Dentro del proceso de selección de la estructura del modelo de Deep learning, utilizaron los datos del año 2024-2, en el cual se realizó prueba de concepto con el fin de validar el comportamiento de los diferentes modelos con los datos de prueba con el fin de seleccionar la arquitectura de MLP más adecuada para realizar la clasificación de los resultados del saber 11°, por ello se utilizaron las siguientes arquitecturas iniciales, así.

Tabla 6. *Modelos MLP Exploratorios.*

Versión Modelo	**Numero Capa**	**Numero Neurona**	**Función Activación**	**Función Perdida**	**Optimizador - Learning Rate**	**Épocas - Batch Size**	**Accuracy Max**	**Precisión Max**	**F1 Score Max**
N°1	24 (Dropout)	600 – 550-500-450-400-300-200-100-50-30 –“Dense Salida 4”- Regularización Pesos	Tanh – ReLu - Sofmax	Categorical Crossentropy	Adam: 0.001	24 - 1024	0.9882	0.9886	0.7222
N°2	26 (Dropout)	600 – 550-500-450-400-350-300-200-100-50-30 - 10 –“Dense Salida 4” - Regularización Pesos	ReLu - Sofmax	Categorical Crossentropy	Adam: 0.001	24 - 512	0.8747	0.8747	0.2340

Versión Modelo	Numero Capa	Numero Neurona	Función Activación	Función Perdida	Optimizador - Learning Rate	Épocas - Batch Size	Accuracy Max	Precisión Max	F1 Score Max
N°3	15 (Dropout)	600 – 550-500-450-400-350-300-250-200-150-100-50-30 - 10 -– “Dense Salida 4” – Regularización Pesos	LeakyReLU - Sofmax	Categorical Crossentropy	Adam: 0.001	24 - 512	0.8747	0.8747	0.2345
N°4	15 (Dropout)	600 – 550-500-450-400-350-300-250-200-150-100-50-30 - 10 -– “Dense Salida 4” - Batch Normalización	LeakyReLU - Sofmax	Categorical Crossentropy	Adam: 0.001	24 - 512	0.9961	0.9935	0.7357

Los resultados evidencian en comprender cuál es el mejor comportamiento según las arquitecturas experimentales desarrolladas, por lo anterior las versión N° 2 y N°3 los resultados evidencia un colapso en las clases mayoritaria, teniendo en cuenta que el gradiente de la red se estancó, minimizando local plano de la red, sin aprender los patrones reales en la clasificación de los resultados del saber 11° teniendo en cuenta que el Accuracy y la precisión fue de “0.87”, pero el F1 Score fue de “0.2345” lo que nos permite evidenciar la pérdida de capacidad para identificar clases minoritarias.

A diferencia de la versión No 4, logra mejores resultados de forma sobresaliente, obteniendo los resultados en Accuracy “0.9961”, Precisión “0.9935” y F1 Score “0.7357”, estos resultados se logran por introducción la función de activación “LeakyRelu” y el uso de Normalización por Lotes (Batch Normalization) elimina el estancamiento del gradiente, logrando aprender las características

de las clases minoritarias demostrando que es un modelo matemáticamente más robusto y balanceado.

Teniendo en cuenta lo anterior se inicia e implementar los modelos principales los cuales y la selección de todos los datos del examen saber 11° desde 2014-2 hasta 2024-2 del ICFES disponibles, con el fin de evaluar la clasificación de los modelos MLP y establecer el mejor modelo, para ello se tendrá en cuenta las siguientes arquitecturas, así.

Tabla 7. *Modelos Principales Producción de Clasificación Resultado Saber 11°.*

Versión Modelo	**Numero Capa**	**Numero Neurona**	**Función Activación**	**Función Perdida**	**Optimizador - Learning Rate**	**Épocas - Batch Size**	**Accuracy Max**	**Precisión Max**	**F1 Score Max**
N° 14 categorías	15 (Dropout)	Input_Shape = 201 600 – 550-500-450-400-350-300-250-200-150-100-50-30-10 –-“Dense Salida 4” - Batch Normalización	Leaky ReLU - Sofmax	Categorical Crossentropy	Adam – 0.0001 clipnorm = 1.0	30 - 512	0.8213	0.8231	0.7758
N° 25 categorías	15 (Dropout)	Input_Shape = 201 600 – 550-500-450-400-350-300-250-200-150-100-50-30-10 –-“Dense Salida 5” - Batch Normalización	Leaky ReLU – Sofmax	Categorical Crossentropy	Adam – 0.0001 clipnorm = 1.0	30 - 512	0.6976	0.7000	0.6516
N° 35 categorías	15 (Dropout)	Input_Shape = 206 600 – 550-500-450-400-350-300-250-200-150-100-50-30-10 –-“Dense Salida 5” - Batch Normalización	Softplus (6)- Leaky ReLU (8) - Sofmax	Categorical Crossentropy	Adam – 0.001 clipnorm = 1.0	30 - 512	0.6997	0.7021	0.6565

Versión Modelo	Numero Capa	Numero Neurona	Función Activación	Función Perdida	Optimizador - Learning Rate	Épocas - Batch Size	Accuracy Max	Precisión Max	F1 Score Max
N° 4 5 categorías	15 (Dropout)	Input_Shape = 206 600 – 550-500-450-400-350-300-250-200-150-100-50-30 -10 -–"Dense Salida 5" - Batch Normalización	Softplus (6)- Leaky ReLU (8) - Sofmax	Categorical Crossentropy - Momentum:0.9	Adam – 0.001 clipnorm = 1.0 *Decay Steps:10.000 – Decay Rate: 0.90 Decaimento:0.90*	30 - 512	0.6981	0.7006	0.6522
N° 5 5 categorías	15 (Dropout)	Input_Shape = 206 600 – 550-500-450-400-350-300-250-200-150-100-50-30 -10 -–"Dense Salida 5" - Batch Normalización	Softplus (6)- Leaky ReLU (8) - Sofmax	Categorical Crossentropy - Momentum:0.8	Adam – 0.001 clipnorm = 1.0 *Decay Steps:10.000 – Decay Rate: 0.7- Decaimento:0.70*	30 - 512	0.7007	0.7046	0.6622

Los modelos en producción desarrollados nos permiten analizar una estructura de datos con mayor granularidad, teniendo en cuenta que se está haciendo uso de cinco categorías, aumentando la complejidad en la clasificación del rendimiento académico del personal que presentaron el saber 11 durante el 2014-2 hasta 2024-2 teniendo en cuenta la heterogeneidad y desbalance de los datos.

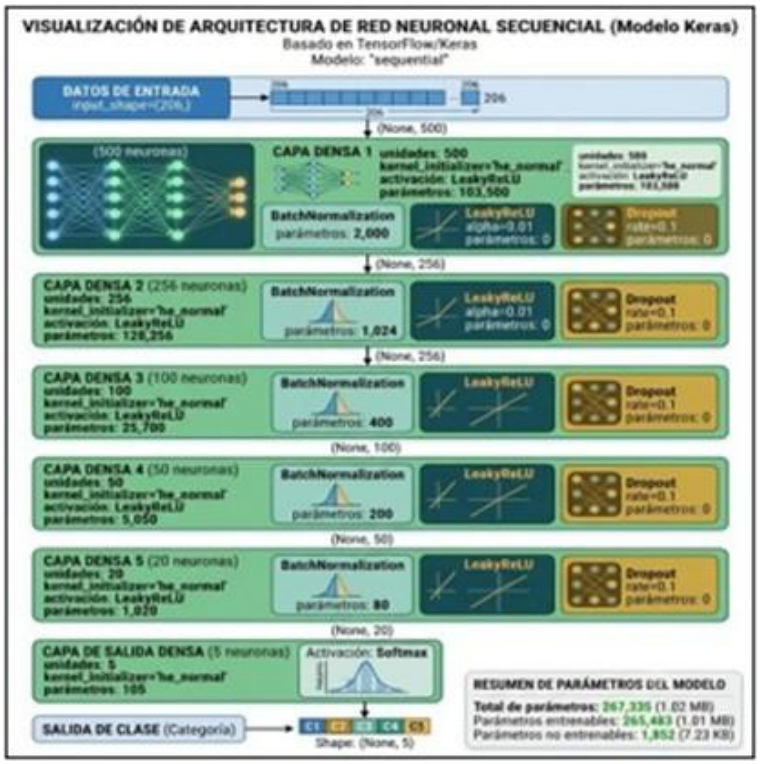
Teniendo en cuenta los resultados obtenidos de las métricas, se evidencia que la arquitectura de la versión No 7 obtuvo los mejores resultados con una exactitud de "0.7041", precisión de "0.7075" y F1 Score de "06675", esta arquitectura se obtuvo mediante el uso de algoritmos genéticos para una optimización heurística, permitiendo realizar diferentes modelamientos hasta identificar la estructura de MLP más adecuada, así;

Modelo Deep Learning de red multicapa versión N° 7.

```
network = models.Sequential()

network.add(layers.Dense(500,input_shape=(206,), kernel_initializer='he_normal'))
network.add(layers.BatchNormalization()) network.add(layers.LeakyReLU(alpha=0.01))
network.add(layers.Dropout(0.1))

network.add(layers.Dense(256))
network.add(layers.BatchNormalization())
network.add(layers.LeakyReLU(alpha=0.01))
network.add(layers.Dropout(0.1))

network.add(layers.Dense(100))
network.add(layers.BatchNormalization())
network.add(layers.LeakyReLU(alpha=0.01))
network.add(layers.Dropout(0.1))

network.add(layers.Dense(50))
network.add(layers.BatchNormalization())
network.add(layers.LeakyReLU(alpha=0.01))
network.add(layers.Dropout(0.1))

network.add(layers.Dense(20))
network.add(layers.BatchNormalization())
network.add(layers.LeakyReLU(alpha=0.01))
network.add(layers.Dropout(0.1))
network.add(layers.Dense(5, activation='softmax')).
```

El modelo de redes multicapa versión No 7 refleja una estructura sencilla de cinco capas ocultas “Hidden Layers”, con estructura tipo diamante, compacta, simétrica y de alta velocidad, evitando la co-adaptacion de las características y el sobreajuste en la clasificación de las cinco clases, adicionalmente es importante tener en cuenta el uso de la tasa de aprendizaje “Learning Rate” de (0.01) utilizado, lo que permitió mejorar el aprendizaje desde el inicio, y el uso del factor decaimiento en un 80% genero descenso estable durante las 29 épocas, así mismo la función de activación de LeakyReLu permitió un flujo gradiente y constante en el área negativa manteniendo activas las neuronas del modelo durante las épocas, obteniendo las mejores métricas en la clasificación de los resultados del saber 11°.

Figura 3. *Modelo N° 7, Arquitectura de Red Neuronal MLP establecida por Algoritmos genéticos.*



Nota: La elaboración de la figura fue mediante apoyo Gemini Google IA.

Diseño de arquitectura CNN 1D: La arquitectura CNN 1D implementada en el proyecto constituye una adaptación experimental de redes convolucionales tradicionales al contexto de datos tabulares educativos. El modelo utiliza capas Conv1D y MaxPooling1D orientadas a identificar patrones locales entre variables socioeconómicas transformadas.

