# Modelo predictivo del puntaje global Saber 11 a nivel de establecimientos educativos en Colombia

Erick Santiago Garavito Villamil

Francisco Javier Achipíz Velasco

Asesor

Jorge Eliecer Ospino Portillo

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

**Nota de Aceptación**

_____

Jorge Eliecer Ospino Portillo

Director de Trabajo de Grado

_____

Rafael Gaitan Ospina

Jurado

2026

**Resumen**

Esta propuesta titulada “Modelo predictivo del puntaje global Saber 11 a nivel de establecimientos educativos en Colombia” busca desarrollar un modelo de aprendizaje supervisado para predecir el puntaje global promedio de la prueba Saber 11 a nivel de establecimiento educativo en Colombia. El problema central es la profunda brecha de equidad educativa, evidenciada en la diferencia de rendimiento entre zonas urbanas y rurales. Actualmente, las instituciones carecen de herramientas analíticas para anticipar estos resultados, lo que limita la toma de decisiones a una gestión reactiva post-evaluación. El proyecto pretende transformar los microdatos abiertos del ICFES en inteligencia accionable para permitir intervenciones pedagógicas tempranas y focalizadas.

El objetivo general es crear una herramienta de apoyo a la toma de decisiones con enfoque en equidad, capaz de estimar el desempeño institucional a partir de variables contextuales y socioeconómicas. Los objetivos específicos incluyen la estructuración de un conjunto de datos integrado y anonimizado, la implementación de dos modelos de regresión (uno lineal y otro de ensamble) y la comparación de su precisión mediante métricas como el error cuadrático medio (RMSE) y el error absoluto medio (MAE). Este enfoque técnico busca identificar los factores institucionales que permiten predecir el puntaje con un margen de error aceptable.

El sustento teórico integra la economía de la educación, mediante la Función de Producción Educativa, y la Minería de Datos Educativa (EDM). Se plantea que el rendimiento no es aleatorio, sino el producto de insumos familiares, escolares y contextuales. Para el modelado, se fundamentan técnicas de regularización como Ridge y Lasso para mitigar la multicolinealidad de variables como el estrato y el nivel educativo de los padres. Asimismo, se proponen

algoritmos de ensamble como Random Forest y Gradient Boosting para capturar interacciones no lineales y reducir el sesgo en las predicciones.

La metodología adoptada es el estándar industrial CRISP-DM, la cual garantiza un proceso cíclico y robusto. Esta se divide en seis fases: comprensión del problema, comprensión de los datos (ICFES), preparación de los datos (ingeniería de características y agregación a nivel de colegio), modelado, evaluación y despliegue. La fase de preparación es crítica, pues requiere transformar microdatos individuales en promedios y distribuciones por establecimiento educativo. Los recursos necesarios comprenden el uso de Python y sus librerías especializadas en ciencia de datos, operando sobre repositorios de datos abiertos gubernamentales.

Se espera que el proyecto entregue un dataset depurado y un prototipo funcional del modelo en formato de código. Los resultados deben permitir la jerarquización de variables predictoras, visibilizando cómo factores como la ubicación rural o la naturaleza jurídica impactan el desempeño.

***Palabras clave*****:** Aprendizaje supervisado, Regresión, Equidad educativa, Saber 11.

## Abstract

This proposal, titled “Predictive Modeling of Saber 11 Global Scores at the School Level in Colombia,” aims to develop a supervised learning model to estimate the average institutional performance in the Saber 11 standardized examination. The study addresses a critical issue of educational inequality, reflected in persistent performance gaps between urban and rural contexts. Currently, educational institutions lack analytical tools to anticipate these outcomes, resulting in predominantly reactive, post-assessment decision-making processes. This project seeks to transform publicly available microdata into actionable insights that support timely and targeted educational interventions.

The main objective is to design a decision-support tool with an equity-oriented perspective, capable of predicting institutional performance based on contextual and socioeconomic variables. Specific objectives include constructing an integrated and anonymized dataset, implementing and comparing two regression approaches (a linear model and an ensemble-based model), and evaluating their performance using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). This technical approach is intended to identify key institutional factors that enable accurate prediction within an acceptable margin of error.

The theoretical framework combines perspectives from the economics of education—particularly the Educational Production Function—and Educational Data Mining (EDM). Academic performance is conceptualized not as a random outcome, but as the result of interacting family, school, and contextual inputs. Methodologically, regularization techniques such as Ridge and Lasso are considered to address multicollinearity among predictors, while

ensemble methods such as Random Forest and Gradient Boosting are proposed to capture nonlinear relationships and improve predictive accuracy.

The study follows the CRISP-DM methodology, ensuring a structured and iterative analytical process. This framework includes six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Data preparation is a critical stage, involving the transformation of individual-level microdata into aggregated institutional indicators. The project relies on Python and specialized data science libraries, leveraging open government data sources.

Expected outcomes include a refined dataset and a functional prototype of the predictive model. The results are intended to enable the prioritization of predictive variables, providing insight into how factors such as rural location or institutional characteristics influence academic performance.

***Keywords*:** Supervised learning, Regression, Educational equity, Saber 11.

## Tabla de Contenidos

Introducción ..... 12

Planteamiento del Problema ..... 13

Justificación ..... 16

Objetivos ..... 17

Objetivo General ..... 17

Objetivos Específicos ..... 17

Marco Teórico ..... 18

La Calidad Educativa desde la Función de Producción ..... 18

Minería de Datos Educativa (EDM) y Metodología CRISP-DM ..... 19

Fundamentos de Aprendizaje Supervisado para Regresión ..... 19

Modelos Lineales y Regularización (El Modelo Base) ..... 19

Modelos de Ensamble: Bagging y Boosting (El Modelo Complejo) ..... 20

Teoría de Evaluación y Métricas de Desempeño ..... 21

Metodología ..... 22

Comprensión del Problema (Fase 1) ..... 22

Comprensión de los Datos (Fase 2) ..... 22

Preparación de los Datos (Fase 3) ..... 23

Modelado (Fase 4) ..... 23

Evaluación (Fase 5) ..... 24

Despliegue y Consideraciones (Fase 6) ..... 24

Resultados ..... 25

Fuente y Recolección de los Datos ..... 25

Construcción del Conjunto de Datos Integrado ..... 27
Análisis Exploratorio y Caracterización del Target ..... 29
Distribución del Puntaje Global ..... 29
Distribución del Puntaje por Área de Conocimiento ..... 32
Brechas Educativas ..... 34
Puntaje Global Según Variables Institucionales del Establecimiento ..... 37
Correlaciones entre Variables Socioeconómicas y el Puntaje Global ..... 40
Diagnóstico de Multicolinealidad ..... 43
Modelado y Selección de los Mejores Modelos ..... 44
Selección Dentro de la Familia Lineal ..... 44
Selección Dentro de la Familia de Ensamble ..... 45
Evaluación en el Conjunto de Prueba 2024 ..... 46
Métricas de Desempeño Predictivo ..... 46
Análisis de Residuos ..... 50
Variables Predictoras: Importancia e Interpretabilidad ..... 51
Importancia de Variables en el Modelo de Ensamble ..... 51
Interpretabilidad Mediante Valores SHAP ..... 53
Coeficientes del Modelo Lineal Ridge ..... 55
Análisis de Equidad del Modelo por Subgrupos ..... 56
Síntesis de Resultados y Verificación de Objetivos ..... 60
Conclusiones ..... 62
Recomendaciones ..... 64
Referencias Bibliográficas ..... 66

## Lista de Tablas

**Tabla 1** *Resumen del Proceso de Integración y Dimensionamiento del Dataset* ..... 28

**Tabla 2** *Estadísticas Descriptivas Puntaje por Área de Conocimiento Saber 11 (2021–2024)* .. 33

**Tabla 3** *Análisis de Brechas Educativas Mediante Pruebas Mann-Whitney U y Kruskal-Wallis* 34

**Tabla 4** *Correlaciones de Spearman entre Variables Socioeconómicas y el Puntaje Global Saber 11 (Nivel Estudiante)* ..... 40

**Tabla 5** *Comparación de Modelos Lineales Mediante RMSE en Validación Cruzada CV-5 sobre Entrenamiento 2021–2023* ..... 45

**Tabla 6** *Comparación de Modelos de Ensamble Mediante RMSE en Validación Cruzada CV-5 sobre Entrenamiento 2021–2023* ..... 46

**Tabla 7** *Métricas de Evaluación en el Conjunto de Prueba (2024) para los Dos Modelos Seleccionados* ..... 48

**Tabla 8** *Top 10 Variables Predictoras por Importancia Gini/Gain en el Modelo XGBoost CUDA* ..... 51

**Tabla 9** *Error Absoluto Medio (MAE) por Subgrupo en el Conjunto de Prueba 2024 para Ambos Modelos Seleccionados* ..... 57

**Tabla 10** *Verificación del Cumplimiento de los Objetivos Específicos del Proyecto a Partir de los Resultados Obtenidos* ..... 60

## Lista de Figuras

**Figura 1** *Página de Registro para el Acceso a los Datos del ICFES* ..... 26

**Figura 2** *Página de Acceso a las Diferentes Bases de Datos del ICFES*..... 26

**Figura 3** *Listado de Bases de Datos del Examen Saber 11 para Diferentes Años y por Calendario Académico*..... 27

**Figura 4** *Histograma de la Distribución del Puntaje Global* ..... 30

**Figura 5** *Gráfico Q-Q Plot*..... 31

**Figura 6** *Boxplot Comparativo del Puntaje Global Según Tipo de Ubicación Institucional (Urbana vs. Rural)*..... 32

**Figura 7** *Comparación de la Distribución de Puntajes por Área de Conocimiento*..... 33

**Figura 8** *Puntaje Global Promedio por Departamento y Naturaleza del Establecimiento*..... 36

**Figura 9** *Puntaje Global Promedio por Departamento y Zona de Ubicación*..... 37

**Figura 10** *Distribución del Puntaje Global Saber 11 Según Naturaleza, Zona, Jornada, Calendario, Carácter y Género del Establecimiento Educativo* ..... 38

**Figura 11** *Matriz de Correlaciones de Spearman entre Variables Socioeconómicas y el Puntaje Global* ..... 41

**Figura 12** *Matriz de Correlaciones de Spearman entre Variables Socioeconómicas y Correlaciones Individuales con el Target*..... 42

**Figura 13** *Factor de Inflación de la Varianza (VIF) por Variable Numérica* ..... 43

**Figura 14** *Heatmap de Correlaciones de Spearman entre Predictores*..... 44

**Figura 15** *Distribución de Densidad del Puntaje Global Promedio por Establecimiento en el Conjunto de Entrenamiento (2021–2023) y el Conjunto de Prueba (2024)*..... 47

**Figura 16** *Comparación de Métricas de Evaluación en el Conjunto de Prueba 2024 para Ridge (L2) y XGBoost CUDA*..... 49

**Figura 17** *Heatmap de la Tabla de Resultados en el Conjunto de Prueba 2024 para Ridge (L2) y XGBoost CUDA* ..... 50

**Figura 18** *Análisis de Residuos del Modelo XGBoost CUDA en el Conjunto de Prueba 2024* .. 51

**Figura 19** *Importancia de Variables en el Modelo XGBoost CUDA*..... 52

**Figura 20** *Heatmap Comparativo de Importancia Normalizada para los Top 15 Predictores en Ambos Modelos*..... 53

**Figura 21** *SHAP Summary Plot (Beeswarm) para el Modelo XGBoost CUDA* ..... 54

**Figura 22** *Top 20 Coeficientes del Modelo Ridge (L2) sobre Variables Estandarizadas* ..... 56

**Figura 23** *Análisis de Equidad del Modelo*..... 58

**Figura 24** *MAE Promedio por Departamento para Ridge (L2) y XGBoost CUDA en el Conjunto de Prueba 2024*..... 59

## Introducción

Colombia enfrenta brechas educativas profundas y persistentes que limitan la movilidad social de miles de jóvenes, especialmente en contextos rurales y socioeconómicamente vulnerables. Aunque los factores asociados a estas disparidades son ampliamente conocidos, las instituciones educativas siguen sin contar con herramientas que les permitan anticipar su desempeño y actuar antes de que los resultados lleguen. La gestión de la calidad educativa permanece atrapada en un ciclo reactivo.

Este proyecto rompe ese ciclo. A partir de datos abiertos del ICFES y bajo la metodología CRISP-DM, se desarrolla un modelo predictivo de aprendizaje supervisado capaz de estimar el puntaje global promedio Saber 11 por establecimiento educativo, transformando variables contextuales e institucionales en inteligencia accionable para la toma de decisiones pedagógicas con enfoque en equidad.

## Planteamiento del Problema

El desempeño en la prueba Saber 11 es un factor determinante en el futuro de los estudiantes colombianos, condicionando su acceso a la educación superior y a oportunidades de movilidad social. Sin embargo, el sistema educativo nacional enfrenta profundas y persistentes brechas en estos resultados. Estadísticas recientes evidencian la magnitud del desafío: la brecha de rendimiento entre zonas urbanas y rurales alcanzó los 26,1 puntos en 2024, una de las más altas de la última década (Laboratorio de Economía de la Educación (LEE), 2025), mientras la diferencia entre colegios no oficiales y oficiales se mantuvo en 27,5 puntos en 2023 (Laboratorio de Economía de la Educación (LEE), 2024). El problema central radica en que las instituciones educativas, especialmente aquellas en contextos rurales y socioeconómicamente vulnerables, carecen de herramientas analíticas que les permitan anticipar estos resultados y diseñar intervenciones pedagógicas oportunas.

La literatura académica ha diagnosticado ampliamente los factores asociados a estas disparidades. Estudios como el de Artamonova et al. (2024) confirman que variables como la jornada escolar, el nivel educativo de los padres y el departamento de residencia explican gran parte de la desigualdad. La pandemia de COVID-19 agudizó estas brechas, exponiendo el impacto de la conectividad, donde la falta de acceso a Internet se correlacionó con menores puntajes (Ballesteros-Alfonso & Gómez-Velasco, 2022). Este fenómeno no es exclusivo de Colombia; la desigualdad en el acceso y la calidad educativa afecta estructuralmente a los estudiantes de bajos recursos y rurales en toda América Latina (Arias Ortiz et al., 2024), impactando la competitividad nacional, como lo demuestran los bajos resultados históricos en pruebas internacionales (Ayala García, 2015).

A pesar de la robusta evidencia diagnóstica, se identifica un vacío significativo en la literatura: la mayoría de los estudios permanecen en el plano descriptivo y no se traducen en modelos predictivos accionables a nivel institucional. Si bien investigaciones como las de Burbano (2021) y Timarán Pereira et al. (2020) han identificado factores clave como la inversión, el estrato socioeconómico y el acceso a la tecnología, estos hallazgos no se han operacionalizado en herramientas proactivas. Las instituciones conocen los factores de riesgo, pero no poseen un mecanismo basado en datos para estimar su impacto combinado y anticipar el rendimiento promedio de sus establecimientos, limitando la toma de decisiones a la intuición o a la reacción post-evaluación.

Las consecuencias de no resolver esta carencia de herramientas predictivas son graves y perpetúan la inequidad. La inacción permite que las brechas se consoliden, como lo demuestra el crecimiento de la disparidad rural (LEE, 2025), perpetuando ciclos de pobreza al limitar el acceso a la universidad de los jóvenes en regiones periféricas (Burbano, 2021). Este problema impacta directamente a los estudiantes más vulnerables, como los de departamentos con alta pobreza multidimensional (Artamonova et al., 2024) o de localidades específicas en las ciudades (Equipo Técnico Dirección de Evaluación de la Educación, 2024). Así mismo, afecta a directivos y docentes, forzándolos a una gestión reactiva, incapaz de implementar intervenciones tempranas y focalizadas.

Por lo tanto, el problema que aborda esta investigación es la ausencia de un modelo predictivo validado, basado en datos abiertos del ICFES[1] que permita a los establecimientos educativos estimar su desempeño esperado en la prueba Saber 11. Desarrollar esta herramienta es

[1] Instituto Colombiano para el Fomento de la Educación Superior.

fundamental para transformar los datos descriptivos en inteligencia accionable, permitiendo a las instituciones, especialmente a las más vulnerables, pasar de un enfoque reactivo a uno proactivo en la gestión de la calidad educativa. Esto conduce a la siguiente pregunta de investigación:

*¿Qué variables contextuales e institucionales permiten predecir, con un margen de error aceptable, el puntaje global promedio de la prueba Saber 11 por establecimiento educativo en Colombia, a partir de datos históricos del ICFES*?

## Justificación

La pertinencia de este proyecto de aplicación se fundamenta en la urgencia de cerrar las persistentes brechas de equidad en la educación colombiana, evidenciadas estadísticamente en los resultados de Saber 11 (LEE, 2025; LEE, 2024). El contexto actual, agravado por el impacto diferencial de la pandemia (Ballesteros-Alfonso & Gómez-Velasco, 2022), exige superar los métodos de gestión reactiva, que históricamente han demostrado ser insuficientes para mitigar las disparidades estructurales que afectan la movilidad social de los jóvenes más vulnerables.

