MODELO PREDICTIVO DEL DESEMPEÑO ACADÉMICO EN EL DISTRITO DE BARRANQUILLA MEDIANTE TÉCNICAS DE CIENCIA DE DATOS

Laura Melisa Acosta Quintero
Código 9027755

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

Directora
Delia Ortega Lenis

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

FICHA RESUMEN

MODELO PREDICTIVO DEL DESEMPEÑO ACADÉMICO EN EL DISTRITO DE BARRANQUILLA MEDIANTE TÉCNICAS DE CIENCIA DE DATOS

1. ÁREA DE TRABAJO: Educación
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE: Laura Melisa Acosta Quintero
4. CORREO ELECTRÓNICO: lmacosta7@javerianacali.edu.co
5. DIRECCIÓN Y TELÉFONO: Calle 96 No. 71-109 Ap1102A, Barranquilla. 3015351845
6. DIRECTORA: Delia Ortega Lenis
7. VINCULACIÓN DE LA DIRECTORA: Profesora del Departamento de Salud Pública y Epidemiología, Pontificia Universidad Javeriana Cali
8. CORREO ELECTRÓNICO DE LA DIRECTORA: delia.ortega@javerianacali.edu.co
9. PALABRAS CLAVE: ciencia de datos, minería de datos educativa, desempeño académico, aprendizaje supervisado, XGBoost, Saber 11°, educación en Barranquilla.
10. FECHA DE INICIO: 6 de octubre de 2025
11. DURACIÓN ESTIMADA: 8 meses
12. RESUMEN: El presente proyecto aplicado desarrolló un modelo predictivo del desempeño académico de los estudiantes del Distrito de Barranquilla en el Examen Saber 11°, a partir de los microdatos públicos del repositorio DataICFES correspondientes al período 2014-2 a 2024-2. Tras un proceso de consolidación, diagnóstico, limpieza e ingeniería de variables sobre 183.984 registros, se seleccionaron 17 predictores socioeducativos mediante criterios estadísticos (ANOVA, η^2) y algorítmicos (Información Mutua y Random Forest *Feature Importance*). El análisis no supervisado mediante FAMD y K-Means segmentó la población en dos perfiles a lo largo de un gradiente socioeducativo dominante: uno aventajado (≈31%) y otro vulnerable (≈69%), con una brecha de 48 puntos en el puntaje promedio. Para el modelado supervisado se entrenaron y compararon cinco algoritmos: regresión lineal, Ridge, Lasso, Random Forest y XGBoost sobre la variable continua Puntaje Global. XGBoost optimizado mediante `RandomizedSearchCV` con `TimeSeriesSplit` resultó el modelo de mejor desempeño, con un error medio de 35 puntos en una escala de 0 a 500 y una varianza explicada cercana al 38% (R^2) sobre una cohorte futura no

observada. El análisis de interpretabilidad mediante valores SHAP identificó el nivel socioeconómico del establecimiento, los hábitos de lectura diaria y la edad de presentación como las variables dominantes, seguidas por la educación de los padres, evidenciando que el contexto institucional y el capital cultural familiar presentan mayor contribución predictiva que la riqueza material directa. Estos hallazgos delimitan las condiciones de uso del modelo como herramienta de apoyo a la política pública educativa del Distrito y confirman la viabilidad de construir, con datos públicos y técnicas estándar, un instrumento predictivo interpretable y contextualizado.

TABLA DE CONTENIDO

Pág.

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

Anexo A. Análisis univariado. Distribución y forma de la variable objetivo Puntaje global y de los 42 predictores del *dataset* de modelado: estadísticos descriptivos, inventario de cardinalidad de las variables categóricas y visualizaciones de la distribución de cada variable, agrupadas por tipo y nivel de cardinalidad.

INTRODUCCIÓN

El desempeño académico de los estudiantes constituye uno de los indicadores más relevantes para evaluar la calidad educativa de un territorio. En el caso del Distrito de Barranquilla, los resultados obtenidos en las pruebas Saber 11° han evidenciado una tendencia de mejora sostenida durante los últimos años; sin embargo, persisten brechas significativas entre instituciones, contextos socioeconómicos y tipos de gestión educativa. Comprender los factores que explican dichas diferencias y anticipar los resultados de aprendizaje es un desafío que exige el uso de enfoques analíticos avanzados y metodologías basadas en evidencia.

En este contexto, la ciencia de datos ofrece herramientas que permiten identificar patrones, modelar comportamientos y generar predicciones a partir de grandes volúmenes de información. Aplicar estos métodos al análisis educativo posibilita transformar los datos en conocimiento útil para la planeación, el seguimiento y la toma de decisiones estratégicas. El presente proyecto aplicado desarrolló un modelo predictivo del desempeño académico de los estudiantes de Barranquilla en el Examen Saber 11°, sustentado en los microdatos oficiales del ICFES y en la aplicación de técnicas de aprendizaje supervisado, complementadas con análisis no supervisado y métodos de interpretabilidad.

La propuesta se desplegó en cuatro acciones, alineadas con los objetivos específicos del proyecto: (i) identificar las variables relevantes asociadas al rendimiento académico, mediante un proceso de consolidación, diagnóstico, limpieza y selección de características que partió de los archivos del repositorio DataICFES y produjo un *dataset* con 17 predictores socioeducativos; (ii) descubrir patrones latentes en los datos mediante reducción de dimensionalidad con FAMD y segmentación con K-Means, identificando dos perfiles a lo largo de un gradiente socioeducativo dominante, consistentes con las variables seleccionadas; (iii) desarrollar modelos predictivos de regresión supervisada sobre la variable continua Puntaje Global, recorriendo una escalera de complejidad desde modelos lineales hasta ensambles no lineales (Random Forest y XGBoost); y (iv) evaluar el desempeño del modelo seleccionado mediante análisis de residuos, validación segmentada por subgrupos relevantes para la política pública, validación por niveles oficiales del ICFES e interpretabilidad mediante valores SHAP.

La metodología se estructuró bajo el estándar CRISP-DM (Cross Industry Standard Process for Data Mining), garantizando un proceso ordenado de comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación y consideraciones de despliegue. El *dataset* final de modelado contiene aproximadamente 134.000 registros del Distrito de Barranquilla para el período 2017–2024, una vez aplicados los filtros de comparabilidad documentados en el cuerpo del documento. La evaluación adoptó una validación temporal estricta, entrenando con las

cohortes 2017–2022 y reservando 2023 y 2024 para validación y prueba, de modo que las métricas reflejen el escenario real de predicción sobre cohortes futuras.

Los resultados obtenidos confirman la viabilidad de construir, con datos públicos y técnicas estándar, un modelo predictivo del desempeño académico contextualizado al Distrito de Barranquilla: el modelo XGBoost seleccionado predice el Puntaje Global con un error medio de 35 puntos sobre una escala de 0 a 500 y explica cerca del 38% de su varianza en una cohorte futura no observada. El análisis de interpretabilidad muestra que el contexto institucional (nivel socioeconómico del establecimiento) y el capital cultural familiar (educación de los padres, número de libros en el hogar, hábitos de lectura) presentan mayor contribución predictiva que la riqueza material directa, hallazgo con implicaciones para la formulación de políticas educativas focalizadas. La validación segmentada evidenció también limitaciones del modelo en el sector oficial y en la población de educación de adultos, así como un patrón de regresión hacia la media en los extremos de la escala, lo que delimita sus condiciones de uso y sienta las bases para los trabajos futuros propuestos.

El documento se organiza de la siguiente forma. El Capítulo 1 presenta la contextualización del proyecto, incluyendo la definición del problema, los objetivos y el marco de referencia teórico y conceptual. El Capítulo 2 expone el diseño metodológico bajo el estándar CRISP-DM. El Capítulo 3 describe el proceso de identificación de las variables relevantes, desde la adquisición de los microdatos hasta la selección final de 17 predictores. El Capítulo 4 reporta el descubrimiento de patrones mediante FAMD y K-Means. El Capítulo 5 presenta el desarrollo de los modelos predictivos de regresión bajo validación temporal. El Capítulo 6 evalúa el desempeño del modelo seleccionado mediante análisis de residuos, validación por subgrupos e interpretabilidad SHAP. Finalmente, el Capítulo 7 presenta las conclusiones del proyecto y propone líneas de trabajo futuro.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Definición del problema

1.1.1. Planteamiento del problema

El desempeño académico de los estudiantes en el Distrito de Barranquilla constituye un tema de creciente relevancia para las políticas públicas locales y nacionales, dado su papel como indicador de la calidad y equidad del sistema educativo. A pesar de los avances en cobertura, infraestructura y programas de calidad, los resultados de las pruebas Saber 11° evidencian brechas persistentes entre sectores educativos, zonas de la ciudad y grupos poblacionales [1].

En los últimos años, Barranquilla ha presentado mejoras sostenidas en su promedio general, superando incluso el promedio nacional en 2022; sin embargo, la diferencia entre instituciones oficiales y no oficiales continúa siendo significativa, con una brecha superior a 30 puntos en el puntaje global [1], [2]. Este panorama muestra que, aunque las políticas de ampliación de cobertura y fortalecimiento docente han generado resultados positivos, persisten desigualdades estructurales que inciden en los logros académicos de los estudiantes.

Los síntomas del problema se observan en indicadores como la repitencia, la reprobación y la deserción escolar, los cuales muestran incrementos en determinados niveles educativos, especialmente en transición y primaria [1]. Estas dinámicas sugieren la existencia de factores no abordados integralmente —como las condiciones del entorno, la calidad del proceso pedagógico o la disponibilidad de recursos tecnológicos— que podrían estar influyendo de manera conjunta en el desempeño académico.

Desde la perspectiva de la ciencia de datos, la problemática radica en la ausencia de un análisis sistemático que aproveche la información oficial disponible (ICFES, MEN, Secretaría Distrital de Educación) para identificar los patrones y relaciones que explican el rendimiento académico [3]. Los informes estadísticos y de gestión actualmente ofrecen descripciones agregadas del fenómeno, pero no permiten predecir comportamientos futuros ni comprender cómo interactúan las distintas variables que inciden en los resultados [4], [5], [6].

De mantenerse esta situación, las decisiones de política educativa continuarán basándose principalmente en análisis descriptivos y no en modelos predictivos fundamentados en evidencia, limitando la capacidad del Distrito para anticipar riesgos, focalizar intervenciones y optimizar la asignación de recursos en programas de calidad educativa [1].

En consecuencia, el problema central se sitúa en la falta de un modelo analítico y predictivo que permita explicar y proyectar el desempeño académico de los estudiantes de Barranquilla mediante el análisis de los microdatos educativos oficiales. Esta carencia dificulta el desarrollo de

estrategias basadas en evidencia que orienten la toma de decisiones institucionales y contribuyan a disminuir las brechas existentes entre estudiantes, instituciones y contextos socioeconómicos.

En este contexto, surge la necesidad de determinar si, a partir de los datos educativos existentes y mediante técnicas de aprendizaje supervisado, es posible predecir el desempeño académico de los estudiantes del Distrito e identificar los factores que explican tales predicciones.

1.1.2. Formulación del problema

En este marco, surge la siguiente pregunta de investigación:

¿Cómo pueden los modelos de aprendizaje supervisado predecir el desempeño académico de los estudiantes del Distrito de Barranquilla e identificar los factores que expliquen dicha predicción?

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 disponibles en las fuentes oficiales presentan relación significativa con el desempeño académico de los estudiantes de Barranquilla?
- ¿Qué patrones pueden identificarse en los datos que permitan explicar diferencias en los niveles de logro académico dentro del Distrito?
- ¿Qué modelo de aprendizaje supervisado resulta más adecuado para estimar el desempeño académico de los estudiantes, considerando las características del contexto local?
- ¿Cómo varía el desempeño de los modelos construidos según las métricas de evaluación, y qué criterios permiten seleccionar el modelo más adecuado para el contexto educativo del Distrito?

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar modelos de aprendizaje supervisado con el propósito de predecir el desempeño académico de los estudiantes del Distrito de Barranquilla, aportando evidencia para la toma de decisiones educativas.

1.2.2. Objetivos Específicos

- Identificar las variables relevantes presentes en las fuentes oficiales de información educativa que presenten relación significativa con el desempeño académico de los estudiantes de Barranquilla, a partir de análisis exploratorios y estadísticos descriptivos.

- Descubrir patrones y relaciones en los datos que permitan explicar las diferencias en los niveles de logro académico entre los estudiantes del Distrito, empleando técnicas de minería de datos y análisis multivariado.
- Desarrollar modelos predictivos de aprendizaje supervisado para estimar el desempeño académico de los estudiantes del Distrito de Barranquilla.
- Evaluar el desempeño de los modelos predictivos desarrollados, con el propósito de seleccionar aquel con mayor capacidad explicativa y predictiva para el contexto local.

1.3. Marco de Referencia

1.3.1. Marco Teórico

El estudio del desempeño académico y su predicción mediante técnicas de ciencia de datos constituye un campo interdisciplinario que combina los fundamentos de la educación, la estadística y la inteligencia artificial. En la última década, el uso de métodos analíticos avanzados ha permitido superar los enfoques tradicionales descriptivos para ofrecer modelos capaces de explicar y anticipar los resultados de los estudiantes en pruebas estandarizadas o entornos de aprendizaje continuo.

Fundamentos del desempeño académico. El desempeño académico es un fenómeno multicausal que integra factores personales, familiares, institucionales y sociales, los cuales interactúan de forma compleja para determinar los resultados de aprendizaje de los estudiantes [7]. Tradicionalmente, los estudios sobre logro escolar se apoyaban en modelos lineales y correlacionales que identificaban relaciones directas entre variables —como nivel socioeconómico, escolaridad de los padres o clima escolar—, sin capturar adecuadamente las interacciones no lineales presentes en los contextos educativos.

El desarrollo de sistemas nacionales de evaluación a gran escala, como las pruebas Saber 11° en Colombia, ha proporcionado un conjunto amplio de variables cuantitativas y categóricas que describen tanto las competencias de los estudiantes como su contexto socioeconómico e institucional [8]. Estas bases de datos constituyen un terreno fértil para la aplicación de técnicas de minería de datos y aprendizaje automático, orientadas a construir modelos predictivos que identifiquen patrones complejos y permitan fundamentar decisiones en materia de calidad educativa.

Minería de datos educativa. La minería de datos educativa se define como el proceso de extracción de conocimiento útil y previamente desconocido a partir de grandes volúmenes de datos educativos, con el fin de comprender mejor los procesos de aprendizaje y apoyar la toma de decisiones [7].

Paradigmas de aprendizaje automático. Las técnicas de aprendizaje automático aplicadas en la EDM se agrupan en dos grandes paradigmas: el aprendizaje supervisado, que predice una variable objetivo conocida a partir de ejemplos etiquetados, y el aprendizaje no supervisado, que descubre patrones o estructuras latentes sin una variable objetivo predefinida [7]. El presente proyecto integra ambos paradigmas con finalidades distintas y complementarias: el aprendizaje no supervisado se emplea para descubrir y caracterizar perfiles de estudiantes que ayuden a explicar las diferencias de puntaje en el Examen Saber 11° del Distrito de Barranquilla, mientras que el aprendizaje supervisado se emplea para estimar el puntaje global como variable continua. El propósito es estimar el puntaje global, la etiqueta continua que reporta el ICFES, y explicar, mediante la caracterización de perfiles de estudiantes, las brechas observadas en los resultados.

Aprendizaje no supervisado y segmentación de estudiantes. El agrupamiento (*clustering*) permite descubrir perfiles de estudiantes con características similares sin imponer categorías predefinidas. En este proyecto se utiliza con una finalidad descriptiva y explicativa: construir una narrativa que dé cuenta de por qué difieren los puntajes en el Examen Saber 11° [9]. Cuando las variables combinan atributos numéricos y categóricos, como ocurre con los datos socioeducativos del ICFES, el análisis factorial de datos mixtos (AFDM) ofrece una reducción de dimensionalidad conceptualmente más adecuada que el análisis de componentes principales aplicado sobre variables categóricas codificadas, pues equilibra la contribución de ambos tipos de variables en la construcción de los ejes factoriales [10]. Sobre ese espacio factorial, el algoritmo K-Means agrupa a los individuos según su cercanía a los centroides [9], y la calidad de la partición se evalúa con medidas internas como el coeficiente de silueta [11]. Los perfiles resultantes permiten interpretar las brechas de desempeño en función de las condiciones socioeducativas que las acompañan.

Modelos supervisados de regresión y métodos de ensamble. Para estimar el puntaje global, los modelos de regresión abarcan desde formulaciones lineales, regresión por mínimos cuadrados ordinarios y sus variantes regularizadas Ridge y Lasso, valoradas por su interpretabilidad, hasta los métodos de ensamble basados en árboles, que capturan interacciones no lineales con mayor precisión [9]. Entre estos últimos, los bosques aleatorios (Random Forest) agregan múltiples árboles entrenados sobre muestras aleatorias para reducir la varianza, mientras que la potenciación del gradiente (*gradient boosting*) construye los árboles de forma secuencial corrigiendo los errores residuales [9]; su implementación regularizada y escalable, XGBoost, se ha consolidado como uno de los algoritmos de mayor desempeño en datos tabulares [12]. La literatura coincide en que los métodos de ensamble suelen superar en precisión a los modelos lineales, con el costo de una menor transparencia, lo que requiere el uso de técnicas de interpretabilidad.

Evaluación y validación de modelos. En problemas de regresión, el desempeño se cuantifica mediante métricas como el coeficiente de determinación (R^2), el error absoluto medio (MAE) y la

raíz del error cuadrático medio (RMSE) [9]. Adicionalmente, cuando los datos poseen una estructura temporal, como las cohortes anuales del Examen Saber 11°, la validación debe respetar el orden cronológico para evitar el uso de información futura en la estimación del desempeño; la validación cruzada temporal, en la que cada partición de entrenamiento precede a su partición de validación, es el procedimiento adecuado en estos casos [13].

Interpretabilidad y aprendizaje automático explicable. La adopción de modelos de mayor complejidad en contextos educativos exige mecanismos que expliquen sus predicciones de forma comprensible para actores no especializados. Los valores SHAP (*SHapley Additive exPlanations*) atribuyen a cada variable una contribución a la predicción, lo que permite obtener explicaciones tanto globales como individuales y compensar la precisión de los modelos de ensamble con la transparencia requerida para sustentar decisiones de política educativa [14].

Integración de enfoques. El marco teórico que sustenta este proyecto se apoya en ejes complementarios: el paradigma del descubrimiento de conocimiento a partir de datos educativos (KDD/EDM), que justifica el uso de algoritmos de aprendizaje automático para extraer relaciones entre las variables de contexto y el desempeño [15]; el modelo metodológico CRISP-DM, que proporciona una estructura ordenada para el ciclo completo de datos [8]; y un enfoque analítico que combina la segmentación no supervisada de los estudiantes, para explicar las diferencias de puntaje, la predicción supervisada del puntaje y la interpretabilidad de los resultados, privilegiando que los hallazgos puedan ser utilizados por actores institucionales para orientar decisiones educativas.

1.3.2. Marco Conceptual

El presente marco conceptual tiene como propósito definir los principales términos, categorías y constructos que orientan el desarrollo del proyecto aplicado. Estos conceptos permiten comprender tanto el fenómeno educativo abordado como los métodos analíticos y tecnológicos empleados para su estudio, integrando la perspectiva de la educación con los fundamentos de la ciencia de datos.