El principio metodológico detrás de esta arquitectura consiste en permitir que los filtros convolucionales detecten relaciones implícitas entre variables cercanas dentro de la representación tabular. Investigaciones recientes han evidenciado que las CNN pueden generar representaciones competitivas en datos estructurados cuando existe dependencia local entre características [18].

La arquitectura CNN implementada incluyó capas convolucionales seguidas de Max Pooling,

Flatten y capas densas finales orientadas a clasificación multiclase. Asimismo, se incorporaron mecanismos de regularización y normalización similares a los utilizados en las arquitecturas MLP, como se observa en la arquitectura.

Modelo Rede Neuronal Convolucional de una dimensión:

```
def build_model_santo_tomas(n_features):
    model = models.Sequential([
        layers.Input(shape=(n_features, 1)),

        layers.Conv1D(64, kernel_size=5, padding='same'),
        layers.BatchNormalization(),
        layers.LeakyReLU(alpha=0.1),
        layers.MaxPooling1D(pool_size=2),

        layers.Conv1D(103, kernel_size=3, padding='same'),
        layers.BatchNormalization(),
        layers.LeakyReLU(alpha=0.1),

        layers.Conv1D(103, kernel_size=3, padding='same'),
        layers.BatchNormalization(),
        layers.LeakyReLU(alpha=0.01),

        layers.GlobalAveragePooling1D(),

        layers.Dense(206),
        layers.BatchNormalization(),
        layers.LeakyReLU(alpha=0.01),
        layers.Dropout(0.4),

        layers.Dense(103),
        layers.BatchNormalization(),
        layers.LeakyReLU(alpha=0.01),

        layers.Dense(5, activation='softmax')
    ])

    # Definición de métricas dentro de la función
    metrics = [
        'accuracy',
        tf.keras.metrics.Precision(name='precision'),
        tf.keras.metrics.Recall(name='recall'),
        tf.keras.metrics.AUC(name='auc', multi_label=True),
        tf.keras.metrics.F1Score(name='f1_score', average='weighted')
```

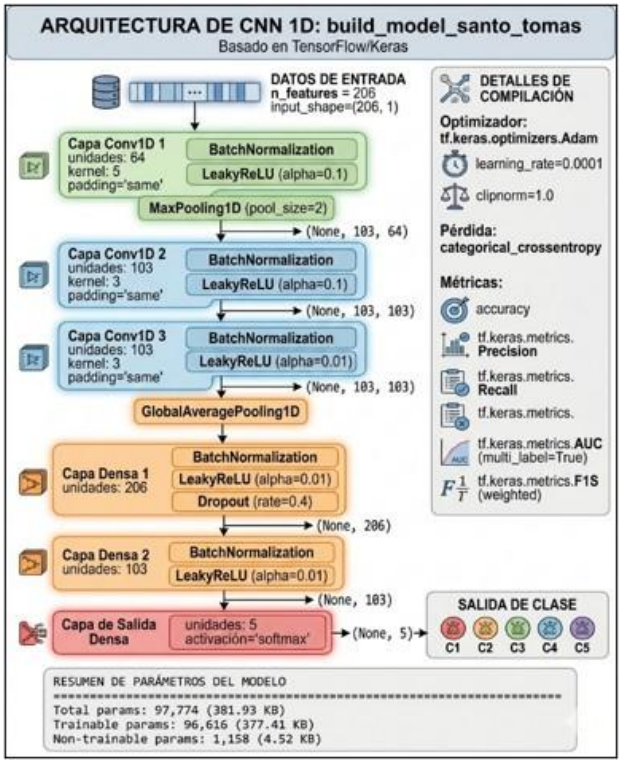
```
    ]
    optimizer = tf.keras.optimizers.Adam(learning_rate=0.0001,
    clipnorm=1.0)

    model.compile(
        optimizer=optimizer,
        loss='categorical_crossentropy',
        metrics=metrics
    )
    return model

n_features = 206
model = build_model_santo_tomas(n_features)
model.summary()
```

La arquitectura de red neuronal convolucional de una dimensión se encuentra compuesta por un bloque de extracción de patrones que hace uso de 64 filtros vectoriales, en el que cada filtro recorre las 206 variables en subgrupo de cinco en cinco con el fin de captar relaciones de vecindad entre las variables continuas, así mismo al realizar la reducción de la dimensionalidad a través del MaxPooling1D mejora el rendimiento computacional a través de la introducción invariancias a pequeñas traslaciones enfocándose en las características más relevantes detectadas en la vecindad.

Las capas abstracción profunda permite representar de forma más amplia el vocabulario de las representaciones abstractas y reduciendo la ventana de inspección, mejorando el campo receptivo con menos datos lo que permitirá identificar las estructuras semánticas entre cada clase.

El bloque de clasificador fue diseñado con capas densas en el cual permite tomas las características promediadas y desarrolla combinaciones no lineales mediante razonamiento lógico para realizar la clasificación probabilista de las cinco categorías a través de la función de activación softmax.

Figura 4. *Arquitectura Redes Convolucional Unidimensional "CNN 1D".*



Nota: La figura fue elaborada mediante apoyo Gemini Google IA.

Diseño de LightGBM: El modelo LightGBM fue seleccionado como representante de algoritmos boosting debido a su elevado desempeño reportado en problemas tabulares complejos. LightGBM se fundamenta en árboles de decisión optimizados mediante gradiente boosting, permitiendo reducción secuencial del error mediante ensamblaje iterativo de clasificadores débiles.

[49] establecen que los modelos boosting representan una de las estrategias más robustas para clasificación tabular debido a su capacidad para manejar relaciones no lineales, alta dimensionalidad y variables heterogéneas. LightGBM incorpora además estrategias de optimización computacional orientadas a reducir tiempo de entrenamiento y consumo de memoria.

El diseño experimental del modelo contempló ajuste iterativo de hiperparámetros relacionados

con profundidad de árboles, learning rate, regularización y número de estimadores, como se evidencia en el modelo de machine learning.

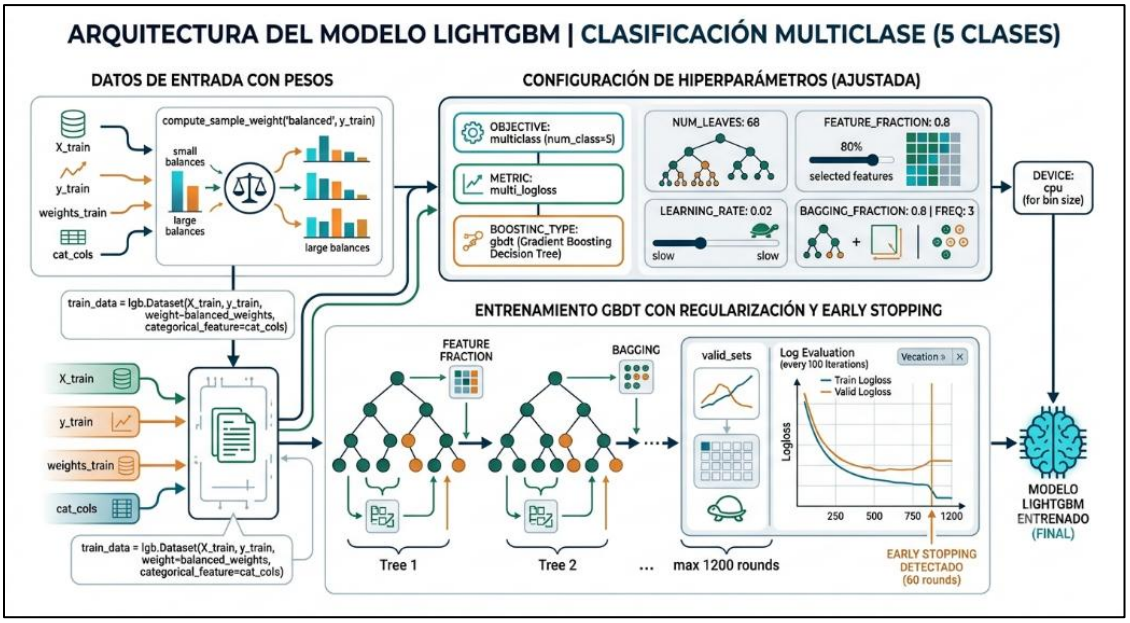
Arquitectura del Modelo LightGBM

```
weights_train = compute_sample_weight ('balanced', y_train)

params = {
   'objective': 'multiclass',
   'num_class': 5,
   'metric': 'multi_logloss',
   'boosting_type': 'gbdt',
   '68': 68
   'learning_rate': 0.02,
   'feature_fraction': 0.8,
   'bagging_fraction': 0.8,
   'bagging_freq': 5,
   'verbose': -1,
   'device': 'cpu'
}

train_data = lgb.Dataset(X_train, label=y_train, weight=weights_train,
categorical_feature=cat_cols)
test_data = lgb.Dataset(X_test, label=y_test, categorical_feature=cat_cols,
reference=train_data)

model_lgbm = lgb.train(
   params,
   train_data,
   num_boost_round=1200,
   valid_sets=[train_data, test_data],
   valid_names=['train', 'valid'],
   callbacks=[
      lgb.early_stopping(stopping_rounds=60),
      lgb.log_evaluation(period=100)
   ]
)
```

El algoritmo de gradiente descendente boosting sobre árbol de decisiones segmenta y ramifica de forma secuencial, en el cual hace uso de la función de entropía cruzada que permite evaluar la probabilidad de las predicciones, con una tasa de aprendizaje de learning rate de (0.02) asegurando

la convergencia, así mismo implementa que cada árbol se construya con subconjuntos de variables usando el ochenta por ciento (80%) de los atributos, y por cada cinco iteración “Baggin Freq” se genera un nuevo subconjunto, generando una diversidad de variables y disminuir el sobreajuste “overfiting”, de igual forma la arquitectura del LightGBM implemento un balanceador de pesos “compute sample weigth” con el fin de penalizar las clases minoritarias y desbalanceada del dataset (Muy Bajo – Bajo – Muy Alto), con el fin afinar la clasificación de las categorías a través de penalización.

Figura 5. *Arquitectura de Modelo LigthGBM.*



Nota: La elaboración de la figura fue Elaboración propia mediante uso de Gemini Google IA.

Optimización mediante algoritmos genéticos: La investigación incorporó adicionalmente algoritmos genéticos orientados a optimización de hiperparámetros. Estos algoritmos se fundamentan en principios evolutivos inspirados en selección natural y supervivencia adaptativa.

Los algoritmos genéticos permiten explorar espacios complejos de búsqueda mediante mecanismos

de selección, cruce y mutación. Su incorporación en el proyecto permitió evaluar configuraciones arquitectónicas potencialmente superiores a aquellas obtenidas mediante ajuste manual tradicional.

4.3 Implementación técnica

Las arquitecturas MLP utilizaron capas densas decrecientes, activaciones no lineales, normalización por lotes y regularización Dropout. Batch Normalization se incorporó para estabilizar el entrenamiento y reducir cambios internos en la distribución de activaciones, mejorando la convergencia del modelo [52].