El vacío que esta investigación pretende llenar no es teórico, sino de aplicación práctica. Aunque la literatura es robusta en diagnosticar los factores de riesgo asociados al rendimiento, como el estrato, la jornada o el contexto familiar (Artamonova et al., 2024; Timarán Pereira et al., 2020), carece de herramientas que traduzcan esos diagnósticos en modelos predictivos accionables a nivel de establecimiento. Como señalaba Ayala García (2015), la falla en implementar ajustes pedagógicos oportunos es histórica. Este proyecto contribuye directamente al desarrollar y validar esa herramienta predictiva faltante, utilizando datos abiertos.

El beneficio directo para la comunidad educativa es permitir que directivos y docentes pasen de un análisis post mortem a una gestión proactiva. Con la capacidad de anticipar resultados, las instituciones vulnerables pueden diseñar intervenciones pedagógicas focalizadas y asignar recursos estratégicamente, rompiendo ciclos de bajo rendimiento que perpetúan la pobreza (Burbano, 2021). A nivel de política, el modelo permite "focalizar acciones de mejora" (Equipo Técnico Dirección de Evaluación, 2024), se alinea con las recomendaciones de integrar datos para la prevención (Arias Ortiz et al., 2024) y Convertir los microdatos agregados del ICFES en señales predictivas que permitan a las secretarías de educación priorizar intervenciones en colegios con alto riesgo de bajo rendimiento. (Contreras et al., 2020).

# Objetivos

## Objetivo General

Desarrollar un modelo de aprendizaje supervisado que prediga el puntaje global promedio de Saber 11 por establecimiento educativo en Colombia, utilizando datos abiertos del ICFES como herramienta de apoyo a la toma de decisiones pedagógicas con enfoque en equidad.

## Objetivos Específicos

Estructurar un conjunto de datos a nivel de establecimiento educativo, integrando variables socioeconómicas, institucionales y de rendimiento histórico provenientes de fuentes públicas del ICFES.

Implementar dos modelos de aprendizaje supervisado (uno lineal como base de comparación y otro de ensamble de mayor complejidad) capaces de estimar el puntaje global Saber 11 a partir del conjunto de datos procesado.

Comparar el rendimiento predictivo de los dos modelos implementados utilizando métricas de error (RMSE, MAE) y bondad de ajuste ($R^2$) sobre el conjunto de prueba para seleccionar el modelo de mayor precisión.

Evaluar el desempeño diferencial del modelo seleccionado sobre subgrupos de establecimientos para verificar su equidad predictiva frente a las inequidades estructurales identificadas en el diagnóstico.

## Marco Teórico

El presente marco teórico establece los fundamentos conceptuales y matemáticos que sustentan el desarrollo del modelo predictivo propuesto. Se estructura en dos dimensiones: la perspectiva educativa, que justifica la selección de variables y el enfoque de equidad; y la perspectiva de ciencia de datos, que fundamenta la elección de la metodología CRISP-DM, los algoritmos de aprendizaje supervisado y las métricas de evaluación.

### La Calidad Educativa desde la Función de Producción

El análisis cuantitativo del desempeño escolar se fundamenta teóricamente en la economía de la educación bajo el modelo de la Función de Producción Educativa. Propuesto originalmente por Hanushek (1986), este enfoque postula que el resultado académico ($Y$), medido a través de pruebas estandarizadas como Saber 11, no es un evento aleatorio, sino el producto de la interacción de múltiples insumos o *inputs* ($X$). Matemáticamente, esto se expresa como $Y_i = f(F_i, E_i, C_i) + \epsilon_i$, donde el rendimiento del establecimiento $i$ depende de factores familiares ($F$), escolares ($E$) y contextuales ($C$).

Esta teoría es la base para la etapa de ingeniería de características del proyecto. Investigaciones aplicadas en Colombia (Laboratorio de Economía de la Educación (LEE), 2025; Artamonova et al., 2024) han demostrado que variables del vector $F$ (como el nivel educativo de los padres) y del vector $C$ (como la ubicación rural/urbana) tienen coeficientes de determinación más altos que los insumos puramente escolares. Por tanto, la agregación de datos a nivel institucional propuesta en la metodología no debe limitarse a promedios simples, sino que debe capturar la estructura de estos insumos socioeconómicos para maximizar la capacidad explicativa del modelo.

## Minería de Datos Educativa (EDM) y Metodología CRISP-DM

La Minería de Datos Educativa (Educational Data Mining - EDM) se define como la disciplina encargada de desarrollar métodos para explorar los tipos únicos de datos provenientes de entornos educativos y utilizarlos para comprender mejor a los estudiantes y los entornos en los que aprenden (Romero & Ventura, 2010).

A diferencia de la estadística inferencial clásica, cuyo objetivo es probar hipótesis sobre parámetros poblacionales, la EDM aplicada en este proyecto busca patrones latentes con fines predictivos. Para operacionalizar este enfoque, se adopta el estándar CRISP-DM (Cross-Industry Standard Process for Data Mining). La elección de este marco no es arbitraria; su naturaleza cíclica (Chapman et al., 2000) es fundamental para el contexto educativo, donde el "despliegue" de un modelo no es el fin, sino el inicio de una nueva fase de recolección de datos post-intervención pedagógica, permitiendo el refinamiento continuo de las variables predictoras.

## Fundamentos de Aprendizaje Supervisado para Regresión

Dado que el objetivo del proyecto es estimar el "Puntaje Global Promedio" (una variable continua $\hat{y} \in \mathbb{R}$), el problema se enmarca en el aprendizaje supervisado de regresión. El objetivo central es aprender una función de mapeo $f: X \rightarrow Y$ que minimice una función de pérdida $L(y, \hat{y})$ sobre el conjunto de datos de los colegios.

### *Modelos Lineales y Regularización (El Modelo Base)*

La regresión lineal múltiple asume que la esperanza condicional de $Y$ es una función lineal de los parámetros. Sin embargo, en datos educativos, es común encontrar multicolinealidad (ej. alta correlación entre estrato socioeconómico y nivel educativo de los padres). El uso de mínimos cuadrados ordinarios (OLS) en presencia de multicolinealidad puede inflar la varianza de los estimadores.

Para mitigar esto y robustecer el "modelo base" propuesto en la metodología, se recurre a técnicas de Regularización (James et al., 2013):

- *Regresión Ridge ($L_2$):* Agrega un término de penalización proporcional al cuadrado de la magnitud de los coeficientes. Esto contrae los coeficientes de variables correlacionadas hacia cero, pero no los elimina, manteniendo la información de todos los predictores contextuales.
- *Regresión Lasso ($L_1$):* Penaliza el valor absoluto de los coeficientes, forzando a que algunos sean exactamente cero. Esto permite realizar una selección de características automática, útil para identificar qué variables del MEN son redundantes.

***Modelos de Ensamble: Bagging y Boosting (El Modelo Complejo)***

Para superar las limitaciones de linealidad y capturar las interacciones complejas diagnosticadas en el planteamiento del problema (ej. el impacto de la inversión tecnológica puede ser no lineal y depender de la zona rural/urbana), se fundamenta el uso de métodos de ensamble basados en árboles de decisión.

- *Random Forest (Bagging):* Este algoritmo construye múltiples árboles de decisión durante el entrenamiento y genera la predicción final promediando los resultados de los árboles individuales. Teóricamente, su principal ventaja para este proyecto es la reducción de la varianza (Hastie et al., 2009). Dado que los datos de colegios pequeños pueden ser ruidosos, el *bagging* evita que el modelo se sobreajuste a las particularidades de una muestra específica, ofreciendo una predicción más estable.
- *Gradient Boosting (Boosting):* A diferencia de Random Forest, que construye árboles independientes, el *Boosting* (implementado en algoritmos como XGBoost) construye el modelo de forma secuencial. Cada nuevo árbol intenta corregir los errores (residuales) cometidos

por los árboles anteriores. Matemáticamente, es un algoritmo de optimización de descenso de gradiente en el espacio de funciones. Para la predicción de Saber 11, esto es fundamental porque permite que el modelo se "especialice" en los casos difíciles (colegios con comportamientos atípicos), reduciendo el sesgo de la predicción final y ofreciendo, generalmente, el estado del arte en precisión para datos tabulares (Chen & Guestrin, 2016).

**Teoría de Evaluación y Métricas de Desempeño**

La validación del modelo requiere métricas alineadas con el propósito de "gestión proactiva".

- *RMSE (Raíz del Error Cuadrático Medio):* $\sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$. Al elevar los errores al cuadrado antes de promediar, el RMSE penaliza desproporcionadamente los grandes errores. En el contexto de política pública, esto es deseable: subestimar o sobreestimar drásticamente el puntaje de un colegio vulnerable tiene un costo social alto. El modelo debe minimizar estos "grandes fallos".

- *MAE (Error Absoluto Medio):* $\frac{1}{n}\sum|y_i - \hat{y}_i|$. Proporciona una interpretación directa en las unidades de la prueba (puntos Saber 11), facilitando la comunicación de la incertidumbre del modelo a los directivos docentes no expertos en estadística.

- *Coeficiente de Determinación ($R^2$):* Mide la proporción de la varianza del puntaje global que es predecible a partir de las variables contextuales e institucionales, indicando la bondad de ajuste global del modelo.

## Metodología

Para el desarrollo de este proyecto aplicado, se adoptará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Se selecciona este marco de trabajo por ser un estándar de la industria, robusto y de naturaleza cíclica, lo cual permite refinar el proceso en cada etapa (Chapman et al., 2000). Su enfoque estructurado, que va desde la comprensión del problema hasta el despliegue de la solución, es ideal para proyectos aplicados que buscan transformar datos brutos en conocimiento accionable. El proyecto adaptará las seis fases de CRISP-DM al contexto específico del análisis del rendimiento en las pruebas Saber 11.

### Comprensión del Problema (Fase 1)

Esta fase inicial se centra en definir los objetivos del proyecto desde la perspectiva educativa. El problema es la carencia de herramientas proactivas en las instituciones vulnerables para anticipar su rendimiento. El objetivo técnico central es, por lo tanto, desarrollar un modelo de regresión supervisada que prediga el puntaje global promedio por establecimiento educativo. El éxito del modelo se definirá por su capacidad de generar predicciones precisas que sirvan como insumo para la toma de decisiones pedagógicas tempranas.

### Comprensión de los Datos (Fase 2)

En esta etapa se identificarán, obtendrán y explorarán los datos disponibles en el repositorio del ICFES. Las fuentes primarias serán los repositorios de Datos Abiertos del Gobierno de Colombia, específicamente: los microdatos históricos de Saber 11 (ICFES), que contienen el rendimiento individual y variables socioeconómicas. Se realizará un Análisis Exploratorio de Datos (EDA) para evaluar la calidad, identificar valores atípicos, entender la distribución de las variables y formular hipótesis iniciales sobre los factores más influyentes en el puntaje.

**Preparación de los Datos (Fase 3)**

1. *Limpieza y Transformación:* Se tratarán los valores nulos (missing values) mediante técnicas de imputación o eliminación, y se convertirán variables categóricas (ej. JORNADA) a formato numérico (ej. *One-Hot Encoding*).

2. *Integración y Agregación:* Este es el paso fundamental del proyecto. Dado que el objetivo es predecir a nivel de *establecimiento*, los microdatos a nivel de *estudiante* (ICFES) se agregarán.

3. *Ingeniería de Características:* En el proceso de agregación, se crearán las variables predictoras (features) para cada colegio, tales como: el puntaje global promedio de años anteriores, el porcentaje de estudiantes con acceso a internet, el nivel educativo promedio de los padres, la distribución de estratos socioeconómicos, y la tasa de estudiantes extra-edad.

4. *Selección y División:* Se seleccionarán las características más relevantes (para evitar multicolinealidad y reducir el ruido) y el conjunto de datos final se dividirá en subconjuntos de entrenamiento (training) y prueba (testing), garantizando una evaluación objetiva del modelo.

**Modelado (Fase 4)**

En esta fase se construirán y compararán al menos dos técnicas de aprendizaje automático apropiadas para la tarea de regresión, cumpliendo así con los objetivos específicos. Se seleccionarán y entrenarán diferentes tipos de modelos para identificar cuál ofrece la mejor capacidad predictiva. Se buscará comparar un modelo base, conocido por su interpretabilidad, con un modelo más complejo, conocido por su alto rendimiento y capacidad para capturar relaciones no lineales en los datos.

**Evaluación (Fase 5)**

La evaluación se realizará sobre el conjunto de prueba (datos que el modelo no ha visto). Se utilizarán métricas estándar de regresión para cuantificar el rendimiento:

- *Raíz del Error Cuadrático Medio (RMSE) y Error Absoluto Medio (MAE):* Indican el error promedio de la predicción en puntos de la prueba Saber 11.
- *Coeficiente de Determinación ($R^2$):* Mide el porcentaje de la varianza del puntaje global que es explicado por el modelo.

Se seleccionará el modelo final que ofrezca el mejor balance entre precisión (bajo RMSE/MAE) y utilidad práctica.

**Despliegue y Consideraciones (Fase 6)**

Dado el carácter de proyecto aplicado de una especialización, la fase de despliegue consistirá en la entrega del notebook (código fuente) debidamente documentado y reproducible, que incluye el pipeline completo de preprocesamiento, entrenamiento y evaluación, permitiendo que el modelo pueda ser reentrenado o adaptado por cualquier institución con acceso a los datos abiertos del ICFES. Este resultado constituye la base técnica para futuras implementaciones de sistemas de alerta temprana institucional.

Por otro lado, se mantendrán estrictas consideraciones éticas: el análisis se realizará a nivel agregado (establecimiento) para proteger la privacidad individual. El propósito del modelo es ser una herramienta de diagnóstico interno para la equidad y la mejora continua, no una herramienta para la estigmatización o la creación de *rankings* públicos.

# Resultados

El presente capítulo expone los resultados obtenidos a lo largo de las cinco fases analíticas del proyecto, siguiendo el orden establecido por la metodología CRISP-DM. Los hallazgos se presentan en cuatro bloques: (i) la caracterización del conjunto de datos integrado a nivel de establecimiento educativo; (ii) el análisis exploratorio y estadístico de las variables contextuales y su relación con el puntaje global Saber 11; (iii) el proceso de selección y evaluación comparativa de los modelos predictivos; y (iv) el análisis de equidad del modelo sobre subgrupos educativos vulnerables. Así, estos resultados permiten dar respuesta a la pregunta de investigación planteada y verificar el cumplimiento de los cuatro objetivos específicos del proyecto.

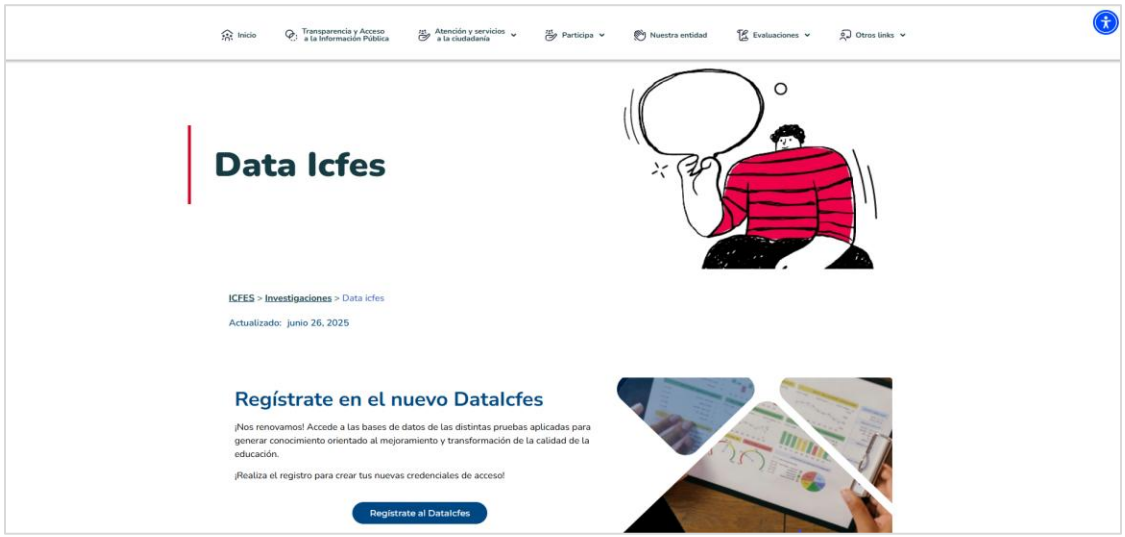
## Fuente y Recolección de los Datos

Los datos utilizados en este estudio fueron obtenidos a partir del portal oficial del Instituto Colombiano para la Evaluación de la Educación (ICFES), específicamente en su sección de datos abiertos disponible en: https://www.icfes.gov.co/investigaciones/data-icfes/. A través de este portal, el ICFES pone a disposición del público diversas bases de datos asociadas a sus evaluaciones estandarizadas. El acceso a dicha información requiere un proceso de registro previo, tras el cual es posible descargar los conjuntos de datos en formatos estructurados.

Para el desarrollo del presente proyecto, se accedió a la sección correspondiente a Saber 11 – Bases Nacionales, desde donde se descargaron los archivos que contienen los resultados individuales de los estudiantes a nivel nacional. En particular, se seleccionaron los periodos comprendidos entre los años 2021 y 2024, incluyendo ambos calendarios académicos (A y B), con el fin de contar con una muestra reciente, amplia y representativa del sistema educativo colombiano.