Desempeño académico. Resultado del proceso educativo expresado en el nivel de logro alcanzado por los estudiantes frente a los objetivos curriculares establecidos; refleja la interacción entre factores individuales, familiares, institucionales y contextuales que inciden en el aprendizaje [3]. En el caso colombiano, los resultados de las pruebas Saber 11° constituyen un indicador nacional estandarizado del desempeño al finalizar la educación media [16].

Factores asociados al logro educativo. Variables que explican las diferencias en el rendimiento académico, incluyendo dimensiones socioeconómicas, institucionales, pedagógicas y personales.

La literatura señala el entorno familiar, el nivel educativo de los padres, la infraestructura escolar y el clima institucional entre los principales predictores [7], [17].

Ciencia de datos. Campo interdisciplinario que integra estadística, programación, inteligencia artificial y conocimiento de dominio para extraer información significativa a partir de grandes volúmenes de datos [18]. En educación posibilita pasar de la simple descripción a la explicación y la predicción de fenómenos.

Minería de datos educativa (EDM). Rama especializada de la ciencia de datos que busca descubrir patrones, tendencias y relaciones significativas en bases de datos provenientes de contextos educativos [3], empleando algoritmos de regresión, agrupamiento y detección de asociaciones para analizar el desempeño académico y apoyar la gestión institucional.

Aprendizaje supervisado. Subdisciplina del aprendizaje automático en la que el modelo aprende a partir de ejemplos etiquetados, donde se conocen tanto las variables predictoras como la variable objetivo [15], [18]. En este proyecto, el puntaje global reportado por el ICFES es la etiqueta continua que permite entrenar los modelos de regresión.

Modelo predictivo. Representación matemática o computacional que utiliza datos históricos para anticipar resultados [8]. En educación se emplea para estimar el desempeño académico e identificar los factores que más contribuyen a él; su valor radica tanto en su precisión como en su capacidad explicativa.

Calidad educativa. Grado en que el sistema educativo garantiza aprendizajes pertinentes, equitativos y sostenibles; constructo multidimensional influido por el desempeño académico, la equidad, la eficiencia institucional y la pertinencia curricular [16].

Sistema Nacional de Evaluación Estandarizada (SNEE). Conjunto de evaluaciones con que el sistema educativo colombiano mide de manera objetiva y comparable los aprendizajes a lo largo de la trayectoria escolar, permitiendo el monitoreo de la calidad a nivel nacional y comparaciones temporales y territoriales [19].

Examen Saber 11°. Examen de Estado de la Educación Media, obligatorio y censal, seleccionado en este proyecto como fuente principal de información por su cobertura amplia y representativa. Su pertinencia se sustenta en la Ley 1324 de 2009, que asigna al ICFES la evaluación de la educación [20], y en el Decreto 869 de 2010, que lo reglamenta [21]. Además de los puntajes, el ICFES aplica un cuestionario socioeconómico que recoge características familiares, condiciones del hogar y trayectorias escolares [19].

DataICFES. Repositorio público administrado por el ICFES que centraliza y ofrece acceso libre a microdatos anonimizados de las evaluaciones estandarizadas, con el fin de facilitar el análisis de la calidad educativa y promover una gestión basada en evidencia [19].

Estructura y comparabilidad del Examen Saber 11°. A partir del segundo semestre de 2014, el examen se estructuró en cinco pruebas genéricas: Matemáticas, Lectura Crítica, Ciencias Naturales, Sociales y Ciudadanas, e Inglés, lo que permite coherencia longitudinal y comparabilidad entre periodos posteriores a 2014-2 [19].

Puntaje global. Promedio ponderado de las cinco competencias evaluadas, con media 250, desviación estándar 50, mínimo 0 y máximo 500 [19]. Constituye la variable objetivo continua que el proyecto busca estimar y explicar.

Aprendizaje no supervisado. Subdisciplina del aprendizaje automático que descubre estructuras o patrones latentes en los datos sin una variable objetivo etiquetada [18]. En este proyecto se empleó para descubrir y caracterizar perfiles socioeducativos que permiten explicar las diferencias de puntaje entre los estudiantes.

Agrupamiento (*clustering*). Conjunto de técnicas de aprendizaje no supervisado que particionan las observaciones en grupos (clústeres) de modo que los individuos de un mismo grupo sean más similares entre sí que respecto a los de otros grupos [9]. Permite identificar perfiles sin imponer categorías predefinidas. En el proyecto, esos perfiles dan sustento a la narrativa que explica por qué difieren los puntajes.

Análisis factorial de datos mixtos (AFDM). Método de reducción de dimensionalidad que generaliza el análisis de componentes principales y el de correspondencias múltiples para tratar de forma simultánea variables cuantitativas y categóricas, equilibrando la influencia de ambos tipos en la construcción de los ejes factoriales [10]. Es pertinente cuando los datos combinan atributos numéricos y categóricos, como los del ICFES, y constituye una alternativa más adecuada que aplicar componentes principales sobre variables categóricas codificadas. En inglés se denomina *Factor Analysis of Mixed Data* (FAMD).

Algoritmo K-Means. Algoritmo de agrupamiento por particiones que divide las observaciones en k grupos, asignando cada una al clúster cuyo centroide (la media de sus miembros) se encuentra más próximo y actualizando los centroides de forma iterativa hasta la convergencia [9]. Resulta apropiado tras una reducción de dimensionalidad; en el proyecto se aplicó sobre las coordenadas factoriales del AFDM para construir los perfiles socioeducativos que explican las diferencias de puntaje.

Formalmente, K-Means busca la partición que minimiza la suma de distancias cuadráticas intra-clúster (inercia), como se expresa en (1)~~(1)~~:

$$\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2 \quad (1)$$

donde μ_j es el centroide del clúster C_j y $\|\cdot\|$ la distancia euclidiana. De esta formulación se derivan sus condiciones de uso: depende de la distancia euclidiana, es sensible a la escala de las variables, por lo que exige estandarización previa, y tiende a recuperar grupos de forma aproximadamente esférica y tamaño comparable; en este proyecto se aplicó sobre las coordenadas factoriales del AFDM, ya estandarizadas.

Coeficiente de silueta. Medida interna de validación del agrupamiento que cuantifica, para cada observación, qué tan cohesionada está respecto a su propio clúster frente al clúster vecino más cercano; toma valores entre −1 y 1, donde valores altos indican una asignación adecuada [11]. Su promedio se usa para elegir el número de clústeres k. De manera complementaria, la validez de la segmentación se evaluó con el coeficiente eta cuadrado (η^2), que mide la proporción de varianza del puntaje global explicada por la pertenencia a los clústeres, evidenciando que los perfiles efectivamente separan los puntajes.

Selección de características. Identificación del subconjunto de variables predictoras más relevantes para mejorar la generalización del modelo, reducir el sobreajuste y facilitar la interpretación [9]. En el proyecto se combinaron la correlación de Pearson y el análisis de varianza para medir la asociación con el puntaje, la información mutua para capturar dependencias no lineales y la importancia derivada de un modelo de bosques aleatorios; el factor de inflación de la varianza (VIF) se utilizó para descartar multicolinealidad.

Regresión lineal y regularización. La regresión lineal por mínimos cuadrados ordinarios (OLS) estima la variable objetivo como una combinación lineal de los predictores y ofrece alta interpretabilidad, pero es sensible a la multicolinealidad y al sobreajuste [9]. La regularización añade una penalización sobre la magnitud de los coeficientes: Ridge (penalización L2) reduce su varianza ante variables correlacionadas, y Lasso (penalización L1) puede anular coeficientes, actuando además como método de selección de variables. En el proyecto, estos modelos lineales actuaron como referencia interpretable.

La regresión lineal estima el objetivo como una combinación lineal de los predictores, como se muestra en (2)~~(2)~~:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_j \tag{2}$$

y OLS ajusta los coeficientes minimizando la suma de errores cuadráticos. La regularización añade a esa suma una penalización sobre la magnitud de los coeficientes. Ridge (penalización L2) minimiza la expresión dada en (3)~~(3)~~:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

mientras que Lasso (penalización L1) minimiza la expresión dada en (4)~~(4)~~:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

donde λ ≥ 0 controla la intensidad de la penalización. La penalización L2 reduce la varianza de los coeficientes sin anularlos, en tanto que la L1 puede llevarlos exactamente a cero, actuando como método de selección de variables. Como la penalización depende de la escala de los predictores, estos se estandarizan antes del ajuste.

Bosques aleatorios (Random Forest). Modelo de ensamble que promedia las predicciones de múltiples árboles de decisión entrenados sobre muestras aleatorias de los datos y de las variables, lo que reduce la varianza y el riesgo de sobreajuste respecto a un árbol único [9]. Captura relaciones no lineales e interacciones, y su sobreajuste se controla con hiperparámetros como la profundidad máxima y el tamaño mínimo de hoja.

En el enfoque de agregación (*bagging*), la predicción es el promedio de B árboles entrenados sobre muestras *bootstrap* de los datos y subconjuntos aleatorios de variables, como expresa (5)~~(5)~~:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b \, (x) \tag{5}$$

El entrenamiento en paralelo y la agregación reducen la varianza respecto a un árbol único.

Potenciación del gradiente y XGBoost. La potenciación del gradiente (*gradient boosting*) construye los árboles de forma secuencial, de modo que cada nuevo árbol corrige los errores residuales del conjunto acumulado [9]. XGBoost (*Extreme Gradient Boosting*) es una implementación escalable y regularizada de este enfoque que incorpora penalizaciones sobre la complejidad del modelo, lo que la ha consolidado como uno de los algoritmos de mayor desempeño en datos tabulares [12].

A diferencia del *bagging*, la potenciación (*boosting*) construye los árboles de forma secuencial, sumando funciones que corrigen el error residual acumulado; XGBoost añade a la función de pérdida un término de regularización sobre la complejidad de cada árbol, según se muestra en (6)~~(6)~~:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i) \qquad L = \sum_i l_i + \sum_m \Omega(f_m) \qquad \Omega(f) = \gamma T + 1/2\,\lambda \parallel w \parallel^2 \tag{6}$$

donde l es la pérdida, $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penaliza el número de hojas T y la magnitud de sus pesos w, γ y λ regulan la complejidad.

Así, el *bagging* reduce principalmente la varianza (árboles independientes entrenados en paralelo) y el *boosting* reduce el sesgo (árboles secuenciales que corrigen los errores previos); esta capacidad de modelar relaciones no lineales e interacciones, junto con la regularización, explica su alto desempeño en datos tabulares.

Modelo de referencia (*baseline*). Modelo ingenuo de comparación que establece el desempeño mínimo que un modelo entrenado debe superar para ser útil; en regresión suele predecir la media de la variable objetivo del entrenamiento [9]. En el proyecto permitió cuantificar la ganancia real del modelo seleccionado en términos de error y de varianza explicada.

Sobreajuste y compromiso sesgo–varianza. El sobreajuste ocurre cuando un modelo memoriza el ruido del entrenamiento y generaliza mal sobre datos nuevos; se diagnostica comparando el desempeño en entrenamiento, validación y prueba [9]. Se vincula al compromiso entre sesgo y varianza, gestionado mediante regularización, control de hiperparámetros y una validación adecuada.

Validación cruzada temporal. Esquema de validación para datos con componente temporal en el que cada partición de entrenamiento precede cronológicamente a su partición de validación, evitando usar información futura para predecir el pasado [13]. A diferencia de la validación cruzada aleatoria, respeta el orden de las cohortes; en el proyecto se adoptó por tratarse de datos longitudinales del Examen Saber 11°.

Optimización de hiperparámetros. Búsqueda de la combinación de hiperparámetros, parámetros de configuración que no se aprenden de los datos, como la profundidad de los árboles o la tasa de aprendizaje, que maximiza el desempeño en validación [9]. En el proyecto se realizó con búsqueda aleatoria combinada con validación cruzada temporal, evitando contaminar la estimación final sobre el conjunto de prueba.

Métricas de evaluación en regresión. Indicadores que cuantifican el error de un modelo de regresión [9]: el coeficiente de determinación (R^2) mide la proporción de varianza explicada; el error absoluto medio (MAE) promedia las magnitudes de los errores; la raíz del error cuadrático medio (RMSE) penaliza con mayor severidad los errores grandes; y el error porcentual absoluto medio (MAPE) los expresa en términos relativos. Estas métricas se definen en (7)~~(7)~~, (8)~~(8)~~, (9)~~(9)~~ y (10)~~(10)~~:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_i |y_i - \hat{y}_i| \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \tag{9}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_i \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{10}$$

donde y es el valor observado, ŷ la predicción, ȳ la media y n el número de observaciones. En la escala del puntaje global (0 a 500), el MAE y el RMSE se interpretan directamente en puntos y el R^2 como la fracción de varianza explicada, mientras que el MAPE, al dividir por y, es sensible a los puntajes bajos. El criterio principal para seleccionar el modelo final fue el R^2 sobre la cohorte de validación (2023), complementado con la estabilidad bajo validación temporal y el diagnóstico de residuos.

Diagnóstico de residuos. Conjunto de pruebas que examinan los errores de predicción para diagnosticar sesgos sistemáticos y caracterizar el comportamiento del error del modelo [9]. Estas pruebas de normalidad de los residuos (Shapiro–Wilk), homocedasticidad (Breusch–Pagan) y ausencia de autocorrelación (Durbin–Watson), se emplean como herramientas descriptivas del comportamiento del error y no como supuestos necesarios del modelo, dado que el modelo finalmente seleccionado (XGBoost) no requiere normalidad, homocedasticidad ni independencia de los residuos en el sentido de la regresión lineal clásica; la interpretación se complementó con

el tamaño del efecto, dado que con muestras grandes los contrastes tienden a rechazar la hipótesis nula ante desviaciones de escasa relevancia práctica.

Interpretabilidad y valores SHAP. Capacidad de explicar las predicciones de un modelo en términos comprensibles para actores no especializados. Los valores SHAP (*SHapley Additive exPlanations*) asignan a cada variable una contribución a la predicción de cada observación, ofreciendo explicaciones globales e individuales [14]. En el proyecto permitieron identificar de forma transparente qué factores socioeducativos inciden con mayor fuerza en el puntaje estimado.

1.3.3. Antecedentes

Los antecedentes que se presentan a continuación resumen estudios relevantes que abordan la predicción del desempeño académico desde diversas metodologías y contextos. Cada uno aporta elementos teóricos o técnicos para el desarrollo del presente proyecto, ya sea por las técnicas empleadas, los enfoques metodológicos o la manera en que gestionan los datos educativos.

Evaluación de modelos predictivos basados en CRISP-DM para el desempeño estudiantil en Colombia [8]. Acosta-Solano, Lancheros, Umaña y Coronado evaluaron, bajo la metodología CRISP-DM, varios modelos de aprendizaje automático para predecir el desempeño de estudiantes de educación media de la región Caribe colombiana en el Examen Saber 11° durante el periodo 2017–2019 [8]. La tarea se planteó como clasificación; tras seleccionar los atributos por ganancia de información, con variables como el tipo de establecimiento y el acceso a internet, el modelo más preciso fue el *Logistic Model Tree*, con una exactitud cercana al 73,6 %. Su aporte más relevante es la validación empírica de CRISP-DM en el mismo contexto regional y de prueba que aborda este proyecto, lo que respalda su adopción metodológica. La diferencia es que aquel trabajo clasifica el desempeño en categorías, mientras que el presente proyecto lo modela como un puntaje continuo mediante regresión, e integra el modelado con un enfoque explicativo y de apoyo a la política pública distrital.

Minería de datos para identificar factores vinculados al logro académico [7]. Martínez-Abad y Chaparro aplicaron técnicas de minería de datos, principalmente árboles de decisión, para detectar los factores asociados al rendimiento en evaluaciones a gran escala, sobre un diseño transversal con 18.935 estudiantes de educación media de 99 instituciones de Baja California (México) [7]. Concluyeron que los factores de orden personal resultaron más determinantes que los de orden escolar o social. Su aporte fundamental al presente proyecto es el énfasis en la interpretación de las variables, la analítica explicativa por encima de la meramente predictiva, orientación que aquí se retoma mediante los valores de Shapley. La diferencia está en el contexto local, el Distrito de Barranquilla, y en que este proyecto cuantifica la contribución de cada factor sobre un puntaje continuo, y no sobre una clasificación del logro.

Revisión sistemática sobre predicción del rendimiento en educación básica y media [17]. Rodrigues, dos Santos, Costa y Moreira realizaron una revisión sistemática de la literatura sobre la predicción del rendimiento académico en educación primaria y secundaria [17]. Entre sus hallazgos destacan que el interés investigativo sigue la tendencia del nivel universitario y que la exploración de datos semiestructurados con redes neuronales profundas es aún incipiente en estos niveles. Su aporte es el panorama de buenas prácticas internacionales y la justificación de la pertinencia de los modelos supervisados para contextos escolares. La diferencia es que constituye una síntesis bibliográfica, mientras que el presente proyecto realiza una implementación práctica, contextual y replicable sobre datos reales del Saber 11°.

Predicción del desempeño en pruebas estandarizadas con aprendizaje automático y valores de Shapley [22]. Suaza-Medina, Peñabaena-Niebles y Jubiz-Diaz, de la Universidad del Norte (Barranquilla), aplicaron nueve algoritmos de clasificación y valores de Shapley para identificar las variables que influyen en el desempeño en el Examen Saber 11° de las regiones rezagadas de Colombia [22]. Los modelos de mayor exactitud fueron las máquinas de potenciación del gradiente (*Extreme Gradient Boosting, Light Gradient Boosting y Gradient Boosting*), y los valores de Shapley señalaron como variables más influyentes el índice de nivel socioeconómico, el género, la región, la ubicación del establecimiento y la edad. Es el antecedente más cercano por dominio y técnica, y respalda directamente el uso conjunto de la potenciación del gradiente y los valores de Shapley sobre datos del Saber 11°. La diferencia central es la formulación de la tarea: dicho estudio clasifica el desempeño, mientras que el presente proyecto estima el puntaje global como variable continua mediante regresión, una tarea más granular; además, circunscribe el análisis al Distrito de Barranquilla frente al alcance nacional de las regiones rezagadas, antepone una segmentación no supervisada para explicar las diferencias de puntaje y aplica una validación temporal estricta entre cohortes.

Segmentación de estudiantes mediante agrupamiento [23]. Talib, Abd Majid y Sahran emplearon el agrupamiento K-Means junto con máquinas de soporte vectorial para identificar perfiles de comportamiento en estudiantes de educación superior; obtuvieron tres clústeres, desempeño bajo, medio y alto, y entrenaron un modelo de predicción por clúster que alcanzó mayor exactitud que el uso de las calificaciones originales como etiqueta [23]. Su aporte es metodológico: evidencia el valor del aprendizaje no supervisado para revelar perfiles de estudiantes a partir de sus características. Las diferencias son de contexto y de propósito: aquel trabajo se sitúa en la educación superior y utiliza los clústeres como insumo de un clasificador, mientras que el presente proyecto opera sobre la educación media y emplea la segmentación con una finalidad descriptiva y explicativa, construir una narrativa de por qué difieren los puntajes del Saber 11°, sin convertirla en una etapa de clasificación.