Por su parte, Dropout se utilizó para mitigar el sobreajuste mediante la desactivación aleatoria de neuronas durante el entrenamiento [53]. El optimizador Adam fue empleado por su eficiencia computacional y su capacidad de adaptación automática de tasas de aprendizaje en problemas de gran escala [56].

Tabla 8. *Arquitecturas de Modelos de Deep Learning y Machine Learnng.*

Modelo	Notebook	Técnica	Hiperparámetros Bases	Función experimental
MLP LeakyReLU	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_1_Leaky_ReLu	Red densa profunda	**No Capas**: 14 **No Neuronas**: 600-550-500-450-400-350-300-250-200-150-100-50-30-10 **Capa Salida**: Softmax - 5 **Optimizador:** Adams: Learning Rate: 0.0001 **Función Activación:** LeakyRelu – Alpha 0.3 - 0.2 - 0.1 - 0.01 **Dropout:** 0.2 – 0.3	Primera arquitectura estable
MLP Softplus	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_2_Softplus	Red densa profunda	**No Capas:** 14 - BatchNormalization **No Neuronas:** 600-550-500-450-400-350-300-250-200-150-100-50-30-10 **Capa Salida**: Softmax - 5 **Función Optimización**: Adams: **Learning Rate**: 0.0001 **Función Activación:** Softplus – LeakyReLU – Alpha 0.3 - 0.2 - 0.1 - 0.01 – 0.001 **Dropout**: 0.2 – 0.3 – 0.1	Comparación de activación suave

Modelo	Notebook	Técnica	Hiperparámetros Bases	Función experimental
MLP Softplus + decay_Momentums	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_3_Softplus_Decay_Momentums		**No Capas:** 14 - BatchNormalization **No Neuronas:** 600-550-500-450-400-350-300-250-200-150-100-50-30-10 **Capa Salida**: Softmax - 5 **Función Optimización**: Adams: **Learning Rate**: 0.0001 - decay_steps=10000 - decay_rate=0.9 - beta_1=0.9 - beta_2=0.999 - epsilon=1e-07 **Función Activación:** Softplus – LeakyReLU Alpha = 0.3 -0.2 - 0.01 – 0.001 **Dropout**: 0.2 – 0.3	Reducción progresiva Gradientes y corrección de configuración manual Hiperparametros
MLP Softplus + decay	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_4_Softplus_Learning_Decay	Red densa con ajuste de tasa	**No Capas:** 14 - BatchNormalization **No Neuronas:** 600-550-500-450-400-350-300-250-200-150-100-50-30-10 **Capa Salida**: Softmax - 5 **Función Optimización**: Adams: **Learning Rate**: 0.0001 - decay_steps=10000 - decay_rate=0.7 - beta_1=0.8 - beta_2=0.999 - epsilon=1e-07 **Función Activación:** Softplus – LeakyReLU Alpha = 0.01 – 0.001 – 0.0001 **Dropout**: 0.2 – 0.1	Evaluación de estabilidad,
MLP Softplus + decay – 09 Capas	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_5_Softplus_Learning_Decay	Red densa reducción de Hiperparametros	**No Capas:** 09 - BatchNormalization **No Neuronas:** 600--500-400-300-200-100-50-30-10 **Capa Salida**: Softmax - 5 **Función Optimización**: Adams: **Learning Rate**: 0.001 - decay_steps=10000 - decay_rate=0.8 **Función Activación:** Softplus – LeakyReLU Alpha = 0.01 – 0.001 – 0.0001 **Dropout**: 0.2 – 0.1	Reducción capacidad de cómputo y validar sus resultados de clasificación.
CNN 1D	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_8_CNN_1D	Convolución unidimensional	**No Bloques:** 02 **ConvD1**: 64 -103 -3 **Kernel Size** = 5 – 3 - 3 **MaxPooling:** 2 **Función Activación:** LeakyReLU GlobalAveragePooling1D **Bloque Clasificador** **Capa:** Dense - BatchNormalization **Función Activación:** LeakyReLU **Dropout**: 04	Adaptación profunda para datos tabulares
LigthGBM	05_Categorias_Data_Saber_11_Unificada_Paquet_Version_9_LightGBM	Ensamble boosting	**Boosting type:** GBDT **Numero Leaves:** 68 **Learning Rate**: 0.02 **Feature Fraction**: 0.8 # Porcentaje Columna Selección **Bagging Fraction**: 0.8 # Porcentaje Fila Selección **Bagging Freq**: 5 # Numero Iteraciones Muestra **Numero Rondas**: 1200	Modelo base robusto para comparación

La implementación técnica del proyecto fue desarrollada completamente en lenguaje Python utilizando notebooks reproducibles ejecutados en Google Colab. La estructura modular del proyecto permitió organizar las diferentes fases del pipeline mediante notebooks independientes especializados en procesamiento, modelado y validación.

La fase inicial de implementación correspondió a la unificación histórica de los microdatos Saber 11. Para ello se desarrollaron notebooks específicos orientados a lectura secuencial de archivos TXT originales, homologación de columnas y consolidación histórica de data sets.

Posteriormente, los datos consolidados fueron almacenados en formato Parquet utilizando PyArrow. El uso de almacenamiento columnar permitió mejorar la eficiencia computacional y reducir tiempos de lectura, aspecto fundamental considerando el volumen masivo de información procesada.

La implementación del pipeline de preprocesamiento fue realizada utilizando Scikit-learn y bibliotecas complementarias. Las transformaciones fueron encapsuladas dentro de pipelines reproducibles con el propósito de garantizar consistencia entre entrenamiento y validación.

Los modelos MLP fueron implementados utilizando TensorFlow y Keras. Cada arquitectura fue definida mediante secuencias de capas Dense, BatchNormalization y Dropout. Las configuraciones experimentales incluyeron variaciones en sus arquitecturas e hiperparametros, entre ellas.

- Número de capas;
- Funciones de activación;
- Learning rate;
- Decay dinámico;
- Estrategias de regularización.

La implementación CNN 1D fue realizada utilizando capas Conv1D y Max Pooling 1D adaptadas a representaciones tabulares. Las entradas fueron reorganizadas en estructuras

compatibles con convoluciones unidimensionales, permitiendo aplicar filtros convolucionales sobre secuencias de características transformadas.

Por otra parte, LightGBM fue implementado utilizando la biblioteca oficial LightGBM, integrando parámetros orientados a clasificación multiclase y optimización boosting.

Finalmente, la implementación estadística contempló generación automática de métricas, almacenamiento de resultados experimentales en archivos CSV y análisis iterativo.

4.4 Descripción del prototipo o sistema desarrollado

El sistema desarrollado corresponde a un prototipo analítico experimental orientado a clasificación del desempeño académico utilizando modelos supervisados de inteligencia artificial sobre datos educativos masivos. El prototipo integra procesamiento tabular, transformación automática, entrenamiento supervisado y validación experimental reproducible.

Desde la perspectiva funcional, el sistema recibe variables socioeconómicas, familiares, institucionales y académicas correspondientes a estudiantes colombianos evaluados mediante la prueba Saber 11. Posteriormente, el pipeline ejecuta procesos automáticos de limpieza, codificación y transformación de datos antes de alimentar las arquitecturas supervisadas.

El prototipo fue diseñado bajo principios de modularidad, permitiendo incorporar nuevas arquitecturas, modificar hiperparámetros y reutilizar componentes metodológicos sin afectar la estructura general del sistema. Esta característica resulta especialmente importante en proyectos de investigación aplicada debido a que facilita replicabilidad y extensión futura.

Asimismo, el sistema fue estructurado como plataforma experimental orientada a comparación metodológica entre arquitecturas profundas y modelos boosting. Esta aproximación permite evaluar diferencias de desempeño bajo condiciones experimentales homogéneas y facilita análisis

comparativos reproducibles.

La solución implementada demuestra además la viabilidad técnica de aplicar modelos avanzados de inteligencia artificial al contexto educativo colombiano utilizando infraestructura accesible y tecnologías abiertas. Desde la perspectiva científica, el prototipo constituye una contribución relevante para el campo de *Educational Data Mining* en Colombia, especialmente debido al uso de datos educativos masivos y arquitecturas profundas adaptadas a clasificación supervisada multiclase.

Finalmente, la arquitectura desarrollada presenta potencial de expansión futura hacia escenarios relacionados con:

- Interpretabilidad mediante SHAP;
- Sistemas de alerta temprana;
- Predicción continua;
- Analítica institucional;
- Dashboards educativos inteligentes.

Esto convierte al sistema desarrollado no solamente en un prototipo experimental de investigación, sino también en una base potencial para futuras soluciones de analítica educativa aplicada en instituciones académicas y entidades gubernamentales.

5 Resultados y validación.

5.1 Resultados experimentales o de aplicación

La presente investigación tuvo como propósito principal evaluar el desempeño de diferentes arquitecturas de aprendizaje automático y aprendizaje profundo para la clasificación del

desempeño académico en las pruebas Saber 11 utilizando variables socioeconómicas, institucionales y familiares derivados de los microdatos oficiales del ICFES. En consecuencia, el proceso experimental fue diseñado como un entorno comparativo orientado a analizar el comportamiento predictivo de múltiples modelos supervisados sobre un conjunto de datos educativos masivos correspondiente al periodo 2014-2 a 2024-2.

El entorno experimental implementado permitió entrenar y validar diferentes arquitecturas utilizando aproximadamente 7.2 millones de registros estudiantiles con 45 variables socioeconómicas, convirtiendo el proyecto en uno de los ejercicios de minería de datos educativos de mayor escala desarrollados sobre datos públicos del sistema educativo colombiano. La magnitud del conjunto de información procesado exigió el uso de estrategias robustas de transformación, almacenamiento y entrenamiento supervisado, particularmente debido a la heterogeneidad estructural de las variables socioeconómicas y académicas presentes en los microdatos.

Los experimentos realizados permitieron evaluar el comportamiento de tres grandes familias de modelos: arquitecturas profundas tipo MLP (*Multilayer Perceptron*), redes convolucionales unidimensionales CNN 1D y modelos boosting basados en árboles mediante LightGBM. Adicionalmente, se incorporaron mecanismos de optimización evolutiva mediante algoritmos genéticos orientados al ajuste de hiperparámetros y exploración de configuraciones arquitectónicas alternativas.

El diseño experimental contempló múltiples iteraciones independientes para cada arquitectura mediante la prueba de Shapiro-Wilk [60] con el propósito de evaluar estabilidad estadística y capacidad de generalización.

Esta estrategia metodológica resulta especialmente importante en modelos de aprendizaje profundo debido a que procesos de inicialización aleatoria, ajuste de pesos y convergencia iterativa pueden generar variabilidad significativa entre ejecuciones [51].