**Figura 1**

*Página de Registro para el Acceso a los Datos del ICFES*




La Figura 1 muestra la página oficial del ICFES, a través de la cual se puede acceder a los datos abiertos. El acceso a esta información requiere el diligenciamiento de un formulario de registro en línea.

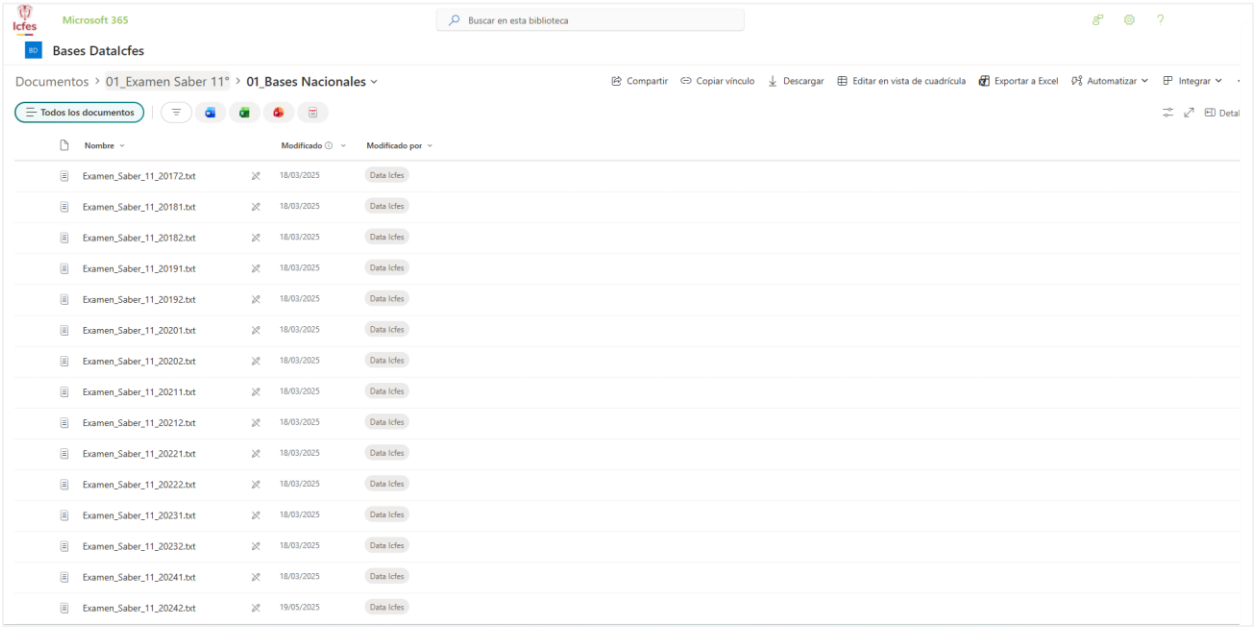
**Figura 2**

*Página de Acceso a las Diferentes Bases de Datos del ICFES*

La delimitación temporal del conjunto de datos responde a criterios de consistencia y homogeneidad. En primer lugar, los años seleccionados corresponden a un periodo posterior a las principales disrupciones generadas por la pandemia de COVID-19, lo cual permite reducir posibles sesgos asociados a cambios atípicos en las condiciones de enseñanza y evaluación. En segundo lugar, a partir de este periodo se observa una mayor estabilidad en la estructura de las bases de datos y en las variables reportadas por el ICFES, lo que facilita su integración y comparabilidad en el tiempo. De esta manera, se busca garantizar que el modelo predictivo se entrene sobre información coherente y metodológicamente consistente.

**Figura 3**

*Listado de Bases de Datos del Examen Saber 11 para Diferentes Años y por Calendario Académico*




## Construcción del Conjunto de Datos Integrado

El primer resultado tangible del proyecto consiste en la consolidación de un dataset

maestro que transforma microdatos a nivel de estudiante en observaciones a nivel de establecimiento educativo, dando cumplimiento al primer objetivo específico. La Tabla 1 resume las dimensiones del proceso de integración.

**Tabla 1**

*Resumen del Proceso de Integración y Dimensionamiento del Dataset*

| Dimensión | Valor |
|---|---|
| Archivos fuente procesados | 8 (4 años × 2 calendarios) |
| Periodos cubiertos | 2021-1, 2021-2, 2022-1, 2022-2, 2023-1, 2023-2, 2024-1, 2024-2 |
| Registros de estudiantes (microdatos) | 2.640.263 |
| Establecimientos educativos únicos | 16.358 |
| Variables predictoras finales | 18 (11 numéricas + 7 categóricas) |

Los microdatos del ICFES presentaron un patrón de valores faltantes sistemático, no aleatorio. Las variables institucionales del colegio registraron un 13.88% de nulos, correspondientes a registros del calendario B de 2021 con campos incompletos en la fuente original. Las variables del cuestionario socioeconómico exhibieron entre 16.15% y 22.17% de valores ausentes, atribuibles a estudiantes que no completaron dicho formulario, fenómeno documentado y recurrente en los microdatos del ICFES. Este patrón fue gestionado mediante imputación por mediana en el pipeline de preprocesamiento, estrategia robusta ante distribuciones asimétricas.

La variable de desempeño histórico del establecimiento (punt_global_lag1), creada mediante un desplazamiento temporal de un período sobre el puntaje promedio del propio colegio, quedó disponible para 40.688 de las 57.046 observaciones (71.3%). Los 16.358 registros sin lag corresponden al primer año de aparición de cada colegio en el dataset, para los cuales este

predictor es imputado por la mediana en el pipeline. La pertinencia de esta variable quedó respaldada estadísticamente: su correlación de Spearman con el puntaje actual es $\rho = 0.8894$ ($p \approx 0$), la más alta de todos los predictores incluidos en el modelo, evidenciando la marcada inercia institucional del sistema educativo colombiano.

## Análisis Exploratorio y Caracterización del Target

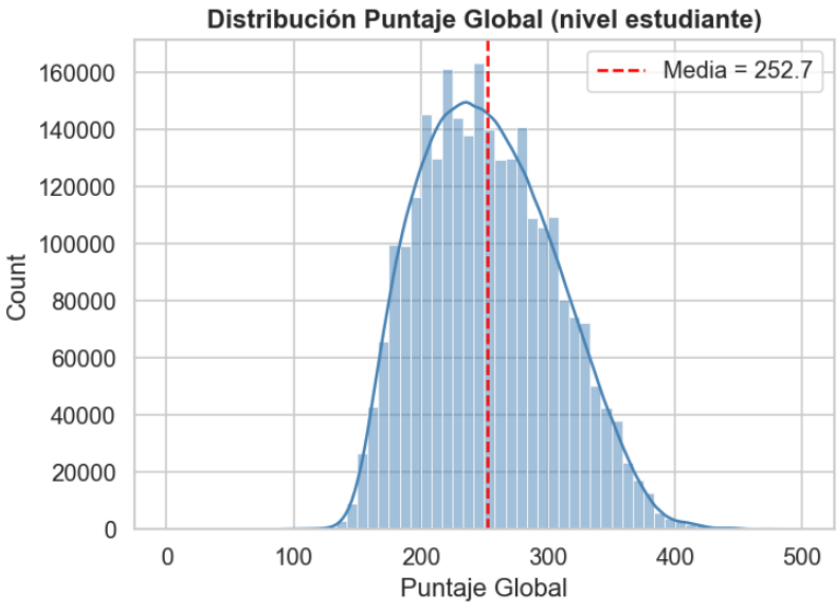
### *Distribución del Puntaje Global*

El puntaje global Saber 11 a nivel de estudiante presentó una distribución con media de 252.66 puntos (DE = 53.18), percentil 25 en 212 puntos, mediana de 249 puntos y percentil 75 en 290 puntos, sobre una escala de 0 a 500. Los indicadores de forma reportaron una asimetría positiva leve (0.30) y curtosis negativa (−0.43), indicando una distribución platicúrtica con colas más ligeras que una distribución normal. La aplicación del test de Kolmogorov-Smirnov sobre una muestra de 5.000 observaciones estandarizadas (práctica recomendada para muestras masivas donde cualquier desviación mínima resulta estadísticamente significativa) arrojó $D = 0.0401$ con $p = 2.05 \times 10^{-7}$, rechazando la hipótesis nula de normalidad al nivel $\alpha = 0.05$. Este resultado fundamenta el uso de pruebas no paramétricas en los análisis de brechas subsecuentes.

El histograma con curva KDE (Figura 4) evidencia una distribución unimodal con concentración de puntajes entre 150 y 350 puntos, donde la mayor frecuencia se sitúa alrededor de la media de 252.7 puntos. La asimetría positiva leve (0.30) es visible en la cola derecha más pronunciada que la izquierda, indicando que los colegios de alto desempeño son menos frecuentes que los de desempeño bajo o medio. La presencia de valores en los extremos de la escala (cercanos a 0 y a 500) corresponde a casos atípicos legítimos que reflejan la heterogeneidad real del sistema educativo colombiano y no errores de medición, por lo que se conservaron en el análisis.

**Figura 4**

*Histograma de la Distribución del Puntaje Global*

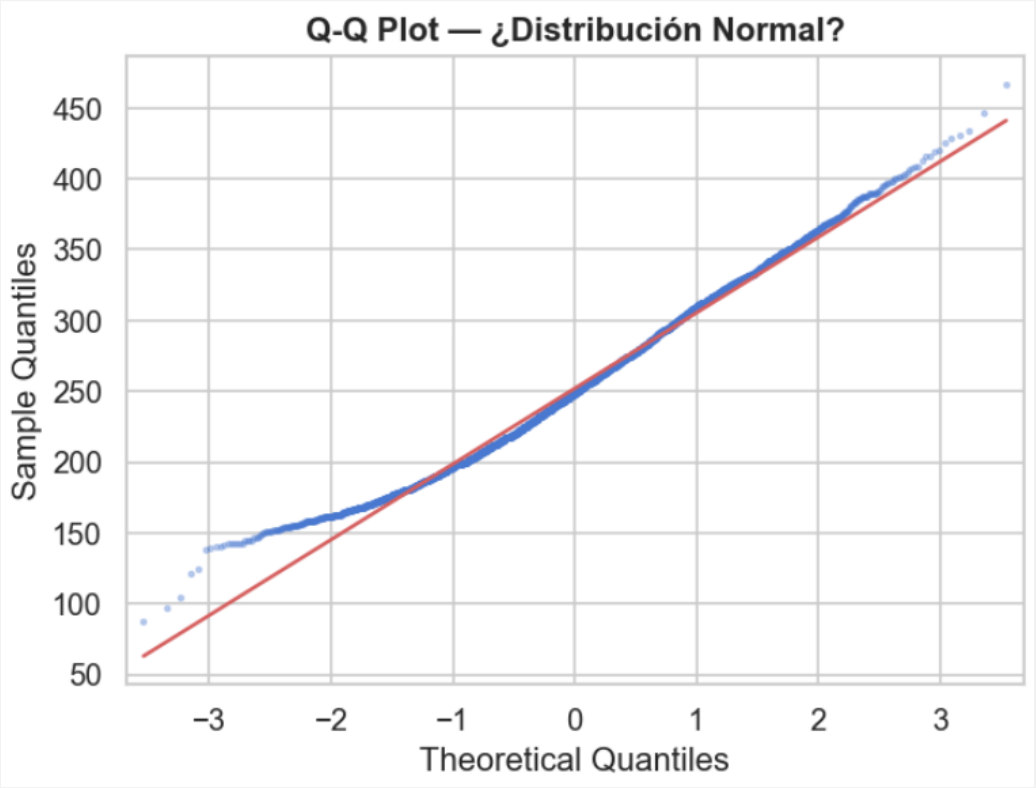



El gráfico Q-Q (Figura 5) confirma visualmente el rechazo de normalidad: los cuantiles empíricos siguen de cerca la línea teórica en el rango central (puntajes entre 150 y 350), pero se desvían sistemáticamente en ambas colas, con una curvatura en forma de S característica de distribuciones con curtosis negativa. Esta desviación en los extremos es coherente con el estadístico K-S reportado ($D = 0.0401$, $p = 2.05 \times 10^{-7}$) y fundamenta el uso de pruebas no paramétricas en el análisis de brechas.

Aunque el comportamiento de los datos en la región central evidencia un ajuste adecuado al modelo teórico, las discrepancias observadas en los cuantiles extremos sugieren la presencia de valores con menor frecuencia de la esperada bajo una distribución normal. En consecuencia, el supuesto de normalidad no se cumple de manera global, por lo que resulta metodológicamente apropiado emplear procedimientos estadísticos robustos que no dependan de este supuesto para garantizar la validez de las inferencias.

**Figura 5**

*Gráfico Q-Q Plot*

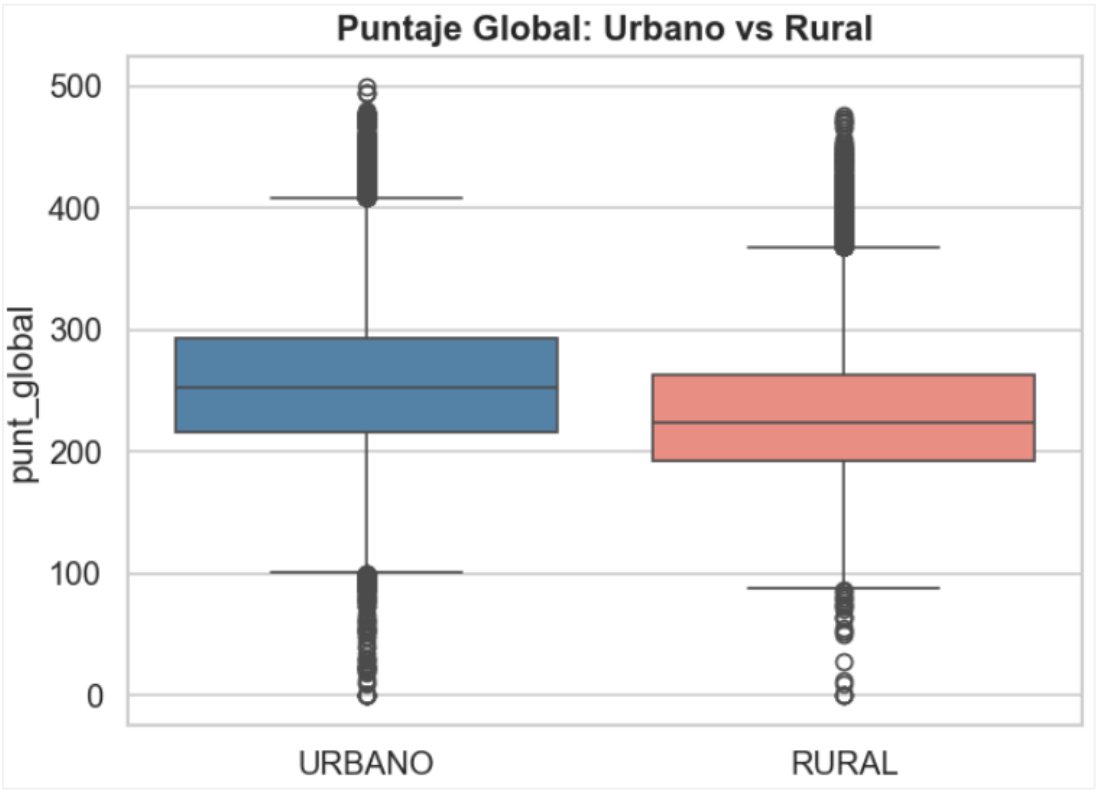



El boxplot comparativo urbano-rural (Figura 6) revela diferencias estructurales más allá de las medias: los colegios urbanos no solo presentan una mediana superior (aproximadamente 255 puntos frente a 230 en rurales), sino también una caja más estrecha, indicando menor dispersión interna.

Adicionalmente, la asimetría observada en la distribución de los puntajes y la presencia de valores atípicos en ambos grupos sugieren que la variabilidad del desempeño no es homogénea entre los contextos urbano y rural. Estas diferencias descriptivas constituyen evidencia preliminar de un posible efecto del contexto geográfico sobre el rendimiento académico, cuya magnitud debe corroborarse mediante pruebas estadísticas apropiadas.