Mejora de la predicción del desempeño estudiantil con el algoritmo XGBoost [24]. Asselman, Khaldi y Aammou incorporaron el algoritmo *Extreme Gradient Boosting (*XGBoost*)* para mejorar el método *Performance Factors Analysis*, un enfoque de rastreo del conocimiento (*knowledge tracing*) propio de los sistemas de hipermedia educativa adaptativa, y mostraron que el ajuste de hiperparámetros incrementa de forma apreciable la exactitud de la predicción [24]. Su aporte al presente proyecto es el respaldo empírico a la elección de XGBoost y a la relevancia de la optimización de hiperparámetros, así como a su capacidad para jerarquizar la influencia de las variables. La diferencia es de dominio y de tarea: aquel trabajo predice el acierto en ejercicios dentro de entornos de aprendizaje adaptativo, mientras que el presente proyecto estima, mediante regresión, el puntaje global continuo del Saber 11° a partir de variables socioeducativas, con fines de apoyo a la política pública distrital.

2. DISEÑO METODOLÓGICO

Este capítulo describe el diseño metodológico del proyecto: las fuentes de información, el procedimiento de obtención, depuración y preparación de los datos, y los modelos y técnicas de ciencia de datos seleccionados para responder a cada objetivo específico. No se reportan aquí las actividades detalladas ni los resultados; estos se desarrollan en los capítulos posteriores, uno por cada objetivo específico. El propósito de este capítulo es ofrecer una visión integral de cómo se abordó el problema, de modo que se puedan comprender las decisiones metodológicas antes de examinar su ejecución y sus hallazgos.

2.1. Enfoque metodológico y entorno de trabajo

El proyecto se estructuró siguiendo el modelo de referencia CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que organiza el trabajo analítico en fases sucesivas e iterativas: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Esta estructura proporciona un marco ordenado que articula los cuatro objetivos específicos del proyecto y garantiza la trazabilidad del ciclo completo de los datos.

El desarrollo se realizó en lenguaje *Python* sobre cuadernos de cómputo modulares, organizados en correspondencia con los objetivos específicos: un cuaderno para la construcción, limpieza y selección de variables; uno para el descubrimiento de patrones; uno para el modelado predictivo; y uno para la evaluación, validación e interpretabilidad. Las principales librerías empleadas fueron pandas y *NumPy* para la manipulación de datos; *Matplotlib* y *Seaborn* para la visualización; *scikit-learn* para el preprocesamiento, la selección de características, los modelos de regresión y los esquemas de validación; *statsmodels* para las pruebas estadísticas y el cálculo del factor de inflación de la varianza; *prince* para el análisis factorial de datos mixtos; XGBoost para el modelo de potenciación del gradiente; y SHAP para la interpretabilidad.

Con el fin de favorecer la reproducibilidad de la lógica metodológica y de los procedimientos descritos, se fijó una semilla aleatoria común en todos los procesos estocásticos. Asimismo, todos los transformadores y modelos se ajustaron exclusivamente sobre los datos de entrenamiento y se aplicaron luego sobre los conjuntos de validación y prueba, lo que evita la fuga de información (*data leakage*) entre etapas.

2.2. Obtención de los datos

La fuente principal de información corresponde a los microdatos oficiales del Examen Saber 11°, disponibles en el repositorio DataICFES, administrado por el Instituto Colombiano para la Evaluación de la Educación (ICFES); su acceso requiere un proceso previo de registro. El repositorio aloja, para cada aplicación del examen, un archivo independiente en formato de texto

plano estructurado por período académico, con codificación UTF-8 y separador de punto y coma. A la fecha de consulta se encontraban disponibles los archivos correspondientes a los períodos comprendidos entre 2010-1 y 2024-2.

El análisis se delimitó al período comprendido entre 2014-2 y 2024-2. Esta decisión obedece a que, a partir de 2014-2, el Examen Saber 11° adoptó una estructura estable de cinco pruebas genéricas y una escala de calificación fija, lo que garantiza la comparabilidad de los puntajes entre cohortes. Geográficamente, los datos se descargaron para el departamento del Atlántico y se restringieron al Distrito de Barranquilla mediante un filtro sobre la variable que identifica el municipio de ubicación del establecimiento, de acuerdo con el diccionario oficial del examen.

La construcción del conjunto de datos maestro siguió un procedimiento ordenado, alineado con la fase de preparación de los datos de CRISP-DM. Cada archivo de período se leyó de forma individual, se homologaron los nombres de las columnas, se aplicó el filtro geográfico y se concatenaron los registros en una única base estructurada. Esta integración produjo el conjunto maestro sobre el cual se ejecutaron las fases posteriores de diagnóstico, limpieza y preparación.

2.3. Depuración y limpieza de los datos

Sobre el conjunto maestro se realizó un diagnóstico estructural y de calidad que comprendió el reconocimiento y tipado de las variables, su clasificación por familia semántica y un análisis de nulidad. Este último distinguió entre la nulidad estructural, que se entiende como ausencia de una variable en los períodos en que no se recolectó, y la nulidad genuina, que son valores faltantes dentro de períodos en que la variable sí existía, distinción determinante para definir el tratamiento posterior de cada caso.

Exclusión de variables. Se descartaron tres grupos de variables: las que inducen fuga de información (*data leakage*), como los percentiles y los niveles cualitativos de desempeño derivados directamente del puntaje; los identificadores administrativos y las variables descriptivas de alta cardinalidad o casi constantes tras el filtro geográfico; y los programas o indicadores con cobertura temporal insuficiente. Adicionalmente, se aplicó un corte temporal que excluyó los períodos anteriores a 2017-1, dado que concentraban una proporción elevada de nulidad estructural que comprometía la consistencia del conjunto.

Normalización y variable objetivo. Se aplicaron reglas de homologación para unificar categorías inconsistentes y se colapsaron los niveles de las variables ordinales con jerarquía natural. Se definió como variable objetivo el puntaje global (punt_global), de naturaleza numérica continua en la escala 0 a 500, lo que enmarca el problema como una tarea de regresión que estima directamente el puntaje de cada estudiante a partir de sus características socioeducativas.

Tratamiento diferenciado de faltantes. Cada variable se asignó a una estrategia de tratamiento según su origen y patrón de nulidad: a los registros con nulidad estructural se les asignó la etiqueta "Sin información", evitando imputaciones con la media o la moda global; algunas variables se imputaron mediante reglas del dominio (por ejemplo, la jornada se completó con una estrategia escalonada por establecimiento, período y calendario, y se empleó la moda dominante cuando superaba un umbral de concentración); y los casos atípicos en el grado y en la edad de presentación se inspeccionaron y corrigieron por contexto, preservando deliberadamente a la población adulta (educación para adultos, CLEI) en lugar de eliminarla con un filtro.

2.4. Preparación de los datos para el modelado

A partir del conjunto depurado se prepararon los datos según los requerimientos de cada técnica. Las variables ordinales se codificaron mediante mapeos numéricos que respetan su jerarquía natural; las variables nominales se transformaron mediante codificación *one-hot*; y las variables binarias se representaron como indicadores 0/1. Para los métodos basados en distancias, las variables se estandarizaron (media cero y desviación estándar uno). El preprocesamiento se encapsuló en transformadores y tuberías (pipelines) de *scikit-learn*, ajustados únicamente con los datos de entrenamiento, lo que blinda el flujo contra la fuga de información.

Partición temporal. Dado el carácter longitudinal de los datos, cohortes anuales sucesivas, con el efecto de la pandemia en 2020 y cambios en el cuestionario entre períodos, la partición no se hizo de forma aleatoria sino mediante un *split* temporal estricto por año de presentación: el conjunto de entrenamiento abarca las cohortes de 2017 a 2022, el de validación la cohorte de 2023 y el de prueba la cohorte de 2024. Este esquema respeta el orden cronológico, evita utilizar información del futuro para predecir el pasado y simula un escenario realista de despliegue, en el que el modelo se entrena con cohortes históricas y se aplica sobre una cohorte nueva.

2.5. Modelos y técnicas por objetivo específico

Cada objetivo específico se abordó con un conjunto de técnicas seleccionadas por su pertinencia frente al fenómeno estudiado y a la naturaleza de los datos. A continuación, se describen los métodos empleados en cada uno. Las actividades detalladas y los resultados correspondientes se presentan en los capítulos respectivos.

2.5.1. Objetivo específico 1: identificación de variables relevantes

El primer objetivo se abordó mediante análisis exploratorio y selección de características. El análisis univariado examinó la distribución de la variable objetivo y de cada predictor. El análisis bivariado midió la asociación de cada predictor con el puntaje global mediante la prueba apropiada según el tipo de variable: el coeficiente de correlación de Pearson para el predictor

numérico y el análisis de varianza de un factor (ANOVA), acompañado del tamaño del efecto eta cuadrado (η^2), para los predictores categóricos.

La selección de características combinó criterios estadísticos y algorítmicos. En el plano estadístico se descartaron las variables sin asociación relevante con el puntaje. En el plano algorítmico se calcularon dos medidas complementarias: la información mutua, que captura dependencias, incluso no lineales, entre cada predictor y el puntaje, y la importancia de variables derivada de un modelo de bosques aleatorios. Ambas medidas se normalizaron y se promediaron en una puntuación combinada, sobre la cual se aplicó un umbral de corte para retener las variables con poder predictivo medio-alto. Finalmente, se verificó la ausencia de multicolinealidad entre las variables seleccionadas mediante el factor de inflación de la varianza (VIF) y se inspeccionó la matriz de correlación. El subconjunto resultante constituyó el insumo de las fases de modelado.

2.5.2. Objetivo específico 2: descubrimiento de patrones mediante aprendizaje no supervisado

El segundo objetivo se abordó con técnicas de aprendizaje no supervisado, con una finalidad descriptiva y explicativa: descubrir y caracterizar perfiles de estudiantes que permitan construir una narrativa de por qué difieren los puntajes en el Examen Saber 11° del Distrito de Barranquilla. No se trata de clasificar a los estudiantes en categorías predefinidas, sino de revelar la estructura latente de la población.

Dado que los datos combinan variables cuantitativas y categóricas, la reducción de dimensionalidad se realizó mediante el análisis factorial de datos mixtos (en inglés, *Factor Analysis of Mixed Data*, FAMD), que equilibra la contribución de ambos tipos de variables en la construcción de los ejes factoriales y resulta más adecuado que el análisis de componentes principales aplicado sobre variables categóricas codificadas. Sobre las coordenadas factoriales se aplicó el algoritmo de agrupamiento K-Means. El número de clústeres se determinó combinando el método del codo, basado en la inercia o suma de cuadrados intra-clúster, y el coeficiente de silueta. La validez de la segmentación se evaluó internamente con el coeficiente de silueta y, de manera complementaria, con el coeficiente eta cuadrado (η^2) del puntaje global respecto a la pertenencia a los clústeres, que mide en qué medida los perfiles explican las diferencias de puntaje. El FAMD y el modelo de agrupamiento se ajustaron sobre el conjunto de entrenamiento y se aplicaron luego sobre la totalidad de los registros.

2.5.3. Objetivo específico 3: modelos predictivos de aprendizaje supervisado

El tercer objetivo se abordó como un problema de regresión sobre el puntaje global continuo. Como punto de comparación se definió un modelo de referencia (*baseline)* ingenuo que predice siempre el promedio del puntaje observado en el entrenamiento, y que establece el piso de desempeño que cualquier modelo debe superar para justificar su utilidad.

Sobre el mismo flujo de preprocesamiento se entrenaron cinco modelos que recorren una escalera de complejidad creciente: la regresión lineal por mínimos cuadrados ordinarios (OLS), la regresión Ridge (regularización L2), la regresión Lasso (regularización L1), los bosques aleatorios (Random Forest) y la potenciación del gradiente (XGBoost). Todos los candidatos se sometieron a un proceso de optimización de hiperparámetros con idéntico presupuesto de búsqueda, mediante búsqueda aleatoria (`RandomizedSearchCV`) mecanizada con validación cruzada temporal (`TimeSeriesSplit`), de modo que la comparación entre modelos se realizara en igualdad de condiciones. El modelo de mejor desempeño en el conjunto de validación se seleccionó como modelo final y se estimó su capacidad predictiva sobre el conjunto de prueba, que no intervino en ninguna etapa del entrenamiento ni de la selección de hiperparámetros. El desempeño se cuantificó con cuatro métricas complementarias: el coeficiente de determinación (R^2), el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el error porcentual absoluto medio (MAPE).

2.5.4. Objetivo específico 4: evaluación, validación e interpretabilidad

El cuarto objetivo se abordó mediante tres bloques de análisis sobre el modelo seleccionado. El primero es el diagnóstico de residuos, que examina los errores de predicción para detectar sesgos sistemáticos; comprende un análisis gráfico y tres pruebas estadísticas formales: la prueba de Shapiro-Wilk para la normalidad de los residuos, la prueba de Breusch-Pagan para la homocedasticidad y el estadístico de Durbin-Watson para la autocorrelación. La interpretación de estas pruebas se acompañó del tamaño del efecto, dado que con muestras grandes los contrastes tienden a rechazar los supuestos ante desviaciones de escasa relevancia práctica.

El segundo bloque es la validación por subgrupos relevantes para la política educativa, que verifica que la capacidad predictiva del modelo sea consistente entre poblaciones: por grado educativo (educación regular frente a educación de adultos/CLEI), por perfil socioeducativo (los clústeres obtenidos en el objetivo anterior) y por naturaleza del establecimiento (oficial frente a no oficial). Se incluyó además una validación a lo largo de los niveles de desempeño oficiales del ICFES, empleada como verificación del comportamiento del modelo y no como una tarea de clasificación.

El tercer bloque es la interpretabilidad mediante valores SHAP (*SHapley Additive exPlanations*), calculados con un *explainer* optimizado para modelos basados en árboles. La interpretabilidad global jerarquiza la influencia de cada variable sobre el puntaje predicho mediante el valor SHAP absoluto promedio y un diagrama de dispersión que revela la dirección del efecto; la interpretabilidad local descompone predicciones individuales en las contribuciones de cada variable, vista útil para que actores institucionales comprendan los factores detrás de un pronóstico. Sobre la base de estos análisis se seleccionó el modelo final y se documentaron sus limitaciones.

3. IDENTIFICAR LAS VARIABLES RELEVANTES PRESENTES EN LAS FUENTES OFICIALES DE INFORMACIÓN

Este capítulo presenta las actividades realizadas y los resultados obtenidos en relación con el primer objetivo específico del proyecto: identificar las variables relevantes presentes en las fuentes oficiales de información educativa que presentan relación significativa con el desempeño académico de los estudiantes del Distrito de Barranquilla. El trabajo se desarrolla en cinco etapas: (i) selección y caracterización de la fuente principal de datos; (ii) consolidación de un *dataset* maestro a partir de los archivos individuales por período; (iii) diagnóstico estructural y de calidad sobre el *dataset* consolidado; (iv) limpieza, normalización e imputación diferenciada; y (v) análisis exploratorio bivariado y selección de características para el enfoque de regresión adoptado en el proyecto. Las decisiones metodológicas tomadas en cada etapa se documentan junto con los criterios técnicos que las sustentan.

3.1. Fuente principal de datos: repositorio DataICFES

La fuente principal de información para este proyecto corresponde a los microdatos oficiales del Examen Saber 11°, disponibles en el repositorio DataICFES, administrado por el Instituto Colombiano para la Evaluación de la Educación (ICFES). El acceso a estos datos requiere un proceso previo de registro en el portal oficial del ICFES, en el formulario disponible en el siguiente enlace: https://www.icfes.gov.co/investigaciones/data-icfes/. Una vez realizado el registro y aceptadas las condiciones de uso de la información, el usuario accede al catálogo de datos abiertos, donde los conjuntos de datos se encuentran organizados por tipo de evaluación y período de aplicación.

Para el Examen Saber 11°, el repositorio aloja múltiples archivos en formato *.txt, estructurados por período académico. A la fecha de consulta (9 de febrero de 2026), se encontraban disponibles treinta archivos correspondientes a los períodos comprendidos entre 2010-1 y 2024-2, uno por cada aplicación semestral. Cada archivo contiene los microdatos individuales de los estudiantes que presentaron la prueba en el período respectivo. Para este proyecto se accedió a las bases departamentales del Atlántico [25], organizadas en el repositorio bajo el código DANE 8.

3.2. Delimitación temporal y geográfica

Aunque el repositorio ofrece información desde 2010-1, este proyecto delimita el análisis al período comprendido entre 2014-2 y 2024-2. Esta decisión obedece a que, como se documentó en el marco conceptual, a partir de 2014-2 el Examen Saber 11° adoptó una estructura estable de cinco pruebas genéricas con escalas estandarizadas y comparables en el tiempo. Incluir archivos anteriores introduciría inconsistencias en la interpretación longitudinal, ya que la guía metodológica del ICFES advierte que los resultados son comparables únicamente dentro de

períodos estructuralmente homogéneos [19]. En consecuencia, se excluyeron los nueve archivos correspondientes a períodos previos a 2014-2, conservando 21 archivos para procesamiento.

Adicionalmente, aunque los datos fueron descargados inicialmente para el departamento del Atlántico, se aplicó un filtro metodológico utilizando la variable *cole_mcpio_ubicacion*. De acuerdo con el diccionario oficial del examen [26], esta variable corresponde al municipio donde se encuentra ubicada la sede educativa, lo cual permite identificar con precisión el lugar de formación académica. Este criterio resulta fundamental, dado que el área metropolitana de Barranquilla incluye municipios cercanos como Soledad, Malambo y Puerto Colombia, cuyos estudiantes pueden presentar la prueba en sedes distintas a su municipio de residencia.

3.3. Consolidación del *dataset* maestro

La construcción del *dataset* maestro siguió un procedimiento ordenado, alineado con la fase de Preparación de los Datos del enfoque metodológico CRISP-DM. Cada archivo se leyó individualmente con codificación UTF-8 y separador punto y coma; se estandarizaron los nombres de columnas y se aplicó el filtro geográfico antes de concatenar verticalmente los *DataFrames* parciales. El fragmento siguiente resume la lógica central del procedimiento.

```
# -- Lectura, homologación de columnas y filtro geográfico --------------
lista_df = []
for archivo in archivos_objetivo:
    ruta = RUTA_DATOS_BRUTOS / archivo
    df_periodo = pd.read_csv(ruta, sep=SEP_ICFES, encoding=ENCODING_ICFES,
                             low_memory=False)
    df_periodo.columns = df_periodo.columns.str.strip().str.lower()
    if 'cole_mcpio_ubicacion' in df_periodo.columns:
        df_periodo = df_periodo[df_periodo['cole_mcpio_ubicacion']
                                .str.upper() == 'BARRANQUILLA']
    lista_df.append(df_periodo)
    print(f'  {archivo:35s} -> {len(df_periodo):>6,} registros en Barranquilla')
print(f'\nArchivos procesados: {len(lista_df)}')
```

La consolidación produjo una base estructurada con 183.984 registros individuales y 86 variables. Esta cifra representa una reducción significativa respecto a las aproximadamente 178 variables únicas identificadas en la auditoría inicial sobre los 30 archivos originalmente disponibles, reducción que se explica por la restricción al período 2014-2 a 2024-2, la exclusión de archivos previos a la reestructuración metodológica y la limitación geográfica al Distrito de Barranquilla.

