MINERÍA DE DATOS EDUCATIVOS PARA IDENTIFICAR FACTORES QUE AFECTAN EL RENDIMIENTO ACADÉMICO EN COLOMBIA

Diana Milena Beltrán García
Código 9019473
Joan Sebastián García
Código 9026103

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Cristian Alejandro Torres Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2026

FICHA RESUMEN

TÍTULO DEL PROYECTO: Minería de Datos Educativos para Identificar Factores que Afectan el Rendimiento Académico en Colombia

1. ÉNFASIS: Sistemas y Computación
2. TIPO DE PROYECTO: Investigación
3. ÁREA DE TRABAJO: Minería de Datos Educativos y Aprendizaje Automático
4. ESTUDIANTE(S): Diana Milena Beltrán García, Joan Sebastián García Fernández
5. CORREO ELECTRÓNICO: dbeltran@javerianacali.edu.co, Sebasgafer95@javerianacali.edu.co
6. DIRECCIÓN Y TELEFONO: 3138090147; Calle 22b # 63-24 Bogotá, 3013345004 /3015351845; Calle 150#15-39 Bogotá
7. DIRECTOR: Cristian Alejandro Torres Valencia
8. VINCULACIÓN DEL DIRECTOR: Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: cristian.torres@javerianacali.edu.co
10. CODIRECTOR: No aplica
11. GRUPO O EMPRESA QUE LO AVALA: No aplica
12. OTROS GRUPOS O EMPRESAS: No aplica
13. PALABRAS CLAVE (al menos 5): minería de datos educativos, rendimiento académico, aprendizaje automático, brechas educativas, análisis comparativo.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): ODS 4 – Educación de Calidad
15. FECHA DE INICIO (Desarrollo del proyecto): 1/07/2025
16. RESUMEN: Este proyecto aplicado identificó y analizó los factores que incidieron significativamente en el rendimiento académico de los estudiantes en Colombia, utilizando técnicas de minería de datos educativos (EDM) y modelos de aprendizaje automático. La investigación partió del reconocimiento de que, a pesar de la disponibilidad de datos generados en instituciones educativas, estos no se aprovechaban adecuadamente para fundamentar decisiones pedagógicas y políticas educativas.

    La problemática abordada se centró en la alta heterogeneidad del rendimiento académico de los estudiantes, evidenciada en brechas de aprendizaje y oportunidades entre sectores rurales y urbanos, así como entre instituciones oficiales y no oficiales.

    Para dar respuesta a esta problemática, se recopilaron y depuraron datos académicos provenientes de fuentes abiertas del ICFES. Posteriormente, se identificaron variables

significativas mediante técnicas de minería de datos y se construyeron modelos predictivos orientados a estimar el rendimiento académico.

El proyecto también realizó un análisis comparativo entre contextos geográficos, evidenciando brechas educativas y permitiendo formular recomendaciones basadas en evidencia.

Entre los resultados más relevantes se destacaron la identificación de factores socioeconómicos e institucionales asociados al rendimiento académico, la validación del modelo XGBoost como el algoritmo con mejor desempeño predictivo y estabilidad, así como la identificación de diferencias significativas entre contextos rurales y urbanos.

Estos resultados pueden servir como insumo para la toma de decisiones pedagógicas y administrativas en instituciones educativas, así como para la formulación de políticas orientadas a la equidad educativa en Colombia.

TABLA DE CONTENIDO

LISTA DE GRAFICAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

En los sistemas educativos colombianos persisten desigualdades importantes que se reflejan en el desempeño académico de los estudiantes, evidenciadas en tasas de bajo rendimiento, repitencia y abandono escolar. Estas brechas son más evidentes en contextos con limitaciones socioeconómicas, geográficas e institucionales, como lo señala el informe Análisis de resultados – Prueba Saber 11° 2024 del Observatorio de Gestión Educativa EXE [1].

A pesar de la gran cantidad de información disponible en las instituciones educativas, estos datos no siempre son analizados de forma sistemática ni utilizados de manera efectiva para la toma de decisiones pedagógicas o de política pública.

Ante esta situación, surgió la necesidad de aprovechar el potencial de la Minería de Datos Educativos (EDM) como enfoque analítico para comprender los factores que inciden en el desempeño académico. La EDM permitió analizar grandes volúmenes de datos con el objetivo de identificar patrones relevantes que apoyaran la toma de decisiones basadas en evidencia.

En este contexto, el presente proyecto tuvo como propósito identificar los factores que influyen en el rendimiento académico de los estudiantes mediante técnicas de minería de datos y modelos de aprendizaje automático supervisado. Para ello, se analizaron variables académicas, socioeconómicas, institucionales y del entorno a partir de bases de datos abiertas como ICFES. Posteriormente, se evaluaron modelos predictivos utilizando particiones de entrenamiento y prueba, validación por bloques y métricas de desempeño.

Se identificaron los factores más relevantes asociados al rendimiento académico, se evaluaron modelos con buena capacidad predictiva e interpretativa, y se evidenciaron posibles brechas entre contextos rurales y urbanos. Estos resultados aportaron insumos para la toma de decisiones educativas basadas en evidencia.

El documento se estructura de la siguiente manera: en el capítulo 1 se presenta el marco de referencia, incluyendo el marco teórico y antecedentes; posteriormente se desarrollan los capítulos correspondientes a los objetivos específicos, relacionados con el análisis de datos, modelado y evaluación de resultados.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Definición del problema

En el contexto educativo colombiano, el análisis del rendimiento académico de los estudiantes evidencia diferencias importantes entre distintos grupos poblacionales, las cuales están asociadas a factores socioeconómicos, geográficos e institucionales. De acuerdo con el informe Análisis de resultados – Prueba Saber 11 (2024) del Observatorio de Gestión Educativa EXE, las diferencias en el desempeño entre colegios oficiales y no oficiales, así como entre zonas urbanas y rurales, pueden ser hasta tres veces mayores [1], lo que refleja la persistencia de desigualdades en el sistema educativo.

1.1.1. Planteamiento del problema

En Colombia persisten diferencias significativas en el rendimiento académico de los estudiantes, asociadas a factores socioeconómicos, geográficos e institucionales. Estas desigualdades se reflejan en los resultados de evaluaciones estandarizadas como las pruebas Saber 11°, donde se evidencian brechas entre estudiantes de zonas rurales y urbanas, así como entre instituciones educativas oficiales y no oficiales. La comprensión de los factores que explican estas diferencias constituye un desafío relevante para el sistema educativo, debido a su impacto sobre las oportunidades de aprendizaje, la permanencia escolar y la calidad de la educación.

Esta situación se refleja en resultados académicos deficientes, en la persistencia de brechas de aprendizaje entre distintos grupos poblacionales y en una limitada capacidad institucional para anticipar o intervenir oportunamente sobre los factores que afectan el desempeño de los estudiantes. A nivel estructural, esta problemática se relaciona con múltiples factores, entre ellos la falta de modelos predictivos que integren variables académicas y no académicas, el desconocimiento de patrones presentes en los datos escolares y la escasa incorporación de herramientas analíticas avanzadas en la planificación educativa.

De mantenerse esta situación, existe el riesgo de profundizar las brechas de rendimiento entre estudiantes, deteriorar la eficiencia del sistema educativo y reducir el impacto de las intervenciones institucionales, debido a la falta de evidencia sólida que sustente la toma de decisiones. Además, se limita la posibilidad de diseñar estrategias de enseñanza y apoyo más personalizadas, lo que dificulta avanzar hacia una educación más equitativa y de calidad.

En este contexto, resulta necesario comprender de manera más profunda y basada en datos los factores que inciden en el rendimiento académico de los estudiantes. La Minería de Datos Educativos (EDM), como campo interdisciplinario dentro de la Ciencia de Datos, permite analizar grandes volúmenes de información educativa para identificar patrones, relaciones y variables significativas que impactan el aprendizaje. No obstante, persiste la necesidad de fortalecer la identificación y caracterización de los factores asociados al rendimiento académico mediante

enfoques de Minería de Datos Educativos, con el fin de generar evidencia que contribuya a la toma de decisiones y al diseño de estrategias orientadas al mejoramiento de la calidad educativa.

1.1.2. Formulación del problema

¿Cuáles son los factores que inciden significativamente en el rendimiento académico de los estudiantes, identificados mediante técnicas de Minería de Datos Educativos (EDM), y cómo puede esta información contribuir a la mejora de la calidad de la educación?

1.1.3. Sistematización

Para responder a esta pregunta general, se plantean los siguientes interrogantes que abordan aspectos clave del fenómeno en estudio:

- ¿Qué variables académicas, socioeconómicas, institucionales y/o contextuales muestran mayor correlación con el rendimiento académico según el análisis de datos educativos disponibles?

- ¿Qué modelos de aprendizaje automático ofrecen mayor precisión y explicabilidad para predecir el rendimiento académico estudiantil?

- ¿Qué diferencias se observan en los factores incidentes del rendimiento académico al comparar distintos contextos (por ejemplo, zonas rurales vs. urbanas)?

- ¿Qué implicaciones tienen los hallazgos de la Minería de Datos Educativos para la toma de decisiones pedagógicas y administrativas?

1.2. Objetivos

1.2.1. Objetivo General

Identificar los factores que inciden significativamente en el rendimiento académico de los estudiantes, mediante el uso de técnicas de Minería de Datos Educativos (EDM) y modelos de aprendizaje automático, para aportar evidencia que contribuya a la mejora de la calidad educativa

1.2.2. Objetivos Específicos

- Analizar las variables académicas, socioeconómicas, institucionales y/o contextuales presentes en los datos educativos disponibles, utilizando herramientas de minería de datos, para determinar su grado de asociación con el rendimiento académico.

- Comparar los factores identificados que inciden en el rendimiento académico entre distintos contextos educativos (rural vs. urbano, por ejemplo), para evidenciar posibles brechas o patrones diferenciales.

- Evaluar modelos de aprendizaje automático aplicados a datos educativos, mediante técnicas de validación cruzada y métricas de desempeño, para identificar aquellos con mayor precisión en la predicción del rendimiento académico.

- Examinar las implicaciones de los hallazgos obtenidos a partir del análisis de datos educativos, a través del cruce con literatura académica y revisión de políticas institucionales, para generar recomendaciones que fortalezcan la toma de decisiones en los ámbitos pedagógico y educativo.

1.3. Marco de Referencia

1.3.1. Marco Teórico

La minería de datos educativos (EDM, por sus siglas en inglés) es un campo interdisciplinar que combina la informática, la estadística y la pedagogía para analizar grandes volúmenes de datos generados en entornos educativos. Su propósito es descubrir patrones, relaciones y conocimientos ocultos que permitan mejorar los procesos de enseñanza y aprendizaje [2].

Como señalan Romero y Ventura, la EDM surgió como respuesta a la necesidad de aprovechar la creciente cantidad de datos generados en entornos educativos, incluyendo sistemas de gestión académica, plataformas de aprendizaje virtual y resultados de evaluaciones estandarizadas [2].

La EDM se basa en técnicas tradicionales de minería de datos, como clasificación, regresión, agrupamiento y descubrimiento de reglas de asociación, adaptadas a las particularidades del entorno educativo [2]. Su aplicación permite desarrollar modelos predictivos que identifican factores clave en el rendimiento estudiantil y ofrecen herramientas para la personalización de la enseñanza y la toma de decisiones informadas [3].

Uno de los principales aportes de la EDM es la posibilidad de construir modelos de aprendizaje automático supervisado que, a partir de datos etiquetados, permiten predecir resultados académicos futuros. Estos modelos emplean algoritmos como redes neuronales profundas (DNN), máquinas de soporte vectorial (SVM), XGBoost (Extreme Gradient Boosting) y bosques aleatorios (RF), entre otros [4].

Por ejemplo, en un estudio reciente se evaluó la eficacia de diversos modelos de aprendizaje automático y profundo para predecir el desempeño académico de los estudiantes, encontrándose

que las redes neuronales profundas superaron a otros modelos tradicionales en términos de precisión y capacidad de detección temprana de estudiantes en riesgo [4].

El uso de modelos supervisados también se ha extendido hacia enfoques explicables, como el uso de técnicas de interpretabilidad tipo SHAP (Shapley Additive Explanation), que permiten comprender cómo influyen las variables en las predicciones [5]. Esto es fundamental en entornos educativos, donde la transparencia y la ética en la toma de decisiones son prioritarias.

La investigación de Sanfo destaca la relevancia de factores como la infraestructura escolar, el involucramiento comunitario y la experiencia docente en la predicción de resultados educativos, lo que subraya la importancia de un enfoque holístico e interpretativo [5].

El rendimiento académico constituye un constructo ampliamente utilizado en investigación educativa para describir el nivel de logro alcanzado por los estudiantes respecto a los objetivos de aprendizaje establecidos. Este constructo integra conocimientos, habilidades y competencias desarrolladas durante el proceso formativo, y puede ser evaluado mediante diferentes indicadores de desempeño académico. En el presente estudio, el rendimiento académico se operacionaliza mediante el Puntaje Global obtenido en la prueba Saber 11°, utilizado como indicador cuantitativo del rendimiento académico de los estudiantes.

Desde el punto de vista del rendimiento académico, diversos estudios han demostrado la complejidad y multifactorialidad de este fenómeno. Según el análisis de los resultados de la prueba Saber 11° en Colombia, el rendimiento académico presenta variaciones significativas relacionadas con factores socioeconómicos, geográficos y de género [1].

Las brechas entre estudiantes de zonas rurales y urbanas, así como entre sectores oficiales y no oficiales, reflejan desigualdades estructurales que afectan las oportunidades de aprendizaje [1]. Estos hallazgos respaldan la pertinencia de emplear modelos que integren variables académicas y no académicas, como las condiciones del hogar, el nivel socioeconómico y la infraestructura escolar.

La EDM también ha demostrado su potencial en la generación de sistemas de recomendación y orientación académica. En estudios internacionales, como el realizado por Kord et al., se han aplicado modelos de aprendizaje automático y profundo para predecir el desempeño de los estudiantes y recomendar cursos y trayectorias académicas más adecuadas, con resultados significativos en la mejora del rendimiento y la reducción de la deserción [3].

Esto refuerza la relevancia de herramientas analíticas que no solo identifiquen patrones, sino que también ofrezcan recomendaciones personalizadas basadas en datos confiables. Por otra parte, investigaciones recientes han abordado el uso de redes neuronales profundas para la predicción del rendimiento en cursos específicos, destacando su capacidad para gestionar grandes volúmenes de datos y resolver problemas de clases desbalanceadas mediante técnicas de

remuestreo como SMOTE y ADASYN [4]. Esto permite un análisis más robusto y fiable, adaptado a las características particulares de los datos educativos.

En el contexto colombiano, el acceso a bases de datos públicas como las del ICFES y el DANE brinda una oportunidad única para aplicar estos enfoques de EDM a nivel nacional. Estas bases de datos contienen información detallada sobre resultados académicos, características socioeconómicas y demográficas, y otras variables relevantes que permiten construir modelos predictivos sólidos y representativos del contexto local.

Sin embargo, la efectividad de estos modelos depende de la calidad y consistencia de los datos disponibles, así como de la capacidad técnica para su procesamiento y análisis.

Sin embargo, es importante señalar que la predicción del rendimiento académico presenta limitaciones inherentes asociadas a la naturaleza del fenómeno y de los datos disponibles. Diversos estudios han evidenciado que el rendimiento estudiantil es un proceso multifactorial influenciado por variables cognitivas, socioeconómicas, emocionales y contextuales, muchas de las cuales no son completamente observables o medibles en bases de datos estructuradas [11], [18].

En este sentido, la capacidad predictiva de los modelos de aprendizaje automático aplicados a datos educativos suele ser moderada, especialmente cuando se utilizan conjuntos de datos tabulares. Investigaciones recientes han demostrado que, incluso con el uso de algoritmos avanzados, los modelos presentan limitaciones en la explicación de la variabilidad total del rendimiento académico debido a la presencia de ruido, variables omitidas y relaciones complejas [17].

Adicionalmente, desde la perspectiva de la analítica predictiva, no todos los problemas permiten alcanzar altos niveles de precisión, particularmente en contextos donde intervienen factores sociales y humanos difíciles de cuantificar [19]. Por tanto, los resultados obtenidos en este tipo de estudios deben interpretarse considerando estas restricciones, reconociendo que un desempeño moderado puede ser coherente con la naturaleza del problema y con lo reportado en la literatura.

A modo de síntesis, el marco conceptual que guía este proyecto combina elementos teóricos y metodológicos de la minería de datos, el aprendizaje automático y la analítica educativa, sustentados en literatura académica y estudios previos. La EDM se establece como un enfoque prometedor para enfrentar los desafíos educativos actuales, ofreciendo herramientas para personalizar la enseñanza, optimizar la gestión escolar y reducir las desigualdades de aprendizaje.

1.3.2. Modelos de Aprendizaje Automático para la Predicción del Rendimiento Académico

Dentro del campo de la Minería de Datos Educativos (EDM), se han utilizado diversos modelos de aprendizaje automático con el fin de predecir y analizar el rendimiento académico de los

estudiantes. Estos modelos permiten comprender cómo influyen las variables académicas, socioeconómicas e institucionales sobre los resultados obtenidos en pruebas estandarizadas como Saber 11°, contribuyendo a la toma de decisiones basadas en evidencia [2].

Para el presente proyecto se consideraron cuatro enfoques representativos: **XGBoost (Extreme Gradient Boosting)**, **Random Forest**, **Máquinas de Soporte Vectorial (SVM-RBF)** y **Redes Neuronales Artificiales (MLP)**. La elección de estos modelos responde a su amplio uso en la literatura y a su capacidad para equilibrar precisión, interpretabilidad y robustez frente a distintos tipos de datos educativos [3][4].

- **XGBoost (Extreme Gradient Boosting):**

  Este modelo se caracteriza por su alta precisión y eficiencia computacional. Se basa en la construcción secuencial de árboles de decisión, donde cada árbol corrige los errores del anterior mediante la optimización por gradiente. Esta estructura permite capturar relaciones no lineales y mejorar la capacidad predictiva del modelo, manteniendo mecanismos de regularización que reducen el sobreajuste. En el contexto del puntaje global, XGBoost ofrece un rendimiento robusto y es ampliamente utilizado en problemas de predicción educativa debido a su estabilidad y capacidad para manejar múltiples variables predictoras [6].

- **Random Forest:**

  Este método amplía el principio de los árboles de decisión mediante la creación de múltiples árboles entrenados con subconjuntos de datos. Al combinar sus resultados, se obtiene una predicción más estable y precisa. Además, permite estimar la importancia relativa de cada variable, lo que ayuda a interpretar qué factores tienen mayor incidencia sobre el rendimiento académico [5].

- **Máquinas de Soporte Vectorial (SVM-RBF):**

  Las SVM son modelos potentes para la clasificación y regresión, especialmente con kernel radial (RBF), ya que permiten capturar relaciones no lineales entre las variables. Se destacan por su precisión en escenarios con gran número de atributos, aunque requieren un ajuste cuidadoso de parámetros y no son tan interpretables como los modelos basados en árboles [4].

- **Redes Neuronales Artificiales (MLP):**

  El modelo Multilayer Perceptron (MLP) pertenece a las redes neuronales profundas y se utiliza cuando se busca capturar interacciones complejas entre variables. Aunque ofrecen

alta precisión, presentan menor interpretabilidad, lo que las hace más útiles en fases avanzadas del modelado o para comparar resultados con modelos más explicativos [4].