Tabla 9. *Resultados promedio de desempeño por modelo Metodología Shapiro Wilk.*

Modelo	Mejor F1-Score	F1-score promedio	Desviación estándar	Estadístico W	Valor P (p-value)	Iteraciones
MLP LeakyReLU Softplus	0.66206	0.660080	0.001060	0.97122	5.73000e-01	30
CNN 1D	0.65867	0.641512	0.009106	0.97467	6.73104e-01	30
LightGBM	0.52446	0.523659	0.000349	0.96190	3.46159e-01	30

Los resultados experimentales evidencian diferencias significativas en el comportamiento predictivo de las arquitecturas evaluadas. El modelo MLP obtuvo el mayor desempeño promedio en términos de F1-score, alcanzando un valor aproximado de 0.660080, superando tanto a CNN 1D como a LightGBM. Asimismo, el modelo presentó baja dispersión entre iteraciones experimentales, indicando elevada estabilidad estadística durante el entrenamiento.

Estos resultados son consistentes con investigaciones recientes sobre aprendizaje profundo aplicado a datos tabulares, donde las arquitecturas densas profundas han demostrado elevada capacidad para modelar relaciones complejas y no lineales entre variables heterogéneas [13], [18]. Particularmente, en contextos educativos, los modelos MLP han mostrado ventajas frente a arquitecturas más rígidas debido a su capacidad para capturar interacciones multidimensionales entre factores socioeconómicos, institucionales y familiares.

De igual forma se evidencian que las arquitecturas profundas basadas en MLP presentaron el comportamiento predictivo más robusto sobre el conjunto de datos analizado. Particularmente, las versiones implementadas utilizando funciones de activación LeakyReLU y Softplus alcanzaron desempeños superiores frente a CNN 1D y LightGBM, mostrando además mayor estabilidad

experimental entre iteraciones.

El desempeño competitivo de las arquitecturas MLP puede explicarse por su capacidad para modelar relaciones altamente no lineales entre variables socioeconómicas y desempeño académico. Investigaciones recientes en minería de datos educativa han demostrado que el rendimiento estudiantil responde a interacciones complejas entre factores familiares, económicos, territoriales e institucionales, relaciones que difícilmente pueden ser representadas adecuadamente mediante modelos lineales tradicionales [13], [33].

Por otra parte, los experimentos realizados con CNN 1D mostraron resultados relevantes, aunque inferiores respecto a las arquitecturas densas MLP. Las redes convolucionales unidimensionales lograron identificar patrones locales dentro de las representaciones tabulares transformadas; sin embargo, el comportamiento experimental sugiere que las dependencias presentes en las variables socioeconómicas del examen Saber 11 podrían responder más adecuadamente a relaciones globales capturadas mediante capas densas profundas.

Este comportamiento resulta consistente con investigaciones recientes relacionadas con aprendizaje profundo sobre datos tabulares. [58] señaló que, aunque las arquitecturas convolucionales pueden generar resultados competitivos en ciertos contextos estructurados, los modelos densos y boosting continúan dominando numerosos escenarios tabulares debido a su capacidad para capturar relaciones globales complejas entre variables heterogéneas.

En relación con LightGBM, los resultados obtenidos muestran un desempeño considerablemente menor frente a las arquitecturas profundas implementadas. Aunque los modelos boosting suelen presentar excelentes resultados en problemas tabulares, particularmente en conjuntos de datos medianos y altamente estructurados [50], el comportamiento observado en esta investigación sugiere que la complejidad relacional del fenómeno educativo analizado favorece arquitecturas profundas capaces de modelar representaciones jerárquicas de mayor abstracción.

Figura 6. *Comparación de la distribución de los F1 Score de la Prueba Shapiro-Wilk.*

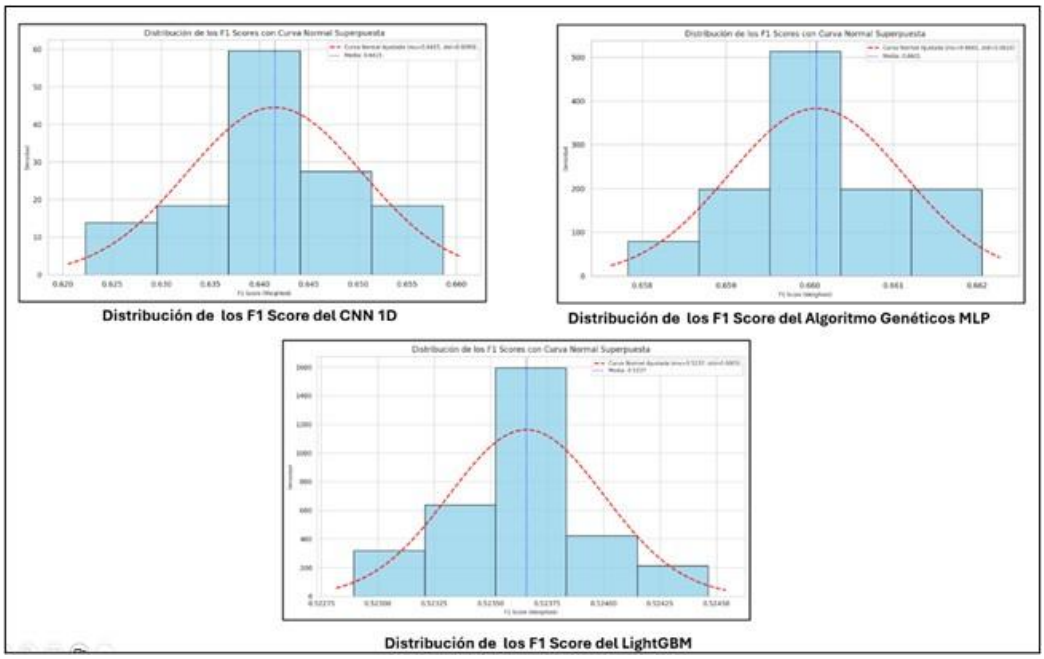
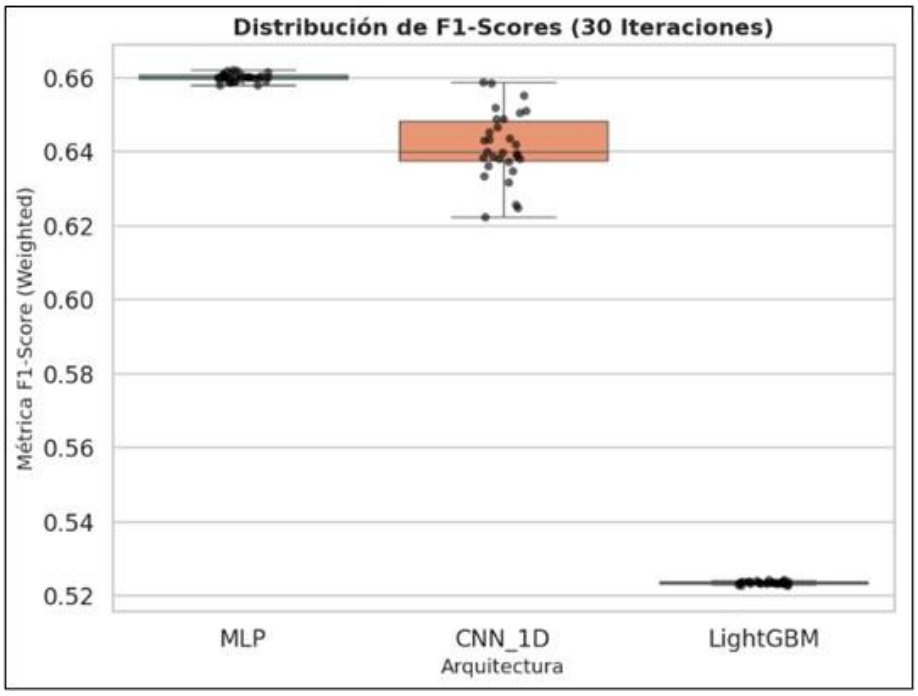


Figura 7. *Distribución F1-Score de las arquitecturas de Inteligencia artificial.*



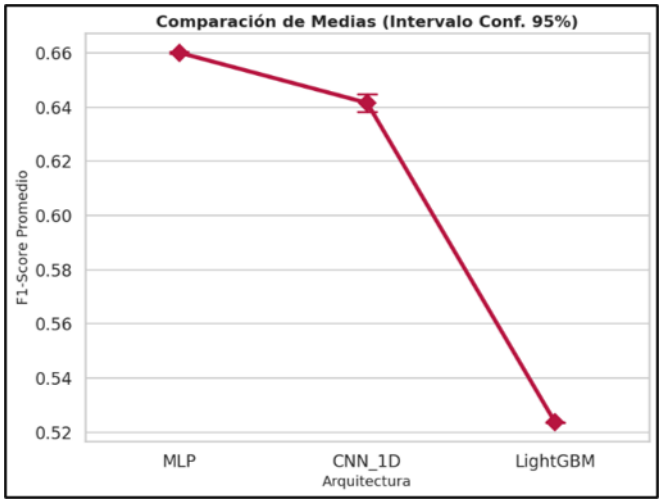
Se evidencia que las métricas F1-Score obtenidas durante la prueba de Shapiro-Wilk de los tres modelos (MLP-CNN 1D – LightGBM) se encuentran de forma normal “Gaussiana” como se observa en la ilustración No 5; así mismo durante el análisis de varianza (ANOVA) se determina que existe diferencias estadísticas entre los tres modelos alcanzando un valor F alto (5853.93888) y Valor p (P-Value): 1.78001e-93, rechazando la hipótesis de igualdad de medias en la métrica de F1-Score evidenciando que los tres modelos no tienen el mismo rendimiento de clasificación, la prueba Post-Hoc de Tukey HSD realizo comparación múltiple de pares como se evidencia en la tabla 7, entre los modelos donde la arquitectura del MLP supero significativamente a la red convolucional unidimensional con una diferencia de (+0.0186), al igual al modelo de LightGBM con una diferencia de (+0.1364), siendo que MLP es el mejor modelo de clasificación de los datos del saber 11°.

Tabla 10. *Comparación múltiple prueba Post-Hoc de Tukey.*

Modelo A	Modelo B	Diferencia Media F1-Score	Valor P Ajustado	Límite Inferior	Límite Superior
CNN_1D	LightGBM	-0.1179	0.0	-0.1211	-0.1146
CNN_1D	MLP	0.0186	0.0	0.0153	0.0218
LightGBM	MLP	0.1364	0.0	0.1332	0.1397

La ilustración 7 nos confirma los resultados obtenidos en el análisis de varianza ANOVA, permitiendo evidenciar las diferencias el rendimiento del promedio de las métricas F1-Score entre las tres arquitecturas de Inteligencia artificial, en el que el modelo MLP logra un rendimiento estadístico significativo a diferencia a los modelos de CNN 1D y LightGBM.