3.4. Naturaleza del conjunto de datos consolidado

Desde el punto de vista estructural, el *dataset* es de tipo tabular: cada fila representa un estudiante que presentó el Examen Saber 11° y cada columna corresponde a una característica asociada al evaluado, a su entorno familiar o a la institución educativa donde cursó la educación media. El conjunto de datos no se limita a resultados del examen, sino que incorpora cuatro

grandes bloques de información: variables académicas de resultado, variables institucionales del establecimiento, variables sociodemográficas individuales del evaluado, y variables del contexto familiar.

Como punto de partida del análisis, la Tabla 1 describe las 86 variables que conforman el *dataset* maestro. Para cada una se especifica el nombre, la denominación en el *dataset*, la definición operacional, el tipo de variable y la escala de medición. La descripción responde a la naturaleza declarativa de las variables según el diccionario oficial del Examen Saber 11° [26] y a la estructura observada en el archivo consolidado.

Tabla 1. Variables del conjunto de datos consolidado del Examen Saber 11° (Barranquilla, período 2014-2 a 2024-2).

Nombre de la variable	Nombre en *dataset*	Definición operacional	Tipo de variable	Escala de medición
Periodo de aplicación	periodo	Período de presentación del examen Saber 11° (codificado AAAAS)	Categórica	Ordinal
ID público estudiante	estu_consecutivo	Identificador público anónimo del evaluado	Categórica	Nominal
Tipo de inscrito	estu_estudiante	Indica si presentó como estudiante o de manera individual	Categórica	Nominal
Tipo de documento	estu_tipodocumento	Tipo de identificación registrada	Categórica	Nominal
Área ubicación sede	cole_area_ubicacion	Zona rural o urbana de la sede	Categórica	Nominal
Colegio bilingüe	cole_bilingue	Indica si el establecimiento es bilingüe	Binaria	Nominal
Calendario académico	cole_calendario	Calendario A, B o Flexible	Categórica	Nominal
Carácter del colegio	cole_caracter	Naturaleza académica/técnica del colegio	Categórica	Nominal
Código DANE establecimiento	cole_cod_dane_establecimiento	Código oficial del establecimiento	Categórica	Nominal
Código DANE sede	cole_cod_dane_sede	Código oficial de la sede	Categórica	Nominal
Código departamento sede	cole_cod_depto_ubicacion	Código DANE del departamento sede	Categórica	Nominal
Código municipio sede	cole_cod_mcpio_ubicacion	Código DANE del municipio sede	Categórica	Nominal
Código ICFES sede	cole_codigo_icfes	Código interno ICFES sede-jornada	Categórica	Nominal

Nombre de la variable	Nombre en *dataset*	Definición operacional	Tipo de variable	Escala de medición
Departamento sede	cole_depto_ubicacion	Nombre del departamento sede	Categórica	Nominal
Género del colegio	cole_genero	Género de la población atendida por el colegio	Categórica	Nominal
Jornada del colegio	Jornada del colegio	Jornada académica	Categórica	Nominal
Municipio sede	cole_mcpio_ubicacion	Municipio donde estudia el estudiante	Categórica	Nominal
Naturaleza del colegio	Naturaleza del colegio	Oficial / No oficial	Binaria	Nominal
Nombre establecimiento	cole_nombre_establecimiento	Nombre del colegio	Categórica	Nominal
Nombre sede	cole_nombre_sede	Nombre de la sede	Categórica	Nominal
Sede principal	cole_sede_principal	Indica si es sede principal	Binaria	Nominal
Desempeño Inglés	desemp_ingles	Nivel cualitativo en Inglés	Categórica	Ordinal
Estudiante agregado	estu_agregado	Indicador técnico del ICFES sobre inclusión en informe nacional	Binaria	Nominal
Código departamento presentación	estu_cod_depto_presentacion	Código DANE departamento de presentación	Categórica	Nominal
Código municipio presentación	estu_cod_mcpio_presentacion	Código DANE municipio de presentación	Categórica	Nominal
Código departamento residencia	estu_cod_reside_depto	Código DANE departamento de residencia	Categórica	Nominal
Código municipio residencia	estu_cod_reside_mcpio	Código DANE municipio de residencia	Categórica	Nominal
Departamento presentación	estu_depto_presentacion	Nombre del departamento de presentación	Categórica	Nominal
Departamento residencia	estu_depto_reside	Nombre del departamento de residencia	Categórica	Nominal
Etnia	estu_etnia	Grupo étnico declarado	Categórica	Nominal
Fecha de nacimiento	estu_fechanacimiento	Fecha de nacimiento	Fecha	Intervalo
Género estudiante	estu_genero	Sexo reportado/registrado del estudiante; el conjunto de datos lo denomina 'género'	Binaria	Nominal

Nombre de la variable	Nombre en *dataset*	Definición operacional	Tipo de variable	Escala de medición
Grado presentado	Grado presentado	Grado matriculado en SIMAT al presentar la prueba	Categórica	Ordinal
INSE individual	estu_inse_individual	Índice Socioeconómico del evaluado	Numérica continua	Intervalo
Municipio presentación	estu_mcpio_presentacion	Municipio donde presentó el examen	Categórica	Nominal
Municipio residencia	estu_mcpio_reside	Municipio de residencia	Categórica	Nominal
Nacionalidad	estu_nacionalidad	Nacionalidad del estudiante	Categórica	Nominal
NSE individual	estu_nse_individual	Nivel socioeconómico categorizado	Categórica	Ordinal
País residencia	estu_pais_reside	País de residencia	Categórica	Nominal
Programa Pilo Paga	estu_pilopaga	Indicador beneficiario Ser Pilo Paga	Binaria	Nominal
Privado de libertad	estu_privado_libertad	Indica si está privado de la libertad	Binaria	Nominal
Número de cuartos del hogar	fami_cuartoshogar	Total de cuartos en la vivienda	Categórica	Ordinal
Educación de la madre	Educación de la madre	Nivel educativo de la madre	Categórica	Ordinal
Educación del padre	Educación del padre	Nivel educativo del padre	Categórica	Ordinal
Estrato vivienda	Estrato vivienda	Estrato socioeconómico	Categórica	Ordinal
Personas en el hogar	fami_personashogar	Número total de personas en el hogar	Categórica	Ordinal
Tiene automóvil	Tiene automóvil	Presencia de automóvil en el hogar	Binaria	Nominal
Tiene computador	Tiene computador	Computador en el hogar	Binaria	Nominal
Tiene internet	Tiene internet	Conectividad en el hogar	Binaria	Nominal
Tiene lavadora	fami_tienelavadora	Electrodoméstico en el hogar	Binaria	Nominal
Servicio de TV	fami_tieneserviciotv	Televisión cerrada	Binaria	Nominal
Puntaje Ciencias Naturales	punt_c_naturales	Puntaje en escala 0–100	Numérica continua	Intervalo
Puntaje Global	punt_global	Puntaje global 0–500	Numérica continua	Intervalo
Puntaje Inglés	punt_ingles	Puntaje en escala 0–100	Numérica continua	Intervalo

Nombre de la variable	Nombre en *dataset*	Definición operacional	Tipo de variable	Escala de medición
Puntaje Lectura Crítica	punt_lectura_critica	Puntaje en escala 0–100	Numérica continua	Intervalo
Puntaje Matemáticas	punt_matematicas	Puntaje en escala 0–100	Numérica continua	Intervalo
Puntaje Sociales y Ciudadanas	punt_sociales_ciudadanas	Puntaje en escala 0–100	Numérica continua	Intervalo
Número de libros en el hogar	Número de libros en el hogar	Cantidad de libros en el hogar	Categórica	Ordinal
Tiene etnia	estu_tieneetnia	Indicador de pertenencia étnica	Binaria	Nominal
Desempeño Ciencias	desemp_c_naturales	Nivel cualitativo en Ciencias Naturales	Categórica	Ordinal
Desempeño Lectura	desemp_lectura_critica	Nivel cualitativo en Lectura Crítica	Categórica	Ordinal
Desempeño Matemáticas	desemp_matematicas	Nivel cualitativo en Matemáticas	Categórica	Ordinal
Desempeño Sociales	desemp_sociales_ciudadanas	Nivel cualitativo en Sociales y Ciudadanas	Categórica	Ordinal
Discapacidad	estu_discapacidad	Reporte de discapacidad	Binaria	Nominal
NSE establecimiento	NSE establecimiento	Índice socioeconómico del colegio	Categórica	Ordinal
Repitencia	estu_repite	Indicador de repitencia	Binaria	Nominal
Percentil Ciencias	percentil_c_naturales	Percentil en la prueba	Categórica	Ordinal
Percentil Global	percentil_global	Percentil del puntaje global	Categórica	Ordinal
Percentil Inglés	percentil_ingles	Percentil en la prueba	Categórica	Ordinal
Percentil Lectura	percentil_lectura_critica	Percentil en la prueba	Categórica	Ordinal
Percentil Matemáticas	percentil_matematicas	Percentil en la prueba	Categórica	Ordinal
Percentil Sociales	percentil_sociales_ciudadanas	Percentil en la prueba	Categórica	Ordinal
Dedicación a internet	Dedicación a internet	Rango ordinal de tiempo de uso de internet (no es un tiempo continuo)	Categórica	Ordinal
Dedicación a lectura	Dedicación a lectura	Rango ordinal de tiempo de lectura diaria (no es un tiempo continuo)	Categórica	Ordinal

Nombre de la variable	Nombre en *dataset*	Definición operacional	Tipo de variable	Escala de medición
Horas trabajadas a la semana	estu_horassemanatrabaja	Rango ordinal de horas de trabajo semanales (no es un tiempo continuo)	Categórica	Ordinal
Tipo de remuneración	estu_tiporemuneracion	Modalidad de pago	Categórica	Nominal
Consumo de proteína animal	fami_comecarnepescadohuevo	Frecuencia de consumo	Categórica	Ordinal
Consumo de cereales	fami_comecerealfrutoslegumbre	Frecuencia de consumo	Categórica	Ordinal
Consumo de lácteos	Consumo de lácteos	Frecuencia de consumo	Categórica	Ordinal
Percepción situación económica	fami_situacioneconomica	Percepción del hogar	Categórica	Ordinal
Consola de videojuegos	fami_tieneconsolavideojuegos	Presencia de consola en el hogar	Binaria	Nominal
Horno microondas	fami_tienehornomicroogas	Presencia de horno microondas	Binaria	Nominal
Motocicleta	fami_tienemotocicleta	Presencia de motocicleta	Binaria	Nominal
Ocupación de la madre	Ocupación de la madre	Tipo de ocupación de la madre	Categórica	Nominal
Ocupación del padre	Ocupación del padre	Tipo de ocupación del padre	Categórica	Nominal
Programa Generación E	estu_generacione	Indicador de Generación E	Categórica	Ordinal

Fuente: elaboración propia con base en el diccionario oficial DataICFES — *Examen Saber 11°* [26].

3.5. Diagnóstico estructural y de calidad del *dataset* maestro

Una vez consolidado el *dataset*, se realizó un diagnóstico de la estructura con el fin de comprender la composición del archivo, detectar problemas de tipado, evaluar la presencia de valores nulos y caracterizar el comportamiento de las variables clave. Este diagnóstico se desarrolló en dos bloques complementarios.

3.5.1. Reconocimiento estructural y tipado

Las 86 columnas del *dataset* se organizaron taxonómicamente según el prefijo semántico definido por el ICFES, lo que permite asignar cada variable a una de seis familias: Estudiante (estu_), Familia (fami_), Colegio (cole_), Puntaje (punt_), Desempeño (desemp_) y Percentil (percentil_), además de la variable temporal periodo. La auditoría de tipos detectó 17 variables que *pandas* infirió

automáticamente como numéricas (int64 o float64) pero que, según el diccionario oficial, son categóricas, incluyendo códigos DANE e ICFES, niveles de desempeño cualitativos (1–4), estu_nse_establecimiento, estu_grado y periodo. Todas estas variables fueron casteadas a *object* para evitar análisis erróneos como el cálculo de promedios sobre códigos nominales. Los percentiles, así como punt_ingles y estu_inse_individual, se convirtieron a *Int64* (entero *nullable*), preservando la posibilidad de representar valores faltantes.

Como complemento al casteo, se construyeron tres variables derivadas de alto valor analítico, cada una con su tipo semántico explícitamente declarado:

- estu_es_repitente. Variable categórica binaria nominal (Si / No) derivada de estu_repite. Aunque *pandas* leería un entero binario como tipo numérico, semánticamente no admite operaciones aritméticas: promediar "ser repitente" no tiene sentido. Por esta razón se castea explícitamente a *object*.

- anio_presentacion. Variable categórica nominal (2017, 2018, …, 2024). Aunque numéricamente es un entero, el efecto del año sobre el puntaje no es monotónico: la pandemia de 2020 introdujo una discontinuidad estructural en la distribución de puntajes y los cambios de cuestionario y logística del ICFES producen efectos categóricos por cohorte, no una tendencia lineal. Tratarla como entero forzaría a los modelos a asumir linealidad temporal que la evidencia refuta. Por ello se castea a object para que se trate como variable de cohorte.

- edad_presentacion. Variable cuantitativa continua (Int64). Diferencia en años entre el año de presentación y el año de nacimiento. A diferencia de las dos anteriores, aquí la magnitud aritmética sí es significativa. Es el único predictor genuinamente numérico continuo del *dataset* analítico.

La distinción anterior es intencional y constituye una corrección metodológica respecto a un enfoque basado únicamente en el *dtype* de pandas: la elección del tipo y de la prueba estadística posterior se basa en la semántica de la variable, no en cómo el motor de lectura la interpreta por defecto.

3.5.2. Diagnóstico de nulidad y distinción entre nulos estructurales y genuinos

El análisis identificó 61 de las 89 columnas (incluyendo derivadas) con presencia de valores nulos, totalizando 1.827.613 celdas faltantes equivalentes a un 11,16% del total. Sin embargo, una inspección por período permitió identificar que la mayoría de estos nulos no corresponden a respuestas omitidas por los estudiantes, sino a ausencia estructural: la pregunta no formaba parte del cuestionario del ICFES en ciertos períodos.

Los nulos del *dataset* se clasificaron en tres categorías:

- **Ausencia en períodos tempranos** (27 variables). Variables introducidas en cuestionarios posteriores. Los niveles de desempeño desemp_*, los percentiles percentil_*, estu_discapacidad, estu_nse_establecimiento y estu_repite aparecen desde 2016-1; las variables de alimentación, tecnología y trabajo (fami_come*, fami_tiene*, estu_dedicacion*, fami_trabajolabor*) desde 2017-1; y fami_numlibros desde 2015-1.
- **Ausencia en períodos puntuales** (2 variables). estu_etnia carece de datos en 2020-2; estu_tieneetnia en 2014-2, 2015-1 y 2023-2.
- **Programas temporales** (2 variables). estu_generacione (Programa Generación E) solo tiene datos en 6 períodos (2018-2 a 2021-2), y estu_pilopaga (Ser Pilo Paga) en 8 períodos (2014-2 a 2018-1).

La descomposición cuantitativa confirma que 1.372.998 nulos (75,1 %) son de origen estructural, y 454.615 (24,9 %) son nulos genuinos. Esta distinción es crítica: imputar nulos estructurales con la media o moda global contaminaría cualquier modelo al mezclar datos inexistentes con datos reales. Por ello, en la fase de limpieza se adoptó un tratamiento diferenciado para cada tipo de nulidad, como se muestra en la Tabla 2. Todos los faltantes fueron tratados o eliminados, teniendo 0 nulos al finalizar la etapa de limpieza, normalización e imputación.

Tabla 2. Diagnóstico y tratamiento de los valores faltantes.

Variable	Registros imputados y/o corregidos	Tipo	Estrategia de tratamiento	Justificación
estu_tieneetnia	33.121 (24,72%)	Estructural y genuino	Etiqueta 'Sin información' en períodos sin la variable y cruces lógicos con la pertenencia a grupos étnicos	Preserva la ausencia estructural y evita imputar con la moda global
cole_bilingue	19.717 (14,72%)	Genuino	Reglas deterministas (oficial+calendario A → 'N'; sedes bilingües conocidas → 'S')	Regla basada en el tipo y el calendario del establecimiento
fami_numlibros	11.877 (8,86%)	Estructural y genuino	Imputación multinivel (sede+período+estrato → estrato+período)	Introducida en 2015-1; moda del contexto de sede y estrato
fami_estratovivienda	11.168 (8,34%)	Genuino	Imputación jerárquica en tres niveles (sede+período+NSE → sede+período → sede)	Moda del contexto institucional más específico disponible

Variable	Registros imputados y/o corregidos	Tipo	Estrategia de tratamiento	Justificación
fami_comecerealfrutoslegumbre	10.624 (7,93%)	Estructural y genuino	Imputación multinivel (sede+período+NSE → sede+período → estrato+período)	Patrón de consumo asociado al contexto socioeconómico de la sede
fami_comelechederivados	10.552 (7,88%)	Estructural y genuino	Imputación multinivel (sede+período+NSE → sede+período → estrato+período)	Patrón de consumo asociado al contexto socioeconómico de la sede
estu_dedicacioninternet	10.202 (7,61%)	Estructural y genuino	Imputación multinivel (sede+período → naturaleza+jornada+período)	Hábito asociado al entorno de la sede y la jornada
fami_comecarnepescadohuevo	10.144 (7,57%)	Estructural y genuino	Imputación multinivel (sede+período+NSE → sede+período → estrato+período)	Patrón de consumo asociado al contexto socioeconómico
fami_tieneserviciotv	10.102 (7,54%)	Genuino	Regla NSE (Tabla 2 del ICFES) e imputación multinivel (sede+período+estrato → estrato+período)	Tenencia de activos coherente con el nivel socioeconómico
estu_dedicacionlecturadiaria	9.879 (7,37%)	Estructural y genuino	Imputación multinivel (sede+período → naturaleza+jornada+período)	Hábito asociado al entorno de la sede y la jornada
fami_tieneinternet	9.700 (7,24%)	Genuino	Regla NSE (Tabla 2 del ICFES) e imputación multinivel	Tenencia de activos coherente con el nivel socioeconómico
fami_educacionmadre	9.670 (7,22%)	Genuino	Imputación multinivel (sede+período+estrato → naturaleza+estrato+período)	Nivel educativo parental homogéneo dentro de la sede y el estrato
fami_educacionpadre	9.521 (7,11%)	Genuino	Imputación multinivel (sede+período+estrato → naturaleza+estrato+período)	Nivel educativo parental homogéneo dentro de la sede y el estrato
estu_tiporemuneracion	7.659 (5,72%)	Estructural y genuino	Imputación multinivel (sede+período → naturaleza+jornada+período)	Asociado al entorno de la sede y la jornada
fami_trabajolaborpadre	7.533 (5,62%)	Estructural y genuino	Imputación multinivel (sede+período+estrato → estrato+naturaleza+período)	Ocupación parental asociada al contexto de la sede y el estrato

Variable	Registros imputados y/o corregidos	Tipo	Estrategia de tratamiento	Justificación
fami_situacioneconomica	7.407 (5,53%)	Estructural y genuino	Imputación multinivel (sede+período+NSE → sede+período → estrato+período)	Asociada al contexto socioeconómico de la sede
estu_horassemanatrabaja	7.389 (5,51%)	Estructural y genuino	Imputación multinivel (sede+período → naturaleza+jornada+período)	Asociado al entorno de la sede y la jornada
fami_tieneautomovil	7.333 (5,47%)	Genuino	Regla NSE (Tabla 2 del ICFES) e imputación multinivel	Tenencia de activos coherente con el nivel socioeconómico
fami_trabajolabormadre	7.313 (5,46%)	Estructural y genuino	Imputación multinivel (sede+período+estrato → estrato+naturaleza+período)	Ocupación parental asociada al contexto de la sede y el estrato
fami_tienecomputador	7.199 (5,37%)	Genuino	Regla NSE (Tabla 2 del ICFES) e imputación multinivel	Tenencia de activos coherente con el nivel socioeconómico
fami_cuartoshogar	7.136 (5,33%)	Genuino	Imputación multinivel (sede+período+estrato → estrato+período)	Asociado al contexto de la sede y el estrato
fami_tienelavadora	7.029 (5,25%)	Genuino	Regla NSE (Tabla 2 del ICFES) e imputación multinivel	Tenencia de activos coherente con el nivel socioeconómico
fami_personashogar	6.889 (5,14%)	Genuino	Imputación multinivel (sede+período+estrato → estrato+período)	Asociado al contexto de la sede y el estrato
cole_caracter	3.087 (2,30%)	Genuino	Reglas y códigos DANE específicos	Carácter académico asignado por regla institucional
edad_presentacion	196 (0,15%)	Genuino (atípicos)	Marcado como faltante y recuperación por rango plausible según el grado (11 vs 26)	Edades imposibles (p. ej., 0 o 123 años) corregidas por contexto de grado
cole_jornada	29 (0,02%)	Genuino	Moda por DANE+período+calendario y, en su defecto, moda institucional	Atributo institucional estable en el tiempo
estu_nse_establecimiento	25 (0,02%)	Estructural y genuino	Moda histórica por sede	Índice contextual estable por establecimiento

Variable	Registros imputados y/o corregidos	Tipo	Estrategia de tratamiento	Justificación
estu_es_repitente	10 (0,01%)	Genuino	Derivada de estu_repite y etiqueta 'Sin información' en remanentes	Consolidación de la variable de repitencia

Fuente: elaboración propia

3.6. Limpieza, normalización e imputación del *dataset*

Con base en los hallazgos del diagnóstico, se diseñó un proceso de limpieza estructurado en tres bloques: normalización de inconsistencias categóricas, definición del rol analítico de cada variable y tratamiento diferenciado de faltantes. El producto de esta fase es el *dataset* limpio.