Diversos estudios han evidenciado que los modelos basados en árboles, como los Árboles de Decisión, Random Forest y XGBoost, tienden a ser los más adecuados cuando se requiere un equilibrio entre interpretabilidad, robustez y desempeño predictivo [3][5][6]. Estos modelos permiten capturar relaciones complejas entre variables y presentan una alta estabilidad en escenarios educativos. Sin embargo, los modelos SVM y MLP pueden ser útiles como contraste para validar la consistencia de los resultados y explorar patrones no lineales que podrían no ser completamente capturados por los métodos basados en árboles.

Por esta razón, el proyecto evaluará la aplicabilidad de estos cuatro modelos con el fin de determinar cuál ofrece **mejor capacidad predictiva y explicativa** en el contexto de los datos del ICFES. Esta comparación permitirá definir si se mantiene un solo modelo final o si se opta por un enfoque combinado.

1.3.3. Métricas y criterios teóricos para la evaluación de modelos predictivos

El proceso de evaluación de los modelos de aprendizaje automático es esencial dentro de la Minería de Datos Educativos, pues permite determinar su capacidad para predecir el rendimiento académico y comparar objetivamente su desempeño.

En este proyecto, el enfoque principal será la regresión, dado que el objetivo es predecir el puntaje global obtenido por los estudiantes en las pruebas Saber 11°.

Sin embargo, de manera complementaria, se explorará también la clasificación del rendimiento en niveles (por ejemplo, bajo, medio y alto), con el fin de validar la consistencia y aplicabilidad práctica de los modelos.

Por tanto, se consideran dos tipos de métricas: las orientadas a regresión (enfoque principal) y las orientadas a clasificación (enfoque exploratorio).

- **Métricas para tareas de regresión (enfoque principal)**

Dado que el objetivo principal del presente estudio consiste en predecir el puntaje global como una variable continua, la evaluación del desempeño de los modelos se realizará mediante métricas comúnmente utilizadas en problemas de regresión. Estas métricas cuantifican la diferencia entre los valores observados y los valores estimados por el modelo.

- **Error Cuadrático Medio (MSE):** (Mean Squared Error, MSE) mide el promedio de los cuadrados de las diferencias entre los valores reales y los valores predichos por el modelo. Se define como:

$$MSE = \frac{1}{n} = \sum_{i=1}^{n} (yi - \hat{y}i)^2$$

donde yi representa el valor real, $\hat{y}i$ el valor predicho por el modelo y n n el número total de observaciones. Esta métrica penaliza con mayor intensidad los errores grandes debido al término cuadrático.

- **Raíz del Error Cuadrático Medio (RMSE):** (Root Mean Squared Error, RMSE) corresponde a la raíz cuadrada del MSE y se expresa como:

$$RMSE = \sqrt{\frac{1}{n} = \sum_{i=1}^{n} (yi - \hat{y}i)^2}$$

Una de sus principales ventajas es que conserva las mismas unidades de la variable objetivo, lo que facilita la interpretación de los resultados en el contexto del problema analizado.

- **Error Absoluto Medio (MAE):** (Mean Absolute Error, MAE) mide el promedio de las diferencias absolutas entre los valores reales y los valores predichos:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

A diferencia del MSE, esta métrica es menos sensible a valores atípicos, lo que permite obtener una estimación más robusta del error promedio.

- **Coeficiente de Determinación (R^2):** Mide la proporción de la varianza de la variable dependiente que es explicada por el modelo. Se calcula mediante**:**

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

donde $\bar{y}$ representa el promedio de los valores observados. Valores de R^2cercanos a 1 indican una mayor capacidad del modelo para explicar la variabilidad de los datos.

En conjunto, estas métricas permiten comparar el rendimiento de diferentes modelos de regresión y seleccionar aquel que ofrezca el mejor equilibrio entre precisión predictiva, estabilidad y capacidad de interpretación.

- **Métricas para tareas de clasificación (validación complementaria)**

En el ámbito de la Minería de Datos Educativos, es común complementar el análisis de regresión con enfoques de clasificación, en los cuales el rendimiento académico se categoriza en niveles discretos (por ejemplo, bajo, medio o alto) a partir de umbrales definidos sobre el puntaje global.

En este contexto, la literatura reporta el uso de diversas métricas de evaluación basadas en la matriz de confusión, que considera cuatro tipos de resultados: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).

- **Exactitud (Accuracy):** Mide la proporción total de predicciones correctas realizadas por el modelo:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precisión (Precision):** Indica la proporción de predicciones positivas que fueron correctamente identificadas:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall o Sensibilidad:** Mide la capacidad del modelo para identificar correctamente los casos positivos:

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:** Corresponde a la media armónica entre precisión y recall, y se utiliza especialmente cuando existe desbalance entre las clases:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **Matriz de confusión**

La matriz de confusión permite visualizar la distribución de aciertos y errores del modelo para cada clase, facilitando la interpretación del desempeño de los algoritmos de clasificación. En el

presente estudio, estas métricas se consideran únicamente como un referente teórico complementario, en la medida en que permiten ilustrar la evaluación de modelos de clasificación y su coherencia con la posible categorización del rendimiento académico en niveles de desempeño.

Cabe destacar que la implementación del modelo se centra en un enfoque de regresión, coherente con el objetivo principal de predecir el puntaje global como una variable continua. Por esta razón, la clasificación del rendimiento en niveles se plantea únicamente como una extensión futura del estudio, orientada a fortalecer la aplicabilidad de los modelos en contextos de toma de decisiones educativas.

- **Validación cruzada y selección de modelos**

Con el fin de garantizar la robustez y reproducibilidad de los resultados, los modelos serán evaluados mediante el método de validación cruzada tipo k-fold.

Este procedimiento consiste en dividir el conjunto de datos en ksubconjuntos o folds. En cada iteración, k−1 subconjuntos se utilizan para entrenar el modelo y el subconjunto restante para su validación, repitiendo el proceso k veces de modo que cada subconjunto actúe como conjunto de prueba una vez.

El promedio de las métricas obtenidas en cada iteración proporciona una estimación más confiable del rendimiento general del modelo y reduce el riesgo de sobreajuste (overfitting).

La comparación basada en estas métricas permitirá identificar el o los modelos más adecuados, priorizando aquellos con buen desempeño predictivo y un alto nivel de interpretabilidad, aspectos fundamentales en aplicaciones de analítica educativa.

- **Consideraciones sobre interpretabilidad**

Además de las métricas de desempeño, es fundamental considerar la interpretabilidad de los modelos, especialmente en contextos educativos, donde la comprensión de los factores que influyen en el rendimiento estudiantil es clave para la toma de decisiones.

En este sentido, se emplean técnicas de explicabilidad basadas en SHAP (Shapley Additive Explanations). Este método, propuesto por Lundberg y Lee (2017), se fundamenta en la teoría de valores de Shapley de la teoría de juegos cooperativos, la cual permite asignar a cada variable una contribución cuantitativa en la predicción realizada por el modelo.

El valor SHAP de una característica i se calcula como el promedio ponderado de su contribución marginal en todas las posibles combinaciones de variables. Formalmente, se define como:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

donde:

- N conjunto total de características del modelo.
- S subconjunto de características que no incluye la variable i.
- $f(S)$ predicción del modelo utilizando únicamente las variables del subconjunto S.
- $f(S \cup \{i\})$ predicción al incluir la variable i.
- ϕ_i contribución de la variable ia la predicción final.

Este enfoque permite descomponer la predicción del modelo como la suma de las contribuciones individuales de cada variable:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$$

donde ϕ_0 corresponde al valor base del modelo (promedio de las predicciones) y ϕ_{i} representa la contribución de cada característica.

El uso de SHAP facilita identificar qué variables influyen más en las predicciones del modelo y en qué dirección lo hacen, lo que resulta particularmente útil para analizar cómo factores académicos, socioeconómicos e institucionales impactan el rendimiento estudiantil. De esta manera, se promueve la transparencia, interpretabilidad y valor pedagógico en la interpretación de los resultados.

Estas consideraciones teóricas constituyen la base para la implementación y evaluación práctica de los modelos en el capítulo correspondiente.

1.3.4. Antecedentes

Uno de los antecedentes más relevantes en el campo de la Minería de Datos Educativos (Educational Data Mining, EDM) es el trabajo de Romero y Ventura, quienes presentan una revisión exhaustiva de técnicas como clasificación, agrupamiento y descubrimiento de patrones aplicadas a contextos educativos. Su estudio evidencia que estas metodologías permiten analizar grandes volúmenes de datos académicos para mejorar la toma de decisiones, personalizar procesos de enseñanza y optimizar el rendimiento estudiantil [2].

De manera complementaria, Baker y Siemens amplían esta perspectiva al integrar la EDM con el enfoque de Learning Analytics, destacando el papel de los modelos predictivos en la comprensión

del comportamiento de los estudiantes y en la generación de sistemas educativos más adaptativos [7]. Estos trabajos establecen las bases conceptuales para el uso de aprendizaje automático en educación, posicionando la predicción del rendimiento académico como una de las aplicaciones más relevantes del campo.

En esta línea, investigaciones recientes han demostrado que los modelos de aprendizaje automático presentan desempeños diferenciados dependiendo del tipo de datos utilizados. En particular, estudios comparativos a gran escala han evidenciado que los modelos de ensamble basados en árboles, como Random Forest y técnicas de boosting, superan consistentemente a otros enfoques en datos tabulares estructurados [8]. Este comportamiento ha sido explicado por su capacidad para modelar relaciones no lineales complejas sin requerir transformaciones extensivas de los datos, lo cual resulta especialmente relevante en contextos educativos donde las variables son heterogéneas y altamente correlacionadas.

Asimismo, trabajos como el de Chen y Guestrin han demostrado que algoritmos como XGBoost optimizan iterativamente los errores residuales mediante técnicas de gradient boosting, logrando altos niveles de precisión en diversos contextos [6]. Por su parte, el modelo Random Forest, propuesto por Breiman, ha sido ampliamente reconocido por su capacidad para reducir la varianza mediante bagging, proporcionando modelos robustos frente al ruido y la heterogeneidad de los datos [15].

En contraste, otros enfoques como las máquinas de soporte vectorial (SVM) y las redes neuronales han mostrado limitaciones en este tipo de escenarios. Vapnik señala que las SVM presentan dificultades de escalabilidad y sensibilidad a la selección de funciones kernel en espacios de alta dimensionalidad [16], mientras que estudios recientes indican que los modelos de deep learning no logran superar a los métodos basados en árboles en datos tabulares sin una ingeniería avanzada de características [17].

En términos de desempeño empírico, la literatura reporta consistentemente que los modelos de ensamble logran menores errores de predicción (RMSE y MAE) y mayores valores de R^2 en comparación con modelos tradicionales [8], [17]. Estos hallazgos se alinean con estudios aplicados en educación, donde se ha observado que algoritmos como XGBoost y Random Forest alcanzan mejores resultados al capturar interacciones complejas entre variables académicas, socioeconómicas e institucionales.

No obstante, el análisis del coeficiente de determinación (R^2) en contextos educativos requiere una interpretación cuidadosa. En términos empíricos, la literatura ha evidenciado que el rendimiento académico es un fenómeno multidimensional influenciado por factores no observables como la motivación, el entorno familiar y la calidad docente [11]. En consecuencia, los valores de R^2 en modelos predictivos educativos suelen ubicarse en rangos moderados, típicamente entre 0.20 y 0.40, incluso al utilizar algoritmos avanzados [9], [18], lo cual depende en gran medida de la calidad de los datos y de las variables disponibles en cada estudio.

Desde una perspectiva metodológica, Shmueli establece una distinción fundamental entre modelos explicativos y predictivos, señalando que en problemas aplicados el objetivo principal no es maximizar la explicación causal, sino minimizar el error de predicción [19]. Este enfoque ha sido adoptado ampliamente en EDM, donde la utilidad de los modelos radica en su capacidad para anticipar resultados y apoyar la toma de decisiones, más que en establecer relaciones causales directas.

Adicionalmente, la literatura ha señalado que la capacidad explicativa de los modelos está limitada por la disponibilidad de variables relevantes. La omisión de factores como el historial académico, variables longitudinales o indicadores conductuales introduce sesgos que reducen la varianza explicada [20]. En este sentido, revisiones sistemáticas destacan que la integración de datos de comportamiento y aprendizaje puede mejorar significativamente el desempeño de los modelos predictivos [12].

Otro aspecto relevante en la evaluación de modelos es su estabilidad y capacidad de generalización. Estudios basados en validación cruzada han demostrado que los modelos de ensamble presentan baja variabilidad entre particiones, lo cual indica un adecuado equilibrio entre sesgo y varianza [10]. Esta característica es fundamental en aplicaciones reales, ya que garantiza que el modelo mantendrá su desempeño en datos no observados.

Asimismo, la eficiencia computacional se ha consolidado como un criterio clave en la selección de modelos, especialmente en escenarios de grandes volúmenes de datos. Domingos señala que la elección de algoritmos debe balancear precisión, interpretabilidad y costo computacional [21]. En este contexto, métodos como XGBoost destacan por su capacidad de paralelización y optimización interna, lo que los hace altamente escalables [6].

En el ámbito educativo colombiano, diversos estudios han evidenciado la existencia de desigualdades estructurales en el rendimiento académico, asociadas a variables como el nivel socioeconómico, la ubicación geográfica y el tipo de institución [1], [9], [14]. Estos factores condicionan de manera significativa los resultados de pruebas estandarizadas como Saber 11, lo que resalta la necesidad de enfoques analíticos que incorporen el contexto social en la interpretación de los datos.

Finalmente, investigaciones recientes han enfatizado la importancia de la interpretabilidad en modelos de aprendizaje automático aplicados a educación. En este sentido, técnicas de inteligencia artificial explicable, como SHAP, permiten analizar la contribución de cada variable en las predicciones, facilitando la comprensión de los modelos y su uso en entornos educativos [13].

En conjunto, los antecedentes revisados permiten establecer que el uso de modelos de aprendizaje automático en educación constituye un enfoque sólido y respaldado por la literatura. Asimismo, evidencian que los modelos de ensamble representan el estado del arte en datos

tabulares, que el rendimiento académico presenta límites estructurales en su capacidad explicativa y que la integración de variables contextuales es fundamental para una interpretación adecuada.

En este contexto, el presente proyecto se fundamenta en estos avances y contribuye mediante la aplicación de técnicas de aprendizaje automático al análisis del rendimiento académico en Colombia, incorporando una perspectiva comparativa entre contextos rurales y urbanos, así como una evaluación integral del desempeño, estabilidad y eficiencia de los modelos.

2. ANÁLISIS Y DEPURACIÓN DE LAS VARIABLES ASOCIADAS AL RENDIMIENTO ACADÉMICO

En este capítulo se analizan los factores que influyen en el rendimiento académico de los estudiantes a partir de la base de datos de los resultados de las pruebas Saber 11. Para ello, se realizó un proceso de exploración estadística que permitió examinar tanto las variables numéricas como las categóricas, evaluando su relación con la variable objetivo PUNTAJE GLOBAL.

El análisis incluyó la verificación de la distribución de los puntajes individuales por área, pruebas de normalidad, correlaciones entre variables numéricas y pruebas no paramétricas (Kruskal-Wallis) para determinar la relación entre las variables categóricas y el puntaje global. A partir de estos análisis se identificaron las variables con mayor grado de asociación al rendimiento académico, las cuales serán utilizadas posteriormente en la fase de modelado predictivo.

En este sentido, el presente capítulo da cumplimiento al Objetivo específico 1, al permitir el análisis de variables académicas, socioeconómicas, institucionales y contextuales, así como la identificación de su nivel de relación con el rendimiento académico de los estudiantes.

2.1. Preparación y limpieza del dataset

Antes de realizar el análisis exploratorio y el desarrollo de los modelos predictivos, fue necesario llevar a cabo un proceso estructurado de preparación y limpieza del dataset correspondiente a los resultados del examen SABER 11, obtenidos a través de la plataforma oficial DataICFES del Instituto Colombiano para la Evaluación de la Educación (ICFES), la cual requiere un proceso previo de registro y solicitud de acceso a las bases de datos: DataICFES – Acceso a datos.

Dado el gran volumen de información contenido en el archivo original, este proceso se desarrolló de manera progresiva en varias etapas, con el fin de mejorar la calidad, consistencia y coherencia de los datos, así como optimizar su manejo computacional.

La *Tabla 1* presenta la evolución del dataset a lo largo de las diferentes fases de limpieza, indicando el número de registros (filas) y variables (columnas) disponibles en cada versión generada durante el proceso. La descripción detallada de las variables utilizadas en el análisis se encuentra consignada en el Anexo A: Diccionario de variables del dataset.

Tabla 1. Evolución del dataset durante el proceso de limpieza

Versión	Filas	Columnas
Documento crudo	7,109,704	51
Limpieza 1	4,500,067	51
Limpieza 2	3,696,144	41
Limpieza Final	3,692,591	40

Como se observa en la *Tabla 1,* el dataset original estaba compuesto por 7.109.704 registros y 51 variables, los cuales incluían información académica, institucional y socioeconómica de los estudiantes evaluados en el examen SABER 11.

Dado el tamaño del conjunto de datos, el proceso de limpieza no se realizó en una única etapa, sino mediante un procedimiento progresivo que permitió generar versiones intermedias del dataset. Esta estrategia facilitó el manejo de la información y evitó problemas asociados a limitaciones de memoria y capacidad de procesamiento en los equipos utilizados para el análisis. En consecuencia, las tareas de depuración se ejecutaron de forma modular, almacenando resultados parciales en cada fase del proceso.

En la primera etapa de limpieza (Limpieza 1) se realizó una depuración inicial orientada a mejorar la calidad básica de los datos. En esta fase se eliminaron registros incompletos y observaciones que presentaban valores faltantes en variables consideradas esenciales para el análisis. Asimismo, se identificaron y corrigieron inconsistencias evidentes en algunos campos. Como resultado de este proceso, el número de registros se redujo a 4.500.067 observaciones, manteniéndose aún las 51 variables originales.

La eliminación de registros se realizó únicamente en aquellos casos donde la proporción de datos faltantes fue inferior al 1%, criterio adoptado con el fin de evitar la pérdida significativa de información y minimizar posibles sesgos derivados de la eliminación masiva de observaciones. Para proporciones entre 1% y 10% se aplicaron estrategias de imputación mediante la moda en variables categóricas y la media en variables numéricas. Variables con niveles superiores fueron evaluadas individualmente según su relevancia analítica.

Posteriormente, en la segunda etapa de limpieza (Limpieza 2) se llevó a cabo un proceso de selección y reducción de variables, con el objetivo de conservar únicamente aquellas que aportaban valor analítico para el estudio. En esta fase se eliminaron columnas correspondientes a identificadores administrativos, variables redundantes y atributos con baja relevancia para el modelamiento predictivo, como:

- IDENTIFICADOR CONSECUTIVO DEL ESTUDIANTE
- CÓDIGO DANE DEL ESTABLECIMIENTO EDUCATIVO
- CÓDIGO DANE DE LA SEDE EDUCATIVA
- CÓDIGO ICFES DE LA INSTITUCIÓN EDUCATIVA
- NOMBRE DEL ESTABLECIMIENTO EDUCATIVO
- NOMBRE DE LA SEDE EDUCATIVA
- IDENTIFICADOR DEL ESTUDIANTE
- ESTADO DE INVESTIGACIÓN DEL ESTUDIANTE
- NACIONALIDAD DEL ESTUDIANTE
- PAÍS DE RESIDENCIA DEL ESTUDIANTE
- SEDE PRINCIPAL DE LA INSTITUCIÓN EDUCATIVA

Como resultado, la dimensionalidad del dataset se redujo de 51 a 41 variables, mientras que el número de registros quedó en 3.696.144 observaciones. Esta selección se realizó considerando su aporte analítico frente a la variable objetivo, así como la necesidad de reducir la dimensionalidad del dataset y evitar redundancias en el modelamiento.