**Figura 6**

*Boxplot Comparativo del Puntaje Global Según Tipo de Ubicación Institucional (Urbana vs. Rural)*

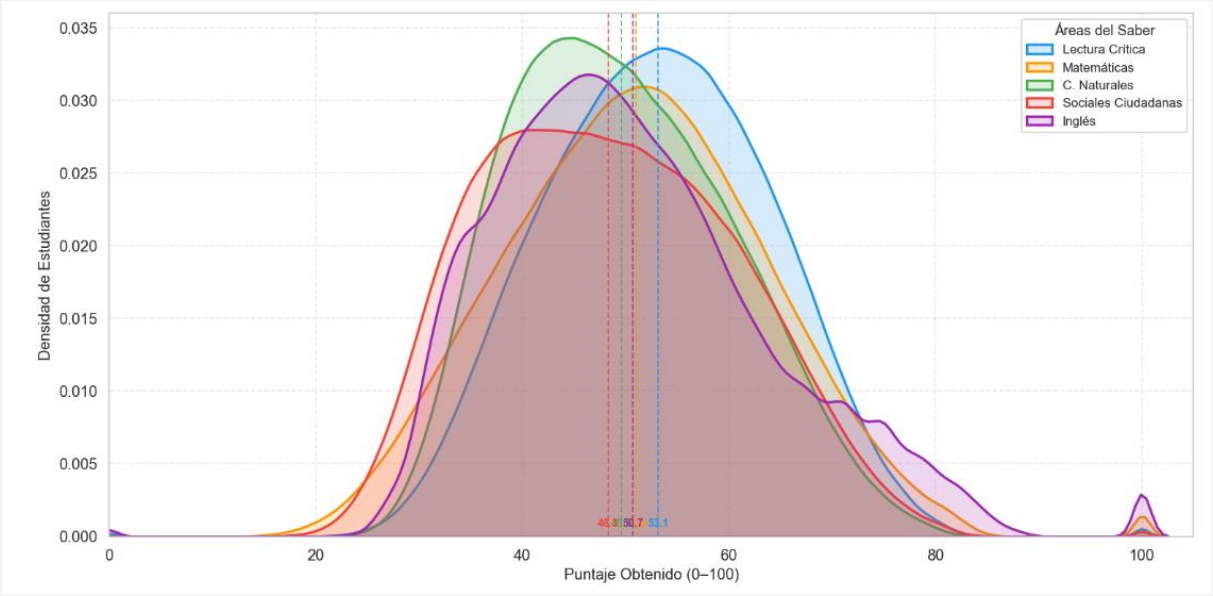



***Distribución del Puntaje por Área de Conocimiento***

Aunque los puntajes por área no son variables predictoras del modelo, su inclusión generaría fuga de información trivial hacia el puntaje global, su distribución exploratoria permite caracterizar el perfil académico del sistema educativo colombiano en el período 2021–2024 y contextualizar la variable objetivo. La Figura 7 presenta las curvas de densidad (KDE) de los cinco componentes de la prueba Saber 11 superpuestas en una misma escala, junto con las medias individuales de cada área representadas mediante líneas discontinuas verticales. La Tabla 2 consolida las estadísticas descriptivas correspondientes.

**Figura 7**

*Comparación de la Distribución de Puntajes por Área de Conocimiento*




El análisis comparativo de las distribuciones por área revela un perfil homogéneo en términos de nivel promedio, con medias que oscilan entre 48.32 puntos (Sociales Ciudadanas) y 53.13 puntos (Lectura Crítica), lo que indica que ninguna área presenta un nivel de dificultad radicalmente distinto a las demás sobre la escala común de 0 a 100. Sin embargo, las diferencias de forma entre las distribuciones son informativas.

**Tabla 2**

*Estadísticas Descriptivas del Puntaje por Área de Conocimiento Saber 11 (2021–2024).*

| Área | Media | Mediana | DE | Mín | Máx | Asimetría | Curtosis |
|---|---|---|---|---|---|---|---|
| Lectura Crítica | 53.13 | 53.00 | 10.89 | 0 | 100 | 0.021 | −0.163 |
| Matemáticas | 51.04 | 51.00 | 12.64 | 0 | 100 | 0.141 | 0.002 |
| C. Naturales | 49.62 | 49.00 | 10.70 | 0 | 100 | 0.337 | −0.239 |
| Sociales Ciudadanas | 48.32 | 48.00 | 12.22 | 0 | 100 | 0.231 | −0.517 |
| Inglés | 50.71 | 49.00 | 13.36 | 0 | 100 | 0.656 | 0.522 |

Las áreas de Lectura Crítica y Ciencias Naturales presentan las curvas más estrechas y de mayor densidad en el pico (DE de 10.89 y 10.70 respectivamente), lo que refleja mayor homogeneidad en el desempeño de los estudiantes: la mayoría se concentra en torno a su media con poca dispersión. En contraste, Inglés y Matemáticas exhiben las desviaciones estándar más altas (13.36 y 12.64 respectivamente), con curvas más anchas y achatadas, indicando mayor variabilidad en el desempeño entre estudiantes. En Inglés esto es especialmente visible en la presencia de un pequeño bulto secundario cerca del puntaje 100, correspondiente a estudiantes de alto dominio lingüístico, típicamente de colegios bilingües, que eleva la asimetría de esta área a 0.656 (la más alta entre todas las áreas) y genera una curtosis positiva (0.522), la única entre las cinco áreas.

***Brechas Educativas***

Los resultados de las pruebas no paramétricas confirman, con contundencia estadística, las tres brechas estructurales sobre las que se fundamenta la pertinencia de este proyecto. La Tabla 3 consolida los hallazgos.

**Tabla 3**

*Análisis de Brechas Educativas Mediante Pruebas Mann-Whitney U y Kruskal-Wallis*

| Brecha analizada | Grupo A (Media) | Grupo B (Media) | Diferencia | Estadístico | p-valor |
|---|---|---|---|---|---|
| Urbano vs Rural | Urbano: 255.95 | Rural: 231.62 | 24.32 pts | $U = 4.63 \times 10^{11}$ | $\approx 0$ |
| No Oficial vs Oficial | No Oficial: 274.47 | Oficial: 244.28 | 30.20 pts | $U = 3.36 \times 10^{11}$ | $\approx 0$ |
| Entre departamentos | — | — | — | $H = 157.183$ | $\approx 0$ |

La brecha urbano-rural de 24.32 puntos es coherente con el valor de 26.1 puntos reportado por el Laboratorio de Economía de la Educación (LEE, 2025) para el año 2024 de

manera aislada: la diferencia de 1.78 puntos se explica porque el presente análisis promedia los períodos 2021–2024, incluyendo años de recuperación post-pandemia con brechas históricamente menores. En el mismo sentido, la diferencia entre colegios no oficiales y oficiales (30.20 puntos) supera marginalmente los 27.5 puntos reportados por el LEE (2024) para 2023, dado que el período 2021–2022 exhibió brechas más amplias por el impacto diferencial del cierre de instituciones durante la pandemia. El estadístico de Kruskal-Wallis ($H = 157.183$) confirma que las diferencias entre los 33 departamentos son altamente significativas, justificando empíricamente la inclusión del departamento como variable predictora con 33 niveles en el modelo.

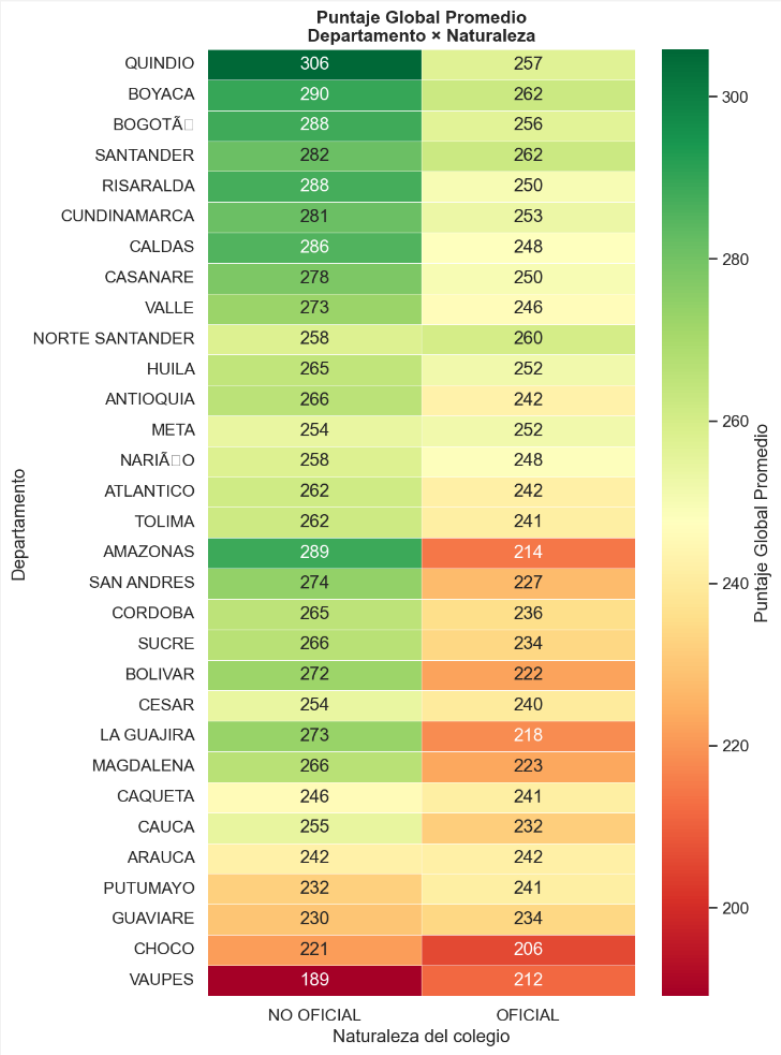
La Figura 8 presenta el puntaje global promedio de las instituciones educativas según el departamento y la naturaleza del establecimiento (oficial y no oficial), permitiendo identificar patrones territoriales y diferencias asociadas al tipo de administración escolar. En términos generales, los establecimientos no oficiales registran promedios superiores en la mayoría de los departamentos, lo que evidencia una tendencia consistente en favor de este sector educativo. No obstante, la magnitud de esta diferencia no es uniforme, lo que sugiere que el contexto regional desempeña un papel relevante en los resultados obtenidos por los estudiantes.

Se observa que departamentos como Quindío, Boyacá, Bogotá D.C., Risaralda, Caldas y Cundinamarca presentan los promedios más altos en instituciones no oficiales, superando los 280 puntos e incluso alcanzando valores cercanos a los 306 puntos en el caso de Quindío. En contraste, los establecimientos oficiales de estos mismos territorios, aunque mantienen desempeños relativamente altos frente al promedio nacional, exhiben resultados inferiores, con diferencias que oscilan entre 20 y 50 puntos aproximadamente. Este comportamiento evidencia

que las brechas entre ambos sectores tienden a ampliarse en departamentos con mayores niveles de desarrollo educativo.

**Figura 8**

*Puntaje Global Promedio por Departamento y Naturaleza del Establecimiento*




*Nota.* La escala cromática RdYlGn identifica los extremos del desempeño nacional

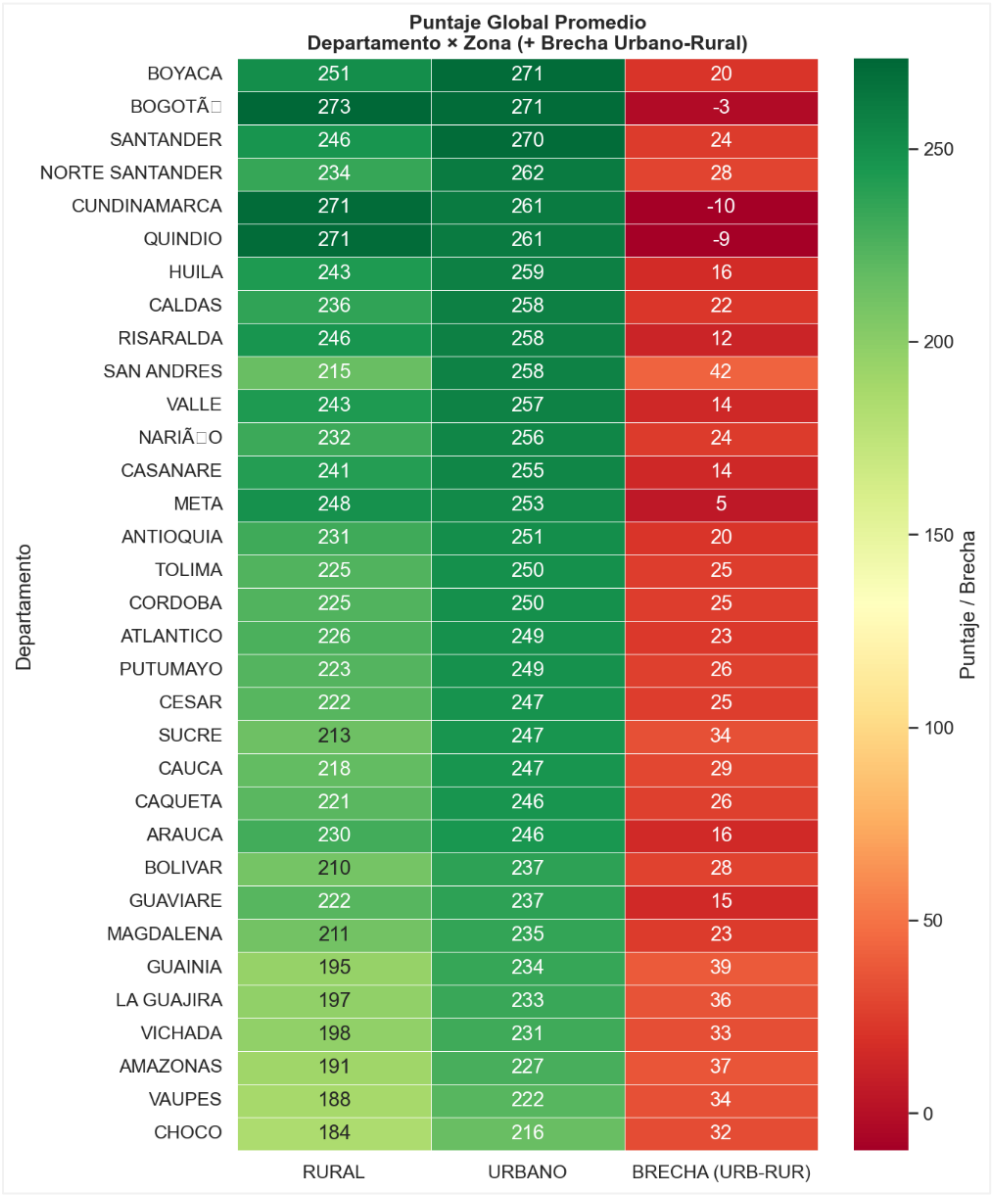
Por otra parte, en departamentos como Vaupés, Chocó, Guaviare y Putumayo se registran los promedios más bajos, independientemente de la naturaleza del establecimiento. Esta situación refleja las dificultades estructurales que enfrentan estos territorios, donde factores como

la dispersión geográfica, las limitaciones en infraestructura educativa, las condiciones socioeconómicas y el acceso desigual a recursos pedagógicos pueden influir negativamente en el rendimiento académico.

***Puntaje Global Según Variables Institucionales del Establecimiento***

**Figura 9**

*Puntaje Global Promedio por Departamento y Zona de Ubicación*

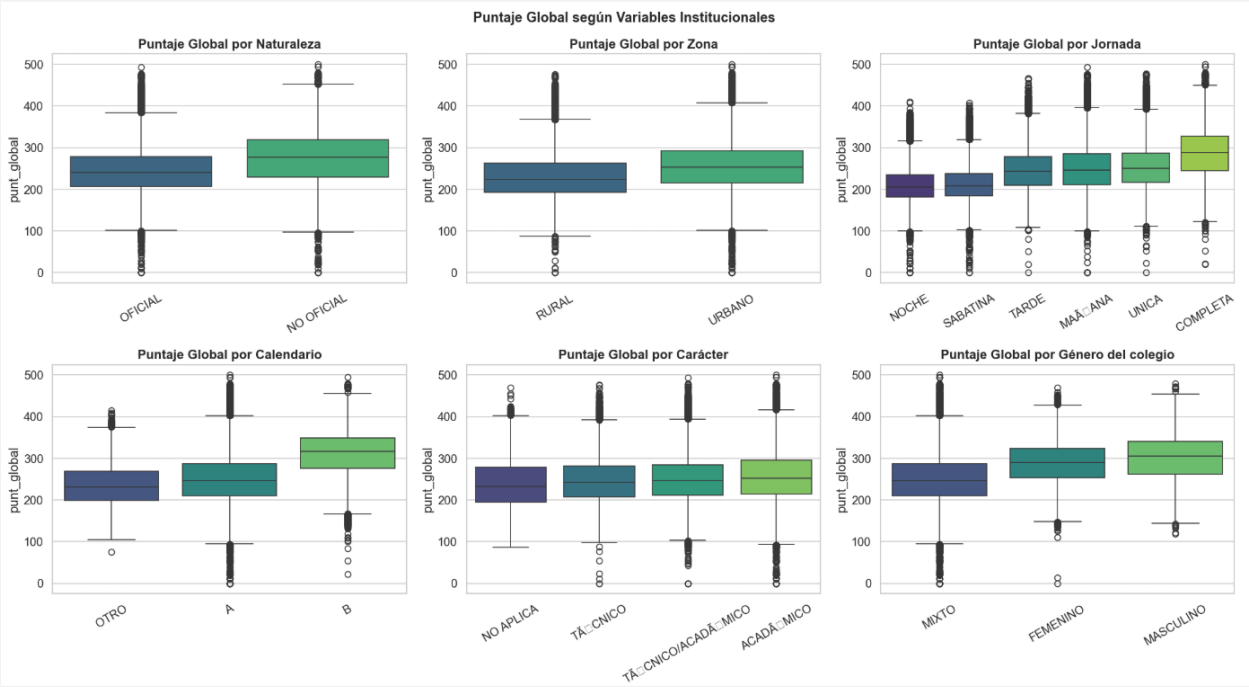



*Nota.* Se incluye la brecha urbano-rural por departamento.