Figura 8. *Comparación de Medias entre arquitecturas de Inteligencia artificial de Clasificación.*



Durante la prueba estadística paramétrica ANOVA de Welch se obtiene una diferencia alta de varianza entre los tres modelos con un valor estadístico de F de (222680.277533) y el Valor P es

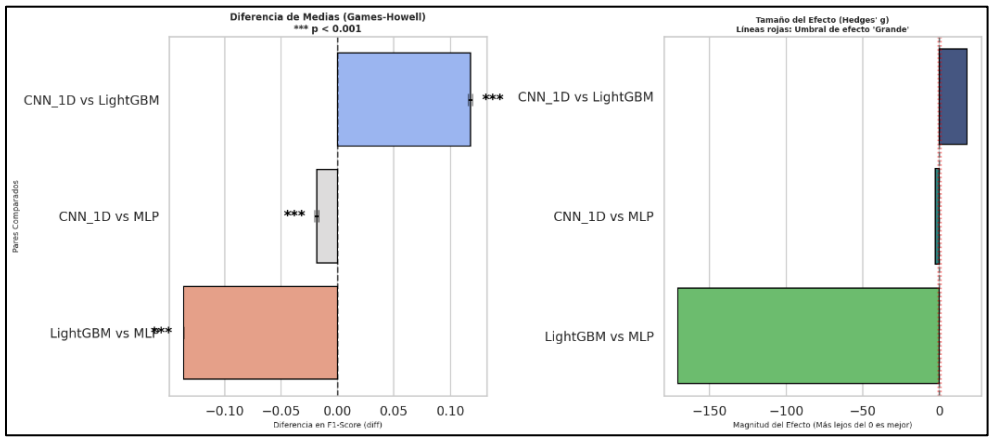
de (4.297480e-86) confirmando sus diferencias en métrica de evaluación de los modelos en la clasificación de los resultados del saber 11°, teniendo en cuenta la comparación múltiple de la prueba Post-Hoc de Game Howell referenciada en la tabla 8, logro determinar que el modelo perceptrón multicapa es el mejor clasificador logrando un F1-Score promedio de 0.660, superando los modelos CNN 1D y LigthGBM, donde los tamaños de efectos (Hadges´s) alcanzaron una superioridad del MLP en la captura de patrones frente a la CNN 1D (-2.82) y LightGBM (-170.6528).

Tabla 11. *Comparación múltiple de Prueba Post-Hoc De Games-Howell.*

Modelo A	Modelo B	Mean A F1 - Score	Mean B F1 - Score	Diferencia	Error Estándar	Estadístico T	Grado Diferencia	P-Valor	Tamaño Efecto
CNN1D	LightGBM	0.641512	0.523659	0.117853	0.001664	70.837543	29.085272	6.661338e-16	18.052640
CNN1D	MLP	0.641512	0.660080	-0.018568	0.001674	-11.094094	29.785555	1.261680e-11	-2.827282
LightGBM	MLP	0.523659	0.660080	-0.136422	0.000204	-669.632024	35.221451	5.551115e-15	-170.652881

Los resultados obtenidos en el ANOVA Welch y prueba Post-Hoc de Game-Howell demuestra que la arquitectura del perceptrón multicapa (MLP) cuenta con una distribución compacta y varianza controlada en el promedio de la métrica F1-Score validando que tiene mejor rendimiento y estabilidad en la clasificación de los resultados del examen saber 11°, aprendiendo de manera consistente en cada interacción, como se logra visualizar en la ilustración 8.

Figura 9. *Comparación Prueba Post-Hoc De Games-Howell.*



Otro aspecto relevante identificado durante el proceso experimental corresponde al impacto significativo del pipeline híbrido de preprocesamiento sobre el desempeño de los modelos. La combinación de RobustScaler, OneHotEncoder y TargetEncoder permitió mejorar considerablemente la calidad de las representaciones de entrada utilizadas por las arquitecturas supervisadas.

Particularmente, el uso de Target Encoding sobre variables categóricas de alta cardinalidad permitió preservar información predictiva relevante asociada a municipios, instituciones educativas y variables territoriales. Este comportamiento coincide con investigaciones previas relacionadas con codificación supervisada en problemas tabulares complejos [42].

Asimismo, los resultados experimentales evidencian la importancia crítica del control de fuga de información (*data leakage*) dentro del pipeline analítico. Durante las primeras fases exploratorias del proyecto se identificó que ciertos esquemas incorrectos de Target Encoding generaban sobreestimaciones artificiales del desempeño predictivo. En consecuencia, se implementaron protocolos estrictos de separación entre entrenamiento, validación y prueba antes del ajuste de codificadores supervisados.

La literatura especializada establece que la fuga de información constituye uno de los principales

riesgos metodológicos en aprendizaje automático, particularmente en contextos donde variables categóricas son transformadas mediante estadísticas derivadas del conjunto completo de datos [44]. La corrección de este problema permitió obtener resultados más robustos y metodológicamente válidos.

Otro hallazgo importante del proceso experimental corresponde al comportamiento diferencial de las funciones de activación evaluadas. Las arquitecturas basadas en LeakyReLU mostraron convergencia estable y elevada capacidad predictiva, mientras que las versiones implementadas con Softplus presentaron entrenamiento más suave y menor oscilación durante procesos iterativos prolongados.

Este comportamiento resulta consistente con fundamentos teóricos de optimización profunda. LeakyReLU favorece flujo continuo de gradientes incluso en regiones negativas, reduciendo riesgo de neuronas inactivas [55]. Por otra parte, Softplus genera superficies de optimización más suaves y continuas, favoreciendo estabilidad numérica durante el entrenamiento profundo.

Los experimentos relacionados con learning rate decay y momentum mostraron mejoras parciales en estabilidad de entrenamiento, particularmente en arquitecturas profundas de mayor complejidad. La incorporación de mecanismos dinámicos de reducción del learning rate permitió disminuir oscilaciones durante etapas avanzadas de convergencia y mejorar comportamiento general del proceso de optimización.

En relación con los algoritmos genéticos implementados, los resultados evidencian que las estrategias evolutivas poseen potencial significativo para optimización automática de hiperparámetros en arquitecturas profundas. Aunque los costos computacionales asociados son elevados debido al tamaño del conjunto de datos utilizado, los experimentos demostraron capacidad para identificar configuraciones arquitectónicas competitivas.

Otro aspecto importante observado durante el desarrollo experimental corresponde a la elevada

sensibilidad de los modelos frente a variables socioeconómicas asociadas a contexto familiar y acceso a recursos educativos. Aunque la presente investigación no tuvo como propósito principal realizar interpretabilidad causal, los resultados preliminares sugieren que variables relacionadas con nivel educativo de los padres, acceso a internet, disponibilidad tecnológica y características institucionales poseen influencia significativa sobre las predicciones generadas por los modelos.

Estos hallazgos coinciden con literatura internacional relacionada con desigualdad educativa y rendimiento académico. [36] demostró mediante metaanálisis que el nivel socioeconómico constituye uno de los predictores más consistentes del desempeño académico en diferentes contextos educativos.

Desde la perspectiva computacional, los resultados obtenidos evidencian que el uso de almacenamiento Parquet y pipelines optimizados permitió manejar eficientemente el procesamiento de millones de registros estudiantiles utilizando infraestructura accesible basada en Google Colab. Este aspecto demuestra la viabilidad técnica de implementar soluciones avanzadas de inteligencia artificial educativa utilizando herramientas abiertas y recursos computacionales relativamente limitados.

Finalmente, los resultados experimentales obtenidos permiten concluir que las arquitecturas profundas tipo MLP representan una alternativa altamente competitiva para clasificación del desempeño académico utilizando variables socioeconómicas y educativas en contextos de Big Data educativo colombiano. Asimismo, los hallazgos evidencian que la combinación entre ingeniería de características, pipelines robustos de transformación y arquitecturas profundas adecuadamente regularizadas puede generar modelos predictivos sólidos y metodológicamente consistentes.

5.2 Métricas de desempeño

La evaluación experimental de los modelos implementados se realizó utilizando métricas ampliamente reconocidas en problemas de clasificación supervisada multiclase. Debido a la naturaleza parcialmente desbalanceada de las categorías de desempeño académico, el principal indicador utilizado para la comparación de arquitecturas fue el F1-score.

Tabla 12. *Métricas de desempeño de Modelos de clasificación.*

Modelo	Mínimo F1-Score	Máximo F1-Score	Media	Desviación estándar
MLP	0.657	0.662	0.660	0.001
CNN 1D	0.627	0.653	0.642	0.009
LightGBM	0.523	0.524	0.524	0.0003

El F1-score constituye la media armónica entre precisión (precision) y exhaustividad (recall), permitiendo evaluar simultáneamente capacidad de clasificación correcta y cobertura de predicciones relevantes. La literatura especializada recomienda esta métrica en escenarios donde las clases presentan distribuciones heterogéneas o cuando resulta necesario equilibrar sensibilidad y precisión predictiva [15].

Los resultados experimentales obtenidos muestran que las arquitecturas MLP alcanzaron el mejor desempeño global del estudio. Los modelos basados en funciones de activación LeakyReLU y Softplus obtuvieron los mayores valores promedio de F1-score durante las iteraciones experimentales desarrolladas.

El modelo MLP presentó un F1-score promedio aproximado de 0.660080, evidenciando elevada capacidad para clasificar correctamente los diferentes niveles de desempeño académico definidos en la investigación. Además del desempeño promedio, el modelo mostró desviaciones estándar relativamente bajas entre iteraciones, indicando estabilidad experimental y consistencia en los procesos de entrenamiento.

La arquitectura CNN 1D alcanzó un F1-score promedio cercano a 0.641512. Aunque este resultado es competitivo, el comportamiento estadístico evidenció mayor variabilidad entre iteraciones respecto a las arquitecturas densas MLP. Esta situación podría estar relacionada con sensibilidad de las convoluciones frente a reorganización estructural de variables tabulares y procesos de convergencia profunda.

Por otra parte, el modelo LightGBM obtuvo un F1-score promedio aproximado de 0.523659. Aunque el desempeño resultó inferior frente a redes profundas, el modelo mostró elevada estabilidad estadística y baja dispersión entre iteraciones, característica típica de algoritmos boosting basados en árboles.

Además del F1-score, la investigación contempló análisis complementarios utilizando precisión, recall y accuracy. Estas métricas permitieron evaluar diferentes dimensiones del comportamiento predictivo de los modelos.

Las matrices de confusión generadas durante el proceso experimental permitieron identificar patrones específicos de clasificación y categorías con mayores niveles de error. Particularmente, se observó que las categorías intermedias de desempeño presentaron mayores niveles de confusión frente a categorías extremas, comportamiento consistente con problemas ordinales multiclase donde existen regiones difusas entre niveles cercanos de desempeño académico.

Desde la perspectiva estadística, también se analizaron medidas de tendencia central y dispersión entre iteraciones experimentales. La evaluación de estabilidad resulta especialmente importante en aprendizaje profundo debido a la naturaleza estocástica de procesos de inicialización y optimización iterativa.

Los análisis realizados muestran que las arquitecturas MLP no solamente alcanzaron mejores métricas promedio, sino también mayor estabilidad experimental, aspecto relevante desde la perspectiva de reproducibilidad científica.

Además, el análisis de estabilidad experimental muestra que el modelo MLP no solamente obtuvo el mejor desempeño promedio, sino también una dispersión relativamente baja entre iteraciones, indicando robustez frente a procesos de inicialización y entrenamiento. Por otra parte, CNN 1D presentó mayor variabilidad estadística, sugiriendo sensibilidad frente a la reorganización estructural de las variables tabulares y a los procesos iterativos de convergencia profunda.