Durante esta etapa también se aplicaron procedimientos de imputación para el tratamiento de valores faltantes en algunas variables categóricas. En particular, para la variable GÉNERO DEL ESTUDIANTE, los valores faltantes representaron aproximadamente una proporción inferior al 1% del total de registros, razón por la cual se decidió aplicar imputación mediante la moda, correspondiente a la categoría F.
Posteriormente, se verificó la distribución resultante obteniendo:

- F: 2.040.825 registros
- M: 1.651.766 registros

La imputación no modificó sustancialmente la distribución original de la variable y permitió preservar la consistencia del conjunto de datos.
Finalmente, en la etapa de limpieza final se realizó una revisión adicional con el propósito de garantizar la coherencia y consistencia del conjunto de datos. En esta fase se verificó nuevamente la presencia de valores faltantes, se eliminaron posibles registros duplicados o inconsistentes y se descartó una variable adicional que no aportaba valor significativo al análisis esta variable fue FECHA DE NACIMIENTO, debido a que su información fue transformada y sintetizada en una nueva variable denominada EDAD, evitando redundancia y facilitando el análisis posterior. Como resultado de este proceso se obtuvo el dataset definitivo, el cual fue utilizado en las etapas posteriores de análisis exploratorio y construcción de modelos predictivos.

2.1.1. Dimensiones del Dataset

Como resultado del proceso de limpieza y preparación descrito en la sección anterior, se obtuvo

un dataset final compuesto por 3.692.591 registros (filas) y 40 variables (columnas), el cual fue utilizado para las etapas posteriores de análisis exploratorio y modelamiento predictivo.
En comparación con el documento original, que contenía 7.109.704 registros y 51 variables, el conjunto de datos final presenta una reducción considerable tanto en el número de observaciones como en la cantidad de atributos. Esta reducción permitió mejorar la calidad de la información disponible y optimizar el manejo computacional del dataset, facilitando su procesamiento en los análisis posteriores.

2.1.2. Carga Inicial, Diagnóstico y Filtrado Básico

En esta primera fase, se carga el dataset original de las pruebas Saber 11 y se realiza una exploración inicial. Esto incluye revisar la estructura de los datos, los tipos de variables y efectuar un análisis de valores faltantes. Además, se aplica un filtrado temporal excluyendo registros irrelevantes y se eliminan aquellos que presentan un puntaje global igual a cero.

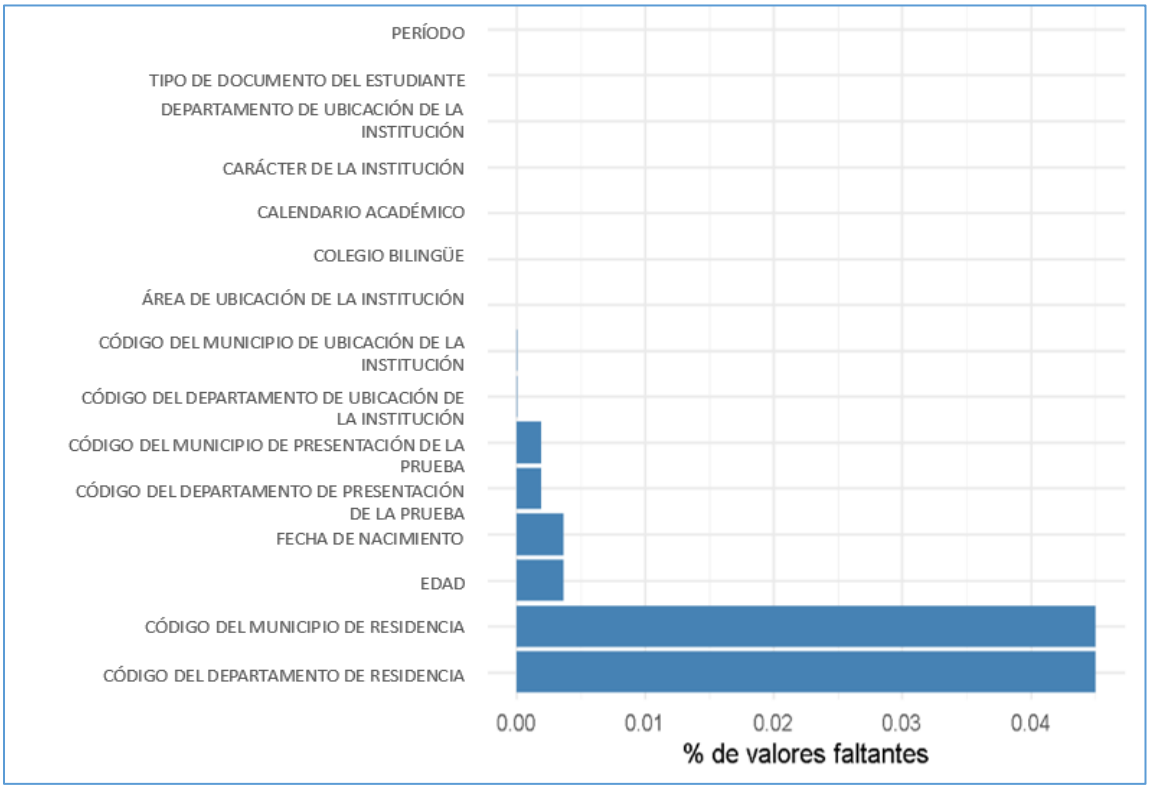

Figura 1. Top 10 de variables con más % de datos faltantes

La *Figura 1* corresponde al diagnóstico inicial realizado sobre el dataset original, previo a las etapas de limpieza descritas en la sección anterior, y presenta las variables con mayor proporción de valores faltantes. Este análisis permitió identificar los campos que requerían tratamiento durante el proceso de depuración y preparación de los datos.

La identificación temprana de estas ausencias facilitó la toma de decisiones relacionadas con la eliminación o imputación de variables, garantizando que la base depurada reflejara datos completos, coherentes y adecuados para los análisis posteriores.

2.1.3. Limpieza Avanzada, Imputación y Normalización

En la segunda fase, se eliminan filas con un alto porcentaje de datos vacíos y se excluyen los años de pandemia (2020-2021). Esta decisión se fundamenta en que las condiciones educativas durante este Período fueron atípicas, lo que podría introducir sesgos estructurales en el análisis y afectar la capacidad de generalización de los modelos.

También se crea la variable de Edad a partir de la fecha de nacimiento y se recodifican todas las columnas con valores vacíos a "SIN DATO". Luego se imputan los valores: las variables numéricas se completan con el promedio y las categóricas con la moda.

La categoría “SIN DATO” se mantuvo en aquellas variables como: ESTRATO VIVIENDA FAMILIAR, NIVEL EDUCATIVO DE LA MADRE, NIVEL EDUCATIVO DEL PADRE, ÁREA DE UBICACIÓN DE LA INSTITUCIÓN, DEPARTAMENTO DE RESIDENCIA DEL ESTUDIANTE donde la imputación podría introducir sesgos, permitiendo conservar la ausencia de información como una categoría informativa dentro del análisis.

Finalmente, se normalizan los nombres de departamentos y municipios para que tengan un formato uniforme, sin caracteres especiales.

Este criterio se definió con el objetivo de preservar la tendencia central de los datos en variables numéricas y mantener la coherencia semántica en variables categóricas, minimizando la distorsión de las distribuciones originales.

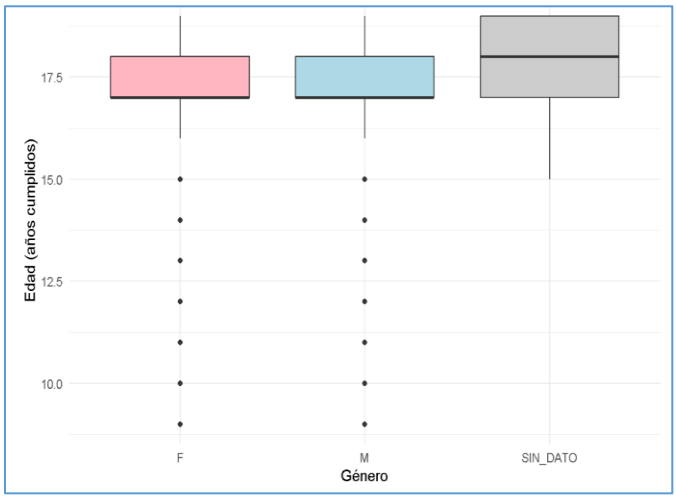

Figura 2. Distribución de Edad de los estudiantes por genero

La *figura 2* evidencia que la distribución de Edades es similar entre hombres y mujeres, con una mediana cercana a los 17 años. Se observa una ligera dispersión mayor en el grupo masculino, lo cual podría deberse a diferencias en la Edad de ingreso o permanencia en la educación media.

Este análisis preliminar resulta útil para verificar la consistencia de la variable “Edad” y garantizar que las imputaciones posteriores no introduzcan sesgos por género. De esta manera, se asegura que los modelos predictivos utilicen una representación demográfica equilibrada.

Tabla 2. Primeros 5 valores de normalización de departamentos frecuencias antes y después

Valor original	**Valor normalizado**	**Frecuencia original**	**Frecuencia normalizada**
ANTIOQUIA	Antioquia	465,973	465,973
BOGOTA	Bogotá D.C.	328,175	614,499
VALLE	Valle del Cauca	317,419	317,419
BOGOTÁ	Bogotá D.C.	286,324	614,499
CUNDINAMARCA	Cundinamarca	251,784	251,784

La *Tabla 2* muestra la estandarización de nombres de departamentos, corrigiendo inconsistencias ortográficas y unificando valores.

Tabla 3. Primeros 5 valores de normalización de municipios frecuencias antes y después

Municipio original	**Valor normalizado**	**Frecuencia original**	**Frecuencia normalizada**
BOGOTÁ D.C.	BOGOTÁ D.C.	610,605	610,605
CALI	CALI	155,687	155,687
BARRANQUILLA	BARRANQUILLA	116,643	116,643
MEDELLIN	MEDELLIN	91,691	183,147 (suma)
SOACHA	SOACHA	52,177	52,177

La *Tabla 3* muestra la estandarización de nombres aplicada a municipios; refleja cómo se consolidaron registros duplicados.

2.1.4. Análisis Exploratorio y Tratamiento de "SIN DATO"

En esta sección se revisan las variables que quedaron con "SIN DATO" y se asegura que tras las imputaciones ya no queden valores faltantes. Se genera un resumen de las variables categóricas

y numéricas para verificar que la limpieza fue exitosa. (Tablas de ejemplos de algunas columnas)

Tabla 4. Resumen de categoría y frecuencia de la columna ESTRATO VIVIENDA FAMILIAR

Categoría	Frecuencia
Estrato 1	1,193,871
Estrato 2	1,281,847
Estrato 3	745,233
Estrato 4	186,599
Estrato 5	65,014
Estrato 6	33,729
SIN DATO	189,851

La *Tabla 4* permite visualizar la distribución de los estratos socioeconómicos reportados por los estudiantes, con conteos absolutos por categoría.

Tabla 5. Resumen de categoría y frecuencia de la columna DISPONIBILIDAD DE AUTOMÓVIL EN EL HOGAR

Categoría	Frecuencia
No	2,692,446
Sí	928,544
SIN DATO	75,154

La *Tabla 5* indica cuántos hogares cuentan con automóvil y cuántos no, incluyendo los valores “SIN DATO”.

Tabla 6. Resumen de categoría y frecuencia de la columna ACCESO A INTERNET EN EL HOGAR

Categoría	Frecuencia
No	1,307,345
Sí	2,279,183
SIN DATO	109,616

La *Tabla 6* presenta la disponibilidad de internet en los hogares, relevante para el análisis socioeconómico.

Como parte del proceso de exploración y calidad de datos, se verificó la presencia de valores faltantes en las variables seleccionadas. El análisis evidenció que las variables PERÍODO, TIPO DE DOCUMENTO DEL ESTUDIANTE, ÁREA DE UBICACIÓN DE LA INSTITUCIÓN, COLEGIO BILINGÜE, CARÁCTER DE LA INSTITUCIÓN EDUCATIVA, CÓDIGO DEL DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN y CÓDIGO DEL MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN no presentan valores faltantes, por lo que no fue necesario aplicar procedimientos de imputación en estas columnas.

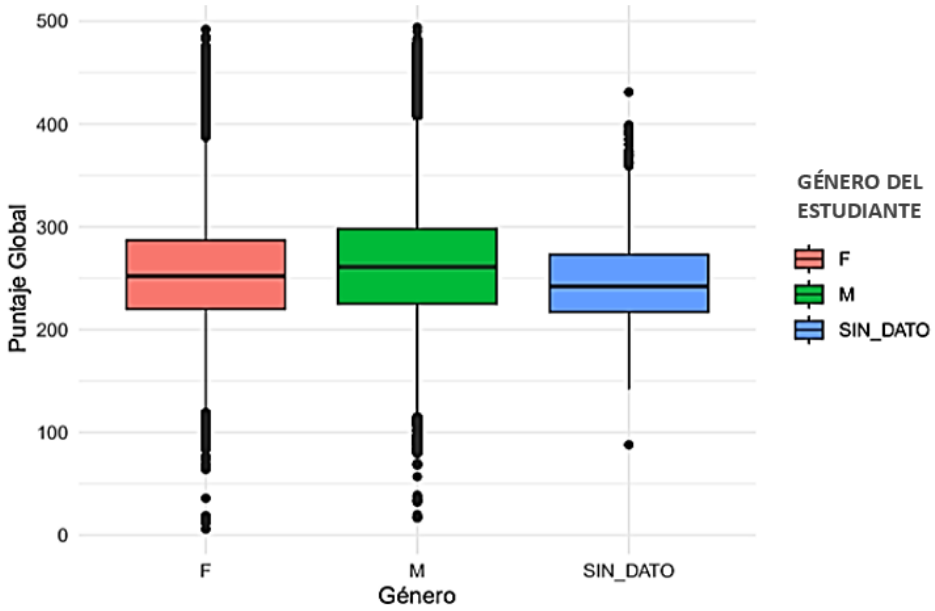

Figura 3. Distribución del puntaje global por genero

La *Figura 3* muestra la distribución del puntaje global según género. En general, se observa una distribución similar entre los grupos, con una mediana ligeramente superior en el grupo masculino.

En términos de dispersión, los tres grupos presentan variabilidad comparable, con presencia de valores atípicos en ambos extremos, especialmente en los grupos femenino y masculino. El grupo SIN DATO muestra una distribución más concentrada, posiblemente debido a su menor tamaño muestral.

En conjunto, los resultados sugieren diferencias leves en el rendimiento académico por género, sin evidenciar brechas significativas en la distribución del puntaje global.

Tabla 7. Valores 'SIN DATO' en variables categóricas

Variable	N_SIN DATO	Porcentaje (%)
COLEGIO BILINGÜE	531,188	14.37
ESTRATO VIVIENDA FAMILIAR	189,851	5.14
ACCESO A INTERNET EN EL HOGAR	109,616	2.97
NIVEL EDUCATIVO DEL PADRE	108,339	2.93
NIVEL EDUCATIVO DE LA MADRE	108,010	2.92
DISPONIBILIDAD DE AUTOMÓVIL EN EL HOGAR	75,154	2.03
NÚMERO DE CUARTOS EN EL HOGAR	74,295	2.01
NÚMERO DE PERSONAS EN EL HOGAR	73,517	1.99
DISPONIBILIDAD DE COMPUTADOR EN EL HOGAR	71,770	1.94
DISPONIBILIDAD DE LAVADORA EN EL HOGAR	71,311	1.93
CARÁCTER DE LA INSTITUCIÓN EDUCATIVA	66,102	1.79
GÉNERO DEL ESTUDIANTE	2,060	0.06
DEPARTAMENTO DE RESIDENCIA DEL ESTUDIANTE	955	0.03
MUNICIPIO DE RESIDENCIA DEL ESTUDIANTE	955	0.03
PERÍODO	0	0.00
TIPO DE DOCUMENTO DEL ESTUDIANTE	0	0.00
ÁREA DE UBICACIÓN DE LA INSTITUCIÓN	5	0.00
CALENDARIO ACADÉMICO	2	0.00
CÓDIGO DEL DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	0	0.00
CÓDIGO DEL MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	0	0.00
DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	2	0.00
GÉNERO DE LA INSTITUCIÓN EDUCATIVA	2	0.00
JORNADA ESCOLAR	2	0.00
MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	2	0.00

NATURALEZA DE LA INSTITUCIÓN	2	0.00
CÓDIGO DEL DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	0	0.00
CÓDIGO DEL MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	0	0.00
CÓDIGO DEL DEPARTAMENTO DE RESIDENCIA	0	0.00
CÓDIGO DEL MUNICIPIO DE RESIDENCIA	0	0.00
DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	75	0.00
MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	75	0.00
ESTUDIANTE PRIVADO DE LA LIBERTAD	0	0.00

La *Tabla 7* cuantifica los registros con "SIN DATO" en variables como bilingüismo, estrato, educación de los padres, etc.

Tabla 8. Resumen de tratamiento en valores 'SIN DATO' en variables.

Variable	Acción	Valor imputado / Filas eliminadas
GÉNERO DEL ESTUDIANTE	Imputado por moda	F
DEPARTAMENTO DE RESIDENCIA DEL ESTUDIANTE	Filas eliminadas	955 filas
MUNICIPIO DE RESIDENCIA DEL ESTUDIANTE	Filas eliminadas	0 filas
PERÍODO	Filas eliminadas	0 filas
TIPO DE DOCUMENTO DEL ESTUDIANTE	Filas eliminadas	0 filas
ÁREA DE UBICACIÓN DE LA INSTITUCIÓN	Filas eliminadas	5 filas
CALENDARIO ACADÉMICO	Filas eliminadas	0 filas
CÓDIGO DEL DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	Filas eliminadas	0 filas
CÓDIGO DEL MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	Filas eliminadas	0 filas
DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	Filas eliminadas	0 filas
GÉNERO DE LA INSTITUCIÓN EDUCATIVA	Filas eliminadas	0 filas

JORNADA ESCOLAR	Filas eliminadas	0 filas
MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	Filas eliminadas	0 filas
NATURALEZA DE LA INSTITUCIÓN	Filas eliminadas	0 filas
CÓDIGO DEL DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	Filas eliminadas	0 filas
CÓDIGO DEL MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	Filas eliminadas	0 filas
CÓDIGO DEL DEPARTAMENTO DE RESIDENCIA	Filas eliminadas	0 filas
CÓDIGO DEL MUNICIPIO DE RESIDENCIA	Filas eliminadas	0 filas
DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	Filas eliminadas	75 filas
MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	Filas eliminadas	0 filas
ESTUDIANTE PRIVADO DE LA LIBERTAD	Filas eliminadas	0 filas
ESTRATO VIVIENDA FAMILIAR	Imputado por moda	Estrato 2
ACCESO A INTERNET EN EL HOGAR	Imputado por moda	Sí
NIVEL EDUCATIVO DEL PADRE	Imputado por moda	Secundaria (Bachillerato) completa
NIVEL EDUCATIVO DE LA MADRE	Imputado por moda	Secundaria (Bachillerato) completa
DISPONIBILIDAD DE AUTOMÓVIL EN EL HOGAR	Imputado por moda	No
NÚMERO DE CUARTOS EN EL HOGAR	Imputado por moda	Tres
NÚMERO DE PERSONAS EN EL HOGAR	Imputado por moda	3 a 4
DISPONIBILIDAD DE COMPUTADOR EN EL HOGAR	Imputado por moda	Sí
DISPONIBILIDAD DE LAVADORA EN EL HOGAR	Imputado por moda	Sí
CARÁCTER DE LA INSTITUCIÓN EDUCATIVA	Imputado por moda	ACADÉMICO
EDAD	Imputado por promedio	17

La *Tabla 8* detalla las acciones aplicadas para imputar o eliminar valores faltantes, especificando el método usado.