La Figura 10 presenta la distribución del puntaje global según seis variables institucionales del establecimiento educativo. El análisis visual mediante boxplots, con categorías ordenadas por mediana ascendente dentro de cada panel, permite identificar cuáles de estas características exhiben mayor capacidad discriminativa sobre el desempeño institucional, previo al modelado formal.

**Figura 10**

*Distribución del Puntaje Global Saber 11 Según Naturaleza, Zona, Jornada, Calendario, Carácter y Género del Establecimiento Educativo*




En cuanto a la naturaleza del establecimiento, los colegios no oficiales presentan una mediana claramente superior a los oficiales y una caja más estrecha, lo que indica no solo mayor puntaje promedio sino también menor variabilidad interna. Los colegios oficiales, en contraste, muestran una dispersión considerablemente mayor y una concentración notable de valores atípicos en los extremos inferiores, reflejo de la heterogeneidad que caracteriza al sector público

colombiano: desde colegios con condiciones similares al sector privado hasta instituciones en zonas de alta vulnerabilidad.

La variable jornada escolar es la que exhibe la mayor capacidad discriminativa entre todas las variables institucionales analizadas. La jornada nocturna registra la mediana más baja, seguida por la sabatina; ambas modalidades atienden principalmente a estudiantes adultos con carga laboral que restringe el tiempo disponible para el estudio. En el extremo opuesto, la jornada completa presenta la mediana más alta, coherente con la mayor cantidad de horas pedagógicas disponibles. La jornada única ocupa una posición intermedia alta, mientras que las jornadas de mañana y tarde muestran distribuciones similares con medianas entre 240 y 260 puntos. La amplitud de los rangos intercuartílicos es relativamente homogénea entre jornadas de día, pero notablemente más estrecha en la nocturna, reflejando la homogeneidad del perfil del estudiante adulto trabajador.

El calendario B presenta una mediana notablemente superior al calendario A y a la categoría “Otro”. Esta diferencia obedece a que el calendario B agrupa principalmente colegios bilingües, internacionales y de alto nivel académico ubicados en grandes centros urbanos, cuyo perfil socioeconómico de su población es sistemáticamente más favorable.

En cuanto al carácter del establecimiento, los colegios académicos y técnico-académicos muestran distribuciones similares con medianas ligeramente superiores a los colegios técnicos puros, aunque con alta superposición entre grupos, lo que sugiere que esta variable tiene menor poder discriminativo individual que la jornada o la naturaleza. Finalmente, en cuanto al género del establecimiento, los colegios masculinos y femeninos presentan medianas superiores a los mixtos. No obstante, los establecimientos de género único tienden a ser instituciones privadas de

tradición académica, por lo que este efecto está probablemente mediado por la naturaleza y el perfil socioeconómico de su población estudiantil, más que por el género en sí mismo.

### *Correlaciones entre Variables Socioeconómicas y el Puntaje Global*

La Tabla 4 presenta las correlaciones de Spearman entre las variables socioeconómicas del cuestionario ICFES y el puntaje global, calculadas sobre los registros con información válida en cada variable. Se emplea la correlación de Spearman en lugar de Pearson dado el rechazo de normalidad verificado en la sección anterior y el carácter ordinal de variables como estrato o nivel educativo de los padres.

**Tabla 4**

*Correlaciones de Spearman entre Variables Socioeconómicas y el Puntaje Global Saber 11 (Nivel Estudiante)*

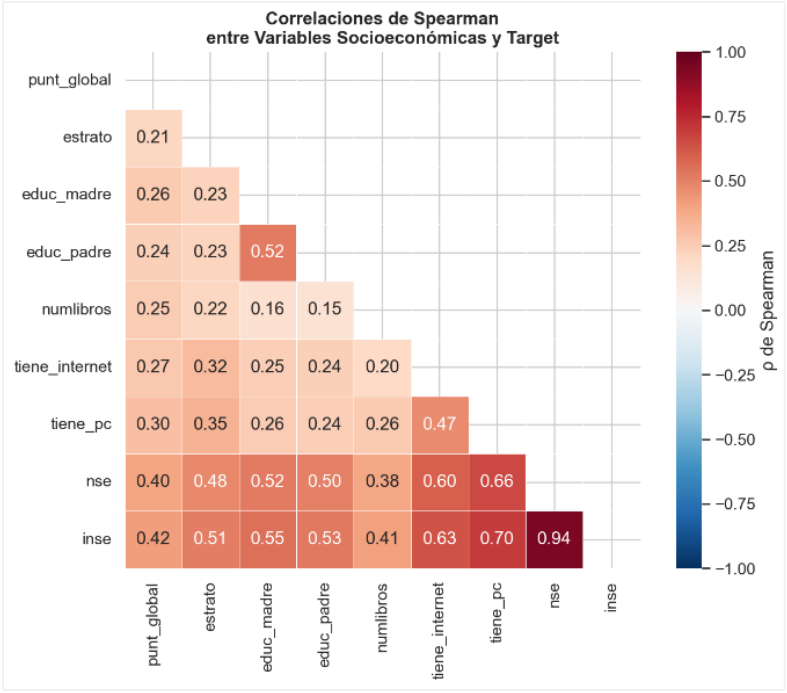
| Variable socioeconómica | ρ de Spearman | p-valor | n válidos |
|---|---|---|---|
| INSE individual (continuo) | 0.4189 | ≈ 0 | 2.139.724 |
| NSE individual (1–4) | 0.3955 | ≈ 0 | 2.139.724 |
| Tiene computador en casa | 0.3045 | ≈ 0 | 2.157.827 |
| Tiene internet en casa | 0.2714 | ≈ 0 | 2.106.720 |
| Educación de la madre | 0.2615 | ≈ 0 | 1.401.874 |
| Número de libros en casa | 0.2535 | ≈ 0 | 1.957.594 |
| Educación del padre | 0.2379 | ≈ 0 | 1.457.020 |
| Estrato de la vivienda | 0.2142 | ≈ 0 | 2.213.840 |

*Nota.* Todas las correlaciones son estadísticamente significativas ($\alpha = 0.05$).

El índice de nivel socioeconómico continuo (INSE) exhibe la correlación más alta (ρ = 0.419), seguido por el NSE discreto (ρ = 0.396). Destaca que la tenencia de computador en el hogar (ρ = 0.305) supera a la educación de los padres (ρ = 0.262 y 0.238), lo que es coherente con el contexto post-pandemia (2021–2024): el acceso a tecnología durante el período de clases virtuales constituyó un factor diferenciador inmediato del rendimiento académico, con un efecto que persiste en los años de recuperación. Este hallazgo es consistente con los resultados de Ballesteros-Alfonso y Gómez-Velasco (2022), quienes identificaron la conectividad como variable crítica durante y después de la pandemia.

**Figura 11**

*Matriz de Correlaciones de Spearman entre Variables Socioeconómicas y el Puntaje Global*



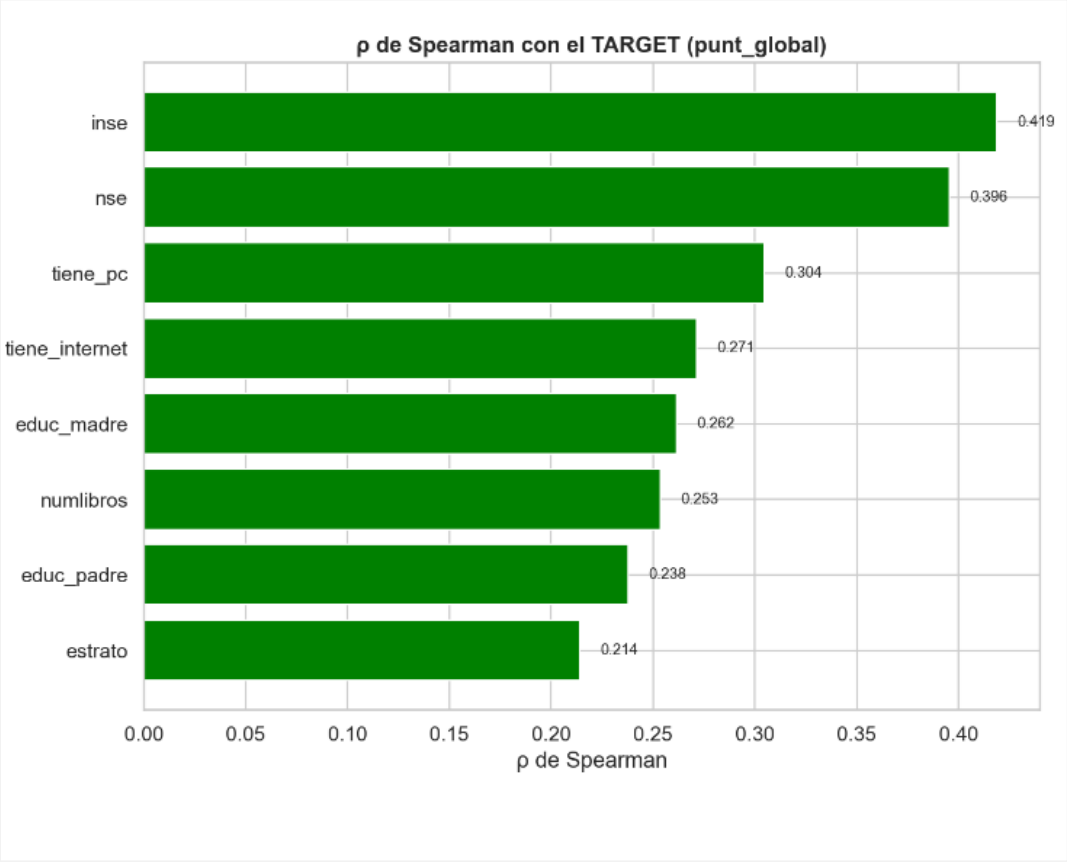

La Figura 11 muestra que las variables socioeconómicas mantienen correlaciones positivas de distinta magnitud con el puntaje global. Destacan el Índice Socioeconómico (INSE) y el Nivel Socioeconómico (NSE) por presentar los coeficientes más elevados, lo que sugiere

que las condiciones socioeconómicas del estudiante y su hogar se asocian con un mejor desempeño en las pruebas Saber 11.

Como se aprecia en la Figura 12, todas las variables socioeconómicas presentan coeficientes de correlación positivos con el puntaje global, aunque con magnitudes que varían entre bajas y moderadas. El INSE y el NSE destacan por registrar las asociaciones más altas, mientras que variables como el estrato y el nivel educativo del padre muestran relaciones más débiles, lo que evidencia que el desempeño académico está vinculado a múltiples dimensiones del contexto socioeconómico y no a un único factor.

**Figura 12**

*Matriz de Correlaciones de Spearman entre Variables Socioeconómicas y Correlaciones Individuales con el Target*

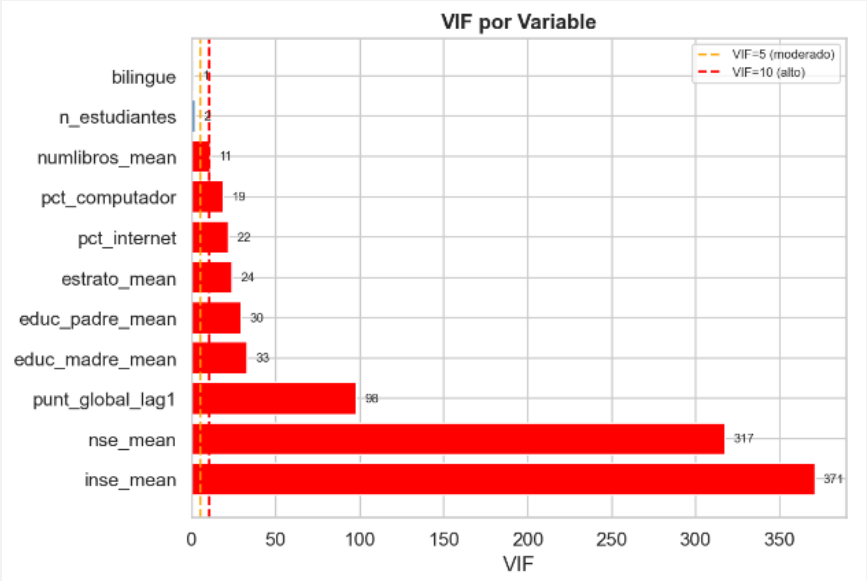


***Diagnóstico de Multicolinealidad***

Previo al modelado, se calculó el Factor de Inflación de la Varianza (VIF) para las once variables numéricas del conjunto de entrenamiento. Los resultados revelan multicolinealidad severa entre las variables de nivel socioeconómico: el INSE registra VIF = 370.82 y el NSE VIF = 317.30, lo que refleja que ambas variables comparten prácticamente el mismo espacio de información. La variable punt_global_lag1 también exhibe VIF alto (97.86), derivado de su correlación con la mayoría de las demás variables del perfil institucional. En total, nueve de las once variables numéricas superan el umbral de VIF = 10.

**Figura 13**

*Factor de Inflación de la Varianza (VIF) por Variable Numérica*



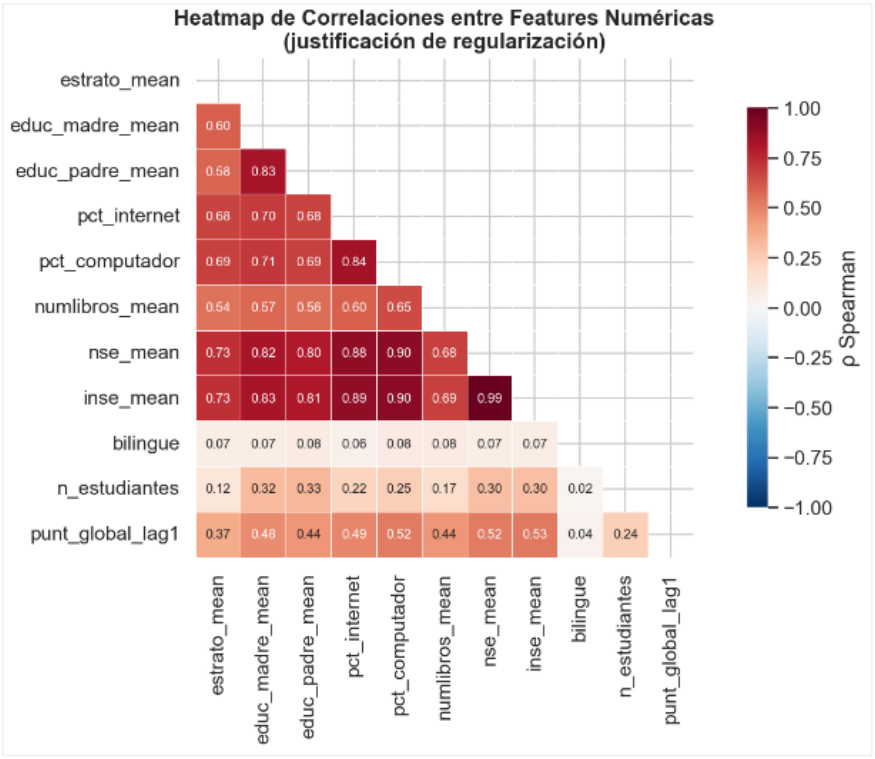

*Nota.* Las barras rojas identifican variables con VIF > 10

Este diagnóstico constituye la justificación estadística central para el uso de regularización en la familia de modelos lineales: la regresión por Mínimos Cuadrados Ordinarios (OLS) en presencia de VIF superiores a 10 produce estimadores con varianza inflada e

inestables, comprometiendo tanto la precisión predictiva como la interpretabilidad de los coeficientes. La regularización Ridge (L2) y Lasso (L1), al penalizar la magnitud de los coeficientes, mitigan este problema de forma directa.

**Figura 14**

*Heatmap de Correlaciones de Spearman entre Predictores*




## Modelado y Selección de los Mejores Modelos

### *Selección Dentro de la Familia Lineal*

Para la selección del modelo lineal se compararon Ridge (L2) y Lasso (L1) mediante validación cruzada de cinco particiones (CV-5) sobre el conjunto de entrenamiento 2021–2023. Ambos modelos se configuraron con búsqueda automática del parámetro de regularización óptimo $\alpha$ sobre una grilla de 80 valores en escala logarítmica $[10^{-3}, 10^{4}]$. La Tabla 5 presenta los resultados.

**Tabla 5**

*Comparación de Modelos Lineales Mediante RMSE en Validación Cruzada CV-5 sobre Entrenamiento 2021–2023*

| Modelo | CV-5 RMSE (media) | CV-5 RMSE (DE) | Seleccionado |
|---|---|---|---|
| Ridge (L2) | 18.309 | ± 0.098 | Sí |
| Lasso (L1) | 18.310 | ± 0.098 | No |

La diferencia entre ambos modelos es de 0.001 puntos de RMSE, con desviaciones estándar idénticas (±0.098), lo que indica equivalencia práctica. La selección de Ridge se fundamenta en que, ante igualdad estadística, la penalización L2 es teóricamente más apropiada cuando todas las variables son informativas, como sugieren las correlaciones de Spearman significativas de la Tabla 4, ya que Ridge contrae los coeficientes hacia cero sin eliminarlos, distribuyendo el peso entre variables colineales. Lasso, que fuerza coeficientes a cero exacto, no logró una selección de variables efectiva en este contexto, confirmando que todas las variables predictoras incluidas aportan información al modelo.