La elevada estabilidad observada en LightGBM coincide con investigaciones previas que destacan la robustez estadística de algoritmos boosting basados en árboles de decisión, especialmente en problemas tabulares estructurados [50]. Sin embargo, el menor desempeño predictivo alcanzado por este modelo sugiere que las relaciones existentes entre las variables socioeconómicas y el desempeño académico poseen complejidad no lineal que puede ser mejor capturada mediante arquitecturas profundas.

5.3 Análisis e interpretación de resultados

Los resultados obtenidos en la investigación evidencian que las variables socioeconómicas y educativas poseen capacidad predictiva significativa sobre el desempeño académico en las pruebas Saber 11. Este hallazgo coincide con investigaciones internacionales relacionadas con minería de datos educativa y aprendizaje automático aplicado a predicción del rendimiento estudiantil [13], [49].

Uno de los hallazgos más importantes corresponde a la superioridad observada en las arquitecturas profundas tipo MLP frente a modelos boosting y CNN 1D. Este comportamiento sugiere que las relaciones existentes entre las variables socioeconómicas y el desempeño.

Tabla 13. *Comparación de hallazgos frente a literatura científica.*

Estudio	Dataset	Mejor modelo reportado	Coincidencia con esta investigación

Suaza-Medina et al. (2024)	Saber 11	Machine Learning + SHAP	Sí
Zhu et al. (2025)	Scientific Literacy	Deep Learning	Sí
Fernández-Delgado et al. (2014)	Benchmark ML	Ensemble/Deep Models	Parcial
Gorishniy et al. (2023)	Tabular Data	MLP avanzados	Sí

Nota: Fue elaborada con base en literatura científica

Las arquitecturas MLP demostraron elevada capacidad para modelar interacciones complejas entre características familiares, institucionales y territoriales. Esto resulta consistente con estudios recientes que establecen que el rendimiento académico responde a sistemas complejos de interacción social y económica más que a relaciones lineales simples.

Por otra parte, aunque las CNN 1D presentaron resultados competitivos, el comportamiento observado sugiere que las dependencias locales capturadas mediante convoluciones podrían no representar completamente la complejidad estructural de los datos educativos analizados.

En relación con LightGBM, los resultados muestran que, aunque los modelos boosting poseen elevada robustez en problemas tabulares, las arquitecturas profundas lograron capturar patrones de mayor complejidad dentro del conjunto de datos analizado.

Otro aspecto relevante corresponde al impacto significativo del preprocesamiento híbrido implementado. La combinación de RobustScaler, OneHotEncoder y TargetEncoder permitió mejorar sustancialmente la representación computacional de las variables de entrada.

Además, los resultados obtenidos evidencian la importancia metodológica del control de *data leakage*. La corrección de fugas de información permitió obtener métricas más realistas y metodológicamente válidas, fortaleciendo consistencia científica del proyecto.

Los hallazgos obtenidos sugieren que modelos avanzados de inteligencia artificial pueden convertirse en herramientas relevantes para sistemas de alerta temprana educativa, identificación

de patrones de desigualdad y apoyo a toma de decisiones institucionales basadas en evidencia.

Por otra parte, la comparación de los resultados obtenidos frente a literatura científica reciente evidencia coherencia metodológica y experimental entre los hallazgos de esta investigación y estudios internacionales relacionados con minería de datos educativa y aprendizaje profundo sobre datos tabulares.

Particularmente, [59] identificaron que modelos de aprendizaje automático poseen elevada capacidad predictiva sobre el desempeño académico en pruebas estandarizadas colombianas cuando se incorporan variables socioeconómicas y técnicas avanzadas de interpretación. Del mismo modo, Zhu et al. (2025) reportaron que arquitecturas profundas pueden capturar relaciones complejas entre variables educativas y desempeño académico con resultados superiores frente a modelos estadísticos tradicionales.

Asimismo, los resultados obtenidos son consistentes con los hallazgos [58], quienes demostraron que arquitecturas profundas densas continúan siendo altamente competitivas en problemas tabulares complejos, particularmente cuando las relaciones entre variables poseen elevada no linealidad y multidimensionalidad.

5.4 Validación frente a los objetivos planteados

Los resultados obtenidos permiten afirmar que los objetivos planteados en la investigación fueron alcanzados satisfactoriamente.

En primer lugar, se logró consolidar y procesar exitosamente un conjunto masivo de microdatos educativos correspondientes al periodo 2014-2 a 2024-2, integrando información socioeconómica, institucional y académica de aproximadamente 7.2 millones de estudiantes colombianos.

También, fue posible diseñar e implementar un pipeline robusto de preparación y transformación de datos utilizando estrategias híbridas de preprocesamiento compatibles con

escenarios de Big Data educativo.

El proyecto también logró implementar y comparar múltiples arquitecturas supervisadas de inteligencia artificial, incluyendo MLP, CNN 1D y LightGBM, evaluando su comportamiento sobre un mismo conjunto experimental.

Los resultados obtenidos permitieron identificar que las arquitecturas profundas tipo MLP representan la alternativa más competitiva dentro de los modelos evaluados, alcanzando los mejores desempeños predictivos y mayor estabilidad experimental.

Adicionalmente, la investigación permitió validar la viabilidad técnica y metodológica de aplicar aprendizaje profundo sobre datos educativos tabulares masivos en el contexto colombiano, contribuyendo al fortalecimiento del campo de *Educational Data Mining* en el país.

Tabla 14. *Principales hallazgos de la investigación.*

Hallazgo	Evidencia experimental	Implicación
MLP obtuvo mejor desempeño	Mayor F1-score promedio	Relaciones no lineales complejas
CNN 1D fue competitivo	F1 superior a 0.64	Existencia de patrones locales parciales
LightGBM fue estable pero inferior	Baja desviación estándar	Limitaciones frente a complejidad educativa
Pipeline híbrido mejoró desempeño	Mejor convergencia y estabilidad	Relevancia del preprocesamiento
Variables socioeconómicas poseen alta capacidad predictiva	Clasificación multiclase consistente	Impacto estructural de desigualdad educativa

Los hallazgos obtenidos permiten concluir que las variables socioeconómicas y educativas poseen capacidad predictiva significativa sobre el desempeño académico en pruebas estandarizadas colombianas. Además, los resultados experimentales sugieren que arquitecturas profundas densas presentan ventajas relevantes para modelar relaciones complejas presentes en contextos educativos masivos.

Desde la perspectiva metodológica, la investigación demuestra que la combinación entre pipelines robustos de transformación, control de fuga de información y arquitecturas profundas adecuadamente regularizadas puede generar modelos predictivos sólidos y reproducibles en escenarios de Big Data educativo, minería de datos educativa e inteligencia artificial puede generar herramientas relevantes para comprensión y análisis del desempeño académico desde perspectivas predictivas y basadas en evidencia.

6 Conclusiones y recomendaciones.

6.1 Conclusiones generales del proyecto

La presente investigación permitió desarrollar, implementar y validar un sistema de clasificación supervisada orientado a la predicción del desempeño académico en las pruebas Saber 11 a partir de variables socioeconómicas, familiares, institucionales y educativas derivadas de los microdatos oficiales del ICFES correspondientes al periodo 2014-2 a 2024-2. El proyecto integró técnicas avanzadas de minería de datos educativa, aprendizaje automático y aprendizaje profundo sobre un conjunto masivo de aproximadamente 7.2 millones de registros estudiantiles, constituyéndose en uno de los ejercicios de analítica educativa más amplios desarrollados sobre

datos públicos del sistema educativo colombiano.

Los resultados obtenidos evidencian que las variables socioeconómicas y contextuales poseen una capacidad predictiva significativa sobre el desempeño académico de los estudiantes colombianos. Este hallazgo confirma la existencia de relaciones estructurales complejas entre condiciones familiares, acceso a recursos educativos, contexto institucional y rendimiento académico, resultados consistentes con investigaciones internacionales sobre desigualdad educativa y Educational Data Mining [11], [13], [49].

Uno de los principales aportes de la investigación corresponde a la validación experimental de arquitecturas profundas aplicadas a clasificación educativa sobre datos tabulares masivos. Los resultados obtenidos demuestran que las arquitecturas MLP basadas en capas densas profundas alcanzaron el mejor desempeño predictivo dentro de los modelos evaluados, superando a CNN 1D y LightGBM. Este comportamiento experimental sugiere que las relaciones presentes entre las variables socioeconómicas y el desempeño académico poseen naturaleza altamente no lineal y multidimensional, favoreciendo arquitecturas capaces de modelar interacciones complejas entre características heterogéneas.

Asimismo, la investigación permitió evidenciar que la calidad del pipeline de preprocesamiento constituye un componente crítico en proyectos de aprendizaje automático aplicados a datos educativos. La implementación híbrida de RobustScaler, OneHotEncoder y TargetEncoder permitió mejorar considerablemente la representación computacional de las variables de entrada y optimizar el comportamiento de los modelos supervisados. Particularmente, el uso adecuado de Target Encoding sobre variables de alta cardinalidad facilitó preservar capacidad predictiva relevante asociada a variables territoriales e institucionales.

Otro hallazgo metodológico relevante corresponde al impacto del control de fuga de información (*data leakage*) dentro de los procesos de transformación supervisada. La investigación

confirmó que configuraciones incorrectas de codificación categórica pueden producir sobreestimaciones artificiales del desempeño predictivo, afectando validez experimental y reproducibilidad científica. En consecuencia, el proyecto demuestra la importancia de implementar protocolos estrictos de separación entre entrenamiento, validación y prueba antes de realizar transformaciones supervisadas.

Desde la perspectiva computacional, el estudio demuestra la viabilidad técnica de desarrollar sistemas avanzados de inteligencia artificial educativa utilizando herramientas abiertas y entornos accesibles como Python y Google Colab. La utilización de almacenamiento Parquet y pipelines optimizados permitió manejar eficientemente millones de registros estudiantiles sin requerir infraestructura especializada de alto costo, aspecto relevante para futuras investigaciones académicas en contextos latinoamericanos.

La investigación también aporta evidencia empírica al campo de *Educational Data Mining* en Colombia, particularmente en relación con el uso de aprendizaje profundo aplicado a datos educativos masivos. Aunque investigaciones previas habían utilizado modelos tradicionales de aprendizaje automático sobre datos Saber 11, este proyecto amplía el estado del arte mediante la incorporación comparativa de arquitecturas profundas, convolucionales y boosting sobre una ventana histórica de diez años.

Otro aporte importante corresponde al diseño de una arquitectura analítica modular y reproducible. La estructuración del pipeline mediante notebooks independientes facilita trazabilidad metodológica, replicabilidad experimental y extensión futura hacia nuevos modelos, métricas y estrategias de interpretación. Esto convierte la solución desarrollada no solamente en un prototipo experimental, sino también en una plataforma potencial para futuras investigaciones de analítica educativa aplicada.