2.1.5. Exploración Final del Dataset

Finalmente, se realiza una exploración del dataset resultante, verificando la presencia de valores "SIN DATO" y NA tras el proceso de limpieza. Asimismo, se analizan distribuciones del puntaje global y la Edad de los estudiantes mediante visualizaciones descriptivas (véanse Figuras 4, 5 y 6).

Los resultados muestran que, después del proceso de limpieza, la mayoría de las variables no presentan valores “SIN DATO”, lo que evidencia un adecuado nivel de consistencia en el dataset. No obstante, la variable COLEGIO BILINGÜE conserva un 14.35% de registros sin información. Esta situación se debe a que no fue posible imputar dichos valores sin introducir sesgos, ni eliminar los registros asociados sin afectar significativamente el tamaño de la muestra.

Posterior al proceso de limpieza, se verificó que las transformaciones aplicadas no alteraran significativamente la distribución de las variables, garantizando la consistencia del dataset para su uso en etapas posteriores.

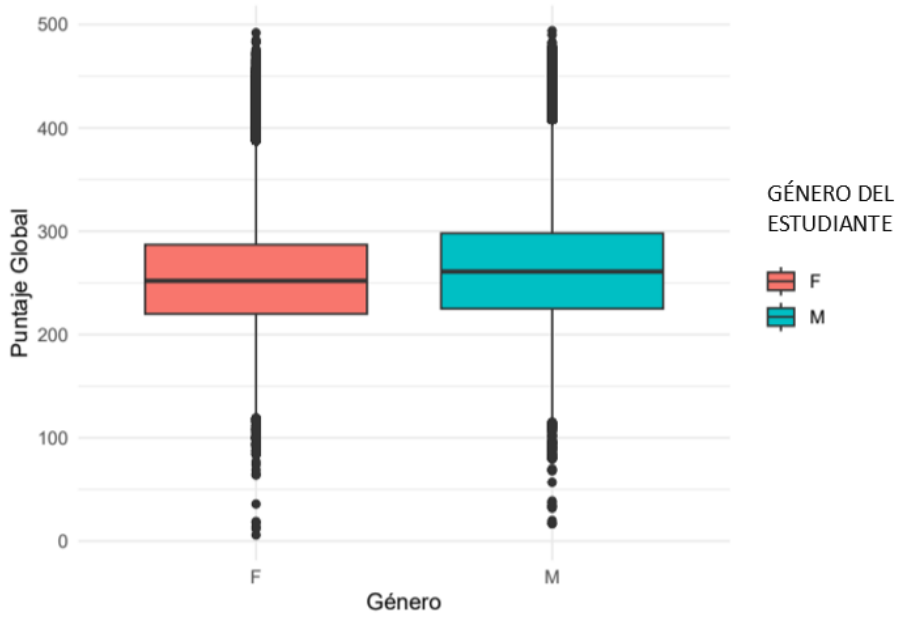

Figura 4. Distribución del puntaje global por genero después del tratamiento a valores “SIN DATO”

La *Figura 4* muestra la distribución del puntaje global por género después del tratamiento de los valores “SIN DATO”.

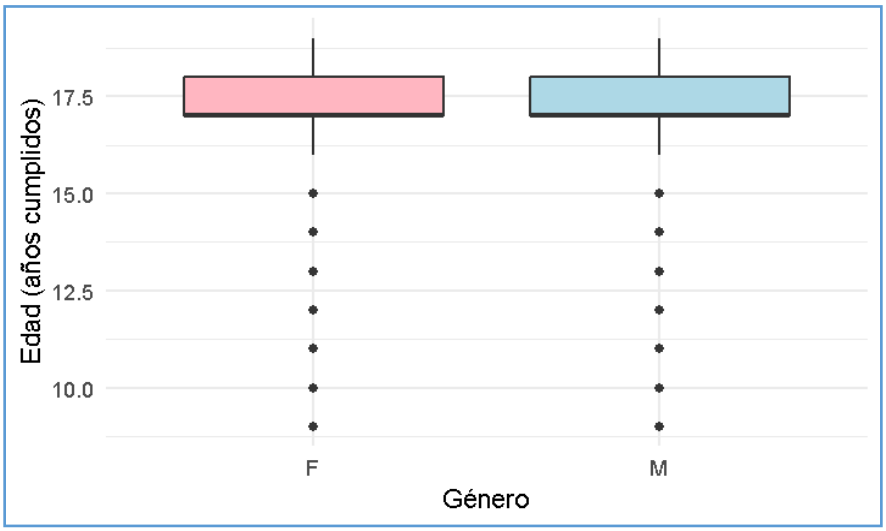

Figura 5. Distribución de Edad de los estudiantes por genero después del tratamiento "SIN DATO"

La *Figura 5* presenta un diagrama de cajas que muestra la distribución de la Edad de los estudiantes por género después del proceso de limpieza de datos. Se observa que la mayor parte de las Edades se concentra en un rango estrecho alrededor de los valores centrales, con una dispersión similar entre ambos grupos. También se identifican algunos valores atípicos que se alejan del rango principal de la distribución.

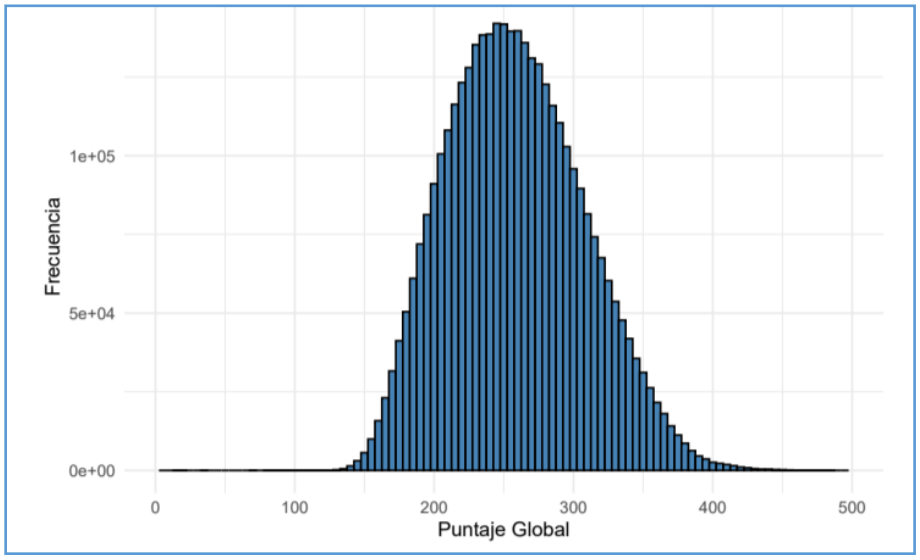

Figura 6. Distribución del puntaje global

La *Figura 6* presenta el histograma del puntaje global de los estudiantes. Los resultados muestran una media de 248 puntos y una mediana de 245 puntos, lo que indica una distribución aproximadamente centrada en valores intermedios. Los puntajes observados oscilan entre 110 y 410 puntos, con mayor concentración de estudiantes alrededor de los valores medios.

2.2. Análisis estadístico y exploratorio

2.2.1. Análisis de distribución y normalidad de los puntajes

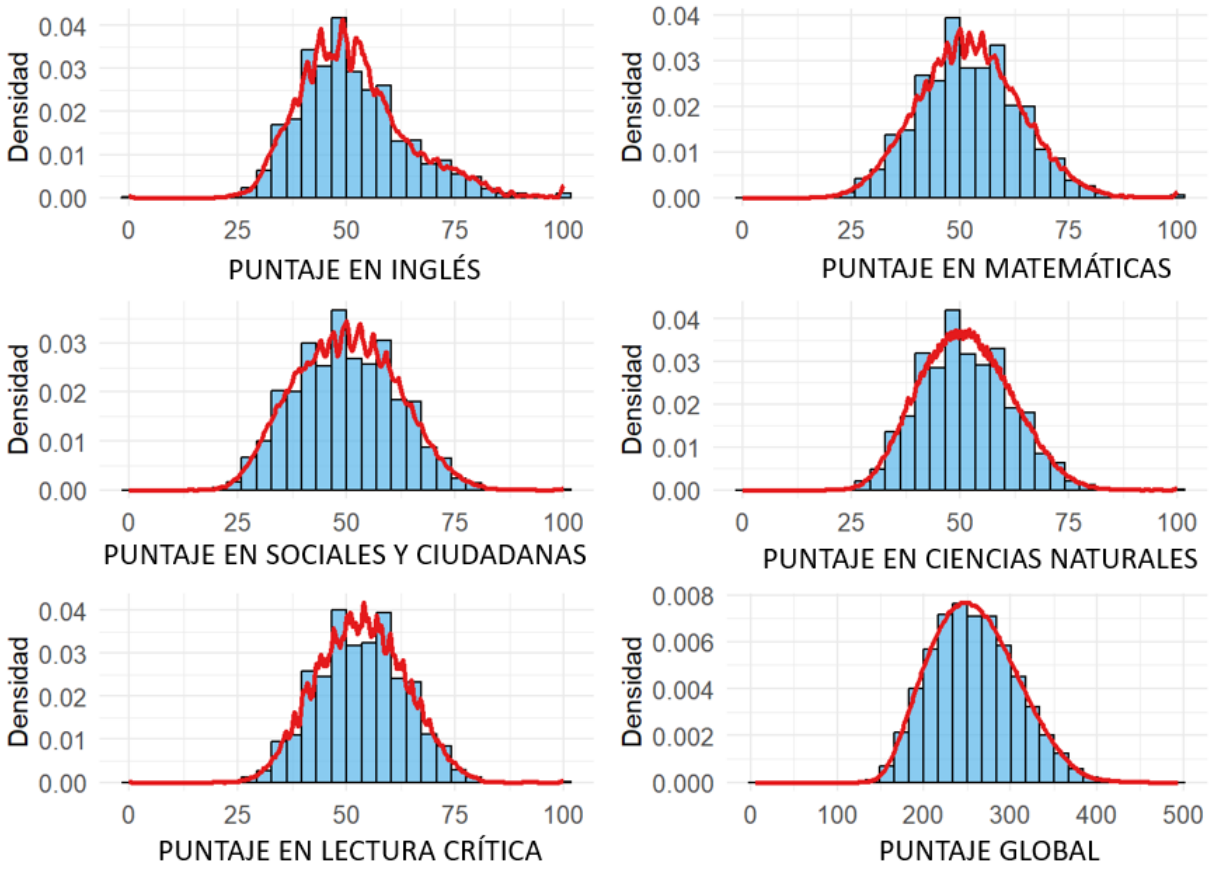

Figura 7.Histogramas de variables numéricas vs PUNTAJE GLOBAL

La *Figura 7* presenta la distribución de los puntajes obtenidos en las diferentes áreas evaluadas y del Puntaje Global. En general, las variables muestran una forma aproximadamente simétrica, con ligeras asimetrías y concentraciones alrededor de los valores medios.

Las áreas de Matemáticas, Ciencias Naturales y Lectura Crítica presentan distribuciones relativamente concentradas, mientras que Inglés exhibe una mayor dispersión. Por su parte, el Puntaje Global mantiene una distribución consistente con la agregación de los resultados obtenidos en las diferentes competencias evaluadas.

Estos resultados permiten una primera aproximación al comportamiento de los datos y evidencian la existencia de variabilidad suficiente para el desarrollo de análisis estadísticos y modelos predictivos.

Tabla 9.Resultados de la prueba de normalidad (Shapiro-Wilk)

Variable	p_valor	Normalidad
PUNTAJE EN INGLÉS	0	No (no es normal)
PUNTAJE EN MATEMÁTICAS	0	No (no es normal)
PUNTAJE EN SOCIALES Y CIUDADANAS	0	No (no es normal)
PUNTAJE EN CIENCIAS NATURALES	0	No (no es normal)
PUNTAJE EN LECTURA CRÍTICA	0	No (no es normal)
PUNTAJE GLOBAL	0	No (no es normal)

Los resultados de la prueba de Shapiro–Wilk indican que ninguna de las variables analizadas presenta una distribución normal ($p < 0.05$ en todos los casos).

Este resultado es consistente con el gran tamaño muestral del estudio, ya que en muestras de gran magnitud pequeñas desviaciones respecto a la normalidad suelen resultar estadísticamente significativas. No obstante, dado que no se cumple el supuesto de normalidad, para los análisis bivariados posteriores se emplearon métodos no paramétricos, específicamente la correlación de Spearman y la prueba de Kruskal–Wallis.

Tabla 10.Resumen estadístico de los puntajes

Variable	Min	Q1	Mediana	Media	Q3	Max
PUNTAJE EN INGLÉS	0	43	50	51.47	58	100
PUNTAJE EN MATEMÁTICAS	0	44	52	52.15	60	100
PUNTAJE EN SOCIALES Y CIUDADANAS	0	41	50	49.97	58	100
PUNTAJE EN CIENCIAS NATURALES	0	44	51	51.30	58	100
PUNTAJE EN LECTURA CRÍTICA	0	46	53	53.34	60	100
PUNTAJE GLOBAL	6	222	255	258.36	292	494

La *Tabla 10* presenta las principales medidas descriptivas de los puntajes obtenidos por los estudiantes. En general, las medias de las áreas evaluadas oscilan entre 50 y 53 puntos, mientras que el Puntaje Global presenta una media de 258.36 puntos.

El amplio rango de valores observado evidencia una alta heterogeneidad en el desempeño académico de la población estudiada, aspecto favorable para la construcción de modelos predictivos y para la identificación de factores asociados al rendimiento académico.

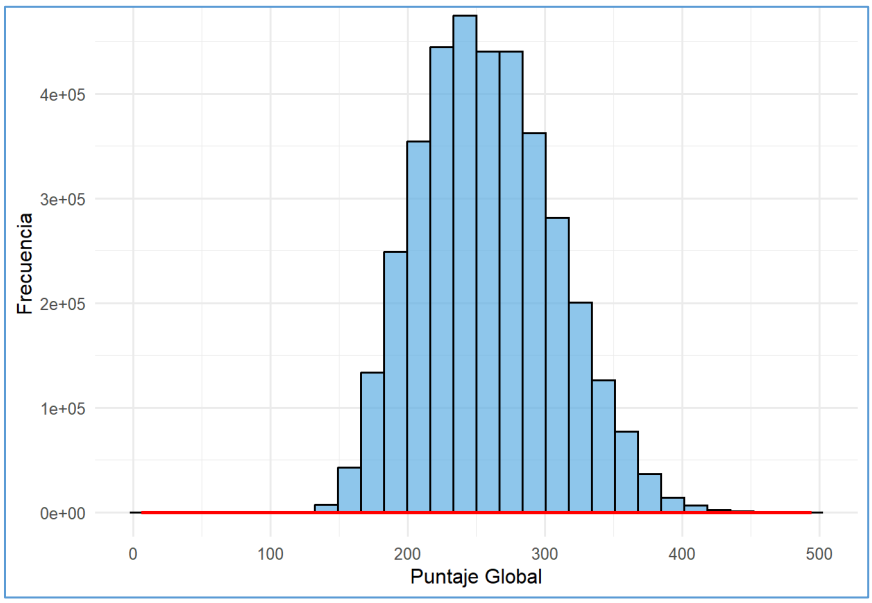

Figura 8.histograma de distribución general del puntaje global.

La *Figura 8* muestra la distribución del Puntaje Global. Se observa una concentración importante de estudiantes entre los 200 y 300 puntos, correspondiente a niveles medios de desempeño académico.

Asimismo, se evidencia una ligera asimetría y una dispersión suficiente para diferenciar distintos niveles de rendimiento entre los estudiantes analizados. Esta variabilidad favorece la capacidad explicativa y predictiva de los modelos desarrollados en etapas posteriores.

En síntesis, los resultados descriptivos muestran que los puntajes presentan distribuciones no normales, aunque suficientemente estables y representativas para la aplicación de técnicas estadísticas robustas y algoritmos de aprendizaje automático.

2.2.2. Correlación entre variables numéricas

Tabla 11. Correlación entre las variables numéricas

Variable	Correlación (r)	Tamaño del efecto
Puntaje Global	10.000	Grande
Puntaje en Sociales y Ciudadanas	0.9098	Grande
Puntaje en Ciencias Naturales	0.9080	Grande
Puntaje en Matemáticas	0.8899	Grande
Puntaje en Lectura Crítica	0.8868	Grande
Puntaje en Inglés	0.7641	Grande
Edad	-0.1862	Pequeño

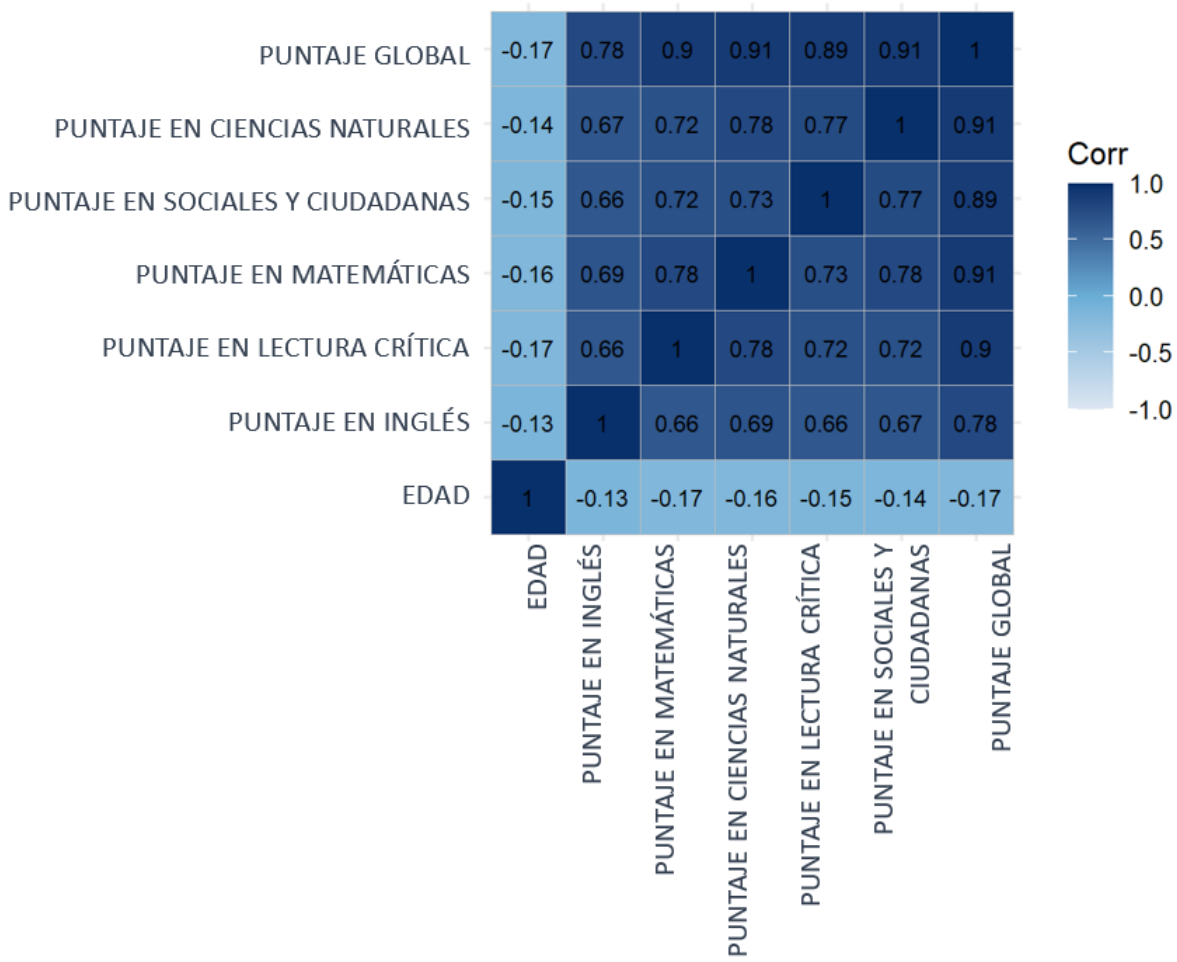

Figura 9.Matriz de correlación de variables numéricas vs PUNTAJE GLOBAL

Considerando que las variables numéricas no presentaron distribución normal según la prueba de Shapiro–Wilk, se empleó el coeficiente de correlación de Spearman (ρ) para evaluar la intensidad y dirección de las asociaciones entre variables.