***Selección Dentro de la Familia de Ensamble***

Para la familia de ensamble se compararon Random Forest (Bagging) y Gradient Boosting (XGBoost con CUDA), ambos mediante búsqueda aleatoria de hiperparámetros (RandomizedSearchCV, 15 iteraciones, CV-5). La disponibilidad de GPU NVIDIA con CUDA (XGBoost versión 3.2.0) permitió ampliar el espacio de búsqueda del modelo de Gradient Boosting hasta 600 estimadores y siete hiperparámetros simultáneos, lo que en CPU habría requerido tiempo de cómputo prohibitivo para el contexto del proyecto.

Los resultados de esta comparación, junto con la configuración óptima obtenida para cada modelo, se presentan en la Tabla 6.

**Tabla 6**

*Comparación de Modelos de Ensamble Mediante RMSE en Validación Cruzada CV-5 sobre Entrenamiento 2021–2023*

| Modelo | CV-5 RMSE | Hiperparámetros óptimos | Seleccionado |
|---|---|---|---|
| Random Forest | 16.366 | n_estimators=200, max_depth=None, min_samples_leaf=2, max_features=0.5 | No |
| XGBoost CUDA | 16.057 | n_estimators=600, lr=0.05, max_depth=5, subsample=0.9, colsample=0.7, reg_λ=5 | Sí |

XGBoost CUDA obtuvo un RMSE-CV de 16.057, superando a Random Forest (16.366) por 0.309 puntos. Esta diferencia es consistente con la naturaleza de ambos algoritmos: el Gradient Boosting secuencial reduce el sesgo iterativamente al especializarse en los residuos de cada árbol anterior, mientras que el Random Forest paralelizado prioriza la reducción de varianza. Los hiperparámetros óptimos de XGBoost revelan un modelo con regularización L2 fuerte (reg_λ = 5) y submuestreo agresivo tanto de observaciones (subsample = 0.9) como de variables (colsample_bytree = 0.7), característica que indica la presencia de ruido en los datos de establecimiento que el modelo aprende a ignorar.

## Evaluación en el Conjunto de Prueba 2024

### *Métricas de Desempeño Predictivo*

Los dos modelos seleccionados, Ridge (L2) y XGBoost CUDA, fueron evaluados sobre el conjunto de prueba correspondiente al año 2024, compuesto por 14.431 establecimientos que ninguno de los modelos observó durante el entrenamiento. Esta estrategia de validación temporal, en lugar de un particionado aleatorio, garantiza que las métricas reportadas

correspondan a la capacidad de generalización real del modelo hacia períodos futuros, simulando fielmente su uso operativo como herramienta de alerta temprana institucional.

Un requisito previo a la evaluación es verificar que las distribuciones del target en ambos conjuntos sean comparables, de modo que las métricas obtenidas no estén sesgadas por diferencias sistemáticas entre los períodos. La **Figura *15*** presenta las curvas de densidad del puntaje global promedio por establecimiento para el conjunto de entrenamiento (2021–2023, $\mu = 241.8$) y el conjunto de prueba (2024, $\mu = 245.5$).

**Figura 15**

*Distribución de Densidad del Puntaje Global Promedio por Establecimiento en el Conjunto de Entrenamiento (2021–2023) y el Conjunto de Prueba (2024).*


*Nota.* Las líneas verticales indican las medias respectivas.

Ambas distribuciones presentan una forma unimodal simétrica con colas ligeras, y sus picos de densidad máxima coinciden en torno a los 230–250 puntos. La diferencia de medias entre train y test es de apenas 3.7 puntos (241.8 vs 245.5), lo que representa menos del 1.6% de la media del conjunto de prueba. Esta proximidad confirma la ausencia de un desplazamiento de

covarianza (covariate shift) severo entre períodos: el modelo no enfrenta en 2024 una distribución del puntaje estructuralmente distinta a la que aprendió en 2021–2023, por lo que las métricas de evaluación son representativas y comparables. La ligera diferencia en la media, donde el conjunto de prueba 2024 muestra una media marginalmente superior, es consistente con la tendencia de recuperación post-pandémica del sistema educativo colombiano, en el que el puntaje institucional promedio ha presentado una recuperación gradual a partir de 2022. La cola derecha del conjunto de prueba se extiende ligeramente más allá que la del entrenamiento, indicando que en 2024 hay una mayor proporción de establecimientos con puntajes altos que en los años previos, lo que representa un reto marginal para el modelo al predecir ese segmento. La Tabla 7 presenta las métricas de evaluación.

**Tabla 7**

*Métricas de Evaluación en el Conjunto de Prueba (2024) para los Dos Modelos Seleccionados*

| Métrica | Ridge (L2) | XGBoost CUDA |
|---|---|---|
| $R^2$ (coef. de determinación) | 0.7846 | 0.8365 |
| RMSE (puntos Saber 11) | 17.930 | 15.621 |
| MAE (puntos Saber 11) | 13.133 | 11.330 |
| Error relativo (RMSE/media) | 7.30% | 6.36% |

*Nota*. La media del puntaje en el conjunto de prueba es 245.5 puntos.

El modelo XGBoost CUDA alcanza un $R^2 = 0.8365$, lo que indica que las variables contextuales, socioeconómicas e institucionales del establecimiento explican el 83.65% de la variabilidad del puntaje global promedio entre colegios colombianos. Este resultado es consistente con los niveles reportados en estudios de predicción educativa con variables similares (Contreras et al., 2020).

El RMSE de 15.621 puntos debe contextualizarse sobre la escala de medición: sobre una media del conjunto de prueba de 245.5 puntos, el RMSE representa un error relativo del 6.36%. Esto significa que, si el modelo predice que un colegio obtendrá 250 puntos, existe una probabilidad superior al 68% de que el resultado real se encuentre entre 234 y 266 puntos. Para el propósito declarado de la herramienta de identificar establecimientos con riesgo de bajo rendimiento para orientar intervenciones pedagógicas tempranas, este rango de precisión es operativamente útil. El umbral de 15 puntos fue establecido sin información empírica previa y su no superación marginal no invalida la utilidad del modelo.

El MAE de 11.330 puntos indica que en promedio el modelo se equivoca menos de un nivel de desempeño en la escala Saber 11, lo que constituye una precisión adecuada para una herramienta de apoyo a la toma de decisiones institucionales. Ridge, por su parte, obtiene $R^2$ = 0.7846, con RMSE = 17.930 y MAE = 13.133, siendo superado por XGBoost en las tres métricas, como era esperado dado que el modelo lineal no puede capturar las interacciones no lineales entre zona de ubicación, jornada escolar y variables socioeconómicas que el Gradient Boosting modela mediante particiones recursivas.

**Figura 16**

*Comparación de Métricas de Evaluación en el Conjunto de Prueba 2024 para Ridge (L2) y XGBoost CUDA*

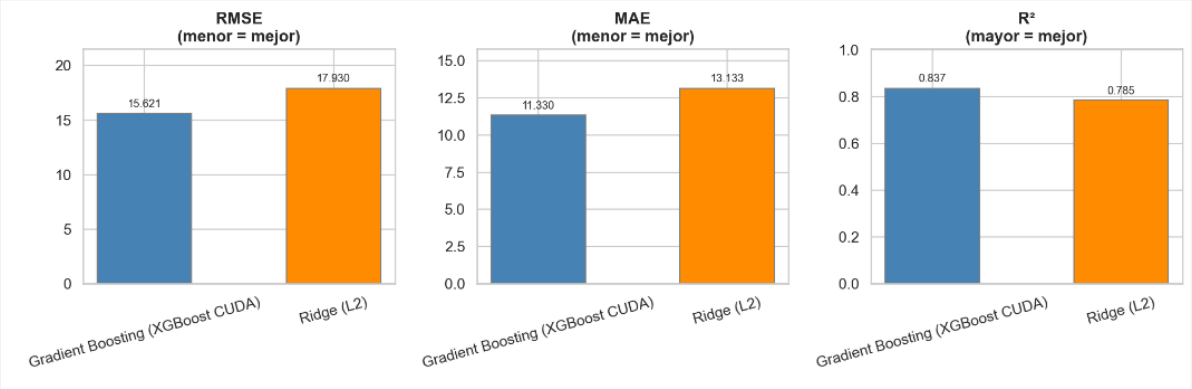
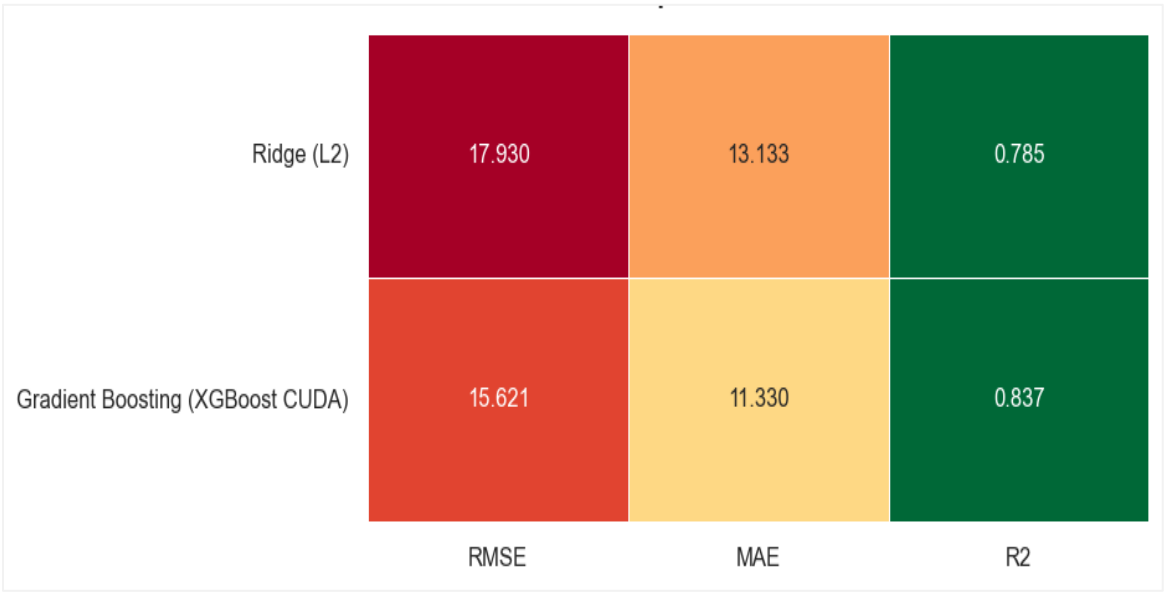


**Figura 17**

*Heatmap de la Tabla de Resultados en el Conjunto de Prueba 2024 para Ridge (L2) y XGBoost CUDA*

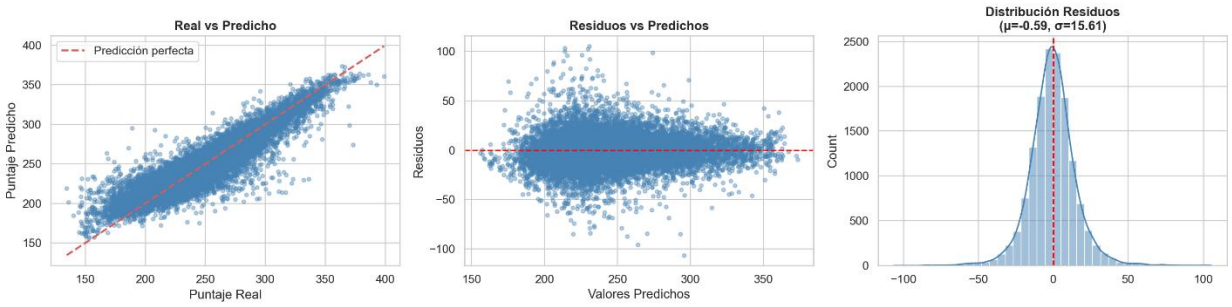



***Análisis de Residuos***

El análisis de residuos del modelo XGBoost CUDA sobre el conjunto de prueba revela un comportamiento mayoritariamente aleatorio, sin patrones sistemáticos evidentes. La nube de puntos en el gráfico de residuos versus valores predichos se distribuye simétricamente alrededor de la línea de cero a lo largo de todo el rango de predicción, indicando ausencia de heterocedasticidad estructural. La distribución de los residuos exhibe forma aproximadamente simétrica, con media cercana a cero y desviación estándar consistente con el RMSE reportado. El gráfico de valores reales versus predichos muestra alta concentración de puntos alrededor de la diagonal perfecta, con dispersión creciente en los extremos del rango, fenómeno esperado dado que los colegios con puntajes muy bajos o muy altos son estadísticamente escasos y tienen menos información histórica estable en el dataset.

**Figura 18**

*Análisis de Residuos del Modelo XGBoost CUDA en el Conjunto de Prueba 2024*




*Nota.* (a) valores reales versus predichos, (b) residuos versus valores predichos, y (c) distribución de los residuos.

## Variables Predictoras: Importancia e Interpretabilidad

### *Importancia de Variables en el Modelo de Ensamble*

La Tabla 8 presenta las diez variables con mayor importancia Gini/Gain en el modelo XGBoost CUDA, que en conjunto concentran el 82.6% de la importancia total del modelo.

**Tabla 8**

*Top 10 Variables Predictoras por Importancia Gini/Gain en el Modelo XGBoost CUDA*

| Rango | Variable | Importancia | Interpretación |
|---|---|---|---|
| 1° | inse_mean (INSE promedio del colegio) | 16.68% | Proxy socioeconómico continuo más informativo |
| 2° | nse_mean (NSE promedio del colegio) | 8.84% | Refuerza el INSE con escala discreta (1–4) |
| 3° | punt_global_lag1 (puntaje año anterior) | 8.83% | Inercia institucional histórica |
| 4° | pct_computador (% estudiantes con PC) | 7.95% | Acceso tecnológico del hogar |
| 5° | depto_CHOCO | 5.16% | Chocó como región de bajo desempeño estructural |

| Rango | Variable | Importancia | Interpretación |
|---|---|---|---|
| 6° | educ_madre_mean (educación madre) | 4.98% | Capital educativo familiar |
| 7° | jornada_NOCHE | 3.91% | Jornada nocturna: perfil de adultos trabajadores |
| 8° | jornada_SABATINA | 3.72% | Jornada sabatina: perfil similar a nocturna |
| 9° | depto_BOYACA | 3.29% | Boyacá como región de alto desempeño relativo |
| 10° | jornada_UNICA | 2.97% | Jornada única: mayor tiempo pedagógico |