3.6.1. Normalización de inconsistencias categóricas

Para resolver las inconsistencias detectadas en el diagnóstico, se aplicaron reglas de homologación específicas. Los valores 'NSE1' y '1.0' (y equivalentes para los demás niveles) se unificaron al rótulo numérico estandarizado. La variable estu_tieneetnia se homologó a las categorías 'Si', 'No' y 'Sin información'. Las variables fami_personashogar y fami_cuartoshogar, que mezclaban rótulos textuales y rangos, se mapearon a un único esquema ordinal coherente.

3.6.2. Definición de la variable objetivo

El proyecto adopta un enfoque de regresión, dado que la variable objetivo punt_global es numérica continua en escala 0 a 500. Este enfoque permite estimar directamente el puntaje de cada estudiante a partir de sus características socioeducativas.

3.6.3. Tratamiento diferenciado de variables y faltantes

Cada variable se asignó a una de las siguientes categorías de acción, según su origen y su patrón de nulidad:

- Variables descartadas. Identificadores, redundancias geográficas, derivados de los puntajes (percentiles y desempeños cualitativos, que generan *data leakage*) y programas temporales con cobertura insuficiente.
- Variables conservadas con etiqueta especial. Variables con nulidad estructural cuyo contenido es valioso para los períodos en que sí se recolectaron. En sus registros nulos se asignó la etiqueta 'Sin información', evitando imputación con media o moda global.
- Variables imputadas mediante reglas del dominio. cole_jornada se imputó mediante una estrategia escalonada por establecimiento, período y calendario; cole_bilingue se completó asignando 'N' a colegios oficiales calendario A; y cole_caracter se validó contra

el SINEB del MEN para reclasificar registros marcados como 'NO APLICA' que correspondían a colegios con carácter académico definido.

- Variables corregidas por validación cruzada. Los registros con estu_grado = 12 se inspeccionaron caso por caso: en instituciones consistentemente reportadas en grado 11 se interpretó como error de digitación; en colegios calendario B se conservó. Los *outliers* extremos en edad_presentacion se identificaron como errores de registro y se trataron diferenciadamente.

Como producto de este proceso, el *dataset* limpio resultante conserva los 183.984 registros originales, pero reduce el conjunto de variables analíticas de 86 a 41 columnas. La Tabla 3 sintetiza las variables eliminadas y la justificación de cada decisión.

Tabla 3. Variables eliminadas durante el proceso de limpieza y justificación.

Variable o grupo	Razón de eliminación
estu_consecutivo, estu_tipodocumento	Identificadores administrativos sin valor predictivo.
Códigos DANE/ICFES de establecimiento, sede, departamento y municipio	Identificadores nominales redundantes con la delimitación geográfica ya aplicada.
Variables de geolocalización tras filtro a Barranquilla (cole_nombre_*, cole_depto_ubicacion, cole_mcpio_ubicacion, estu_*_presentacion, estu_*_reside, estu_pais_reside, estu_nacionalidad)	Constantes o casi-constantes tras el filtro geográfico al Distrito de Barranquilla.
estu_fechanacimiento	Reemplazada por la variable derivada edad_presentacion.
estu_repite	Reemplazada por la variable binaria derivada estu_es_repitente.
estu_agregado, estu_estudiante, estu_privado_libertad	Indicadores técnicos o categorías con frecuencia ínfima.
estu_generacione, estu_pilopaga	Programas temporales con cobertura limitada (6 y 8 períodos).
percentil_global, percentil_c_naturales, percentil_ingles, percentil_lectura_critica, percentil_matematicas, percentil_sociales_ciudadanas	Derivados directos del puntaje (*data leakage*).
desemp_c_naturales, desemp_ingles, desemp_lectura_critica, desemp_matematicas, desemp_sociales_ciudadanas	Niveles cualitativos derivados de los puntajes (*data leakage*).

Fuente: elaboración propia.

3.7. Análisis univariado

Con el *dataset* ya depurado se examinó la distribución individual de los 42 predictores, 41 categóricos y 1 numérico, y de la variable objetivo del enfoque de regresión, Puntaje global. El

análisis univariado persigue tres fines: verificar que la limpieza eliminó las anomalías detectadas en el diagnóstico estructural, caracterizar la forma, dispersión y asimetría de cada variable para anticipar el comportamiento de las pruebas bivariadas, e identificar categorías minoritarias o concentraciones extremas que pudieran condicionar la modelación.

La variable objetivo Puntaje Global (N = 133.985), como se muestra en la Figura 1. Distribución de la variable Puntaje global sobre el *dataset* de modelado.~~Figura 1. Distribución de la variable Puntaje global sobre el *dataset* de modelado.~~, tiene media 254,65 y mediana 251, con desviación estándar 53,50 sobre el rango teórico 0 a 500. La distribución es ligeramente asimétrica a la derecha (asimetría = 0,30) y algo más plana que una normal (curtosis = −0,46). Estos valores son coherentes con la referencia teórica del ICFES para 20142 (media 250, desviación 50): el rango intercuartílico va de 213 (P25) a 293 (P75).

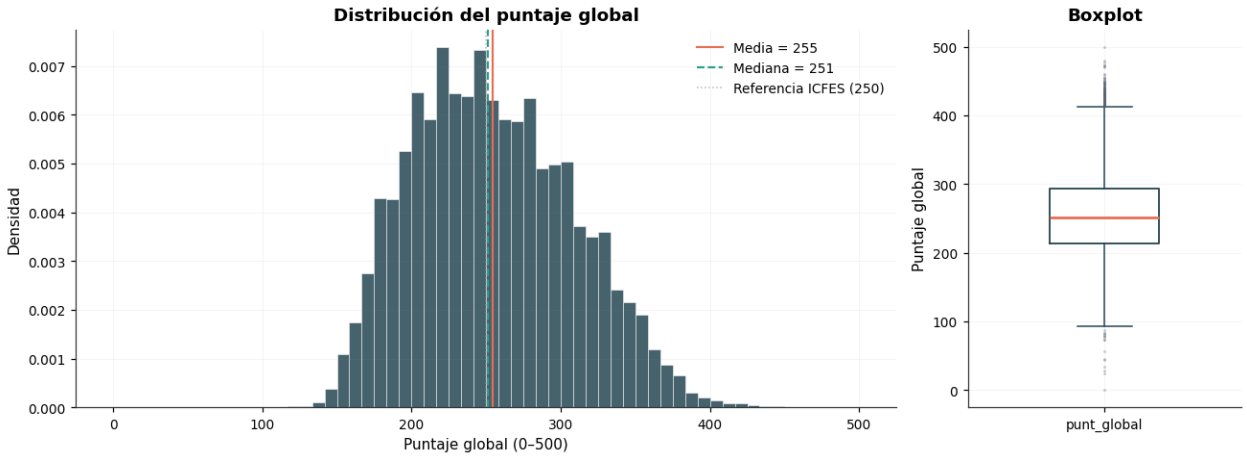


Figura 1. Distribución de la variable Puntaje global sobre el *dataset* de modelado.

El único predictor genuinamente cuantitativo continuo es la edad de presentación (rango 12 a 72 años). La distribución, como se muestra en la Figura 2 es marcadamente leptocúrtica (curtosis = 146,29; asimetría = 9,84): la mediana se sitúa en 17 años, el rango intercuartílico es de apenas 1 año (Q1 = 17, Q3 = 18) y la cola superior se extiende de manera atenuada pero persistente hasta los 72 años. Esa cola corresponde principalmente a los estudiantes de educación para adultos identificados en el diagnóstico estructural, cuya presencia justifica la inclusión simultánea de la edad de presentación y el grado de presentación como predictores diferenciados.

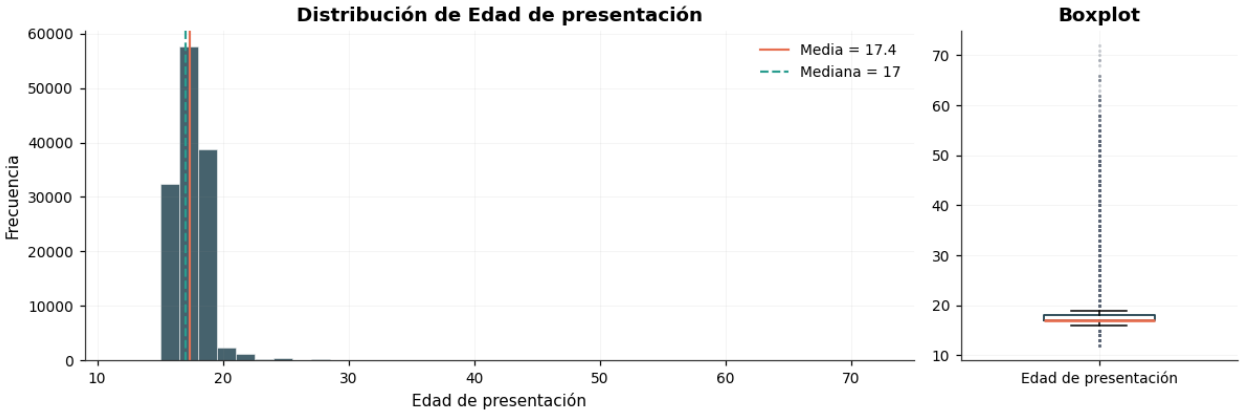


Figura 2. Distribución de la variable Edad de presentación sobre el *dataset* de modelado.

3.7.1. Variables categóricas predictoras

El inventario de cardinalidad sobre los 41 predictores categóricos arrojó la siguiente composición: 18 variables binarias (2 categorías), 15 de cardinalidad baja (3 a 5 categorías) y 7 de cardinalidad media (6 a 10 categorías). El bloque binario agrupa los indicadores de tenencia de bienes y servicios del hogar (Tiene computador, Tiene internet, entre otros) y características institucionales como Naturaleza del colegio, Calendario académico o Colegio bilingüe. El bloque de cardinalidad media incluye Jornada del colegio (6 categorías), los niveles educativos de los padres (Educación de la madre y Educación del padre, ambos con 6 categorías tras la reagrupación), Estrato vivienda (6 estratos), las variables de trabajo de los padres (Ocupación de la madre y Ocupación del padre, 7 categorías cada una) y Año de presentación (8 períodos).

Los hallazgos del análisis univariado se resumen a partir de la moda (categoría más frecuente) de cada variable (véase el Anexo A). En el plano institucional, la jornada predominante es la única, que concentra el 36,9% de los registros. En el grado escolar, el 89,1% de los estudiantes cursa educación regular (grado 11/12) frente al 10,9 % de educación de adultos (grado 25/26). En el contexto familiar, el nivel educativo modal de madres y padres es secundaria (42,8% y 43,6%, respectivamente), el estrato modal es el 1 (36,9% de los hogares) y el 46,8% de los estudiantes reporta tener entre 0 y 10 libros en casa. En cuanto a hábitos, el 37,5% de los estudiantes dedica 30 minutos o menos a la lectura diaria y el 33,3% usa internet entre 1 y 3 horas al día. Estas frecuencias caracterizan al estudiante típico del Distrito de Barranquilla en el período de modelado y establecen el contexto sociodemográfico sobre el que operarán los modelos predictivos.

3.8. Análisis exploratorio bivariado

Sobre el conjunto de datos limpio se midió la asociación entre cada predictor candidato y Puntaje Global con la prueba apropiada según el tipo de variable: calificación de Pearson para el único

predictor numérico (Edad de presentación) y ANOVA de un factor con tamaño de efecto Eta cuadrado (η^2) para los 41 predictores categóricos. Se privilegió η^2 sobre el estadístico F o el p-valor porque, con una muestra de 133.985 registros, casi cualquier prueba alcanza significancia estadística, aunque el efecto sea trivial; el tamaño del efecto da una interpretación independiente del tamaño muestral.

Para el predictor numérico, la clasificación de Pearson entre Edad de presentación y Puntaje Global fue $r=-0{,}233$ ($p<0{,}001$), una asociación negativa débil pero significativa: a mayor edad de presentación, menor puntaje promedio, en línea con la presencia de población adulta y de trayectorias escolares interrumpidas.

En el ranking de predictores categóricos por η^2, las cuatro variables con efecto grande ($\eta^2 \geq 0{,}14$) fueron NSE establecimiento ($\eta^2=0{,}229$), Jornada del colegio ($\eta^2= 0{,}191$), Educación de la madre ($\eta^2=0{,}159$) y Educación del padre ($\eta^2=0{,}144$). El predictor Grado presentado quedó en el rango de efecto mediano ($\eta^2=0{,}103$). La Figura 3 muestra los *boxplots* de Puntaje Global para los seis predictores con mayor η^2, lo que permite leer el gradiente de puntaje a través de las categorías de cada variable.

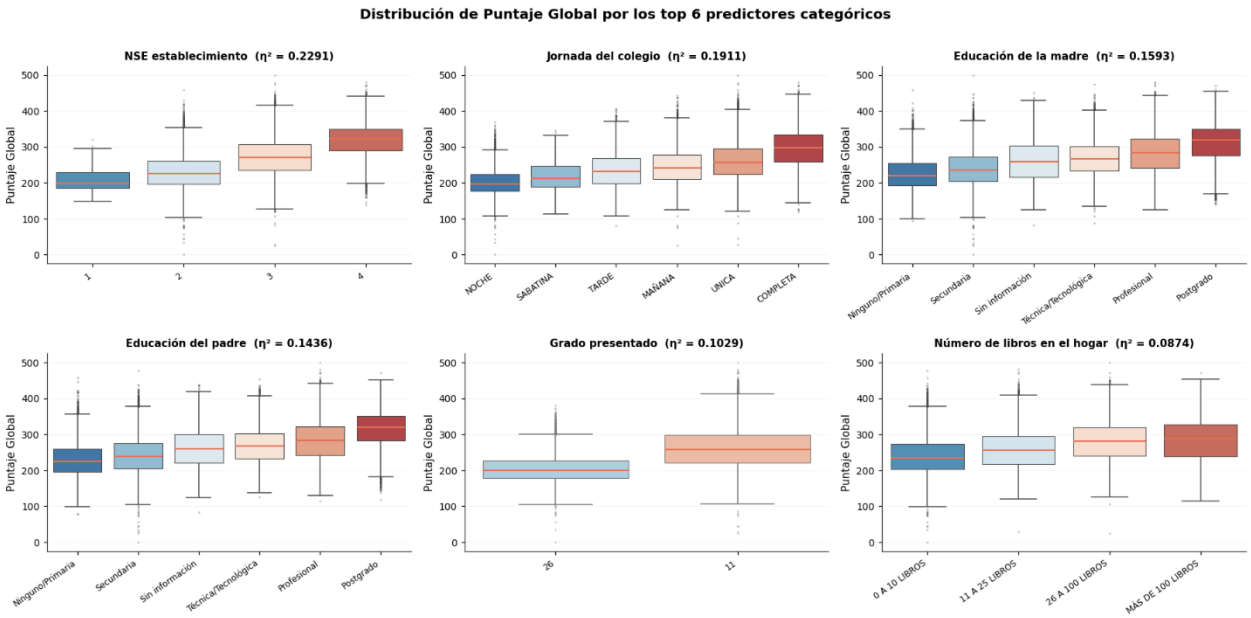


Figura 3. *Boxplots* de Puntaje Global para los seis predictores categóricos con mayor η^2.

El contraste por jornada escolar evidencia diferencias sustanciales entre las categorías de Jornada del colegio, la segunda variable del ranking de selección combinado (sección 3.9). El contraste por Grado presentado opone los dos contextos educativos consolidados, educación regular (grado

11/12) frente a educación de adultos (grado 25/26), y confirma que el grado capta un mecanismo institucional relevante, no una variable trivial.

3.9. Selección de características

La selección combinó criterios estadísticos y algorítmicos. En el plano estadístico, el análisis bivariado de la sección 2.8 ya había filtrado las variables sin asociación relevante con Puntaje Global ($\eta^2<0,01$ o $|r|<0,1$). En el plano algorítmico se aplican dos métodos complementarios, ejecutados sobre el conjunto de datos completo con un propósito exploratorio de ranking (la validación con partición temporal se reserva para el modelado). Como verificación de que la selección definitiva no depende de información de las cohortes futuras, el ranking se recalculó utilizando únicamente el conjunto de entrenamiento (cohortes 2017–2022): las 17 variables seleccionadas se conservan sin excepción y la correlación de rangos con el ranking sobre el conjunto completo es muy alta (ρ de Spearman = 0,996); la única diferencia es que una variable situada en el límite del umbral cruza el corte en la versión de entrenamiento, lo que confirma la robustez de la selección frente a una posible fuga indirecta de información. Los dos métodos algorítmicos aplicados son:

- **Información Mutua (*Mutual Information*).** Mide la dependencia estadística general entre cada predictor y la variable objetivo, capturando relaciones no lineales que la correlación de Pearson no detecta.
- **Random Forest *Feature Importance*.** Mide la importancia de cada variable como reducción de impureza (varianza) en árboles aleatorios. Se entrenó un Random Forest *Regressor* sobre el *dataset* limpio con codificación ordinal de las variables categóricas.