La *Tabla 11* presenta los coeficientes de correlación entre el Puntaje Global y las variables numéricas del conjunto de datos. Asimismo, la *Figura 9* ilustra dichas relaciones mediante una matriz de correlación.

Se observa que el Puntaje Global presenta correlaciones positivas muy altas con los puntajes obtenidos en Sociales y Ciudadanas (ρ = 0.9098), Ciencias Naturales (ρ = 0.9080), Matemáticas (ρ = 0.8899) y Lectura Crítica (ρ = 0.8868). De igual manera, Inglés presenta una correlación positiva alta (ρ = 0.7641).

Estos resultados son consistentes con la construcción del Puntaje Global, el cual integra el desempeño alcanzado por los estudiantes en las diferentes áreas evaluadas por el examen Saber 11.

Por otra parte, la variable Edad presenta una correlación negativa débil ($\rho = -0.1862$), lo que sugiere que los estudiantes con mayor edad tienden a obtener puntajes globales ligeramente inferiores.

De acuerdo con los criterios propuestos por Cohen (1988), las correlaciones observadas entre el Puntaje Global y las áreas evaluadas corresponden a tamaños del efecto grandes ($|\rho| \geq 0.50$), evidenciando asociaciones de alta magnitud. En contraste, la variable Edad presenta un tamaño del efecto pequeño ($|\rho| < 0.30$).

En conjunto, los resultados confirman la coherencia interna del instrumento de evaluación, ya que el Puntaje Global refleja adecuadamente el desempeño académico general de los estudiantes. No obstante, las altas correlaciones entre las variables académicas sugieren la posible presencia de multicolinealidad, aspecto que debe considerarse durante las etapas posteriores de modelamiento predictivo.

Los resultados obtenidos en este capítulo permitieron caracterizar de manera integral el comportamiento de las variables asociadas al rendimiento académico, proporcionando una base sólida para las fases posteriores de modelado y análisis predictivo.

De esta manera, se da cumplimiento al Objetivo Específico 1, relacionado con la exploración, depuración y análisis estadístico de los datos utilizados en la investigación.

3. ANÁLISIS COMPARATIVO DE FACTORES SEGÚN CONTEXTOS EDUCATIVOS

En este capítulo se realiza el análisis comparativo de los factores asociados al rendimiento académico de los estudiantes, con énfasis en el análisis de variables categóricas relacionadas con el contexto educativo, familiar e institucional y su asociación con el rendimiento académico.

A diferencia del capítulo anterior, donde se exploró la relación entre variables numéricas y el Puntaje Global, en esta sección se evalúan las diferencias de desempeño entre grupos definidos por características categóricas, con el fin de identificar si existen variaciones significativas en el rendimiento académico según dichos contextos.

Para ello, se aplican pruebas no paramétricas, específicamente el test de Kruskal–Wallis, debido a la ausencia de normalidad en las variables de puntaje. Posteriormente, se complementa el análisis con representaciones gráficas tipo diagrama de caja, que permiten visualizar la distribución del Puntaje Global entre las diferentes categorías.

El análisis desarrollado en este capítulo se enfoca en identificar diferencias estadísticamente significativas entre grupos, como base para comprender las diferencias observadas en el desempeño académico según distintos contextos educativos.

Es importante señalar que el alcance de este análisis corresponde a una comparación descriptiva del rendimiento académico entre grupos definidos por variables contextuales, institucionales y sociodemográficas. En este sentido, los resultados permiten identificar diferencias y patrones de comportamiento entre categorías, evidenciando posibles brechas asociadas a distintos contextos educativos. No obstante, debido al carácter observacional de los datos y a la metodología empleada, los hallazgos deben interpretarse como asociaciones observadas entre variables y no como relaciones de causalidad.

3.1. Comparación entre variables categóricas

Tabla 12. Prueba de Kruskal-Wallis

Variable	p_value	significativo
ÁREA DE UBICACIÓN DE LA INSTITUCIÓN	0	Sí
COLEGIO BILINGÜE	0	Sí
CALENDARIO ACADÉMICO	0	Sí
CARÁCTER DE LA INSTITUCIÓN EDUCATIVA	0	Sí
DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	0	Sí
GÉNERO DE LA INSTITUCIÓN EDUCATIVA	0	Sí
JORNADA ESCOLAR	0	Sí
MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	0	Sí
NATURALEZA DE LA INSTITUCIÓN	0	Sí

TIPO DE DOCUMENTO DEL ESTUDIANTE	0	Sí
DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	0	Sí
MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	0	Sí
DEPARTAMENTO DE RESIDENCIA DEL ESTUDIANTE	0	Sí
MUNICIPIO DE RESIDENCIA DEL ESTUDIANTE	0	Sí
GÉNERO DEL ESTUDIANTE	0	Sí
NÚMERO DE CUARTOS EN EL HOGAR	0	Sí
NIVEL EDUCATIVO DE LA MADRE	0	Sí
NIVEL EDUCATIVO DEL PADRE	0	Sí
ESTRATO VIVIENDA FAMILIAR	0	Sí
NÚMERO DE PERSONAS EN EL HOGAR	0	Sí
DISPONIBILIDAD DE AUTOMÓVIL EN EL HOGAR	0	Sí
DISPONIBILIDAD DE COMPUTADOR EN EL HOGAR	0	Sí
ACCESO A INTERNET EN EL HOGAR	0	Sí
DISPONIBILIDAD DE LAVADORA EN EL HOGAR	0	Sí
ESTUDIANTE PRIVADO DE LA LIBERTAD	0	Sí

La *Tabla 12* presenta los resultados de la prueba no paramétrica de Kruskal–Wallis aplicada a las variables categóricas del conjunto de datos, con el fin de evaluar diferencias en el Puntaje Global entre grupos.

Los resultados evidencian diferencias estadísticamente significativas en todas las variables analizadas ($p < 0.05$), lo que indica que el rendimiento académico varía según las categorías de factores como el contexto sociodemográfico, institucional y familiar.

En particular, variables asociadas al tipo de institución educativa, la jornada escolar y el nivel socioeconómico presentan diferencias relevantes en la distribución del Puntaje Global, evidenciando patrones diferenciales de desempeño académico entre grupos. Estos resultados permiten identificar posibles brechas asociadas a las condiciones institucionales y socioeconómicas de los estudiantes.

Estos hallazgos justifican la inclusión de las variables categóricas en la etapa de modelado predictivo, donde se evaluará su aporte conjunto con las variables numéricas.

3.2. Análisis descriptivo de variables categóricas en relación con el Puntaje Global

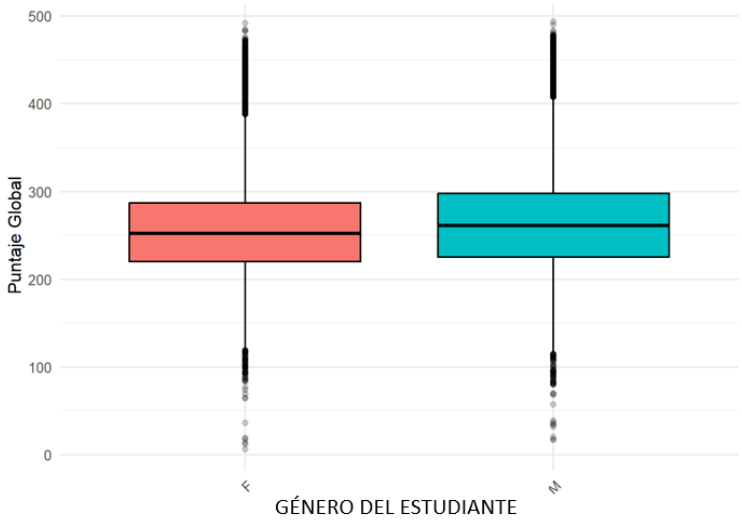

Figura 10. Diagrama de caja de la distribución del puntaje global según el género del estudiante

La *Figura 10* evidencia diferencias leves en el rendimiento promedio entre hombres y mujeres. Se observa una mediana ligeramente superior en los estudiantes de género masculino, aunque ambos grupos presentan una dispersión similar en los puntajes. Esto sugiere que existen diferencias en el rendimiento académico según el género; sin embargo, estas diferencias son relativamente moderadas y no permiten explicar por sí solas la variabilidad observada en los puntajes.

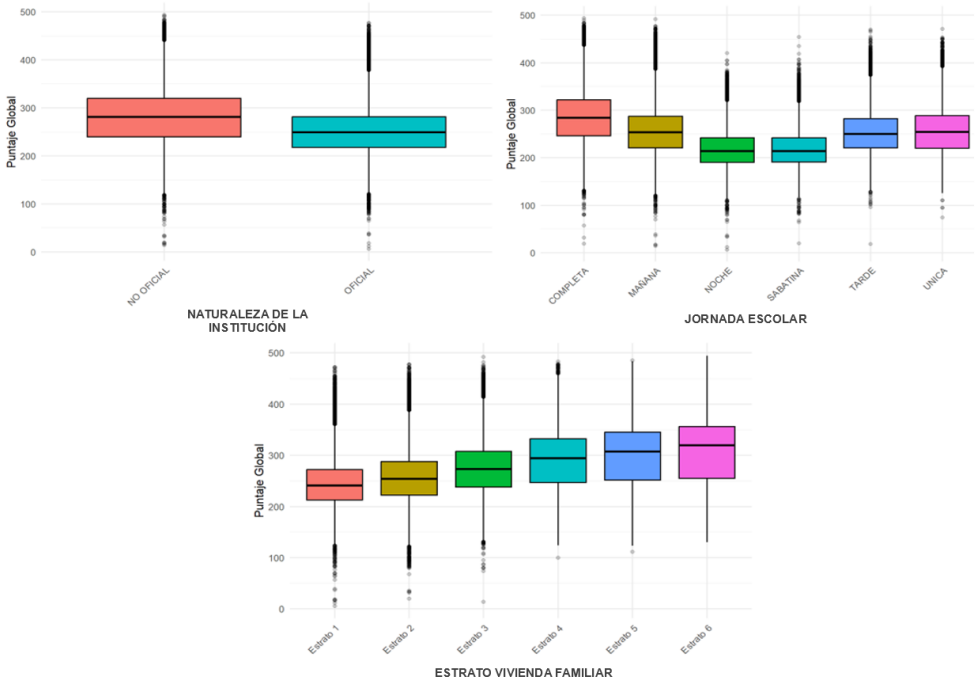

Figura 11. Diagrama de caja de la distribución del puntaje global según colegio, jornada escolar y estrato

socioeconómico

La *Figura 11* presenta la distribución del puntaje global en función de la naturaleza de la institución, la jornada escolar y el estrato socioeconómico. Se observa que los estudiantes de colegios no oficiales tienden a obtener puntajes más altos en comparación con los de colegios oficiales, lo cual refleja las brechas estructurales en recursos y calidad educativa.

En cuanto a la jornada escolar, los estudiantes que asisten a jornada completa muestran un rendimiento superior frente a las jornadas nocturna y sabatina, posiblemente debido a una mayor disponibilidad de tiempo académico y acompañamiento docente.

Finalmente, en la variable estrato socioeconómico se evidencia una relación positiva entre el nivel del estrato y el puntaje global, observándose que, a mayor estrato, mayor es el rendimiento promedio. Este comportamiento evidencia una asociación positiva entre el nivel socioeconómico y el rendimiento académico observado en los estudiantes.

En conjunto, los resultados obtenidos en este capítulo evidencian la existencia de diferencias estadísticamente significativas en el rendimiento académico según variables de tipo institucional, familiar y socioeconómico. Estas diferencias evidencian que el rendimiento académico presenta variaciones relevantes según características del contexto educativo, familiar y socioeconómico de los estudiantes.

Asimismo, los análisis gráficos permiten complementar los resultados de las pruebas estadísticas, mostrando patrones consistentes de variación del rendimiento entre grupos. Esto refuerza la importancia de considerar variables categóricas en la explicación del desempeño académico.

De esta manera, los resultados obtenidos permiten dar respuesta al Objetivo Específico 2, al comparar el rendimiento académico entre grupos definidos por variables contextuales, institucionales y sociodemográficas, evidenciando posibles brechas y patrones diferenciales entre distintos contextos educativos. Asimismo, las variables identificadas serán integradas en la etapa de modelamiento predictivo del estudio para evaluar su aporte en la predicción del rendimiento académico.

4. VALIDACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO

El presente capítulo tiene como objetivo validar y comparar diferentes modelos de aprendizaje automático para la predicción del puntaje global obtenido por los estudiantes en las pruebas Saber 11°. Esta validación se centra en analizar el desempeño predictivo, la estabilidad de los resultados y la viabilidad computacional de cada modelo, utilizando distintos esquemas de partición de los datos.

El proceso de modelado y evaluación se desarrolló utilizando el lenguaje de programación R y el entorno de desarrollo RStudio. Para ello, se emplearon librerías especializadas en análisis de datos, entrenamiento de modelos y evaluación de desempeño, entre las que se destacan dplyr y caret para el preprocesamiento y partición de datos; ranger, xgboost, kernlab y RSNNS para el entrenamiento de los modelos; y Metrics y ggplot2 para el cálculo de métricas y la visualización de resultados.

El entrenamiento y validación de los modelos se realizó en un equipo de cómputo personal con sistema operativo Windows 11 y 16 GB de memoria RAM. Estas condiciones permitieron ejecutar el proceso de modelado; sin embargo, se evidenciaron limitaciones relacionadas con el tiempo de procesamiento y el consumo de memoria, especialmente en algoritmos de mayor complejidad computacional, como SVM y redes neuronales, al incrementarse el volumen de datos.

4.1. Evaluación inicial de modelos con el 10% de los datos

Con el propósito de realizar una primera aproximación al desempeño de los modelos predictivos y evaluar su viabilidad computacional, se llevó a cabo una fase de validación inicial utilizando una muestra del 10% del conjunto de datos previamente depurado en el Objetivo 1.

Este subconjunto fue seleccionado de manera aleatoria, garantizando representatividad de las variables académicas, socioeconómicas e institucionales. Posteriormente, se realizó una partición de los datos en un 80% para entrenamiento y un 20% para prueba, siguiendo un esquema estándar de validación en problemas de aprendizaje supervisado.

4.1.1. Variables utilizadas en el modelado

Para el entrenamiento de los modelos se emplearon variables previamente identificadas como relevantes en el análisis exploratorio, integrando dimensiones académicas, institucionales y socioeconómicas:

Variable objetivo:

PUNTAJE GLOBAL

Variables predictoras:

PERÍODO, ÁREA DE UBICACIÓN DE LA INSTITUCIÓN, COLEGIO BILINGÜE, CALENDARIO ACADÉMICO, CARÁCTER DE LA INSTITUCIÓN EDUCATIVA, DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN, GÉNERO DE LA INSTITUCIÓN EDUCATIVA, JORNADA ESCOLAR, NATURALEZA DE LA INSTITUCIÓN, DEPARTAMENTO DE RESIDENCIA DEL ESTUDIANTE, GÉNERO DEL ESTUDIANTE, ESTUDIANTE PRIVADO DE LA LIBERTAD, NÚMERO DE CUARTOS EN EL HOGAR, NIVEL EDUCATIVO DE LA MADRE, NIVEL EDUCATIVO DEL PADRE, ESTRATO VIVIENDA FAMILIAR, NÚMERO DE PERSONAS EN EL HOGAR, DISPONIBILIDAD DE AUTOMÓVIL EN EL HOGAR, DISPONIBILIDAD DE COMPUTADOR EN EL HOGAR, ACCESO A INTERNET EN EL HOGAR, DISPONIBILIDAD DE LAVADORA EN EL HOGAR

De acuerdo con los requerimientos de cada algoritmo, se aplicaron transformaciones específicas, tales como codificación de variables categóricas (factores o variables dummy) y escalado de datos en modelos sensibles a la magnitud de las variables (SVM y MLP).

4.1.2. Modelos evaluados

Se Se evaluaron cuatro modelos de aprendizaje automático supervisado con el propósito de predecir el puntaje global de los estudiantes en las pruebas SABER 11:

- XGBoost
- Random Forest (Ranger)
- Support Vector Machine (SVM)
- Redes Neuronales Artificiales (MLP)

El desempeño de los modelos fue evaluado mediante las métricas RMSE, MAE y coeficiente de determinación $\boldsymbol{R^2}$, calculadas sobre el conjunto de prueba. Adicionalmente, se registró el tiempo de entrenamiento como indicador de eficiencia computacional.

Para garantizar la reproducibilidad de los experimentos, en todos los modelos se utilizó una semilla aleatoria fija ($\boldsymbol{seed = 42}$) y una partición de datos de entrenamiento y prueba en proporción 80/20. Asimismo, debido al alto volumen de información disponible, la validación inicial se realizó utilizando una muestra correspondiente al 10 % del dataset limpio.

Configuración de hiperparámetros utilizada

- **Random Forest (Ranger)**

  El modelo Random Forest fue implementado mediante la librería ranger, utilizando 300 árboles de decisión ($num.trees = 300$), con selección aleatoria de 10 variables por división ($mtry = 10$) y un tamaño mínimo de nodo de 10 observaciones

($min.node.size = 10$). La importancia de variables se calculó mediante el criterio de impureza ($importance = "impurity"$). Además, se habilitó el uso de todos los núcleos disponibles del procesador para optimizar el tiempo de entrenamiento ($num.threads = detectCores()$).

- **XGBoost**

  El modelo XGBoost fue configurado para tareas de regresión utilizando la función objetivo $objective = "reg: squarederror"$. Se empleó una profundidad máxima de 6 niveles por árbol ($max.depth = 6$), una tasa de aprendizaje de 0.3 ($eta = 0.3$), y 100 iteraciones de entrenamiento ($nrounds = 100$). Adicionalmente, se configuró un muestreo del 80 % tanto para filas como para columnas ($subsample = 0.8$ y $colsample.bytree = 0.8$), con un peso mínimo por nodo hijo de 1 ($min.child.weight = 1$) y un parámetro de regularización gamma igual a 0 ($gamma = 0$).

- **Support Vector Machine (SVM)**

  El modelo SVM fue implementado mediante la función ksvm de la librería kernlab, utilizando una formulación de regresión epsilon-SVR ($type = "eps - svr"$) con kernel lineal ($kernel = "vanilladot"$). Se estableció un parámetro de regularización $C = 0.1$, una tolerancia epsilon de 1 ($epsilon = 1$) y una memoria caché de 40 MB ($cache = 40$). Previamente al entrenamiento, las variables fueron escaladas mediante estandarización, por lo que el parámetro $scaled = FALSE$evitó un reescalado interno adicional.