Los resultados experimentales obtenidos permiten además concluir que las arquitecturas

profundas tipo MLP representan una alternativa altamente competitiva para modelar fenómenos educativos complejos cuando se dispone de grandes volúmenes de datos estructurados. Este hallazgo coincide con tendencias recientes en inteligencia artificial aplicada a datos tabulares, donde modelos densos profundos continúan mostrando elevado potencial predictivo frente a escenarios de alta dimensionalidad y relaciones no lineales complejas [18].

Finalmente, la investigación evidencia que la integración entre minería de datos educativa, aprendizaje profundo y análisis socioeconómico puede contribuir significativamente a la comprensión del rendimiento académico desde una perspectiva basada en evidencia. Los hallazgos obtenidos poseen relevancia potencial para instituciones educativas, entidades gubernamentales e investigadores interesados en sistemas predictivos, analítica educativa y políticas públicas orientadas al mejoramiento de la calidad educativa en Colombia.

6.2 Cumplimiento de los objetivos

Los resultados obtenidos durante el desarrollo del proyecto permiten concluir que los objetivos planteados inicialmente fueron alcanzados satisfactoriamente tanto desde la perspectiva metodológica como experimental.

En relación con el objetivo general, se logró desarrollar e implementar un sistema basado en técnicas de aprendizaje automático y aprendizaje profundo para la clasificación del desempeño académico en las pruebas Saber 11 utilizando variables socioeconómicas y educativas. La solución desarrollada permitió entrenar, validar y comparar diferentes arquitecturas supervisadas sobre un conjunto masivo de datos educativos reales provenientes del ICFES.

Respecto al primer objetivo específico relacionado con la consolidación y preparación de datos, el proyecto logró integrar exitosamente los microdatos oficiales del examen Saber 11

correspondientes al periodo 2014-2 a 2024-2. Este proceso incluyó homologación estructural de variables, limpieza, transformación, normalización y almacenamiento eficiente en formato Parquet. Asimismo, se desarrolló un pipeline robusto de preparación de datos orientado a garantizar consistencia estadística y compatibilidad longitudinal entre periodos históricos.

En coherencia con el objetivo asociado a la implementación de modelos supervisados, se desarrollaron múltiples arquitecturas pertenecientes a diferentes paradigmas de inteligencia artificial, incluyendo MLP, CNN 1D y LightGBM. Cada modelo fue entrenado y evaluado bajo condiciones experimentales homogéneas, permitiendo realizar comparaciones metodológicamente válidas entre arquitecturas profundas y modelos boosting.

Por otra parte, el objetivo relacionado con la evaluación comparativa de desempeño fue alcanzado mediante la utilización de métricas de clasificación multiclase, particularmente F1-score, precisión, recall y análisis de estabilidad estadística. Los resultados experimentales permitieron identificar diferencias relevantes entre arquitecturas y establecer que las redes profundas MLP presentaron el comportamiento predictivo más robusto dentro de los modelos evaluados.

Además, se logró cumplir el objetivo relacionado con la validación experimental y reproducibilidad metodológica. La implementación de múltiples iteraciones independientes permitió analizar estabilidad, dispersión estadística y capacidad de generalización de los modelos desarrollados. Adicionalmente, la incorporación de mecanismos de control frente a *data leakage* fortaleció la validez científica de los resultados obtenidos.

Otro aspecto importante corresponde al cumplimiento del objetivo relacionado con la integración de técnicas modernas de procesamiento tabular y aprendizaje profundo aplicadas al contexto educativo colombiano. El proyecto no solamente implementó modelos supervisados avanzados, sino que además integró estrategias híbridas de transformación y regularización

compatibles con escenarios de Big Data educativo.

Finalmente, la investigación logró generar una arquitectura modular reproducible que puede servir como base para futuras investigaciones relacionadas con predicción educativa, sistemas de alerta temprana y analítica institucional basada en inteligencia artificial. En este sentido, el proyecto supera el alcance exclusivamente experimental y aporta una estructura metodológica reutilizable dentro del contexto académico colombiano.

6.3 Limitaciones identificadas

A pesar de los resultados obtenidos y de los aportes metodológicos alcanzados, la investigación presenta diversas limitaciones asociadas a aspectos metodológicos, computacionales y estructurales propios del conjunto de datos analizado.

Una de las principales limitaciones corresponde a la naturaleza observacional de los microdatos utilizados. Aunque los modelos desarrollados lograron identificar patrones predictivos relevantes, la investigación no permite establecer relaciones causales directas entre variables socioeconómicas y desempeño académico. En consecuencia, los resultados deben interpretarse desde una perspectiva predictiva y correlacional más que causal.

Otra limitación importante se relaciona con la calidad y heterogeneidad histórica de los datos publicados por el ICFES. Aunque el proyecto implementó procesos de homologación y limpieza, ciertos periodos históricos presentan diferencias estructurales en variables, categorías y formatos de captura. Estas inconsistencias obligaron a excluir determinadas variables o realizar procesos de transformación que podrían afectar parcialmente granularidad de la información original.

De esta forma, la investigación depende exclusivamente de variables disponibles dentro de los

microdatos públicos del examen Saber 11. En consecuencia, factores potencialmente relevantes como variables psicológicas, emocionales, pedagógicas o contextuales no pudieron ser incorporados debido a limitaciones de disponibilidad de información.

Desde la perspectiva computacional, otra limitación importante corresponde al elevado costo de procesamiento asociado al entrenamiento de arquitecturas profundas sobre millones de registros. Aunque Google Colab permitió desarrollar el proyecto utilizando infraestructura accesible, ciertas configuraciones experimentales avanzadas requirieron simplificaciones metodológicas relacionadas con tamaño de batch, número de épocas y complejidad arquitectónica.

En relación con las CNN 1D, una limitación metodológica importante corresponde a la representación secuencial artificial de variables tabulares. Aunque las convoluciones unidimensionales permitieron capturar ciertos patrones locales, las variables socioeconómicas no poseen necesariamente relaciones espaciales o secuenciales naturales equivalentes a aquellas presentes en imágenes o señales temporales. Esto podría limitar parcialmente la capacidad representacional de las arquitecturas convolucionales implementadas.

Por otra parte, aunque la investigación incorporó múltiples iteraciones experimentales, el proyecto no contempló validación cruzada completa debido al costo computacional asociado al volumen de datos procesado. En consecuencia, los resultados obtenidos deben interpretarse dentro de las condiciones experimentales específicas implementadas.

Otra limitación relevante corresponde a la ausencia de técnicas avanzadas de interpretabilidad explicativa como SHAP o LIME. Aunque el proyecto identificó capacidad predictiva significativa en variables socioeconómicas, no se desarrolló una interpretación profunda de importancia individual de características debido a restricciones temporales y computacionales.

Adicionalmente, la investigación se enfocó principalmente en clasificación categórica del desempeño académico y no en predicción continua del puntaje global. Aunque la categorización

facilita interpretación institucional y comparaciones multiclase, la discretización puede generar pérdida parcial de información respecto al comportamiento continuo de los puntajes originales.

Finalmente, otra limitación importante corresponde a la imposibilidad de evaluar el impacto longitudinal posterior del desempeño académico sobre trayectorias universitarias o laborales, dado que los microdatos utilizados corresponden exclusivamente al contexto del examen Saber 11 y no incluyen seguimiento educativo posterior.

6.4 Recomendaciones y trabajos futuros

A partir de los resultados obtenidos y de las limitaciones identificadas, se proponen diversas líneas de trabajo futuro orientadas a ampliar el alcance metodológico y científico de la investigación.

En primer lugar, se recomienda incorporar técnicas avanzadas de interpretabilidad explicativa como SHAP (*SHapley Additive exPlanations*) y LIME (*Local Interpretable Model-agnostic Explanations*). Estas metodologías permitirían identificar con mayor precisión la contribución individual de variables socioeconómicas, familiares e institucionales sobre las predicciones generadas por los modelos, fortaleciendo la utilidad analítica e interpretabilidad educativa de la solución desarrollada.

También, futuras investigaciones podrían incorporar arquitecturas más avanzadas de aprendizaje profundo orientadas específicamente a datos tabulares, incluyendo TabNet, FT-Transformer y arquitecturas híbridas basadas en atención. Investigaciones recientes sugieren que estos modelos poseen capacidad competitiva en problemas tabulares complejos y podrían mejorar el desempeño predictivo sobre datasets educativos masivos [18].

Otra línea de investigación relevante corresponde a la implementación de modelos híbridos

multimodales capaces de integrar simultáneamente variables estructuradas, texto libre y datos temporales. Esto permitiría incorporar información adicional relacionada con contexto institucional, evaluaciones cualitativas y trayectorias educativas longitudinales.

Se recomienda también desarrollar modelos de predicción continúa orientados a estimar directamente el puntaje global del examen Saber 11 mediante técnicas de regresión profunda. Esta aproximación permitiría conservar mayor granularidad de la información original y facilitar análisis más detallados del comportamiento académico.

Desde la perspectiva computacional, futuras investigaciones podrían incorporar entrenamiento distribuido y procesamiento paralelo utilizando infraestructura de alto rendimiento o entornos cloud especializados. Esto permitiría evaluar arquitecturas más complejas y realizar validación cruzada exhaustiva sobre el conjunto completo de datos.

Asimismo, se recomienda extender la investigación hacia sistemas de alerta temprana educativa aplicados a instituciones académicas colombianas. Los modelos desarrollados podrían adaptarse para identificar estudiantes en riesgo académico y apoyar estrategias institucionales de acompañamiento basadas en evidencia.

Otra línea futura importante corresponde al análisis longitudinal de trayectorias educativas. Integrar datos universitarios, resultados Saber Pro y variables laborales permitiría construir modelos predictivos más amplios orientados a analizar continuidad educativa y movilidad social.

Adicionalmente, futuras investigaciones podrían incorporar enfoques de aprendizaje causal y modelos explicativos orientados no solamente a predecir desempeño académico, sino también a identificar mecanismos estructurales asociados a desigualdad educativa.

Desde la perspectiva metodológica, se recomienda implementar estrategias más avanzadas de optimización automática de hiperparámetros mediante búsqueda bayesiana, AutoML y técnicas evolutivas híbridas. Estas aproximaciones podrían mejorar desempeño y eficiencia computacional

de los modelos desarrollados.

Finalmente, se considera relevante extender la aplicación de la arquitectura desarrollada hacia otros contextos educativos latinoamericanos. La modularidad y reproducibilidad del sistema permiten adaptar fácilmente el pipeline a pruebas estandarizadas de otros países, contribuyendo al fortalecimiento regional del campo de *Educational Data Mining* y analítica educativa basada en inteligencia artificial.