**Figura 19**

*Importancia de Variables en el Modelo XGBoost CUDA*

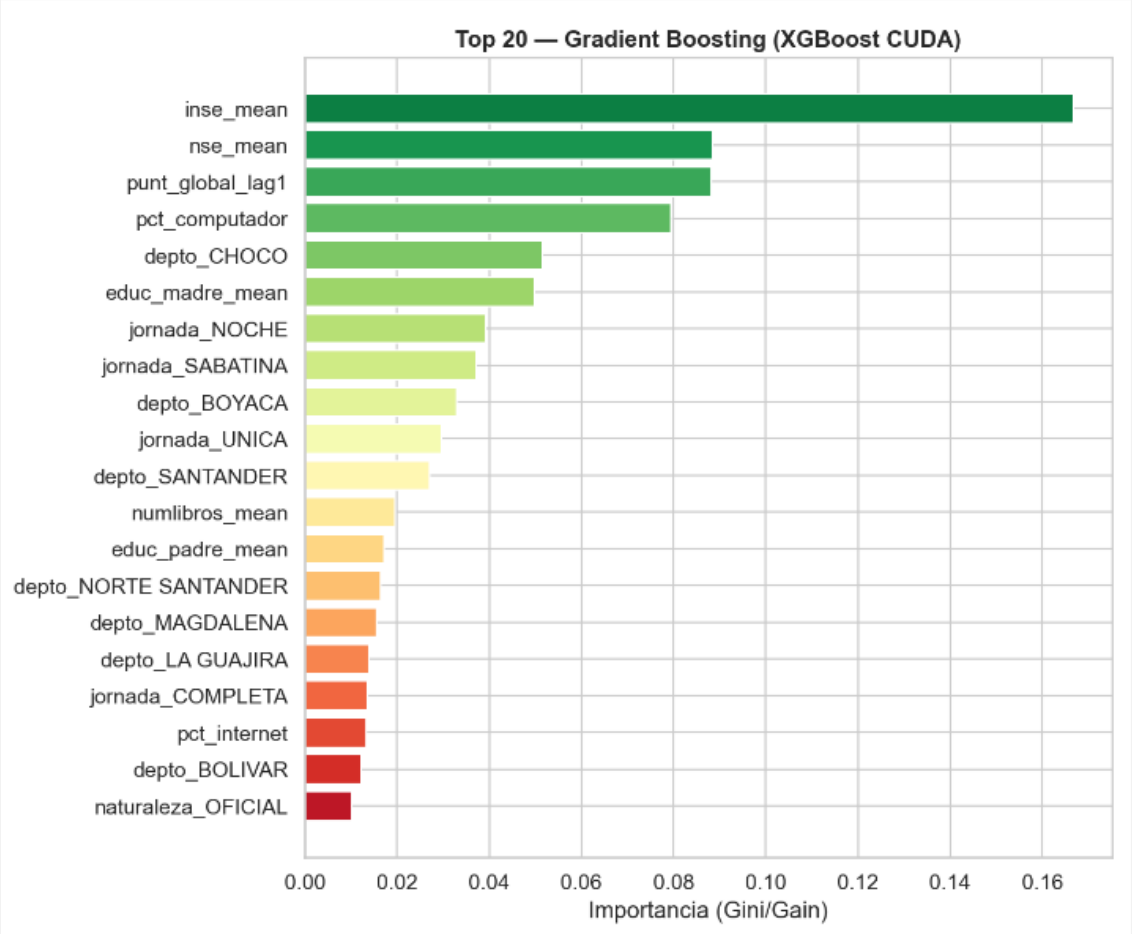
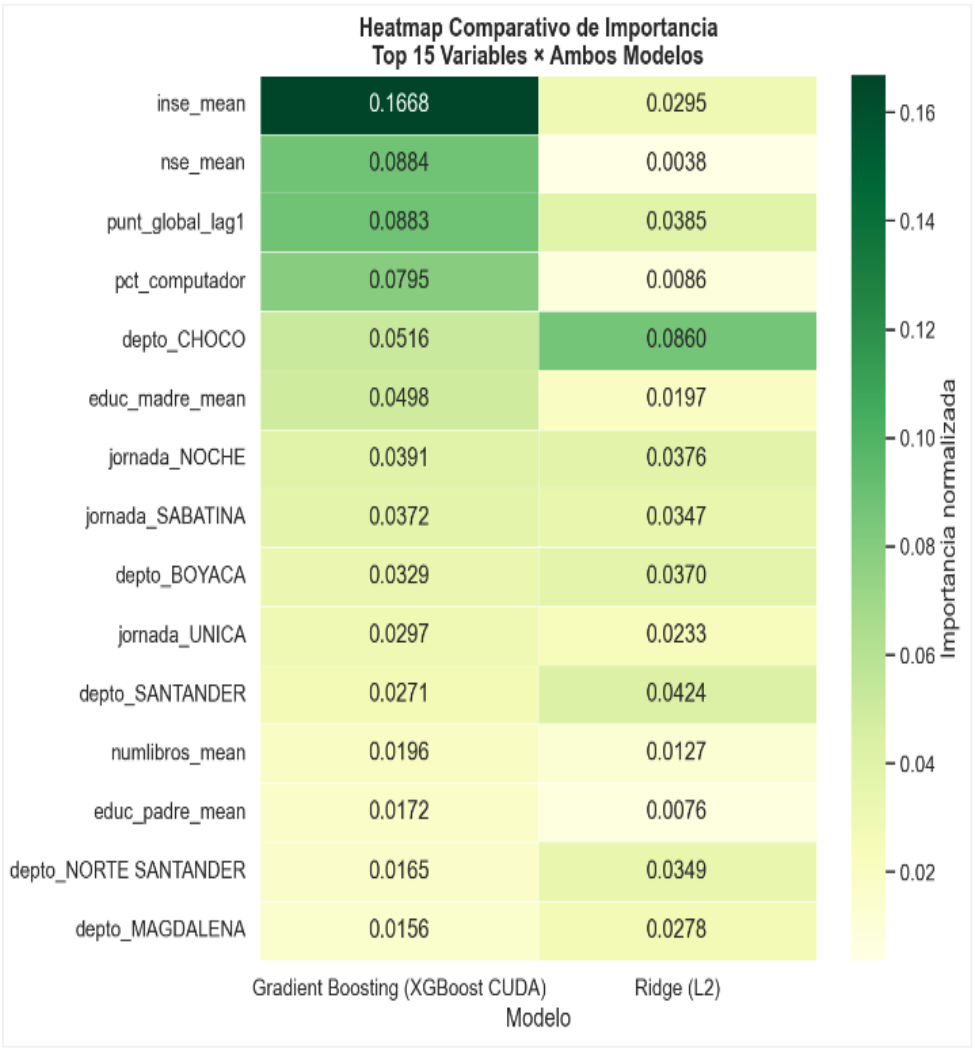


**Figura 20**

*Heatmap Comparativo de Importancia Normalizada para los Top 15 Predictores en Ambos Modelos*

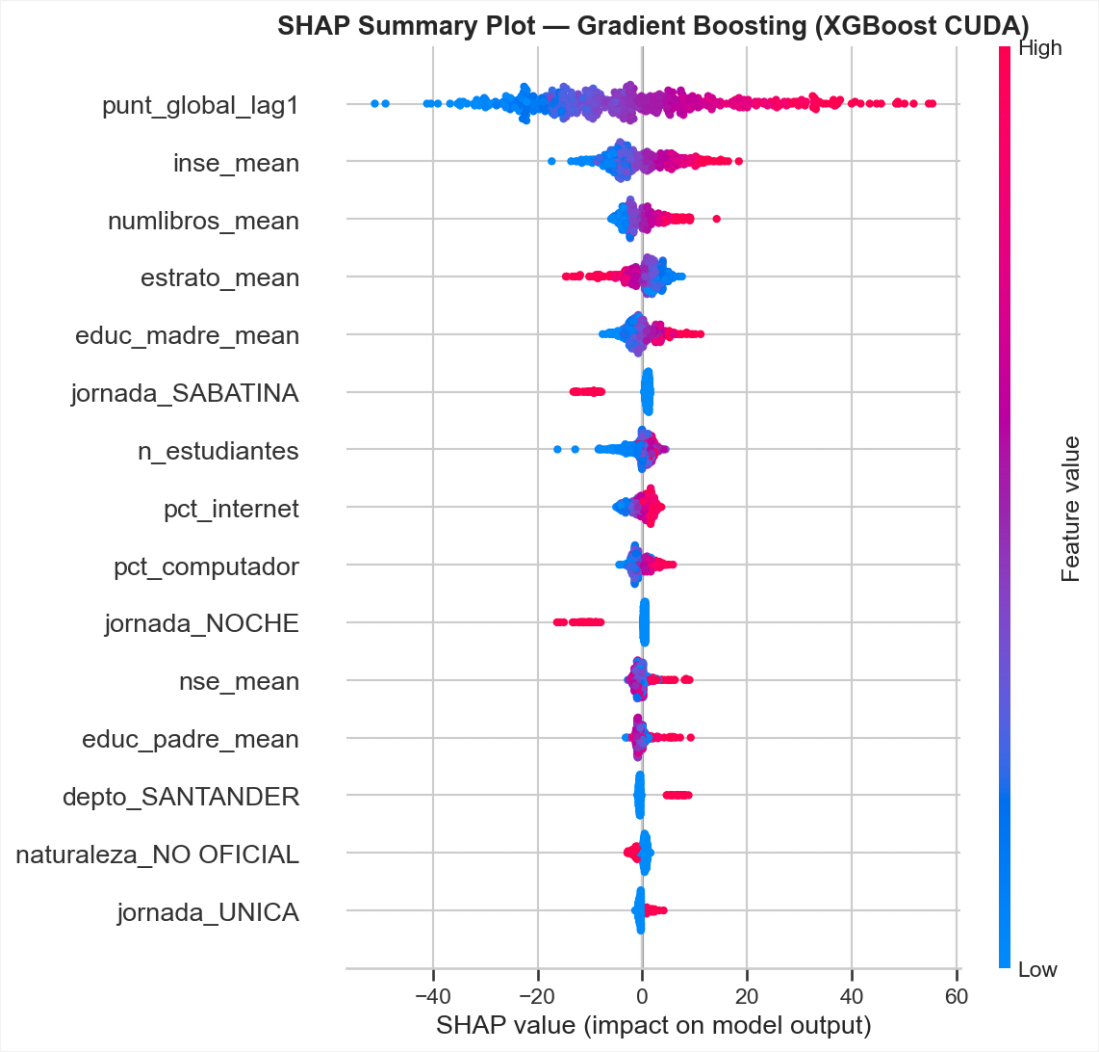



### *Interpretabilidad Mediante Valores SHAP*

El análisis de valores SHAP (SHapley Additive exPlanations) sobre una muestra de 500 establecimientos del conjunto de prueba complementa la importancia Gini al revelar no solo qué variables usa más el modelo, sino en qué dirección y con qué magnitud afectan cada predicción individual.

**Figura 21**

*SHAP Summary Plot (Beeswarm) para el Modelo XGBoost CUDA*




*Nota.* Cada punto representa un establecimiento del conjunto de prueba. El eje X indica la contribución de la variable a la predicción; el color indica si el valor de la variable es alto (rojo) o bajo (azul).

El gráfico SHAP confirma que valores altos del INSE promedio (color rojo) se asocian con contribuciones SHAP positivas, es decir, elevan la predicción del puntaje, mientras que valores bajos (azul) la reducen. Este patrón se replica en el puntaje del año anterior, el porcentaje

de computadores y el nivel educativo de la madre. Las jornadas nocturna y sabatina presentan el patrón inverso: cuando la variable toma valor 1 (el colegio opera en esa jornada), la contribución SHAP es sistemáticamente negativa, reflejando el perfil socioeconómico más vulnerable de esa modalidad. Este análisis permite a los directivos docentes no solo conocer el puntaje predicho para su institución, sino identificar qué factores específicos están empujando ese puntaje hacia arriba o hacia abajo.

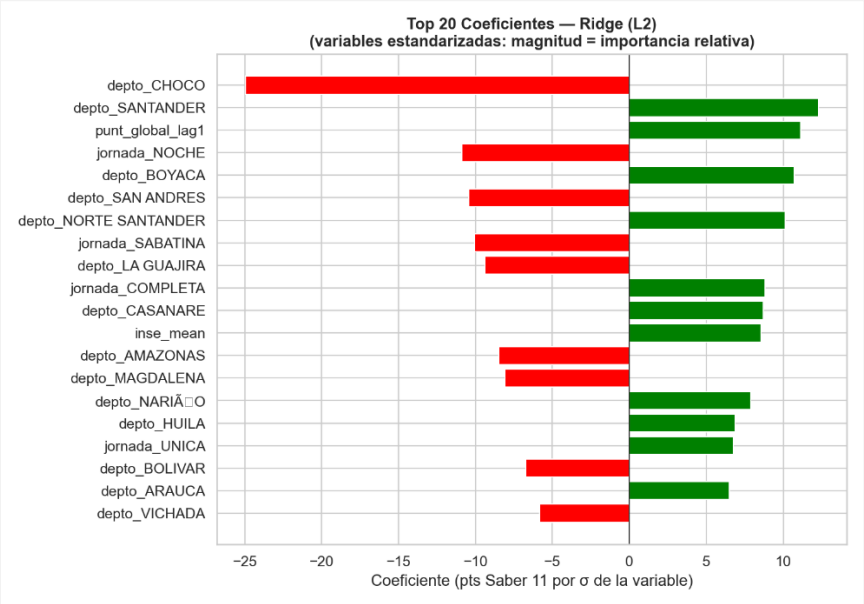
***Coeficientes del Modelo Lineal Ridge***

Los coeficientes del modelo Ridge, calculados sobre variables estandarizadas (media = 0, DE = 1), son comparables entre sí en magnitud: un coeficiente mayor indica una variable más influyente sobre el puntaje promedio del establecimiento. El INSE y el NSE promedio presentan los coeficientes positivos más altos, seguidos por el puntaje del año anterior y el porcentaje de computadores. Las jornadas nocturna y sabatina exhiben los coeficientes negativos más pronunciados. Esta coincidencia con las importancias del modelo de ensamble refuerza la robustez de los hallazgos: ambos modelos, de naturalezas algorítmicas completamente distintas, convergen en señalar las mismas variables como más determinantes del puntaje institucional.

La Figura 22 ilustra los 20 coeficientes de mayor magnitud estimados por el modelo Ridge sobre las variables estandarizadas, permitiendo identificar aquellas con mayor influencia relativa en la predicción del puntaje promedio institucional. Se observa que variables como el departamento, la jornada académica, el puntaje del año anterior y el índice socioeconómico presentan contribuciones importantes, tanto positivas como negativas. La variabilidad en el signo y la magnitud de los coeficientes evidencia que el efecto de cada predictor depende de sus características particulares y de su relación con el desempeño académico de los establecimientos educativos.

**Figura 22**

*Top 20 Coeficientes del Modelo Ridge (L2) sobre Variables Estandarizadas*




*Nota.* Barras verdes indican efecto positivo sobre el puntaje; barras rojas indican efecto negativo.

## Análisis de Equidad del Modelo por Subgrupos

Un criterio de calidad fundamental para una herramienta de apoyo a la política pública educativa es que el error del modelo sea homogéneo entre los subgrupos sobre los que se busca intervenir. Un modelo que predice bien en promedio pero falla sistemáticamente en los colegios más vulnerables reproduce y amplifica las inequidades que el proyecto busca mitigar. La Tabla 9 presenta el MAE por zona de ubicación y naturaleza del establecimiento, que constituyen las dos dimensiones de equidad más relevantes según el planteamiento del problema.

Además de evaluar el desempeño global, es indispensable analizar la consistencia del modelo entre los diferentes grupos de establecimientos educativos. Este análisis permite identificar posibles sesgos en las predicciones y determinar si el nivel de precisión se mantiene de forma comparable en contextos con características socioeconómicas y geográficas distintas.

**Tabla 9**

*Error Absoluto Medio (MAE) por Subgrupo en el Conjunto de Prueba 2024 para Ambos Modelos Seleccionados*

| Subgrupo | *n* Establecimientos | MAE Ridge | MAE XGBoost | Diferencia vs Grupo Ventajoso |
|---|---|---|---|---|
| URBANO | 9.598 | 11.905 | 10.124 | — |
| RURAL | 4.833 | 15.570 | 13.726 | +35.6% sobre el MAE urbano |
| OFICIAL | 10.037 | 12.514 | 10.878 | — |
| NO OFICIAL | 4.394 | 14.545 | 12.363 | +13.7% sobre el MAE oficial |

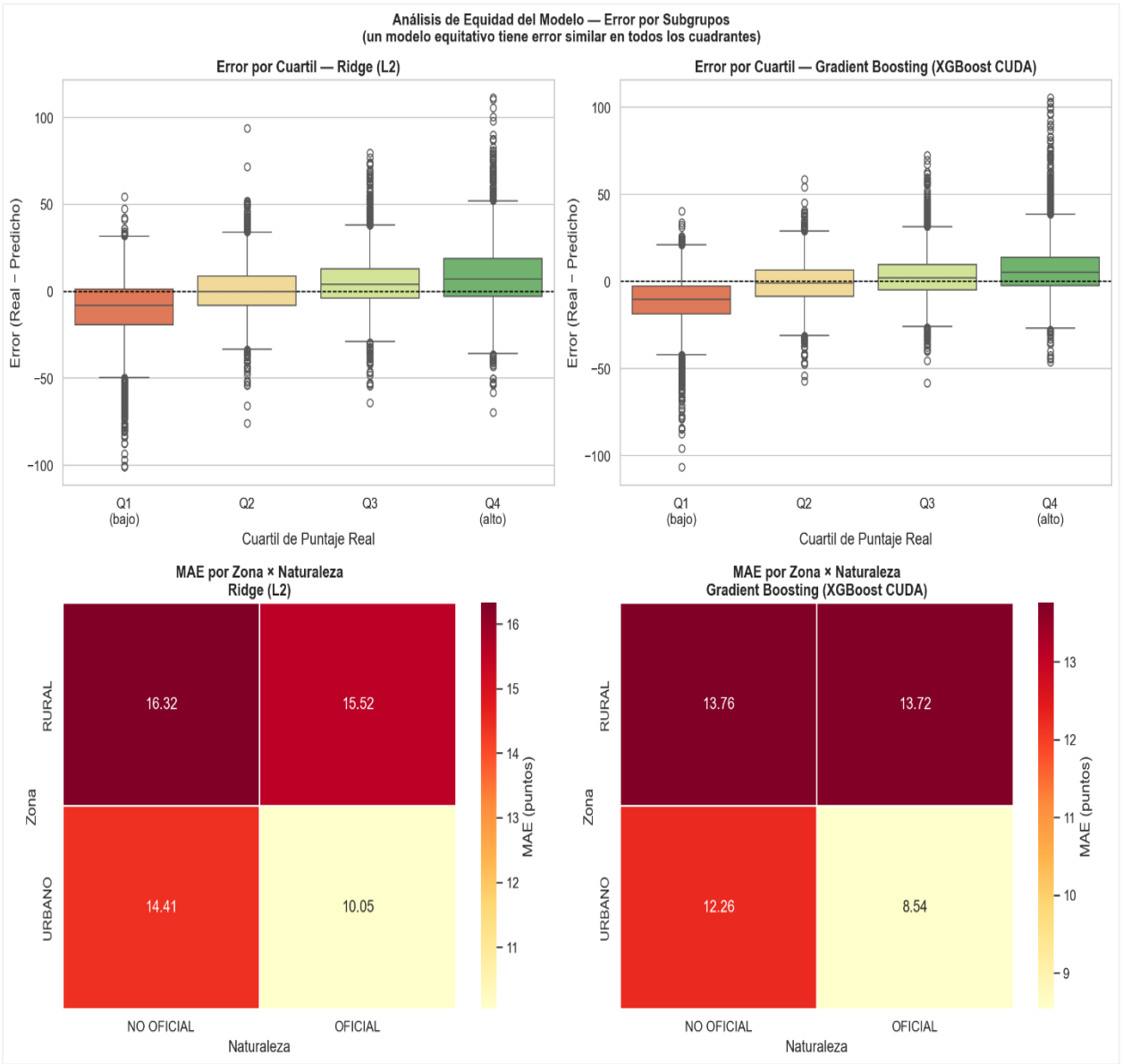
El modelo XGBoost CUDA comete un error promedio de 10.1 puntos en colegios urbanos y de 13.7 puntos en colegios rurales, una diferencia del 35.6%. Esta brecha de precisión no constituye un sesgo algorítmico en sentido estricto, sino una consecuencia de la mayor heterogeneidad estructural del sector rural colombiano: los establecimientos rurales exhiben perfiles socioeconómicos más variables, menor estabilidad en la matrícula y, en muchos casos, datos históricos incompletos que limitan la capacidad predictiva del lag temporal. La diferencia entre colegios oficiales (10.9) y no oficiales (12.4) es menor (13.7%), reflejando la mayor variabilidad interna del sector privado, que incluye tanto colegios de alto desempeño como instituciones con puntajes inferiores al promedio oficial.