- **Red Neuronal Artificial (MLP)**

  El modelo de red neuronal multicapa (MLP) fue implementado utilizando la librería caret con el método "mlp". Se configuró una arquitectura con una capa oculta de 5 neuronas ($size = 5$). Antes del entrenamiento, las variables predictoras fueron normalizadas mediante centrado y escalamiento, debido a la sensibilidad de este tipo de modelos a las magnitudes de los datos. El entrenamiento se realizó sin validación cruzada interna ($trainControl(method = "none")$), utilizando la configuración base proporcionada por la librería.

4.1.3. Resultados de la evaluación inicial

Tabla 13. Resultados de desempeño de los modelos en la validación inicial (10% de los datos).

Modelo	RMSE	MAE	R^2	Tiempo (s)
XGBoost	39.83	31.96	0.343	1.42
Random Forest	40.41	32.32	0.324	70.43
SVM	42.47	34.14	0.254	59.43
MLP	74.02	58.40	0.045	157.63

A partir de la *Tabla 13*, se observa que existen diferencias claras en el desempeño de los modelos tanto en términos de error como de eficiencia computacional. En particular, los modelos basados en árboles presentan menores valores de error en comparación con los demás enfoques, lo que sugiere una mejor capacidad para capturar las relaciones presentes en los datos.

Asimismo, se evidencia que el tiempo de entrenamiento varía significativamente entre modelos, lo cual resulta relevante considerando el tamaño del dataset utilizado en este estudio. Esta diferencia en eficiencia permite identificar qué modelos son más adecuados no solo por su precisión, sino también por su viabilidad en escenarios de gran escala.

Es importante precisar que el tiempo reportado corresponde al tiempo total de entrenamiento de cada modelo en una única ejecución experimental, medido desde el inicio hasta la finalización del ajuste sobre el conjunto de entrenamiento. Todos los modelos fueron evaluados mediante un esquema de partición simple train-test.

Estos resultados iniciales se complementan con el análisis gráfico presentado a continuación, el cual permite examinar el comportamiento de cada modelo en términos de ajuste y distribución del error, facilitando una interpretación más detallada de su capacidad predictiva.

4.1.4. Análisis gráfico de los modelos

El análisis gráfico permite complementar la interpretación de los resultados cuantitativos, evidenciando el comportamiento de cada modelo frente a los datos.

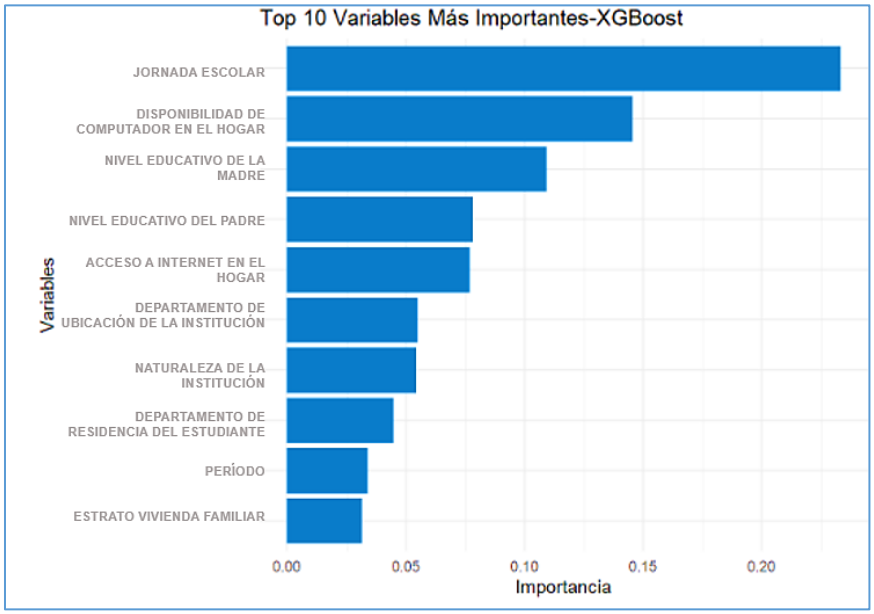

Figura 12. Importancia de variables del modelo XGBoost

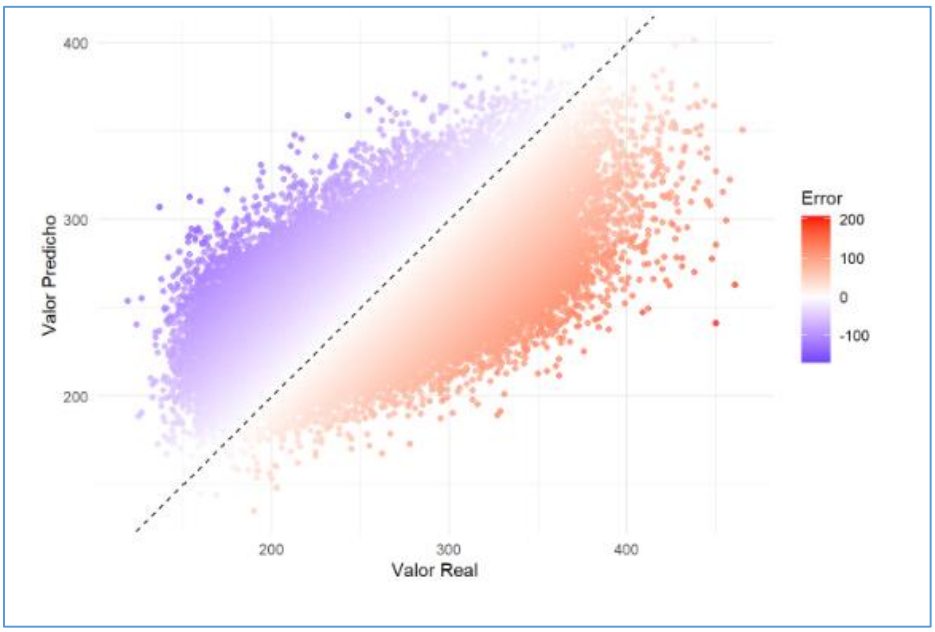

Figura 13. Relación entre valores reales y predichos – XGBoost

En el caso de XGBoost, la *Figura 12* muestra la importancia de variables, donde se identifican como más relevantes aquellas asociadas al contexto socioeconómico y educativo del estudiante. La *Figura 13* evidencia una alineación clara de los valores en torno a la diagonal, indicando una adecuada capacidad del modelo para capturar la tendencia general del puntaje global.

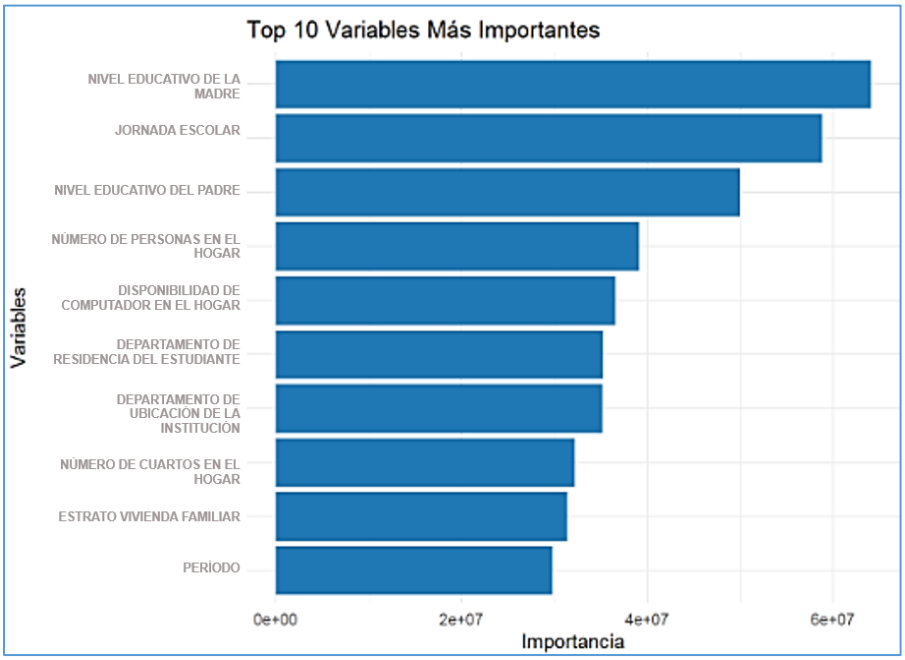

Figura 14. Importancia de variables del modelo Random Forest

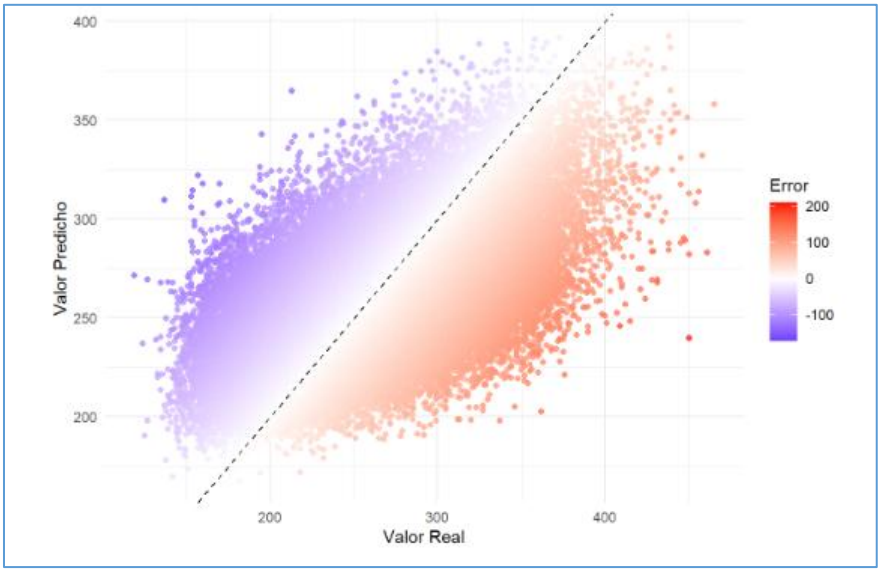

Figura 15. Relación entre valores reales y predichos – Random Forest

Para Random Forest, la *Figura 14* presenta una distribución de importancia de variables consistente con la observada en XGBoost, lo que sugiere estabilidad en la identificación de factores relevantes. En la *figura 15* se observa una ligera mayor dispersión respecto a la diagonal, lo que indica una menor precisión en comparación con XGBoost, aunque manteniendo una buena capacidad de ajuste.

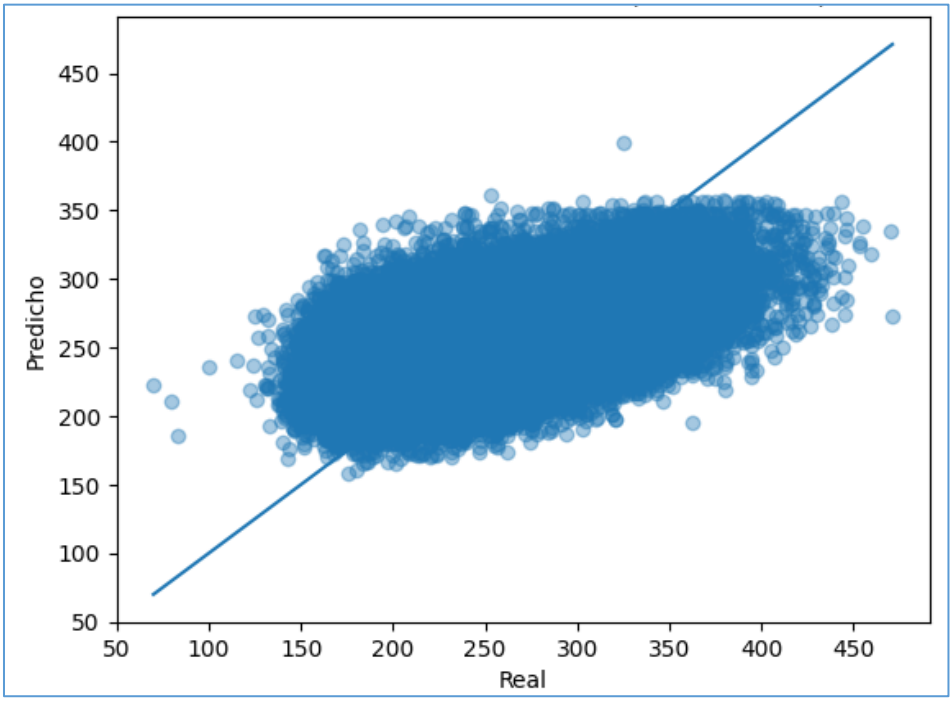

Figura 16. Relación entre valores reales y predichos – SVM

En el modelo SVM, la *Figura 16* muestra una dispersión más amplia de los valores predichos, lo que evidencia dificultades para capturar la complejidad de las relaciones presentes en los datos, especialmente en comparación con los modelos basados en árboles.

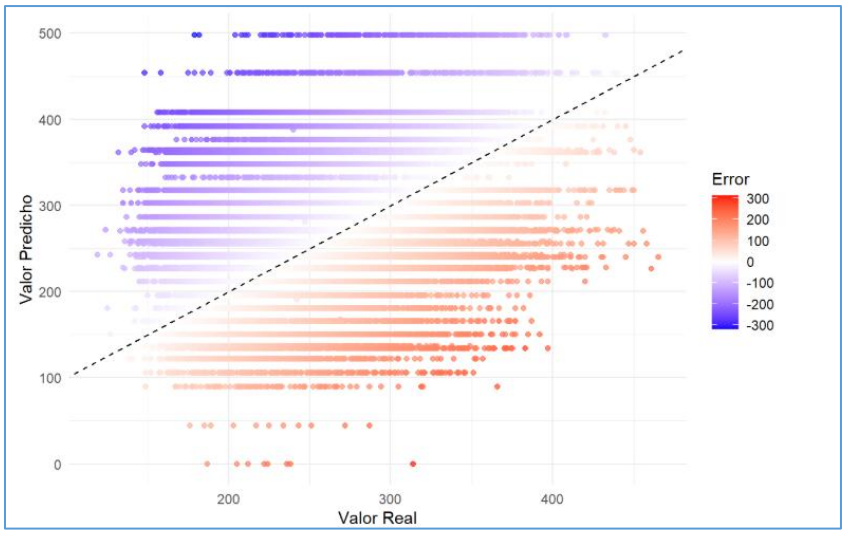

Figura 17. Relación entre valores reales y predichos – MLP

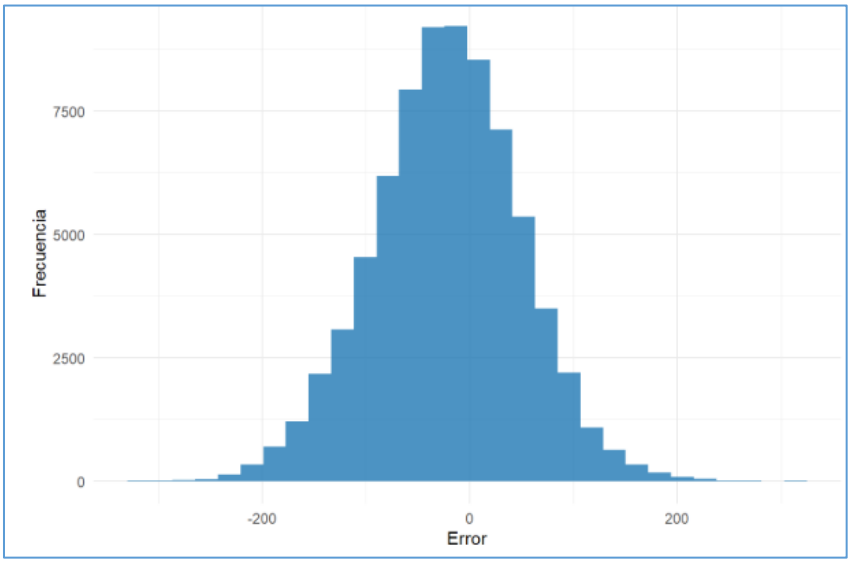

Figura 18. Distribución del error – MLP

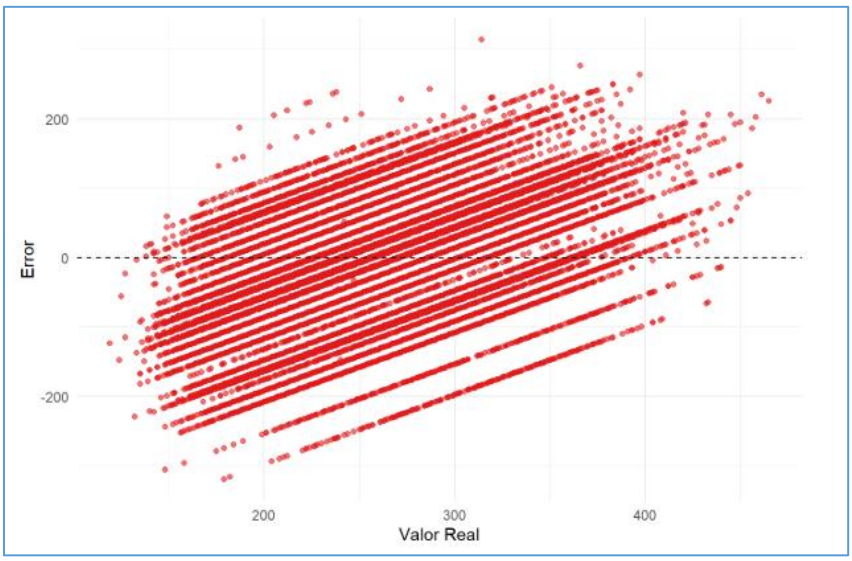

Figura 19. Error vs. valor real – MLP

Por su parte, el modelo MLP presenta un comportamiento significativamente distinto. En la *figura 17* se observa una alta dispersión en la relación entre valores reales y predichos, indicando un bajo nivel de ajuste. La *Figura 18* evidencia una distribución del error con alta variabilidad, mientras que la *Figura 19* no muestra un patrón definido entre el error y el valor real.

Este comportamiento sugiere que el modelo no logra capturar adecuadamente la estructura de los datos bajo las condiciones de entrenamiento utilizadas, lo cual puede estar asociado a la sensibilidad de las redes neuronales al tamaño de la muestra y a la configuración de sus hiperparámetros.

En conjunto, estos resultados permiten concluir que los modelos basados en árboles presentan un mejor desempeño en esta fase inicial, tanto en términos cuantitativos como en su comportamiento gráfico.

4.2. Validación de modelos con mayor volumen de datos mediante bloques

Con el fin de evaluar la estabilidad, capacidad de generalización y consistencia de los modelos previamente analizados, se llevó a cabo una segunda fase de validación utilizando un mayor volumen de datos, correspondiente al 70% del conjunto total.

A diferencia de la evaluación inicial, en esta etapa se implementó un esquema de validación por bloques, el cual permite analizar el comportamiento de los modelos en diferentes particiones del conjunto de datos, reduciendo la dependencia de una única división entrenamiento-prueba.

4.2.1. Metodología de validación por bloques

La validación por bloques consistió en dividir el conjunto de datos en cinco subconjuntos independientes. Para cada uno de ellos, se realizó una partición interna en datos de entrenamiento y prueba, utilizando índices previamente definidos con el fin de garantizar consistencia en la evaluación.

En cada iteración, se seleccionó un bloque específico del dataset, el cual fue dividido en subconjuntos de entrenamiento y prueba para entrenar y evaluar el modelo correspondiente. Este procedimiento se repitió para los cinco bloques, permitiendo obtener múltiples estimaciones del desempeño.