Las dos métricas se normalizaron al rango [0, 1] y se promediaron para obtener una puntuación combinada por variable. Se aplicó un umbral de corte de Score ≥ 0,10, que retiene las variables con poder predictivo combinado mediano-alto y descarta las de aporte marginal. Bajo este criterio quedaron 17 variables seleccionadas de las 42 candidatas, encabezadas por NSE establecimiento (Score = 1,000), Jornada del colegio (0,445), Educación de la madre (0,376), Edad de presentación (0,365) y Grado presentado (0,336).

Sobre el subconjunto seleccionado se calculó el Factor de Inflación de la Varianza (VIF) para detectar redundancia. Todos los predictores presentaron VIF por debajo de 1,79, es decir, dentro del rango de colinealidad baja o nula; no se encontró ninguna variable con VIF superior a 5, por lo que no fue necesario eliminar ni combinar predictores por multicolinealidad. La estructura de correlaciones se confirma adicionalmente con una matriz de Spearman entre las variables seleccionadas, como se muestra en la Figura 4~~Figura 4~~.

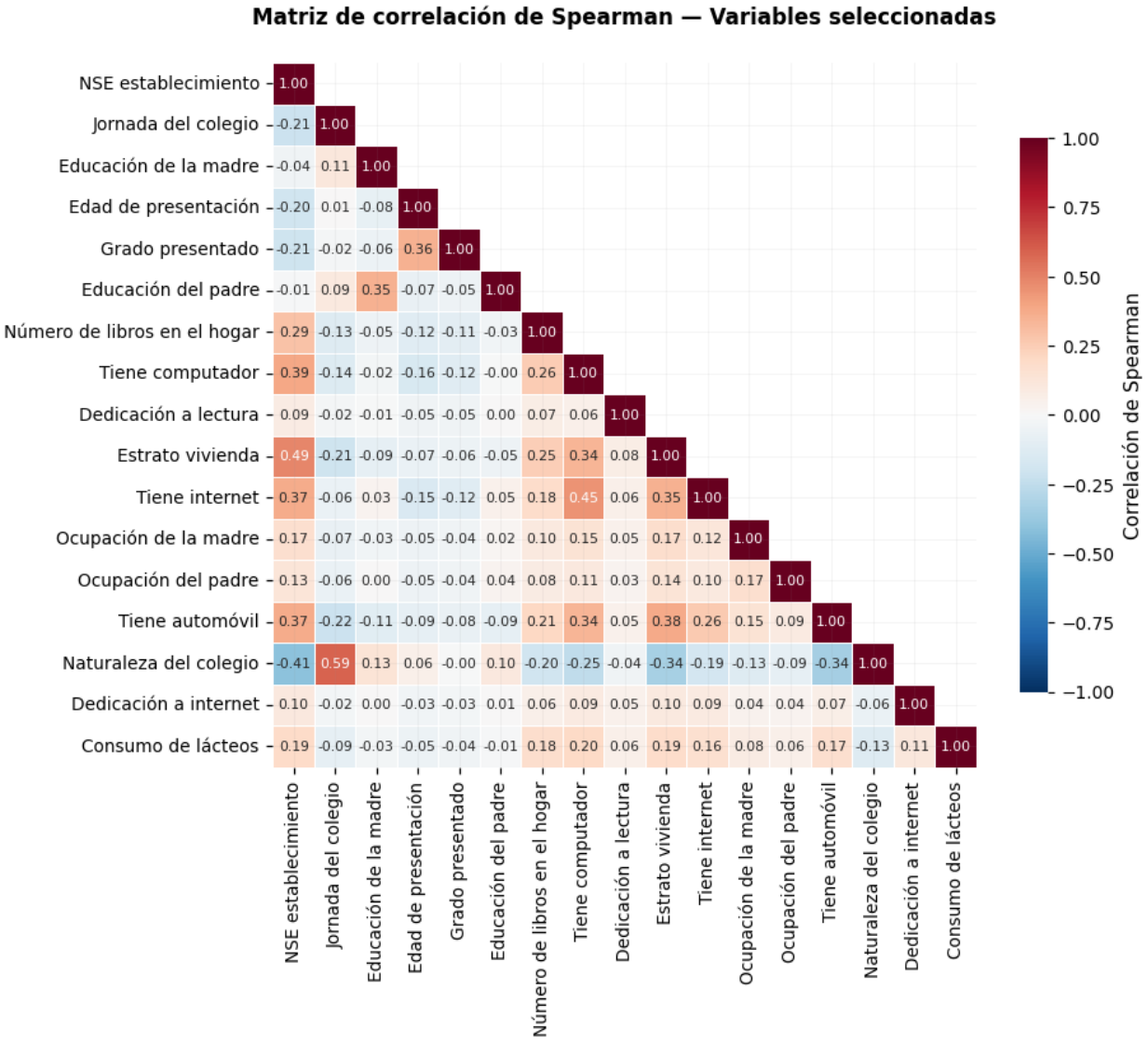


Figura 4. ***Matriz de correlación de*** **Spearman** **–** ***Variables seleccionadas.***

La lectura de la matriz refuerza la conclusión del diagnóstico de multicolinealidad y, a la vez, hace visible la estructura latente de los factores asociados al logro educativo. Las asociaciones más fuertes se concentran en el bloque socioeconómico e institucional: el nivel socioeconómico del establecimiento se relaciona de forma moderada con el estrato de la vivienda (ρ = 0,49) y con la tenencia de bienes del hogar, computador (ρ = 0,39), internet (ρ = 0,37) y automóvil (ρ = 0,37), que a su vez se correlacionan entre sí (ρ = 0,45 entre computador e internet). Este conglomerado expresa, en términos empíricos, el capital económico del hogar descrito en el marco teórico como

uno de los principales determinantes del desempeño, y confirma que estas variables comparten una dimensión común sin llegar a ser redundantes.

En el plano institucional, la correlación más alta de toda la matriz vincula la naturaleza del colegio con su jornada ($\rho = 0{,}59$) y con el nivel socioeconómico del establecimiento ($\rho = -0{,}41$), un patrón coherente con la segmentación entre la oferta oficial y la no oficial del sistema educativo. Más allá de estos bloques, la gran mayoría de los pares presenta coeficientes inferiores a $|0{,}30|$, lo que indica que cada predictor aporta información mayoritariamente propia. Esta baja redundancia, consistente con los valores de VIF reportados, es deseable para el modelado: garantiza que las variables seleccionadas capturen dimensiones complementarias del fenómeno multicausal del desempeño académico, antes que reiterar una misma señal.

El *dataset* de modelado final conserva las 17 variables que superaron el umbral, junto con Año de presentación y la variable objetivo (19 columnas en total). El Año de presentación se retiene para ordenar la línea de tiempo en la validación cruzada temporal del modelado, captura además la discontinuidad de cohortes, incluida la disrupción de 2020. Año de presentación no se emplea como predictor en el modelo final. La

Tabla 4~~Tabla 4~~ del conjunto final de variables constituye el insumo para la modelación predictiva.

Tabla 4. Conjunto final de 17 variables seleccionadas para el enfoque de regresión.

#	Variable	Tipo	Descripción
1	NSE establecimiento	Ordinal (Nivel-2)	Nivel socioeconómico del establecimiento (efecto contextual jerárquico).
2	Educación de la madre	Ordinal	Nivel educativo de la madre.
3	Educación del padre	Ordinal	Nivel educativo del padre.
4	Estrato vivienda	Ordinal	Estrato socioeconómico de la vivienda.
5	Número de libros en el hogar	Ordinal	Número de libros en el hogar.
6	Edad de presentación	Numérica	Edad al presentar la prueba.
7	Grado presentado	Categórica binaria	Grado matriculado (11 = regular, 26 = adultos).
8	Dedicación a lectura	Ordinal	Tiempo de lectura diaria por entretenimiento.
9	Dedicación a internet	Ordinal	Tiempo dedicado a internet.
10	Consumo de lácteos	Ordinal	Frecuencia de consumo de lácteos (proxy nutricional).
11	Tiene computador	Binaria	Computador en el hogar.
12	Tiene internet	Binaria	Conectividad en el hogar.

#	Variable	Tipo	Descripción
13	Tiene automóvil	Binaria	Automóvil en el hogar.
14	Naturaleza del colegio	Nominal	Oficial / No oficial.
15	Jornada del colegio	Nominal	Jornada del colegio (Completa, Mañana, Tarde, Noche, Única, Sabatina).
16	Ocupación del padre	Nominal	Tipo de ocupación del padre.
17	Ocupación de la madre	Nominal	Tipo de ocupación de la madre.

Fuente: elaboración propia con base en los rankings combinados de Información Mutua y Random Forest *sobre el dataset limpio.*

Este conjunto de variables constituye el insumo para los capítulos siguientes: el descubrimiento de patrones mediante aprendizaje no supervisado (Capítulo 4) y el desarrollo de modelos predictivos de regresión supervisada (Capítulo 5).

4. DESCUBRIR PATRONES Y RELACIONES EN LOS DATOS QUE PERMITAN EXPLICAR LAS DIFERENCIAS EN LOS NIVELES DE LOGRO ACADÉMICO

Este capítulo presenta el desarrollo del segundo objetivo específico del proyecto: descubrir patrones y relaciones en los datos que permitan explicar las diferencias en los niveles de logro académico entre los estudiantes del Distrito de Barranquilla. A diferencia del capítulo anterior, donde el análisis se centró en la asociación estadística entre cada predictor individual y el puntaje global, en esta etapa se adopta un enfoque de aprendizaje no supervisado que permite identificar agrupaciones latentes de estudiantes con perfiles socioeducativos similares, sin recurrir a la variable objetivo durante el proceso de segmentación.

El procedimiento se estructura en cuatro pasos: (i) preparación del *set* de modelado mediante codificación e ingeniería de variables; (ii) reducción de dimensionalidad con el Análisis Factorial de Datos Mixtos (FAMD); (iii) determinación del número óptimo de clústeres y entrenamiento del algoritmo K-Means; y (iv) perfilamiento e interpretación de los clústeres obtenidos.

4.1. Preparación del set de modelado

4.1.1. Filtrado temporal y *dataset* de entrada

La fase de modelado parte del archivo *dataset*_limpio.csv, del cual se obtiene un conjunto de trabajo tras un filtrado temporal que excluye los registros previos a 2017. Estos años concentran aproximadamente el 30% de los valores ’Sin información’ en 23 variables del cuestionario socioeconómico, debido a que la estructura del cuestionario es distinta en esos períodos. Eliminarlos preserva la comparabilidad de la información socioeconómica entre los registros conservados (2017 en adelante) y reduce el ruido que afectaría tanto el agrupamiento como los modelos supervisados. El conjunto resultante contiene 133.985 registros y conserva las 17 variables seleccionadas, más las variables temporales Año y Semestre de Presentación y la variable objetivo Puntaje Global.

4.1.2. Ingeniería de variables: indicadores de conocimiento

La imputación multinivel aplicada en la fase de limpieza (Capítulo 3) eliminó la necesidad de tratar ’Sin información’ como una categoría adicional en el *clustering*. Las variables ordinales con jerarquía natural se codificaron mediante mapeos numéricos explícitos que preservan el orden (por ejemplo, Estrato 1 < Estrato 2 < ... < Estrato 6), las variables binarias se codificaron como 0/1, y las nominales sin orden natural se transformaron mediante codificación *one-hot*. Las variables temporales Año y Semestre de presentación se excluyeron del proceso de *clustering* por ser metadatos de la observación, no atributos socioeducativos del estudiante; ambas se conservaron en el *dataset* de salida para su uso en el *split* temporal del Capítulo 5.

4.1.3. Codificación de variables categóricas y escalamiento

KMeans opera con distancias euclidianas y, por tanto, requiere un espacio continuo y métrico. Las variables ordinales y binarias se codificaron numéricamente; el escalamiento de las cuantitativas y la ponderación de las categóricas las realiza internamente el AFDM. El enfoque previo, en cambio, estandarizaba con *StandardScaler* (media 0, desviación estándar 1) para que ninguna domine las distancias por simple efecto de escala. El espacio de trabajo del *clustering* se construyó con el AFDM aplicado a las variables tipadas, sin recurrir a la expansión *one-hot* de las variables nominales.

4.2. Reducción de dimensionalidad mediante AFDM

El Análisis Factorial de Datos Mixtos (AFDM) se aplicó porque, a diferencia del análisis de componentes principales, trata de forma simultánea las variables cuantitativas y las categóricas sin que ninguna domine artificialmente la construcción de los ejes factoriales. Con ello se reduce el ruido al descartar los ejes de menor inercia, se mejora la separabilidad de los clústeres al proyectarlos en un espacio factorial de menor dimensión y se acelera el algoritmo K-Means al reducir el número de ejes a procesar. Como criterio de selección del número de ejes factoriales se buscó retener al menos el 80 % de la inercia explicada acumulada (la inercia es el análogo de la varianza en el marco factorial de datos mixtos).

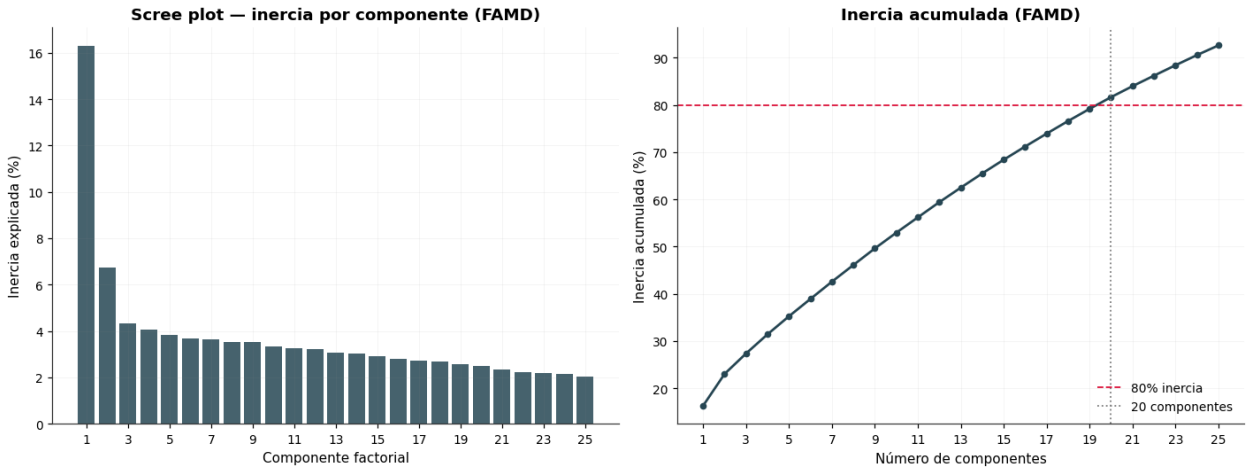


Figura 5. Análisis de inercia explicada por el AFDM.

La Figura 5 muestra que se requieren 20 ejes factoriales para acumular el 80 % de la inercia explicada. El primer eje factorial (F1) absorbe el 16,28% de la inercia y el segundo (F2) el 6,74%; el tercero (F3) aporta el 4,33%, de modo que los tres primeros ejes acumulan el 27,34%. La curva de inercia acumulada alcanza el umbral del 80 % en el eje 20, valor que se adopta como el número de ejes factoriales retenidos. Para referencia, el 70% de la inercia se alcanza en el eje 16 y el 90 % en el eje 24, sobre las 25 componentes diagnósticas examinadas.

El examen de las contribuciones de las variables al primer eje factorial (F1) confirma que esta captura un gradiente socioeducativo claro. Las variables que más lo determinan son el nivel socioeconómico del establecimiento, la jornada, la naturaleza del colegio (oficial frente a no oficial) y la educación de la madre. A un extremo del eje se ubican los estudiantes de colegios no oficiales con jornada completa, alto nivel socioeconómico del establecimiento y educación profesional de la madre; al otro extremo, los estudiantes de colegios oficiales con jornada estándar y contextos socioeconómicos menos favorables.

```
N_COMPONENTES = n_80                          # 20 ejes (>= 80% inercia)
famd = prince.FAMD(n_components=N_COMPONENTES, random_state=SEED)
famd = famd.fit(df_famd[mask_train])          # ajustado SOLO sobre TRAIN
X_famd = famd.transform(df_famd).values       # proyección de todo el dataset
```

4.3. Determinación del número óptimo de clústeres

La selección del número óptimo de clústeres *k* combinó dos métodos complementarios: el Método del Codo, que evalúa la inercia o suma de cuadrados intra-clúster (*within-cluster sum of squares*, WCSS), y el Coeficiente de Silueta (*silhouette*), que cuantifica la separación y cohesión de los clústeres. Para acelerar la fase de exploración se ajustaron modelos *MiniBatchKMeans* sobre una submuestra aleatoria de 30.000 registros (22,4 % del *dataset*), lo que reduce el tiempo de cómputo sin sacrificar la calidad de la estimación.

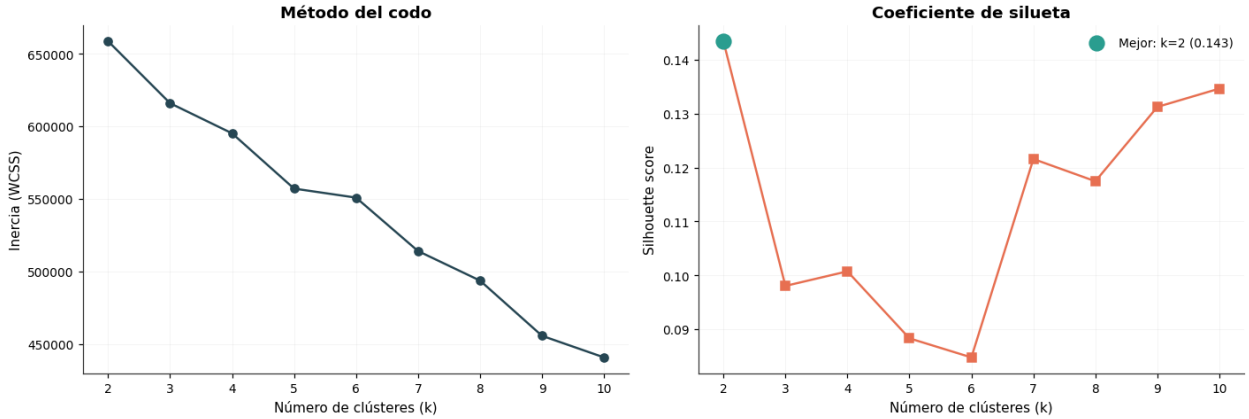


Figura 6. Evaluación del número óptimo de clústeres. Izquierda: inercia (WCSS) en función de k. Derecha: coeficiente de silueta promedio; el máximo absoluto se alcanza en k=2 (*silhouette* = 0,1435).

La Figura 6 presenta los resultados de la exploración para *k* entre 2 y 10. La curva de inercia decrece de manera prácticamente monotónica sin un punto de inflexión claro: la reducción marginal entre *k*=2 y *k*=3 es comparable a la observada en valores posteriores. En contraste, el coeficiente de silueta indica con claridad que el máximo se alcanza en *k*=2 (silueta = 0,1435), seguido por *k*=10 (0,1346) y *k*=9 (0,1313). El valor más bajo se observa en *k*=6 (0,0847).