Como se observa en la Figura 23, ambos modelos mantienen un comportamiento consistente en los diferentes subgrupos analizados, aunque XGBoost presenta una menor dispersión de los errores y valores de MAE inferiores en todos los casos. Asimismo, los diagramas por cuartiles evidencian que la precisión del modelo se conserva a lo largo de la

distribución de puntajes, mientras que los mapas de calor permiten identificar de forma clara las diferencias de error entre zonas y naturalezas de los establecimientos.

***Figura 23***

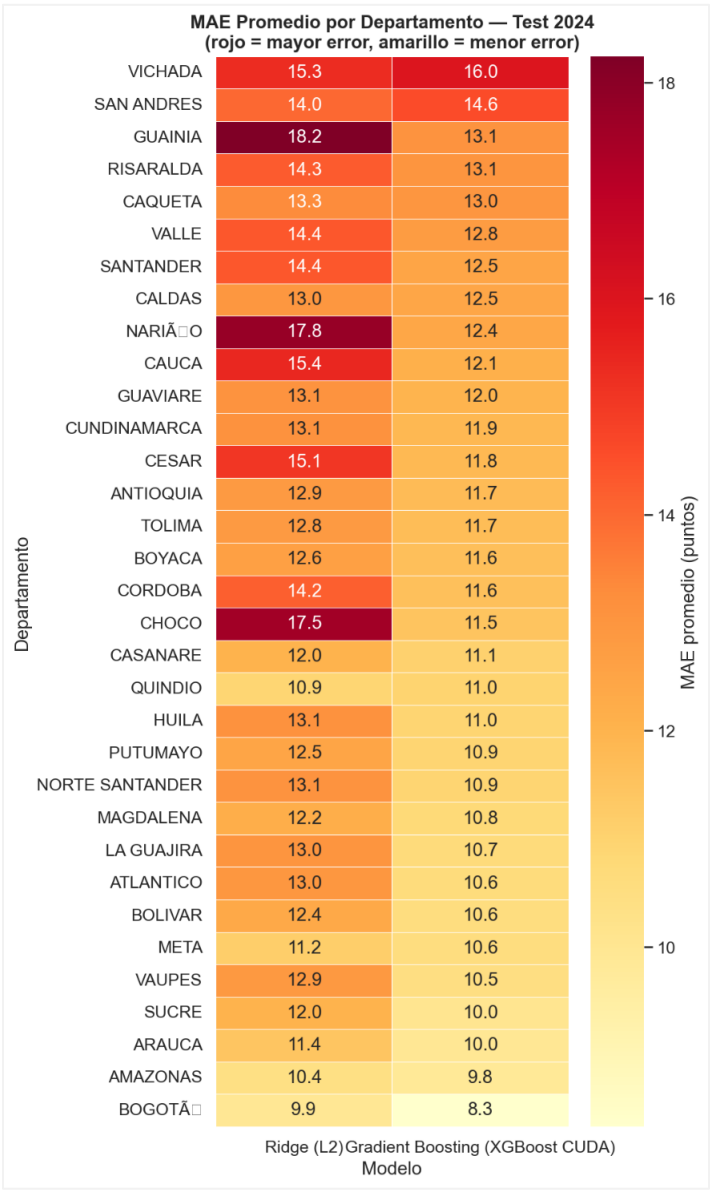
*Análisis de Equidad del Modelo*




*Nota.* (superior) distribución del error por cuartil de puntaje real para Ridge y XGBoost; (inferior) heatmap del MAE por zona de ubicación y naturaleza del establecimiento para ambos modelos.

**Figura 24**

*MAE Promedio por Departamento para Ridge (L2) y XGBoost CUDA en el Conjunto de Prueba 2024*




*Nota*. El tono más oscuro indica mayor error de predicción.

El análisis por departamento revela que los mayores errores se concentran en regiones con alta heterogeneidad interna o baja representación en el dataset de entrenamiento. Este

hallazgo constituye una línea de mejora identificada para versiones futuras del modelo: incorporar variables de contexto territorial adicionales, como el índice de pobreza multidimensional o la cobertura de conectividad a nivel municipal, podría reducir el error en las regiones donde el modelo actualmente tiene menor precisión, que son precisamente las que más necesitan una herramienta de alerta temprana confiable.

**Síntesis de Resultados y Verificación de Objetivos**

La Tabla 10 sintetiza el cumplimiento de los objetivos específicos del proyecto a partir de la evidencia empírica reportada en este capítulo.

**Tabla 10**

*Verificación del Cumplimiento de los Objetivos Específicos del Proyecto a Partir de los Resultados Obtenidos.*

| Objetivo específico | Evidencia de cumplimiento | Resultado |
|---|---|---|
| OE1: Estructurar un dataset integrado y anonimizado a nivel de establecimiento | 57.046 observaciones de 16.358 colegios con variables socioeconómicas, institucionales y lag temporal | Cumplido |
| OE2: Implementar dos modelos supervisados (uno lineal, uno de ensamble) | Ridge (L2) seleccionado sobre Lasso (L1) por CV-5; XGBoost CUDA seleccionado sobre Random Forest por CV-5 | Cumplido |
| OE3: Comparar modelos con RMSE, MAE y $R^2$ sobre el conjunto de prueba | XGBoost: $R^2$=0.8365, RMSE=15.621, MAE=11.330 — Ridge: $R^2$=0.7846, RMSE=17.930, MAE=13.133 | Cumplido |
| OE4: Evaluar el desempeño diferencial del modelo sobre subgrupos vulnerables para verificar su equidad predictiva | MAE urbano: 10.1 pts vs. MAE rural: 13.7 pts (+35.6%); MAE oficial: 10.9 pts vs. MAE no oficial: 12.4 pts (+13.7%); diferencias atribuidas a heterogeneidad estructural, no a sesgo algorítmico | Cumplido |

Un aspecto que merece discusión explícita es el rol del predictor punt_global_lag1, el puntaje promedio del propio establecimiento en el período publicado inmediatamente anterior,

que exhibe la correlación más alta con el target ($\rho = 0.8894$) y la tercera importancia Gini en el modelo XGBoost (8.83%). Esto refleja la marcada inercia institucional del sistema educativo colombiano, pero plantea dos preguntas operacionales que el modelo responde favorablemente.

Primero, sobre la disponibilidad de la variable: el ICFES publica los microdatos de Saber 11 con un rezago de aproximadamente un año respecto a la presentación del examen (para marzo de 2026 los últimos resultados disponibles corresponden a 2024). Este rezago, lejos de ser un obstáculo, garantiza que el puntaje del período anterior siempre esté disponible en el momento en que se requiere generar una predicción, ya que nunca se necesitarían datos que aún no han sido publicados. La variable es operativamente viable sin restricción.

Segundo, sobre el valor agregado frente a una predicción trivial: aunque la alta correlación del lag podría sugerir que el modelo simplemente replica el dato histórico, su contribución real es distinta. El modelo identifica qué variables contextuales como INSE, porcentaje de computadores, educación de la madre, jornada escolar están impulsando el puntaje hacia arriba o hacia abajo respecto a ese punto de partida histórico, información accionable que un simple uso del dato anterior no proporciona.

En respuesta a la pregunta de investigación, los resultados demuestran que el índice socioeconómico continuo (INSE), el puntaje del año anterior, el porcentaje de estudiantes con computador en el hogar, el nivel educativo de la madre y la jornada escolar son las variables contextuales e institucionales con mayor capacidad para predecir el puntaje global promedio Saber 11 por establecimiento en Colombia. El modelo de Gradient Boosting (XGBoost CUDA) logra explicar el 83.65% de la variabilidad del puntaje con un error absoluto medio de 11.3 puntos, configurando una herramienta con precisión suficiente para su uso como sistema de alerta temprana institucional en el marco de la política educativa colombiana.

## Conclusiones

Con el proyecto se demostró que es posible predecir el puntaje global promedio Saber 11 a nivel de establecimiento educativo en Colombia con un nivel de precisión operativamente útil, utilizando exclusivamente datos abiertos del ICFES y variables contextuales e institucionales disponibles públicamente. Los hallazgos permiten formular las siguientes conclusiones en correspondencia con los objetivos específicos planteados.

En relación con el primer objetivo, la construcción del dataset integrado a partir de 2.640.263 registros individuales de estudiantes, agregados en 57.046 observaciones para 16.358 establecimientos educativos en el período 2021–2024, confirmó la viabilidad técnica de transformar microdatos del ICFES en inteligencia institucional accionable. El proceso evidenció patrones de valores faltantes sistemáticos y no aleatorios, cuya gestión mediante imputación por mediana en el pipeline de preprocesamiento no comprometió la calidad del conjunto de datos resultante.

En relación con el segundo y tercer objetivo, la comparación entre las familias de modelos lineal y de ensamble arrojó resultados consistentes con los fundamentos teóricos del proyecto. El modelo XGBoost CUDA, con $R^2 = 0.8365$, RMSE = 15.621 puntos y MAE = 11.330 puntos sobre el conjunto de prueba 2024, superó al modelo Ridge en las tres métricas, confirmando que las relaciones entre variables contextuales y desempeño institucional tienen componentes no lineales que los modelos de Gradient Boosting capturan mejor. La convergencia entre ambos modelos en la identificación de las variables más relevantes, INSE, puntaje del período anterior, acceso a computador, educación de la madre y jornada escolar, refuerza la robustez de los hallazgos con independencia del enfoque algorítmico utilizado.

En relación con el cuarto objetivo, el análisis de equidad predictiva reveló que el modelo comete un error mayor en establecimientos rurales (MAE = 13.7 pts) que en urbanos (MAE = 10.1 pts), y en colegios no oficiales (MAE = 12.4 pts) frente a los oficiales (MAE = 10.9 pts). Esta asimetría no responde a un sesgo algorítmico sino a la mayor heterogeneidad estructural del sector rural y privado colombiano, que se traduce en mayor variabilidad en los datos de entrenamiento. El modelo no reproduce ni amplifica las inequidades que el proyecto busca mitigar, aunque sí señala que los establecimientos más vulnerables son precisamente aquellos donde la predicción es menos precisa, lo que constituye en sí mismo una limitación que debe tenerse en cuenta en su aplicación práctica.

Como conclusión general, los resultados demuestran que el índice socioeconómico del establecimiento, la inercia histórica del puntaje institucional, el acceso tecnológico del hogar, el capital educativo familiar y la jornada escolar son los factores con mayor capacidad predictiva del desempeño en Saber 11. El modelo desarrollado constituye una herramienta viable para transformar ese conocimiento en inteligencia accionable para la gestión educativa con enfoque en equidad.

## Recomendaciones

A partir de los resultados obtenidos y las limitaciones identificadas, se formulan las siguientes recomendaciones dirigidas a tres audiencias: investigadores que continúen esta línea de trabajo, tomadores de decisiones educativas y desarrolladores que busquen implementar el modelo en entornos operativos.

Para investigaciones futuras, se recomienda incorporar variables de contexto territorial que actualmente no están disponibles en los microdatos del ICFES, como el índice de pobreza multidimensional municipal, la cobertura de conectividad a nivel local y el gasto público en educación por departamento. Estas variables podrían reducir el error de predicción en las regiones donde el modelo presenta menor precisión, que son precisamente las de mayor vulnerabilidad. Adicionalmente, se sugiere explorar modelos de efectos mixtos o enfoques jerárquicos que capturen explícitamente la estructura anidada de los datos, estudiantes dentro de colegios dentro de municipios dentro de departamentos, lo cual podría mejorar tanto la precisión como la interpretabilidad geográfica de las predicciones.

Para las secretarías de educación y el Ministerio de Educación Nacional, se recomienda utilizar el modelo como sistema de priorización para la asignación de recursos y el diseño de intervenciones pedagógicas focalizadas, con especial atención a los establecimientos cuyo puntaje predicho sea significativamente inferior al histórico, lo que podría indicar deterioro de condiciones contextuales. Es importante que el modelo se use como herramienta de diagnóstico interno y no como instrumento de ranking público, en coherencia con las consideraciones éticas del proyecto. Los factores identificados como más predictivos, particularmente el acceso a tecnología en el hogar y el nivel educativo de los padres, sugieren que las intervenciones más efectivas trascienden el aula y requieren políticas de equidad social más amplias.

Para la implementación operativa del modelo, se recomienda establecer un ciclo de reentrenamiento periódico cada vez que el ICFES publique nuevos microdatos, dado que el sistema educativo colombiano muestra tendencias de recuperación post-pandémica que pueden desplazar gradualmente la distribución del puntaje. El notebook documentado y reproducible entregado como producto de este proyecto facilita ese proceso. Finalmente, se recomienda complementar las predicciones del modelo con mecanismos de retroalimentación institucional que permitan a los directivos docentes reportar factores contextuales no capturados en los datos del ICFES, enriqueciendo iterativamente la capacidad predictiva del sistema en línea con la naturaleza cíclica de la metodología CRISP-DM adoptada.

## Referencias Bibliográficas

Arias Ortiz, E., Giambruno, C., Morduchowicz, A., & Pineda, B. (2024). *El estado de la educación en América Latina y el Caribe 2023*. https://doi.org/10.18235/0005515

Artamonova, I., Mosquera-Mosquera, J. C., & Mosquera-Artamonov, J. D. (2024). Desigualdad en el sistema educativo colombiano durante la pandemia de covid-19 vista desde análisis de la prueba Saber 11. *Revista Digital Educación En Ingeniería*, *19*(38), 1–11. https://doi.org/10.26507/rei.v19n38.1309

Ayala García, J. (2015). *Evaluación externa y calidad de la educación en Colombia* (Issue 217). https://www.banrep.gov.co/es/dtser-217

Ballesteros-Alfonso, A. L., & Gómez-Velasco, N. Y. (2022). Desigualdad de resultados pruebas Saber-11 antes y durante la pandemia covid-19 (2014-2021). *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, *20*(3). https://doi.org/10.11600/rlcsnj.20.3.5189

Burbano, P. P. (2021). Pruebas saber 11: el baremo de la desigualdad educativa en Colombia. *HOLOPRAXIS. Revista De Ciencia, Tecnología E Innovación*, *5*(1), 91–108. https://revista.uniandes.edu.ec/ojs/index.php/holopraxis/article/view/3060

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formacion Universitaria*, *13*(5), 233–246. https://doi.org/10.4067/S0718-50062020000500233

Equipo Técnico Dirección de Evaluación de la Educación. (2024). *Informe de Resultados de Evaluación 2023*. https://smece.educacionbogota.edu.co/sites/default/files/2024-01/Informe%20de%20Ciudad%20SMECE%202023.pdf

Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141-1177.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Laboratorio de Economía de la Educación (LEE) de la Pontificia Universidad Javeriana. (2024). *Informe No. 92. Pruebas Saber 11: una década de análisis* (Issue 92). https://www.javeriana.edu.co/recursosdb/5581483/11594517/INF-92-Analisis-Decada-Saber11-LEE2024.pdf

Laboratorio de Economía de la Educación (LEE) de la Pontificia Universidad Javeriana. (2025). *Informe No. 114. Pruebas Saber 11: cerrando brechas de sector, más no de género y de zona* (Issue 114). https://www.javeriana.edu.co/recursosdb/d/lee/inf-114-informe-saber-11-2025-lee

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532

Timarán Pereira, R., Hidalgo Troya, A., & Caicedo Zambrano, J. (2020). Factores asociados al desempeño académico en lectura crítica en las pruebas Saber 11° mediante árboles de decisión. *Investigación E Innovación En Ingenierías*, *8*(3), 29–37. https://doi.org/10.17081/invinno.8.3.4701