Con el fin de garantizar la comparabilidad de los resultados, en esta fase se mantuvieron las mismas variables predictoras, la variable objetivo y las métricas de evaluación definidas en la sección 4.1. De esta manera, las diferencias observadas pueden atribuirse al cambio en el volumen de datos y al esquema de validación implementado.

Adicionalmente, se calcularon medidas de tendencia central y dispersión, con el propósito de evaluar la estabilidad de los modelos entre bloques.

4.2.2. Resultados de la validación por bloques

En la *Tabla 14* se presentan los resultados promedio obtenidos para cada modelo bajo el esquema de validación por bloques.

Tabla 14. Resultados promedio de la validación por bloques (70% de los datos).

Modelo	RMSE	MAE	R^2	Tiempo (s)
XGBoost	39.71	31.89	0.346	5.70
Random Forest	40.09	32.04	0.334	170.76
MLP	41.50	33.38	0.286	692.15

Los resultados evidencian un comportamiento consistente con la evaluación inicial, manteniéndose diferencias claras en desempeño y eficiencia entre los modelos.

En la *Tabla 15* se presenta la variabilidad de los resultados a través de la desviación estándar de cada métrica.

Tabla 15. Desviación estándar de las métricas en la validación por bloques.

Modelo	SD RMSE	SD MAE	SD R^2	SD Tiempo
XGBoost	0.073	0.038	0.002	0.28
Random Forest	0.093	0.061	0.003	2.07
MLP	0.419	0.347	0.010	31.36

4.2.3. Análisis de resultados

Los resultados permiten identificar patrones consistentes en el comportamiento de los modelos.

XGBoost presenta el mejor desempeño global, con menores errores y baja variabilidad entre bloques, lo que indica alta estabilidad y buena capacidad de generalización.

Random Forest muestra un rendimiento similar, aunque con mayor costo computacional. A pesar de ello, mantiene resultados estables.

El modelo MLP mejora respecto a la evaluación inicial al aumentar el volumen de datos, lo que sugiere dependencia del tamaño muestral. Sin embargo, continúa siendo el modelo con mayor variabilidad y mayor costo computacional.

En términos generales, los modelos basados en árboles presentan mayor estabilidad frente a diferentes particiones de los datos, mientras que el MLP es más sensible a los cambios en los subconjuntos utilizados.

4.2.4. Comparación con la evaluación inicial

Al comparar con la validación inicial, XGBoost mantiene el mejor desempeño con mejoras marginales. Random Forest conserva un comportamiento estable, aunque con menor eficiencia.

El modelo MLP presenta una mejora al aumentar el tamaño del conjunto de datos, lo que confirma su dependencia del volumen de información y sugiere potencial de mejora en escenarios más grandes.

4.2.5. Consideración sobre el modelo SVM

El modelo SVM no fue incluido en la validación por bloques debido a limitaciones computacionales, ya que su entrenamiento en grandes volúmenes de datos implica alto consumo de memoria y tiempo de procesamiento.

Sin embargo, su evaluación inicial permitió establecer un punto de comparación frente a los demás modelos.

4.3. Validación final del modelo seleccionado

Con base en los resultados obtenidos en la evaluación inicial (Sección 4.1) y en la validación por bloques (Sección 4.2), se seleccionó el modelo XGBoost como la alternativa con mejor desempeño global. Esta selección se realizó considerando simultáneamente el menor error de predicción (RMSE y MAE), la mayor capacidad explicativa (R^2) y la estabilidad evidenciada en la validación por

bloques. Este modelo demostró menores errores de predicción, mayor estabilidad entre particiones y una adecuada eficiencia computacional frente a los demás algoritmos evaluados.

Una vez verificada su consistencia mediante esquemas de validación robustos, se procede a realizar una validación final con el fin de evaluar su desempeño en un conjunto de datos independiente, simulando un escenario de aplicación en condiciones reales.

4.3.1. Metodología de validación final

Para esta etapa se utilizó el dataset completo resultante del proceso de depuración y preparación descrito en el capítulo 2, el cual constituye la base consolidada para el análisis del presente estudio.

A partir de este conjunto de datos, se realizó una partición aleatoria en dos subconjuntos: un 70% destinado al entrenamiento del modelo y el 30% restante reservado exclusivamente para la evaluación. Esta estrategia permite medir la capacidad de generalización del modelo sobre datos no utilizados durante el proceso de aprendizaje.

Previo al entrenamiento, se aplicó la transformación de variables categóricas mediante codificación dummy, con el propósito de adecuar la estructura de los datos a los requerimientos del algoritmo XGBoost. Asimismo, se mantuvieron las mismas variables predictoras definidas en las fases anteriores, garantizando la coherencia metodológica del estudio.

El modelo fue entrenado utilizando los mismos hiperparámetros previamente establecidos, con el fin de asegurar la comparabilidad de los resultados y evaluar su desempeño bajo condiciones controladas.

4.3.2. Resultados de la validación final

Los resultados obtenidos en la evaluación del modelo sobre el conjunto de prueba se presentan a continuación:

Tabla 16. Resultados de la validación final del modelo XGBoost

Modelo	RMSE	MAE	R2	Tiempo_seg
XGBoost	39.72	31.89	0.346	43.74

Los resultados mostrados en la *tabla 16* muestran que el modelo mantiene un desempeño consistente respecto a las fases anteriores, con valores de error similares en RMSE y MAE.

El coeficiente de determinación (R^2) refleja una capacidad moderada para explicar la variabilidad del puntaje global, lo cual es coherente con la naturaleza multifactorial del rendimiento académico, donde intervienen factores académicos, socioeconómicos e institucionales.

4.3.3. Análisis gráfico del desempeño del modelo

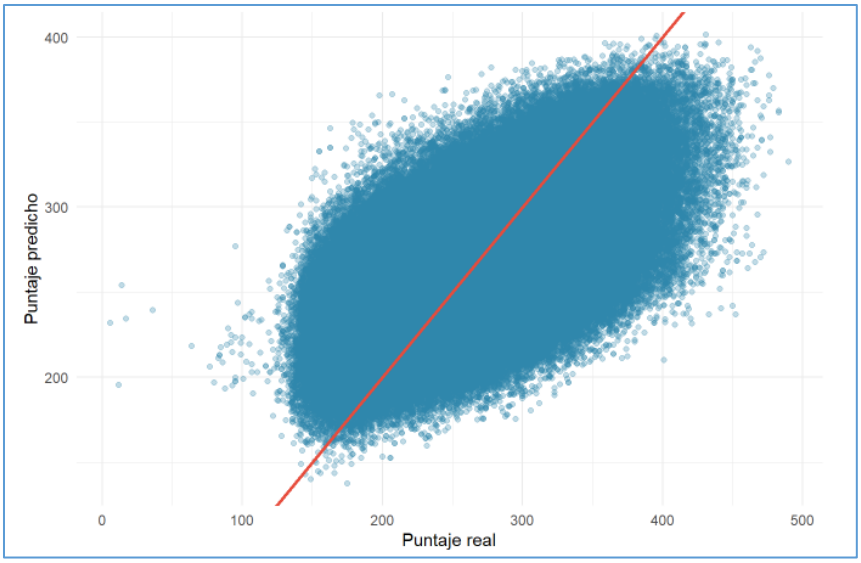

Figura 20. Valores reales vs predichos – Modelo XGBoost

La relación entre los valores reales y los valores predichos (Figura 20) muestra una concentración cercana a la línea diagonal, lo que indica que el modelo logra capturar adecuadamente la tendencia general del puntaje global.

Sin embargo, se observa dispersión en los valores extremos, lo que sugiere limitaciones en la predicción de casos atípicos. Este comportamiento es esperable en fenómenos complejos como el rendimiento académico, donde influyen múltiples variables no observadas.

En conjunto, los resultados gráficos evidencian un nivel adecuado de ajuste y una capacidad razonable de generalización del modelo.

4.3.4. Importancia de variables en el modelo final

Con el fin de identificar los factores que tienen mayor influencia en la predicción del rendimiento académico, se analizó la importancia de variables generada por el modelo XGBoost.

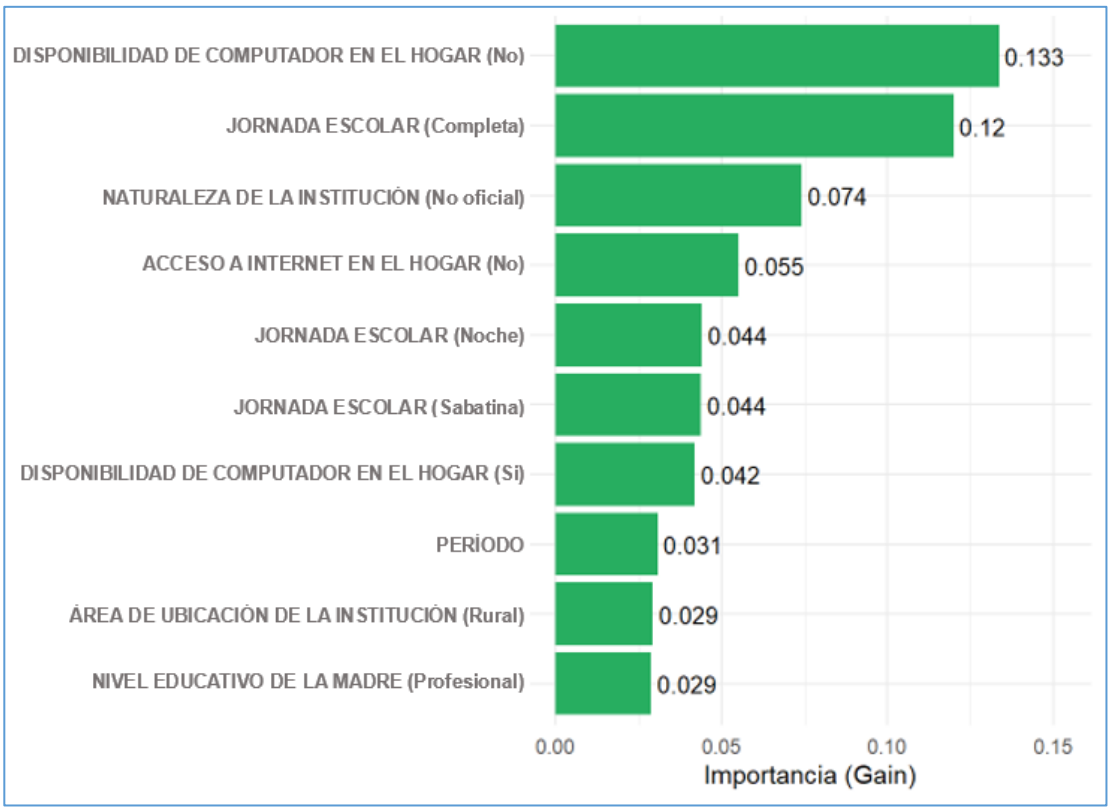

Figura 21. Importancia de variables – Modelo XGBoost

Los resultados de la figura 21 evidencian que las variables con mayor peso en la predicción están asociadas principalmente a condiciones socioeconómicas y características institucionales, tales como el acceso a computador, la disponibilidad de internet, la jornada escolar y la naturaleza del establecimiento educativo.

Estos hallazgos son consistentes con los resultados obtenidos en el análisis exploratorio, donde dichas variables ya habían mostrado una relación significativa con el puntaje global. Asimismo, guardan coherencia con lo planteado en el marco teórico, el cual destaca la influencia del contexto socioeconómico y de las condiciones educativas en el desempeño académico de los estudiantes.

En conjunto, la convergencia entre el análisis estadístico, el modelo predictivo y el sustento teórico refuerza la validez de los resultados obtenidos y respalda la capacidad explicativa del modelo frente al fenómeno estudiado.

4.3.5. Análisis comparativo de errores entre modelos

Con el propósito de complementar la interpretación de las métricas obtenidas en las diferentes fases de validación, se realizó un análisis gráfico comparativo de los errores de predicción de los modelos evaluados. Estas visualizaciones permiten observar de manera más detallada el

comportamiento de los errores, su dispersión y la estabilidad de cada algoritmo frente a la predicción del puntaje global.

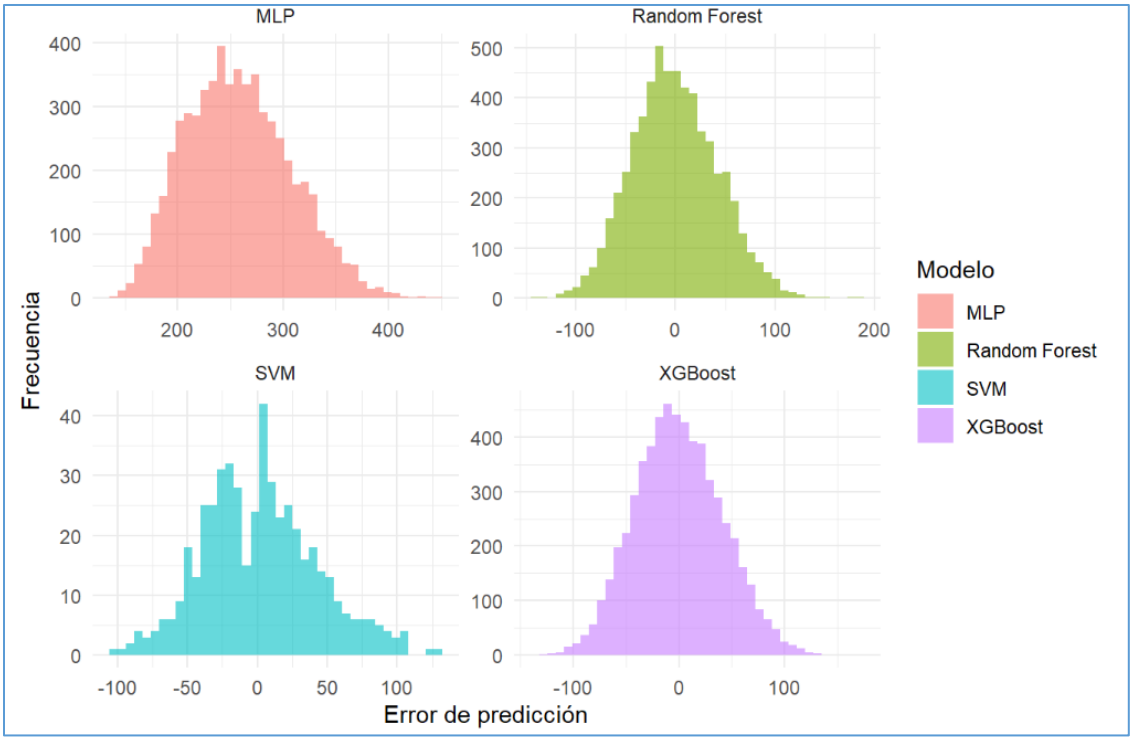

Figura 22. Distribución de errores de predicción por modelo

La Figura 22 muestra la distribución de los errores de predicción para cada uno de los modelos analizados. En los modelos XGBoost y Random Forest se observa una mayor concentración de los errores alrededor de cero, lo que indica una mejor capacidad para aproximarse a los valores reales del puntaje global. Este comportamiento es coherente con los resultados obtenidos previamente en las métricas RMSE y MAE, donde ambos modelos presentaron los menores niveles de error.

En el caso del modelo SVM, la distribución de los errores presenta una dispersión ligeramente mayor, lo que evidencia una menor precisión respecto a los modelos basados en árboles. Sin embargo, su comportamiento continúa siendo relativamente estable y consistente.

Por otra parte, el modelo MLP presenta una distribución mucho más amplia y dispersa, con una mayor presencia de errores alejados de cero. Esto refleja dificultades para capturar adecuadamente las relaciones presentes en los datos y confirma el comportamiento observado durante las etapas anteriores de validación, donde este modelo presentó los mayores errores y menor estabilidad.

Asimismo, la presencia de errores extremos en algunos modelos evidencia que existen observaciones particularmente difíciles de predecir, situación esperable en fenómenos complejos

como el rendimiento académico, donde intervienen múltiples factores académicos, sociales e institucionales.

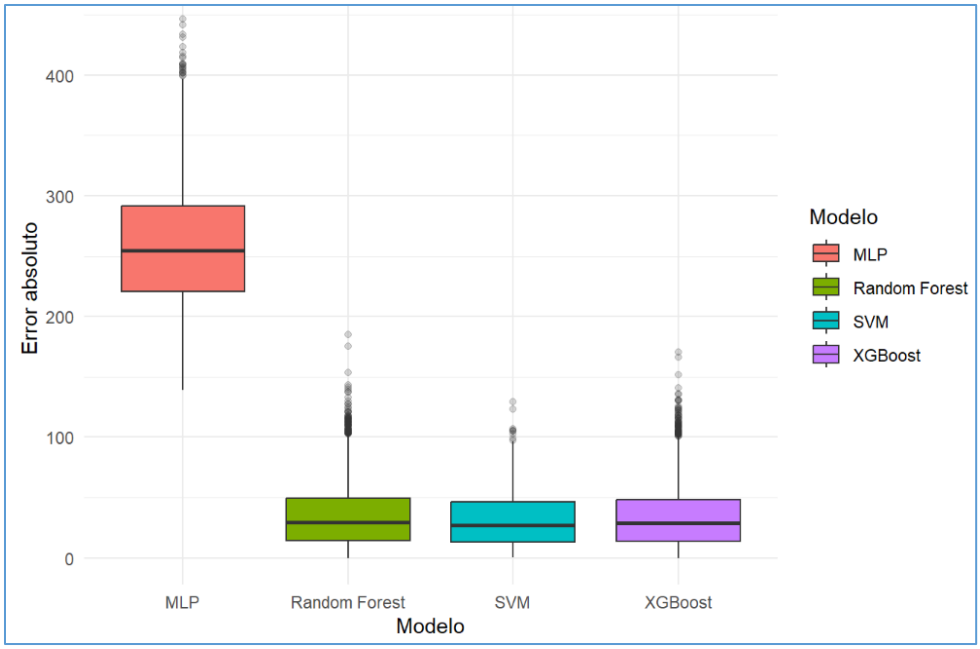

Figura 23. Distribución de errores absolutos por modelo

La Figura 23 permite comparar la distribución de los errores absolutos obtenidos por los modelos evaluados. Se observa que XGBoost y Random Forest presentan menor dispersión y medianas de error más bajas, lo que indica un comportamiento más estable en las predicciones realizadas.

El modelo SVM presenta una variabilidad moderada, mientras que el modelo MLP evidencia una dispersión considerablemente mayor y una presencia más marcada de valores atípicos. Esto sugiere una mayor inestabilidad en sus predicciones y una menor capacidad de generalización bajo las condiciones de entrenamiento utilizadas.

De igual forma, el análisis gráfico permite comprender mejor la diferencia entre las métricas MAE y RMSE obtenidas previamente. En todos los modelos el valor del RMSE es superior al MAE, lo cual indica la existencia de algunos errores de mayor magnitud. Debido a que el RMSE penaliza más fuertemente este tipo de errores extremos, la diferencia entre ambas métricas refleja la presencia de casos donde la predicción resulta más compleja.

En conjunto, los resultados gráficos confirman el mejor desempeño de los modelos basados en árboles, especialmente XGBoost, el cual mostró un comportamiento más estable, menor dispersión del error y una mejor capacidad para representar las tendencias generales del rendimiento académico.

4.3.6. Conclusión de la validación

Los resultados obtenidos en las diferentes fases de validación permitieron identificar a XGBoost como el modelo con mejor desempeño para la predicción del rendimiento académico, al presentar menores niveles de error, mayor estabilidad y una adecuada eficiencia computacional frente a los demás algoritmos evaluados.