Referencias

[1] Nations, U. (2025). Educación para todos | Naciones Unidas. United Nations. Naciones Unidas. https://www.un.org/es/impacto-acad%C3%A9mico/educaci%C3%B3n-para-todos

[2] Naciones Unidas. (2023, January). Educación - Desarrollo Sostenible. Garantizar Una Educación Inclusiva, Equitativa y de Calidad y Promover Oportunidades de Aprendizaje Durante Toda La Vida Para Todos. https://www.un.org/sustainabledevelopment/es/education/

[3] Naciones Unidas. (2015). Educación - Desarrollo Sostenible. https://www.un.org/sustainabledevelopment/es/education/

[4] Constitución Política de Colombia. (1991). Constitución Política de Colombia Edición especial preparada por la Corte Constitucional. 1–125. https://s3.us-east-

2.amazonaws.com/cdn.miraquetemiro.org/Constitucion-politica-de-Colombia---2015_ffa0f55e3e98779b853a4b9e6a457eb8.pdf

[5] Lucia Ramírez De Rincón, M., Angulo, M. V., Colciencias, G., Fernando, D., Losada, H., Monroy, S. E., & Subdirectora, V. (2019). Misión internacional de sabios para el avance de la Ciencia, la Tecnología y la Innovación. Pacto por la Ciencia, la Tecnología y la Innovación: Un sistema para construir el conocimiento del futuro Presidencia de la República Iván Duque Márquez Vicepresidencia de la República.

[6] ICFES. (2023, September 23). Acerca del Examen Saber 11° . https://www.icfes.gov.co/evaluaciones-icfes/saber-11/

[7] PISA, Gustavo Petro Urrego, Figueroa Aurora Vergara, Sánchez Óscar, Blandón Bermúdez Elizabeth, Rafael Hoyos, Castrillón Michelle, Nathaly Córdoba, Montoya Jiménez, & Vera Alejandro López. (2024). Programa para la Evaluacion Internacional de Alumnos (PISA). https://www.mineducacion.gov.co/1780/articles-421217_recurso_03.pdf

[8] Zhu, L., You, H., Hong, M., & Fang, Z. (2025b). Predictive insights into U.S. students' mathematics performance on PISA 2022 using ensemble tree-based machine learning models. International Journal of Educational Research, 130, 102537. https://doi.org/10.1016/j.ijer.2025.102537

[9] Kang, J., & Keinonen, T. (2018). The Effect of Student-Centered Approaches on Students' Interest and Achievement in Science: Relevant Topic-Based, Open and Guided Inquiry-Based, and Discussion-Based Approaches. Research in Science Education, 48(4), 865–885. https://doi.org/10.1007/S11165-016-9590-2

[10] Alkan, B. B., Kuzucuk, S., Odabasi, Ş. Y., & Karakuş, L. (2025a). Educational improvement through machine learning: Strategic models for better PISA scores. PLOS ONE, 20(7), e0326121. https://doi.org/10.1371/JOURNAL.PONE.0326121

[11] Sirin, S. R. (2005a). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453. https://doi.org/10.3102/00346543075003417

[12] Zhu, L., You, H., Hong, M., & Fang, Z. (2025c). Predictive insights into U.S. students' mathematics performance on PISA 2022 using ensemble tree-based machine learning models. International Journal of Educational Research, 130, 102537. https://doi.org/10.1016/j.ijer.2025.102537

[13] Zhu, L., You, H., Hong, M., & Fang, Z. (2025a). Corrigendum to "Predictive Insights into U.S. Students' Mathematics Performance on PISA 2022 Using Ensemble Tree-Based Machine Learning Models" [International Journal of Educational Research 130 (2025) 102537]. International Journal of Educational Research, 131, 102592. https://doi.org/10.1016/J.IJER.2025.102592

[14] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007a). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 2007 26:3, 26(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

[15] Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? In Journal of Machine Learning Research (Vol. 15). http://www.mathworks.es/products/neural-network.

[16] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007b). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 2007 26:3, 26(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

[17] Arabnejad, H., & Barbosa, J. G. (2017). Multi-QoS constrained and Profit-aware scheduling approach for concurrent workflows on heterogeneous systems. Future Generation Computer

Systems, 68, 211–221. https://doi.org/10.1016/j.future.2016.10.003

[18] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2023a). Revisiting Deep Learning Models for Tabular Data. http://arxiv.org/abs/2106.11959

[19] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[20] Baker, R. S., & Inventado, P. S. (2014a). Educational Data Mining and Learning Analytics. Learning Analytics: From Research to Practice, 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4

[21] OECD Economic Surveys. (2024b, September 17). Estudios Económicos de la OCDE: Colombia 2024 | OECD. Estudios Económicos de La OCDE: Colombia 2024. https://www.oecd.org/es/publications/estudios-economicos-de-la-ocde-colombia-2024_e61e16ad-es.html

[22]. Z. Xi and G. Panoutsos, “Interpretable Machine Learning: Convolutional Neural Networks with RBF Fuzzy Logic Classification Rules,” 9th International Conference on Intelligent Systems 2018: Theory, Research and Innovation in Applications, IS 2018 - Proceedings, pp. 448–454, Jul. 2018, doi: 10.1109/IS.2018.8710470.

[23] Bruce, P. C., Shmueli, G., Yahav, I., & Kenneth C. Lichtendahl. (2018). DATA MINING FOR BUSINESS ANALYTICS. https://books.google.com.mx/books?id=ETwuDwAAQBAJ&printsec=copyright#v=onepage&q&f=false

[24] TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX (236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[25] López, R., Mántaras, D. E., & Pere Brunet, Y. (2023). ¿Qué es la inteligencia artificial? A

fondo. 164, 13–21. https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/una-nueva-era-para-el-alzhimer-803/el-

[26] Sebastián Raschka, Yuxi (Hayden) Liu, & Vahid Mirjalili. (2023). Machine Learning con PyTorch y Scikit-Learn. Marcombo S.L.

[27] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

[28] Foster, David., & Aranda González, Virginia. (2023). Deep learning generativo : cómo enseñar a las máquinas a dibujar, escribir, componer y reproducir música (Ediciones ANAYA multimedia, Ed.; 2a. Ed). Anaya Multimedia.

[28] Mejía Trejo, Juan. (2025). Inteligencia Artificial y su repercusión en la educación superior.

[29] Mcculloch, W. S., & Pitts, W. (1943). A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY. BULLETIN OF MATHEMATICAL BIOPHYSICS, 5. https://home.csulb.edu/~cwallis/382/readings/482/mccolloch.logical.calculus.ideas.1943.pdf

[30] Williamson Ben. (2017). Big Data en Educación. El futuro digital del aprendizaje, la politica y la práctica (Roc Filella, Tran.; Ediciones Morata S.L). https://edmorata.es/wp-content/uploads/2020/06/Williamson.BigData.PR_.pdf

[31] Barney Jay. (1991). Firm Resources And Sustained Competitive Advantage. Journal of Management, 17, 99–120a. https://josephmahoney.web.illinois.edu/BA545_Fall%202022/Barney%20(1991).pdf

[32] Solano, J. A., Lancheros Cuesta, D. J., Umaña Ibáñez, S. F., & Coronado-Hernández, J. R. (2022). Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test. Procedia Computer Science, 198, 512–517. https://doi.org/10.1016/J.PROCS.2021.12.278

[33] Suaza-Medina, M., Peñabaena-Niebles, R., & Jubiz-Diaz, M. (2024a). A model for predicting

academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations. Scientific Reports 2024 14:1, 14(1), 1–17. https://doi.org/10.1038/S41598-024-76596-3

[34] You, H., Hong, M., Zhu, L., & Zhenhan, F. (2025). Machine Learning Approaches for Predicting U.S. Students' Scientific Literacy: An Analysis of Key Factors Across Performance Levels and Socioeconomic Statuses. International Journal of Science and Mathematics Education, 1–29. https://doi.org/10.1007/S10763-025-10545-Y/FIGURES/2

[35] Alkan, B. B., Kuzucuk, S., Odabasi, Ş. Y., & Karakuş, L. (2025b). Educational improvement through machine learning: Strategic models for better PISA scores. PLOS ONE, 20(7), e0326121. https://doi.org/10.1371/JOURNAL.PONE.0326121

[36] Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. Journal of Political Economy, 70(5, Part 2), 9–49. https://doi.org/10.1086/258724

[37] Sirin, S. R. (2005b). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453. https://doi.org/10.3102/00346543075003417

[38] Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). Minería de datos: modelos y algoritmos. In Editorial UOC (Ed.), Editorial UOC (Primera Edición, Number 1). Editorial UOC. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part%0Ahttp://www.editorialuoc.com

[39] Instituto Colombiano para la Evaluacion de la Educacion - ICFES. (2024). Diccionario Examen Saber 11°. DataIcfes:Repositorio de Datos Abiertos del Icfes. https://icfesgovco.sharepoint.com/sites/BasesDataIcfes/Documentos%20compartidos/Forms/AllItems.aspx?id=%2Fsites%2FBasesDataIcfes%2FDocumentos%20compartidos%2F01%5FExamen

%20Saber%2011%C2%B0%2F03%5FDocumentaci%C3%B3n%20T%C3%A9cnica%2FDiccionario%20Examen%20Saber%2011%C2%B0%2Epdf&parent=%2Fsites%2FBasesDataIcfes%2FDocumentos%20compartidos%2F01%5FExamen%20Saber%2011%C2%B0%2F03%5FDocumentaci%C3%B3n%20T%C3%A9cnica&p=true&ga=1

[40] ICFES. (2026). DataIcfes: Repositorio de datos abiertos del ICFES.

[41] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

[42] Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter, 3(1), 27–32. https://doi.org/10.1145/507533.507538

[43] Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In Journal of Machine Learning Research (Vol. 12). http://scikit-learn.sourceforge.net.

[44] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579.

[45] Raschka, S., & Mirjalili, V. (2020). Python Machine Learning: aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow. 618. https://www.google.com.pe/books/edition/Python_Machine_Learning/5EtOEAAAQBAJ?hl=es-419&gbpv=0.

[46] Torres, J. (2020). Python Deep Learning: introducción práctica con Keras y TensorFlow 2 (Marcombo, Ed.; Primera). Marcombo. https://elibro.net/en/ereader/usta/281442.

[47] Ruiz Sarrias, O. (2024). Las matemáticas de la IA: introducción al Deep Learning (Los Libros de la Catarata, Ed.; Primera). Los libros de la Catarata. https://elibro.net/es/lc/unir/titulos/284720

[48] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3.

[49] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3). https://doi.org/10.1002/widm.1355

[50] Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785.

[51] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[52] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. http://arxiv.org/abs/1502.03167

[53] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.

[54] Fávero, L. P., & Belfiore, P. (2019). Hypotheses Tests. In Data Science for Business and Decision Making (pp. 199–248). Elsevier. https://doi.org/10.1016/b978-0-12-811216-8.00009-4

[55] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 IEEE International Conference on Computer Vision (ICCV), 1026–1034. https://doi.org/10.1109/ICCV.2015.123.

[56] Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980.

[59] Suaza-Medina, M., Peñabaena-Niebles, R., & Jubiz-Diaz, M. (2024b). A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations. Scientific Reports, 14(1), 25306. https://doi.org/10.1038/s41598-024-76596-3.

[60] Fávero, L. P., & Belfiore, P. (2019). Hypotheses Tests. In Data Science for Business and Decision Making (pp. 199–248). Elsevier. https://doi.org/10.1016/b978-0-12-811216-8.00009-4.