El diagrama de silueta detallado de la Figura 7 confirma este hallazgo: en *k*=2 los dos clústeres presentan distribuciones de silueta amplias y mayoritariamente positivas, mientras que en *k*=3 y *k*=4 aparecen clústeres con valores de silueta cercanos a cero o negativos, señal de observaciones mal asignadas o ambiguas. Por tanto, se selecciona *k*=2 como el número óptimo de clústeres, interpretando que los datos no soportan estadísticamente una segmentación más fina de la población estudiantil de Barranquilla en términos de su perfil socioeducativo agregado. La regla de decisión implementada favorece además el menor *k* por parsimonia cuando la diferencia de silueta frente al mejor valor es inferior al 2%.

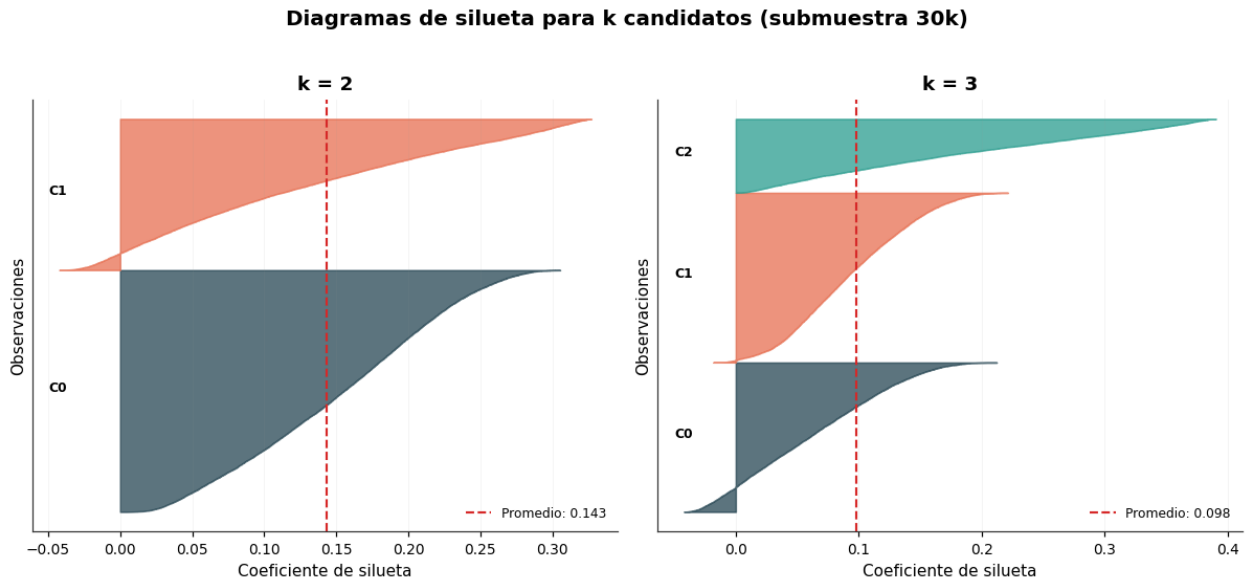


Figura 7. Diagrama de silueta detallado para k=2, k=3 y k=4 (submuestra de 30.000 registros).

4.4. Entrenamiento y visualización del modelo final

El modelo final se entrenó con KMeans estándar (algoritmo de Lloyd), con *k*=2 y diez inicializaciones por el método k-means*++*, ajustando los centroides sobre el conjunto de entrenamiento (n = 100.845) y asignando luego mediante *predict* la totalidad de los 133.985 registros. El tiempo de entrenamiento fue de un segundo y el algoritmo convergió en siete iteraciones. La inercia final sobre el conjunto de entrenamiento fue de 2.242.359,3 y el coeficiente de silueta, estimado sobre una submuestra de 20.000 registros del conjunto de entrenamiento, fue de 0,1496 (0,1527 sobre el *dataset* completo).

La distribución resultante asignó 41.336 estudiantes (30,9 %) al Clúster 0 y 92.649 estudiantes (69,1 %) al Clúster 1. Las proporciones se mantienen estables en los tres conjuntos del *split* temporal: 30,81% / 69,19% en entrenamiento; 30,26% / 69,74% en validación (2023) y 31,65% / 68,35% en prueba (2024).

```
kmeans_final = KMeans(n_clusters=2, init='k-means++', n_init=10,
                      max_iter=300, tol=1e-4, random_state=SEED,
                      algorithm='lloyd')
kmeans_final.fit(X_pca[mask_train])        # centroides ajustados en TRAIN
clusters = kmeans_final.predict(X_pca)     # asignación de todo el dataset
```

La Figura 8 muestra la proyección de los dos clústeres sobre los ejes factoriales F1 y F2. La separación se da fundamentalmente a lo largo del eje F1, lo que confirma que el *clustering* captura el gradiente socioeducativo identificado en el eje *F1*: el Clúster 1 (mayoritario) se ubica en la región de F1 negativo, asociada a colegios oficiales y contextos socioeconómicos menos favorables; el Clúster 0 (minoritario) se concentra en F1 positivo, asociado a colegios no oficiales y contextos socioeconómicos más favorables.

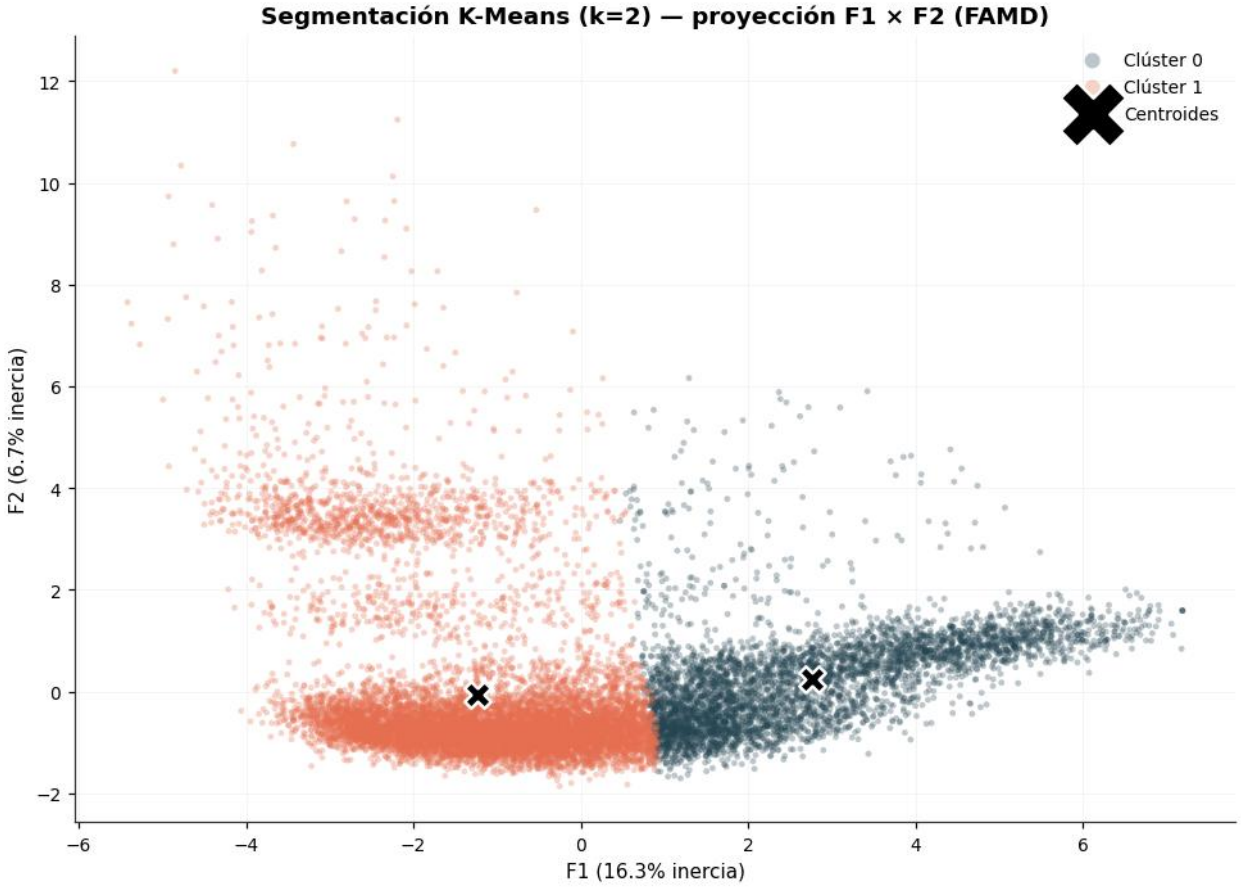


Figura 8. Proyección bidimensional de los clústeres sobre los dos primeros ejes factoriales del AFDM (F1 explica el 16,3 % de la inercia, F2 explica el 6,7 %). Los centroides se marcan con cruces negras. Se observa una separación clara a lo largo del eje F1, eje socioeducativo principal.

Nota metodológica (Ajuste sobre el conjunto de entrenamiento). El AFDM y el algoritmo KMeans se ajustan exclusivamente sobre los registros del conjunto de entrenamiento (`anio_presentacion <= 2022`, n = 100.845) y luego se aplican mediante *transform* y *predict* al

dataset completo. Esta decisión es necesaria porque la columna `cluster_id` derivada del *clustering* se utiliza como variable predictora del modelo supervisado del Capítulo 5, donde se aplica un *split* temporal estricto. Ajustar los centroides sobre el *dataset* completo, incluyendo las cohortes 2023 y 2024, contaminaría las asignaciones de clúster sobre los conjuntos de validación y prueba con información del futuro, lo que constituiría una fuga de información (*leakage*) leve hacia el modelo supervisado. Ajustar los centroides únicamente sobre el conjunto de entrenamiento elimina este riesgo y garantiza que las métricas reportadas en el Capítulo 5 sean defendibles. La comparación entre la versión con *leakage* y la versión corregida muestra que el 98,63% de los registros mantiene su asignación de clúster.

4.5. Perfilamiento e interpretación de los clústeres

Antes del perfilamiento descriptivo se realizó una validación externa de los clústeres contra la variable objetivo Puntaje Global, no utilizada durante el agrupamiento. El análisis de varianza (ANOVA de un factor) arrojó un estadístico F=28.218,57 con un p-valor inferior a 10^{-300}, lo que rechaza la hipótesis de igualdad de medias entre clústeres. El tamaño del efecto $\eta^2 = 0{,}1740$ indica que la pertenencia al clúster explica el 17,4% de la varianza del puntaje global, sin haber observado esta variable durante el ajuste. Según los umbrales de Cohen empleados en el proyecto, este valor se clasifica como efecto grande ($\eta^2 \geq 0{,}14$). Estos resultados son consistentes con la solución de dos clústeres y originan el perfilamiento descriptivo que sigue. No obstante, dado que los valores de silueta son bajos, los clústeres deben interpretarse como una partición útil de un gradiente socioeducativo dominante y no como segmentos naturalmente separados. La Tabla 5 sintetiza el resultado del perfilamiento y permite comparar los dos perfiles socioeducativos identificados.

Tabla 5. Perfil socioeducativo de los dos clústeres identificados por KMeans estándar (k=2).

Atributo	Clúster 0 (~30,9 %)	Clúster 1 (~69,1 %)
Tamaño del grupo	41.336 estudiantes	92.649 estudiantes
Puntaje global (media)	288,1	239,7
Puntaje global (mediana)	290,0	236,0
Puntaje global (desv. estándar)	51,9	47,1
Edad promedio (años)	16,9	17,6
Composición por grado (regular/ adultos)	96,6% / 3,4%	85,8% / 14,2%
Estrato modal	Estrato 3	Estrato 1
Educación de la madre (moda)	Profesional	Secundaria
Educación del padre (moda)	Profesional	Secundaria
Jornada modal	COMPLETA	ÚNICA

Atributo	Clúster 0 (~30,9 %)	Clúster 1 (~69,1 %)
Naturaleza del colegio (moda)	NO OFICIAL	OFICIAL
Nivel socioeconómico del establecimiento (moda)	Nivel 3	Nivel 2
Número de libros en el hogar (moda)	11 a 25	0 a 10
% con computador en casa	91,7 %	40,9 %
% con internet en casa	97,4 %	61,9 %
% con automóvil	60,0 %	10,0 %

Fuente: elaboración propia.

La comparación visual de ambos perfiles se complementa con un *radar chart*, como se muestra en la Figura 9, que normaliza las dimensiones socioeducativas al rango [0, 1] mediante *MinMaxScaler*. El Clúster 0 domina en educación parental, estrato, nivel socioeconómico del establecimiento y puntaje global, mientras que el Clúster 1 presenta valores menores en estas dimensiones.

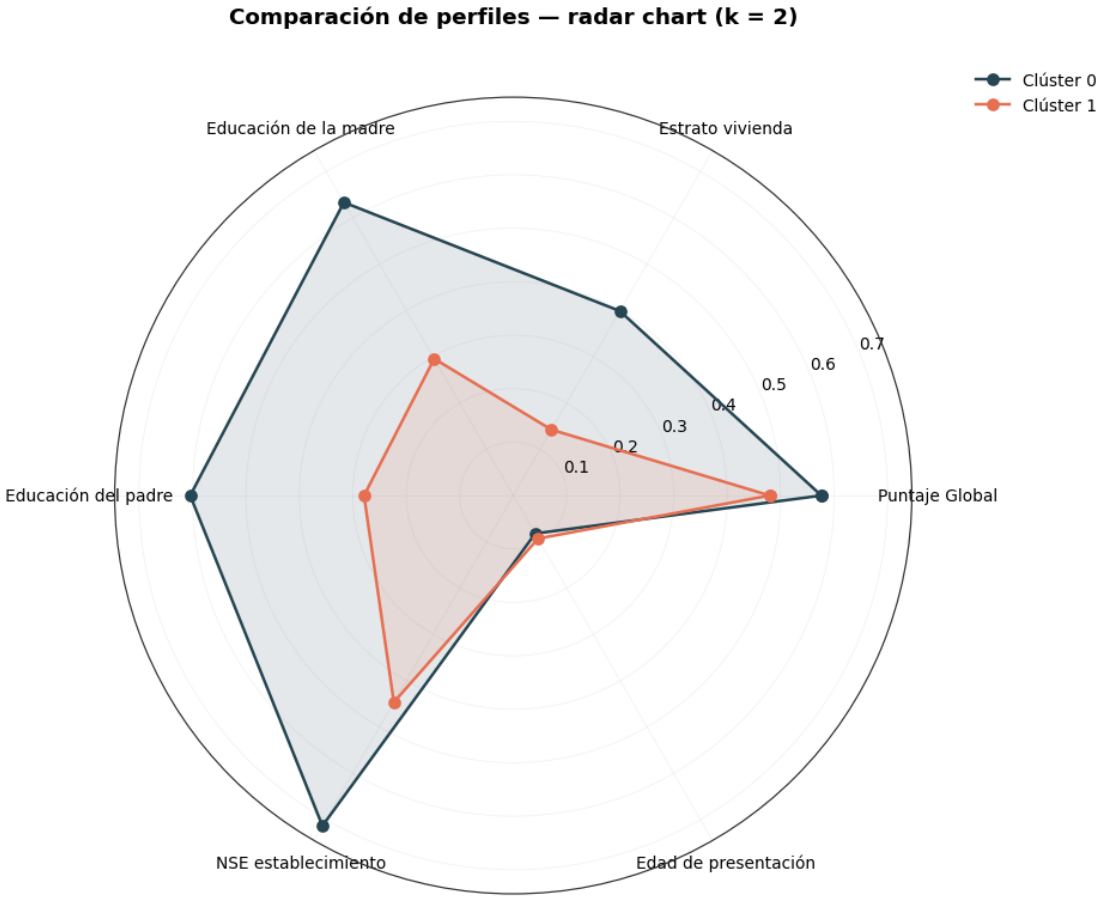


Figura 9. Comparación de perfiles socioeducativos mediante *radar chart*.

4.5.1. Narrativa del Clúster 0: perfil de mayor capital cultural

El Clúster 0 agrupa 41.336 estudiantes (30,9 % del total). Se caracteriza por proceder de hogares con padres y madres con educación profesional como moda, viviendas en estrato 3 (con presencia de estratos 4 a 6 en el resto de la distribución) y asistencia mayoritaria a colegios no oficiales con jornada completa. El nivel socioeconómico del establecimiento es Nivel 3 (modal). La penetración tecnológica del hogar es alta: el 91,7% cuenta con computador, el 97,4% con internet y el 60,0% con automóvil. La composición por grado es predominantemente regular (96,6% en grado 11) con presencia minoritaria de educación para adultos (3,4% en grado 26). El puntaje global promedio es de 288,0 (mediana de 290), valor sustantivamente superior a la media nacional de 250 establecida por el ICFES.

4.5.2. Narrativa del Clúster 1: perfil de mayor vulnerabilidad socioeducativa

El Clúster 1 agrupa 92.649 estudiantes (69,1% del total). El perfil familiar modal corresponde a padres con educación secundaria, viviendas en estrato 1 (con presencia mayoritaria de estratos 1 y 2) y asistencia a colegios oficiales con jornada única. El nivel socioeconómico del establecimiento es Nivel 2 (modal). Las condiciones del hogar muestran brechas pronunciadas respecto al Clúster 0: solo el 40,9% cuenta con computador, el 61,9% con internet y el 10,0% con automóvil. La composición por grado revela un hallazgo relevante: la proporción de estudiantes en modalidad de educación para adultos (grado 26) es de 14,2%, alrededor de cuatro veces la del Clúster 0, lo que indica que la población adulta del CLEI se concentra preferentemente en el perfil socioeducativo más vulnerable. De hecho, de los 14.547 estudiantes adultos del *dataset*, el 90,2% cae en el Clúster 1. El puntaje global promedio es de 239,7 (mediana de 236), por debajo de la media nacional. La diferencia respecto al Clúster 0 alcanza los 48,3 puntos en el puntaje global.

El análisis no supervisado de los datos del Distrito de Barranquilla deja tres hallazgos centrales. En primer lugar, los dos clústeres identificados configuran un eje socioeducativo dominante caracterizado por la educación de los padres, el nivel socioeconómico del establecimiento, la naturaleza del colegio (oficial frente a no oficial), la jornada y la disponibilidad de capital tecnológico en el hogar. En segundo lugar, esta segmentación no supervisada captura de manera consistente las diferencias de rendimiento académico medidas oficialmente por el ICFES, con una brecha de 48,3 puntos en el puntaje global promedio entre ambos grupos y un tamaño de efecto $\eta^2 = 0{,}1740$ estadísticamente robusto. En tercer lugar, el *clustering* revela la concentración preferente de la población de educación para adultos en el clúster de mayor vulnerabilidad socioeducativa, hallazgo de interés para la política pública distrital.

Estos resultados aportan dos elementos clave para la fase de modelado supervisado. Por un lado, justifican empíricamente la importancia de las variables socioeducativas seleccionadas en el Capítulo 3. Por otro lado, alertan sobre un fenómeno que se confirmará en los capítulos

siguientes: el comportamiento del puntaje en el Distrito está dominado por un único gradiente socioeducativo, lo que delimita el techo predictivo de cualquier modelo construido únicamente con estas variables. La variable derivada `cluster_id` se incorpora como predictor adicional en el modelado supervisado del Capítulo 5, en tanto constituye una síntesis multivariada no lineal de las 17 variables seleccionadas, sin riesgo de fuga de información, pues el *clustering* se realizó sin observar la variable objetivo y los centroides se ajustaron únicamente sobre el conjunto de entrenamiento.