La consistencia observada entre la evaluación inicial, la validación por bloques y la validación final evidencia una adecuada capacidad de generalización del modelo, manteniendo un comportamiento estable ante diferentes particiones y volúmenes de datos.

De manera general, los resultados obtenidos muestran que los modelos basados en árboles presentaron un desempeño superior en comparación con enfoques como SVM y MLP, especialmente en términos de estabilidad y precisión predictiva.

Asimismo, el análisis realizado permitió identificar la influencia de variables socioeconómicas e institucionales sobre el rendimiento académico, en coherencia con los hallazgos obtenidos en el análisis exploratorio y con lo planteado en la literatura especializada.

En conjunto, el proceso desarrollado en este capítulo permitió validar un modelo predictivo robusto y consistente, contribuyendo al cumplimiento del objetivo específico relacionado con la validación y comparación de modelos de aprendizaje automático para la predicción del rendimiento académico.

5. ANÁLISIS, INTERPRETACIÓN E IMPLICACIONES DE LOS RESULTADOS

En este capítulo se examinan las implicaciones de los resultados obtenidos a lo largo del análisis, mediante su contraste con la literatura académica y estudios previos en el campo de la Minería de Datos Educativos. El objetivo es interpretar los hallazgos desde una perspectiva teórica y aplicada, con el fin de generar recomendaciones orientadas a la toma de decisiones en el ámbito educativo.

5.1. Interpretación de los resultados del modelo predictivo

Los resultados obtenidos en el capítulo anterior evidencian que el modelo XGBoost alcanzó el mejor desempeño entre los algoritmos evaluados, con valores de R^2 cercanos a 0.34. Si bien este nivel de explicación de la varianza puede considerarse moderado, es consistente con lo reportado en la literatura especializada.

Diversos estudios han demostrado que el rendimiento académico es un fenómeno multifactorial influenciado por variables cognitivas, socioeconómicas, institucionales y emocionales, muchas de las cuales no se encuentran disponibles en bases de datos estructuradas [11], [18]. En este sentido, la capacidad explicativa de los modelos predictivos suele presentar limitaciones inherentes asociadas a la ausencia de variables relevantes y a la complejidad del fenómeno analizado.

Adicionalmente, investigaciones previas han señalado que, en contextos educativos, los valores de R^2 en modelos predictivos suelen ubicarse en rangos moderados, generalmente entre 0.20 y 0.40, incluso al utilizar algoritmos avanzados [9], [18]. Por tanto, los resultados obtenidos en este estudio se encuentran dentro de los rangos esperados, lo que respalda su validez metodológica.

Desde una perspectiva de analítica predictiva, Shmueli [19] establece que el objetivo principal de estos modelos no es maximizar la explicación total de la variabilidad, sino optimizar la capacidad de predicción. En este sentido, métricas como el RMSE y el MAE adquieren mayor relevancia práctica, al reflejar el error promedio en la estimación del puntaje global.

5.2. Comparación con estudios previos

Los hallazgos del presente estudio coinciden con investigaciones recientes que destacan el mejor desempeño de los modelos basados en árboles en datos tabulares estructurados [8], [17]. En particular, algoritmos como XGBoost y Random Forest han demostrado ser más robustos y estables en comparación con modelos como SVM y redes neuronales, especialmente cuando no se dispone de grandes volúmenes de datos o de variables altamente informativas.

Asimismo, la literatura indica que los modelos de deep learning no siempre superan a los enfoques tradicionales en este tipo de datos, debido a su alta dependencia del tamaño muestral y de la

calidad de las variables [17]. Este comportamiento es consistente con los resultados obtenidos, donde el modelo MLP presentó un desempeño inferior y mayor variabilidad.

Por otra parte, los resultados del análisis de importancia de variables confirman la relevancia de factores socioeconómicos e institucionales en el rendimiento académico, lo cual ha sido ampliamente documentado en estudios sobre educación en Colombia [1], [14].

5.3. Implicaciones para la toma de decisiones educativas

Los resultados obtenidos permiten identificar elementos clave para la formulación de estrategias educativas basadas en evidencia.

En primer lugar, la influencia significativa de variables socioeconómicas sugiere la necesidad de fortalecer políticas públicas orientadas a reducir brechas estructurales, especialmente en el acceso a recursos tecnológicos y condiciones adecuadas de estudio.

En segundo lugar, la relevancia de variables institucionales, como la jornada escolar y el tipo de establecimiento, resalta la importancia de mejorar las condiciones educativas y garantizar entornos de aprendizaje equitativos.

Adicionalmente, el uso de modelos predictivos como herramienta de apoyo permite identificar grupos de estudiantes en riesgo académico, facilitando la implementación de estrategias de intervención temprana. En este sentido, estos modelos pueden ser integrados en sistemas de analítica educativa institucional, con el fin de apoyar la toma de decisiones basada en datos y optimizar la asignación de recursos educativos.

5.4. Limitaciones del estudio

A pesar de los resultados obtenidos, es importante reconocer ciertas limitaciones. En particular, la ausencia de variables relacionadas con aspectos emocionales, motivacionales y pedagógicos limita la capacidad explicativa del modelo.

Asimismo, la naturaleza transversal de los datos impide capturar la evolución del rendimiento académico en el tiempo, lo cual podría mejorar significativamente el desempeño de los modelos [12].

5.5. Conclusión del objetivo

Con lo anterior, se logra abordar el objetivo específico 4, al analizar e interpretar los resultados obtenidos a la luz de la literatura académica y del contexto educativo, permitiendo comprender las implicaciones del modelo desarrollado.

Los hallazgos evidencian que, aunque la capacidad explicativa del modelo es moderada, esta se encuentra dentro de los rangos esperados según estudios previos, lo que confirma la complejidad del fenómeno del rendimiento académico.

Asimismo, se identificaron factores clave asociados al desempeño estudiantil y se formularon recomendaciones orientadas a fortalecer la toma de decisiones en el ámbito educativo. En conjunto, estos resultados aportan evidencia empírica y analítica para el diseño de estrategias pedagógicas y el fortalecimiento de políticas educativas basadas en datos.

Finalmente, el estudio contribuye al campo de la Minería de Datos Educativos al integrar técnicas de análisis estadístico y aprendizaje automático en el contexto colombiano, proporcionando una base metodológica replicable para futuras investigaciones orientadas al análisis del rendimiento académico.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. Conclusiones

El presente estudio permitió analizar el rendimiento académico de los estudiantes en las pruebas Saber 11 mediante técnicas de minería de datos educativos y modelos de aprendizaje automático, integrando variables académicas, socioeconómicas e institucionales.

Los resultados obtenidos evidenciaron la existencia de relaciones significativas entre el contexto del estudiante y su desempeño académico, destacando especialmente la influencia de factores socioeconómicos e institucionales sobre el Puntaje Global.

Asimismo, la evaluación de diferentes modelos predictivos permitió identificar que los enfoques basados en árboles, particularmente XGBoost, presentan un mejor desempeño y estabilidad para el análisis de datos educativos estructurados.

De manera general, el estudio demuestra la utilidad de las técnicas de analítica de datos como herramientas de apoyo para el análisis del rendimiento académico y la toma de decisiones basada en evidencia en el ámbito educativo.

6.2. Trabajos futuros

Como trabajo futuro, se recomienda complementar el análisis desarrollado mediante la incorporación de variables relacionadas con factores emocionales, motivacionales y pedagógicos, las cuales no se encontraban disponibles en la base de datos utilizada en esta investigación. Estudios en el campo de la Minería de Datos Educativos han señalado la relevancia de estas variables en el análisis integral del rendimiento académico [11], [18].

Asimismo, sería pertinente desarrollar estudios longitudinales que permitan analizar la evolución del desempeño académico en el tiempo, dado que el presente estudio se realizó a partir de datos de corte transversal [12].

Adicionalmente, futuras investigaciones podrían evaluar modelos de aprendizaje profundo y otras técnicas avanzadas utilizando conjuntos de datos más amplios y con mayor diversidad de variables, con el fin de analizar posibles mejoras en la capacidad predictiva de los modelos [17].
Finalmente, se propone aplicar la metodología desarrollada en otros contextos educativos o niveles de formación, con el propósito de comparar resultados y analizar la generalización del enfoque propuesto.

7. REFERENCIAS BIBLIOGRAFICAS

[1] Instituto Colombiano para la Evaluación de la Educación (ICFES), Análisis de los resultados de la prueba Saber 11° 2024, Bogotá, Colombia, 2024. [En línea]. Disponible en: sitio oficial del ICFES.

[2] C. Romero y S. Ventura, "Data mining in education," WIREs Data Mining and Knowledge Discovery, vol. 3, núm. 1, pp. 12–27, 2013, doi: 10.1002/widm.1075.

[3] A. Kord, A. Aboelfetouh y S. M. Shohieb, "Academic course planning recommendation and students' performance prediction multi-modal based on educational data mining techniques," Journal of Computing in Higher Education, 2025, doi: 10.1007/s12528-024-09426-0.

[4] A. Nabil, M. Seyam y A. Abou-Elfetouh, "Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks," IEEE Access, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.

[5] J.-B. M. B. Sanfo, "Application of explainable artificial intelligence approach to predict student learning outcomes," Journal of Computational Social Science, vol. 8, núm. 1, p. 9, feb. 2025, doi: 10.1007/s42001-024-00344-w.

[6] T. Chen y C. Guestrin, "XGBoost: A scalable tree boosting system," en Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[7] R. S. Baker y G. Siemens, "Educational data mining and learning analytics," en Learning Analytics: From Research to Practice, Nueva York, NY, USA: Springer, 2014, pp. 253–272, doi: 10.1007/978-1-4614-3305-7_1.

[8] M. Fernández-Delgado, E. Cernadas, S. Barro y D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," Journal of Machine Learning Research, vol. 15, pp. 3133–3181, 2014.

[9] E. Hernández-Leal, J. Duque-Méndez y J. A. Rodríguez-Rodríguez, "Educational data mining: A review of case studies," Heliyon, vol. 7, núm. 3, e06623, 2021, doi: 10.1016/j.heliyon.2021.e06623.

[10] G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. Nueva York, NY, USA: Springer, 2021, doi: 10.1007/978-1-0716-1418-1.

[11] S. B. Kotsiantis, K. D. Pierrakeas y P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Computers & Education, vol. 53, núm. 3,

pp. 1093–1104, 2009, doi: 10.1016/j.compedu.2009.05.001.

[12] Z. Papamitsiou y A. A. Economides, “Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence,” Educational Technology & Society, vol. 17, núm. 4, pp. 49–64, 2014.

[13] S. M. Lundberg y S.-I. Lee, “A unified approach to interpreting model predictions,” en Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.

[14] Observatorio de Gestión Educativa ExE, “Resultados Saber 11° en Colombia,” Bogotá, Colombia, 2024. [En línea]. Disponible en: www.datos.gov.co.

[15] L. Breiman, “Random forests,” Machine Learning, vol. 45, núm. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] V. N. Vapnik, Statistical Learning Theory. Nueva York, NY, USA: Wiley, 1998.

[17] L. Grinsztajn, E. Oyallon y G. Varoquaux, “Why do tree-based models still outperform deep learning on tabular data?,” en Advances in Neural Information Processing Systems (NeurIPS), 2022, doi: 10.48550/arXiv.2207.08815.

[18] M. Yağcı, “Educational data mining: Prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, núm. 1, 2022, doi: 10.1186/s40561-022-00192-x.

[19] G. Shmueli, “To explain or to predict?,” Statistical Science, vol. 25, núm. 3, pp. 289–310, 2010, doi: 10.1214/10-STS330.

[20] J. M. Wooldridge, Introductory Econometrics: A Modern Approach, 5th ed. Boston, MA, USA: Cengage Learning, 2015.

[21] P. Domingos, “A few useful things to know about machine learning,” Communications of the ACM, vol. 55, núm. 10, pp. 78–87, 2012, doi: 10.1145/2347736.2347755.

[22] L. Grinsztajn, E. Oyallon y G. Varoquaux, “Deep learning for tabular data: A survey,” arXiv preprint, 2022. [En línea]. Disponible en: https://arxiv.org/abs/2207.08815.

8. ANEXOS

Anexo A. Diccionario de variables del dataset

Información general

El dataset corresponde a los resultados consolidados y procesados de las pruebas SABER 11 en Colombia, utilizados para el proyecto de minería de datos educativos y aprendizaje automático enfocado en la predicción del rendimiento académico y el análisis de factores asociados al desempeño estudiantil.

Variables del Dataset

Variable	Tipo	Descripción
PERÍODO	Categórica	Período de aplicación de la prueba SABER 11.
TIPO DE DOCUMENTO DEL ESTUDIANTE	Categórica	Tipo de documento del estudiante.
ÁREA DE UBICACIÓN DE LA INSTITUCIÓN	Categórica	Área de ubicación de la institución educativa (urbana/rural).
COLEGIO BILINGÜE	Categórica	Indica si la institución educativa es bilingüe.
CALENDARIO ACADÉMICO	Categórica	Calendario académico de la institución educativa.
CARÁCTER DE LA INSTITUCIÓN EDUCATIVA	Categórica	Carácter académico de la institución educativa.
CÓDIGO DEL DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	Numérica/Categórica	Código DANE del departamento de ubicación de la institución.
CÓDIGO DEL MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	Numérica/Categórica	Código DANE del municipio de ubicación de la institución.
DEPARTAMENTO DE UBICACIÓN DE LA INSTITUCIÓN	Categórica	Departamento donde se ubica la institución educativa.
GÉNERO DE LA INSTITUCIÓN EDUCATIVA	Categórica	Género de la institución educativa (mixto, femenino o masculino).

JORNADA ESCOLAR	Categórica	Jornada académica de la institución educativa.
MUNICIPIO DE UBICACIÓN DE LA INSTITUCIÓN	Categórica	Municipio donde se ubica la institución educativa.
NATURALEZA DE LA INSTITUCIÓN	Categórica	NATURALEZA DE LA INSTITUCIÓN (oficial/no oficial).
CÓDIGO DEL DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	Numérica/Categórica	Código DANE del departamento donde presentó la prueba.
CÓDIGO DEL MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	Numérica/Categórica	Código DANE del municipio donde presentó la prueba.
CÓDIGO DEL DEPARTAMENTO DE RESIDENCIA	Numérica/Categórica	Código DANE del departamento de residencia del estudiante.
CÓDIGO DEL MUNICIPIO DE RESIDENCIA	Numérica/Categórica	Código DANE del municipio de residencia del estudiante.
DEPARTAMENTO DE PRESENTACIÓN DE LA PRUEBA	Categórica	Departamento donde el estudiante presentó la prueba.
DEPARTAMENTO DE RESIDENCIA DEL ESTUDIANTE	Categórica	Departamento de residencia del estudiante.
GÉNERO DEL ESTUDIANTE	Categórica	Género del estudiante.
MUNICIPIO DE PRESENTACIÓN DE LA PRUEBA	Categórica	Municipio donde el estudiante presentó la prueba.
MUNICIPIO DE RESIDENCIA DEL ESTUDIANTE	Categórica	Municipio de residencia del estudiante.
ESTUDIANTE PRIVADO DE LA LIBERTAD	Categórica	Indica si el estudiante se encuentra privado de la libertad.
NÚMERO DE CUARTOS EN EL HOGAR	Categórica	Número de cuartos en el hogar del estudiante.
NIVEL EDUCATIVO DE LA MADRE	Categórica	Nivel educativo de la madre.
NIVEL EDUCATIVO DEL PADRE	Categórica	Nivel educativo del padre.

ESTRATO VIVIENDA FAMILIAR	Categórica	Estrato socioeconómico de la vivienda.
NÚMERO DE PERSONAS EN EL HOGAR	Categórica	Número de personas en el hogar.
DISPONIBILIDAD DE AUTOMÓVIL EN EL HOGAR	Categórica	Indica si el hogar posee automóvil.
DISPONIBILIDAD DE COMPUTADOR EN EL HOGAR	Categórica	Indica si el hogar posee computador.
ACCESO A INTERNET EN EL HOGAR	Categórica	Indica si el hogar tiene acceso a internet.
DISPONIBILIDAD DE LAVADORA EN EL HOGAR	Categórica	Indica si el hogar posee lavadora.
DESEMPEÑO INGLES	Categórica	Nivel de desempeño en inglés según clasificación del ICFES.
PUNTAJE EN INGLÉS	Numérica	Puntaje obtenido en la prueba de inglés.
PUNTAJE EN MATEMÁTICAS	Numérica	Puntaje obtenido en matemáticas.
PUNTAJE EN SOCIALES Y CIUDADANAS	Numérica	Puntaje obtenido en sociales y competencias ciudadanas.
PUNTAJE EN CIENCIAS NATURALES	Numérica	Puntaje obtenido en ciencias naturales.
PUNTAJE EN LECTURA CRÍTICA	Numérica	Puntaje obtenido en lectura crítica.
PUNTAJE GLOBAL	Numérica	Puntaje global obtenido en la prueba SABER 11.
EDAD	Numérica	Edad estimada del estudiante al momento de presentar la prueba.
FECHA DE NACIMIENTO	Fecha	Fecha de nacimiento utilizada temporalmente durante el procesamiento.
PUNTAJE GLOBAL	Numérica - Variable Objetivo del Proyecto	Variable objetivo principal utilizada para los modelos predictivos de rendimiento académico.

Transformaciones Aplicadas Durante el Preprocesamiento

1. Eliminación de registros con **PUNTAJE GLOBAL = 0**.
2. Exclusión de registros correspondientes a los años 2020 y 2021.
3. Creación de la variable **EDAD** a partir de la fecha de nacimiento.
4. Eliminación de identificadores institucionales y personales.
5. Imputación de valores faltantes mediante media y moda.
6. Reemplazo de valores vacíos por la categoría **"SIN DATO".**
7. Normalización de nombres de departamentos y municipios.
8. Filtrado de Edades fuera del rango de 9 a 19 años.
9. Conversión de variables de puntaje a formato numérico.
10. Generación del dataset final listo para análisis exploratorio y modelado predictivo.

Anexo B. Código de limpieza y preparación de datos

El código correspondiente a este anexo se encuentra disponible en el siguiente enlace:

Limpieza de datos

Este script contiene el proceso de carga del dataset de las pruebas SABER 11, así como las etapas de limpieza, tratamiento de valores faltantes, depuración de registros y selección inicial de variables utilizadas en el análisis.

Anexo C. Código de análisis exploratorio y estadístico

Análisis exploratorio

Este script incluye el análisis exploratorio de datos (EDA), la evaluación de distribuciones, el análisis de correlaciones entre variables numéricas y la aplicación de pruebas estadísticas de relación con la variable objetivo PUNTAJE GLOBAL.

Anexo D. Código de validación inicial de modelos (10% de los datos)

Validación inicial de modelos

Este script implementa la validación inicial de los modelos de aprendizaje automático utilizando el 10 % del conjunto de datos, incluyendo la partición de los datos, el entrenamiento de los

modelos, la generación de predicciones y el cálculo de métricas de desempeño.

Anexo E. Código de validación por bloques de modelos predictivos

Validación por bloques de modelos

Este script contiene la implementación del esquema de validación por bloques utilizado para evaluar la estabilidad y consistencia del desempeño de los modelos bajo diferentes particiones del conjunto de datos.

Anexo F. Código de validación final del modelo XGBoost

Resultados finales

Este script corresponde al entrenamiento y validación final del modelo XGBoost utilizando un conjunto de prueba independiente, así como a la generación de métricas y visualizaciones empleadas en el análisis final de resultados.