5. DESARROLLAR MODELOS PREDICTIVOS DE APRENDIZAJE SUPERVISADO PARA ESTIMAR EL DESEMPEÑO ACADÉMICO

Este capítulo presenta el desarrollo del tercer objetivo específico: construir modelos predictivos de aprendizaje supervisado que permitan estimar el puntaje global del Examen Saber 11° de los estudiantes del Distrito de Barranquilla a partir de sus características socioeducativas. El proceso sigue una secuencia de complejidad creciente que parte de un modelo de referencia *DummyRegressor*, evalúa modelos lineales y regularizados, y avanza hacia ensambles no lineales (Random Forest y XGBoost). Cada paso responde a una pregunta concreta sobre la naturaleza de la relación entre las variables predictoras y el puntaje, y permite justificar de manera incremental la complejidad del modelo final. La evaluación se realiza bajo un esquema de validación temporal estricto, coherente con la naturaleza longitudinal de los datos.

5.1. Preparación del set de modelado

5.1.1. Variable objetivo y diagnóstico de su distribución

La variable objetivo del proyecto es Puntaje Global, el puntaje global del Examen Saber 11° en escala 0 a 500. La Figura 10 muestra su distribución sobre el *dataset* de modelado: media de 254,65, mediana de 251 y desviación estándar de 53,50. La asimetría es leve (*skewness* ≈ 0,30) y la diferencia entre media y mediana es de apenas 3,7 puntos. Estas características son favorables para el modelado: una distribución aproximadamente normal favorece los métodos de mínimos cuadrados y no se requiere transformación logarítmica del objetivo. El rango intercuartílico (IQR≈80 puntos, de 213 a 293) indica que el 50% central de los estudiantes se concentra en una banda de ±40 puntos alrededor de la mediana. El panel temporal de la figura muestra además que la media del puntaje varía poco entre cohortes (de 249,7 en 2019 a 260,8 en 2024), con un leve repunte en las cohortes recientes.

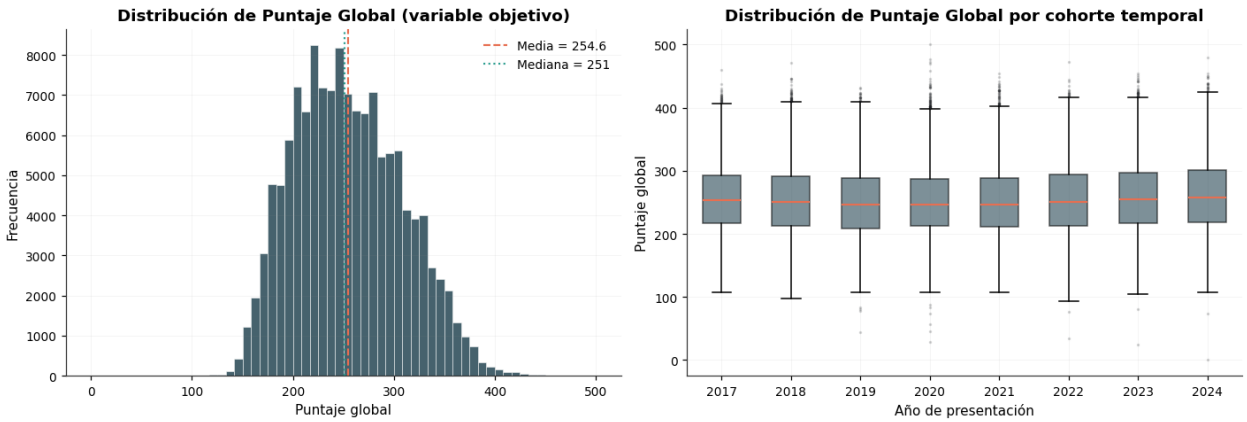


Figura 10. Distribución de la variable objetivo Puntaje Global y su evolución por cohorte temporal.

5.1.2. Split temporal

Dado el carácter longitudinal de los datos (ocho cohortes entre 2017 y 2024, con efecto de la pandemia en 2020 y cambios de cuestionario entre períodos), la partición de los datos no se hizo de forma aleatoria sino mediante un *split* temporal estricto por año de presentación. Un *split* aleatorio mezclaría cohortes futuras con pasadas e inflaría artificialmente el R^2 estimado; el *split* temporal, en cambio, simula el escenario real de despliegue, en el que el modelo aprende del pasado y predice cohortes nuevas. La partición resultante es: entrenamiento con las cohortes 2017 a 2022 (100.845 registros, 75,27%), validación con la cohorte 2023 (16.286 registros, 12,16%) y prueba con la cohorte 2024 (16.854 registros, 12,58%).

La media del puntaje presenta una leve deriva (*drift*) entre conjuntos, 253,02 en entrenamiento, 258,38 en validación y 260,79 en prueba, consistente con el repunte observado en las cohortes recientes. La composición por grado y por clúster se mantiene estable entre los tres conjuntos (alrededor de 89% de grado 11 y 25% de Clúster 0), lo que confirma que la partición no introduce sesgos de composición más allá del efecto cohorte esperado.

5.1.3. Predictores y *pipeline* de preprocesamiento

El conjunto de predictores quedó conformado por 18 variables: las 17 variables socioeducativas seleccionadas en el Capítulo 3 más la variable derivada clúster (perfil socioeducativo del Capítulo 4). Se excluye el año de presentación (anio_presentacion) del conjunto de predictores por dos razones técnicas. La primera es que esta variable define el criterio del *Split*: emplearla además como predictor constituiría una fuga de información (*leakage*), pues el modelo aprendería la frontera artificial que separa los conjuntos de entrenamiento, validación y prueba. La segunda es que codifica un efecto de cohorte, por ejemplo, la disrupción de la pandemia en 2020, que no extrapola a cohortes futuras y que, por tanto, no debe incorporarse como señal predictiva genérica.

El preprocesamiento se encapsuló en un *Pipeline* de *scikit-learn* porque evita la fuga de datos durante la validación cruzada, al ajustar el preprocesador únicamente con los datos de entrenamiento de cada *fold*, y favorece la reproducibilidad de la lógica de preparación al serializarla junto con el modelo. Las 18 variables se clasificaron en cinco grupos según su naturaleza, aplicando a cada uno la codificación más apropiada mediante un *ColumnTransformer*:

```
preprocesador = ColumnTransformer([
    ('num',      StandardScaler(),                cols_numericas),   # 1
    ('bin_sino', OrdinalEncoder(No=0, Si=1),       cols_binarias),    # 3
    ('bin_int',  'passthrough',                    cols_binarias_int),# 3
    ('nom',      OneHotEncoder(drop='first',
                               handle_unknown='ignore'), cols_nominales), # 4
    ('ord',      OrdinalEncoder(categories=ORDENES), cols_ordinales),  # 8
])
pipeline = Pipeline([('prep', preprocesador), ('reg', modelo)])
```

Tras la transformación, el espacio de modelado quedó conformado por 32 *features*: 1 numérica estandarizada con *StandardScaler*; 3 binarias Sí/No codificadas como 0/1; 2 binarias ya enteras (Grado presentado, clúster) incorporadas por *passthrough*; 18 columnas resultantes del *one-hot encoding* de las 4 variables nominales (con drop='first' para evitar colinealidad); y 8 ordinales codificadas con orden explícito mediante *OrdinalEncoder*.

5.2. Modelo de referencia: *Baseline* Naive

El modelo de línea base cumple el rol de establecer un piso de rendimiento que cualquier modelo entrenado deberá superar para justificar su existencia. Como referencia se empleó un *DummyRegressor* con estrategia de media, que predice siempre el promedio del puntaje en el conjunto de entrenamiento (253,02 puntos), ignorando por completo las variables predictoras.

Por construcción, su R^2 sobre entrenamiento es exactamente 0; sobre los conjuntos de validación y prueba puede ser ligeramente negativo cuando la media de esas cohortes difiere de la del entrenamiento, lo cual es esperable bajo *split* temporal con deriva entre cohortes. Sobre la cohorte de prueba (2024), el *baseline* obtuvo $R^2 = -0{,}0196$, MAE = 45,87 puntos y RMSE = 56,06 puntos. Este MAE constituye la cota superior admisible: un modelo que no logre un error menor no estaría aprendiendo nada útil de las variables socioeducativas.

5.3. Modelos candidatos

Sobre el mismo *Pipeline* de preprocesamiento se entrenaron cinco modelos con una configuración inicial razonable y comparable, en particular, el Random Forest se limitó en profundidad (max_depth = 15) y tamaño de hoja (min_samples_leaf = 20) para no sobreajustar, recorriendo una escalera de complejidad creciente: regresión lineal por mínimos cuadrados ordinarios (*OLS*), Ridge (regularización L2), Lasso (regularización L1), Random Forest (ensamble por *bagging*) y XGBoost (ensamble por *gradient boosting*). Todos se evaluaron sobre el conjunto de validación (cohorte 2023). La Tabla 6 sintetiza el desempeño comparado, siguiendo la escalera de complejidad creciente de los modelos.

Tabla 6. Comparación de desempeño de los modelos candidatos en el conjunto de validación.

Modelo	R^2 val	R^2 train	MAE val (puntos)	RMSE val (puntos)	Tiempo (s)
Baseline (Dummy)	−0,0098	0,0000	44,82	54,57	0,00
Lineal (OLS)	0,3475	0,3758	35,16	43,87	0,52
Ridge	0,3475	0,3758	35,16	43,87	0,49
Lasso	0,3452	0,3735	35,22	43,95	0,80
Random Forest	0,3758	0,4722	34,28	42,90	71,61
XGBoost	0,3897	0,4784	33,84	42,43	4,58

Fuente: elaboración propia.

La Figura 11 presenta la comparación visual de R^2 (entrenamiento frente a validación) y de MAE. Los hallazgos centrales son los siguientes:

- Los modelos lineales (OLS y Ridge) producen métricas prácticamente idénticas (R^2 val = 0,3475). Al estandarizar el diseño completo y optimizar la penalización mediante validación cruzada temporal, explorando *alpha* en [0,01; 1000] para Ridge y en [0,0001; 10] para Lasso, la validación selecciona en ambos casos la penalización mínima del rango (Ridge α = 0,01; Lasso α = 0,0001). Esto indica que, con 32 *features* y unas 100.000 observaciones de entrenamiento, el modelo lineal no está sobreajustado y la regularización no aporta mejora. Ridge y Lasso reproducen el desempeño de OLS y Lasso conserva la totalidad de los 32 coeficientes, sin descartar ninguna variable informativa.

- Random Forest, limitado en profundidad (*max_depth* = 15) y tamaño de hoja (*min_samples_leaf* = 20) para evitar el sobreajuste, alcanza un R^2 de 0,4722 en entrenamiento y 0,3758 en validación, por encima de los modelos lineales, aunque por debajo de XGBoost. Esta restricción inicial es deliberada: garantiza que la comparación con XGBoost, que por defecto ya viene regularizado, se realice en condiciones equilibradas y no penalice de forma artificial al Random Forest por un sobreajuste evitable. Su costo computacional sigue siendo el más alto del grupo (71,61 segundos). El ensamble por *bagging* memoriza el conjunto de entrenamiento sin mejorar la generalización.

- XGBoost es el modelo dominante: R^2 val=0,3897, MAE=33,84 puntos y RMSE=42,43 puntos, con una brecha entrenamiento-validación moderada (0,4784 frente a 0,3897) y un tiempo de entrenamiento razonable (4,58 segundos). El *boosting* secuencial captura patrones que el *bagging* de Random Forest no alcanza, gracias a su capacidad de corregir iterativamente los residuos y a su regularización integrada.

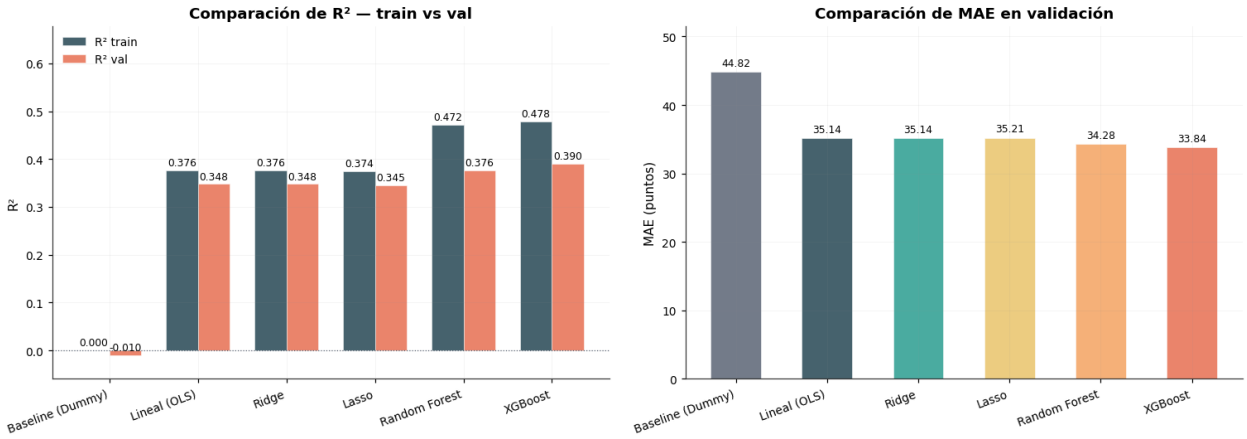


Figura 11. Comparación visual de los modelos candidatos. Panel izquierdo: R^2 en entrenamiento frente a validación. Panel derecho: MAE en validación.

Estos resultados arrojan un hallazgo importante: la transición de modelos lineales a ensambles no lineales produce una ganancia real pero limitada ($\Delta R^2 \approx 0{,}04$ en validación). En el contexto educativo de Barranquilla, la relación entre el contexto socioeconómico y el puntaje Saber 11° tiene una componente lineal dominante, complementada por interacciones no lineales que XGBoost aprovecha mejor. El R^2 cercano a 0,39 también muestra un techo explicativo: factores no medidos, motivación individual, calidad docente específica, eventos familiares, limitan la predicción más allá de cualquier algoritmo.

5.4. Optimización de hiperparámetros del modelo ganador

La selección del modelo final no se basó en optimizar únicamente al candidato más prometedor. Todos los modelos con hiperparámetros —Ridge, Lasso, Random Forest y XGBoost— se sometieron al mismo presupuesto de búsqueda mediante `RandomizedSearchCV` con validación cruzada temporal (`TimeSeriesSplit` de cinco particiones), de modo que ninguno compitiera con una configuración por defecto que pudiera favorecerlo o penalizarlo de forma arbitraria. Para el Random Forest, el espacio de búsqueda abarcó el número de árboles (150, 250 o 350), la profundidad máxima (10, 15 o 20), el tamaño mínimo de hoja (5, 10 o 20) y el número de variables consideradas por división. Optimizado bajo este presupuesto, el Random Forest alcanzó un R^2 de 0,3823 en validación, con un R^2 medio en validación cruzada temporal de 0,4000, frente al R^2 de 0,3895 de XGBoost (validación cruzada de 0,4109). La diferencia entre ambos (0,0072) supera la tolerancia de parsimonia adoptada (0,005), por lo que XGBoost se confirma como modelo final por su mejor desempeño, ya en igualdad de condiciones de optimización.

Confirmado XGBoost como el modelo de mejor balance entre desempeño y costo, se afinaron sus hiperparámetros bajo el mismo presupuesto de búsqueda, como se muestra en la Tabla 7, mediante `RandomizedSearchCV`. Esta estrategia muestrea aleatoriamente combinaciones de un espacio de búsqueda predefinido, lo que resulta más eficiente que *GridSearchCV* en espacios

de alta dimensionalidad. La validación cruzada empleó `TimeSeriesSplit` de 5 *folds* sobre el conjunto de entrenamiento, alineado con la naturaleza temporal de los datos: un *KFold* aleatorio sobre datos longitudinales violaría el supuesto de independencia entre cohortes. Se exploraron 20 combinaciones con 5 *folds* (100 entrenamientos en total), usando R^2 como métrica de *scoring*. La búsqueda se completó en 5,78 minutos y alcanzó un R^2 medio en validación cruzada temporal de 0,4105.

Tabla 7. Espacio de búsqueda y valores óptimos encontrados para los hiperparámetros de XGBoost.

Hiperparámetro	Espacio de búsqueda	Valor óptimo
n_estimators	{200, 400, 600}	400
learning_rate	{0,03, 0,05, 0,1}	0,05
max_depth	{4, 6, 8}	4
subsample	{0,8, 1,0}	0,8
colsample_bytree	{0,8, 1,0}	1,0
reg_lambda	{0,5, 1,0, 2,0}	2,0

Fuente: elaboración propia.

Con los hiperparámetros óptimos, el modelo XGBoost optimizado obtuvo sobre validación R^2=0,3895, MAE=33,89 puntos y RMSE=42,43 puntos, prácticamente idéntico al del modelo sin optimizar ($\Delta R^2 \approx 0$). La configuración elegida es conservadora —*max_depth* de 4, *learning_rate* de 0,05 y regularización L2 (*reg_lambda*) de 2,0—, lo que confirma que el modelo prioriza la generalización sobre la capacidad de ajuste. La mejora marginal indica que el espacio de hipótesis del modelo por defecto ya estaba bien especificado y que el techo predictivo del enfoque está dominado por las variables disponibles, no por la configuración del algoritmo.

5.5. Evaluación final sobre la cohorte de prueba

La estimación de la capacidad predictiva se obtuvo evaluando el modelo optimizado sobre la cohorte de prueba (2024), que no participó en ninguna etapa del entrenamiento ni de la selección de hiperparámetros. La Tabla 8 reporta las métricas del modelo final en los tres conjuntos del *split* temporal.

Tabla 8. Desempeño del modelo XGBoost optimizado en los tres conjuntos del *split* temporal.

Conjunto	Años	R^2	MAE (puntos)	RMSE (puntos)	MAPE (%)
TRAIN	2017–2022	0,4372	31,84	39,67	13,15
VAL	2023	0,3895	33,89	42,43	13,29
TEST	**2024**	**0,3795**	**35,03**	**43,73**	**13,77**

Fuente: elaboración propia.

Son tres puntos por resaltar: Primero, sobre la cohorte de prueba el modelo alcanza R^2=0,3795, MAE=35,03 puntos y RMSE=43,73 puntos, lo que representa una ganancia de 0,399 en R^2 y una reducción de 10,85 puntos en MAE respecto al *baseline* naive sobre el mismo conjunto: el modelo aporta valor predictivo real. Segundo, la brecha de R^2 entre entrenamiento y prueba es de apenas 0,058, dentro del rango aceptable (por debajo de 0,10), lo que descarta sobreajuste relevante y evidencia una generalización robusta a una cohorte futura no observada. Tercero, el descenso suave y monótono del desempeño de entrenamiento a validación y a prueba (R^2 de 0,4372 a 0,3895 y a 0,3795) es coherente con la deriva temporal entre cohortes y confirma que el modelo no memorizó patrones específicos del periodo de entrenamiento.

La estimación final del desempeño del modelo es, por tanto, R^2≈0,38, MAE≈35 puntos y RMSE≈44 puntos sobre la cohorte 2024. El *pipeline* completo, preprocesador más modelo optimizado, se serializó con *joblib* (archivo modelo_regresion_xgboost.joblib) para garantizar su reutilización. Estos valores constituyen el referente final para la evaluación detallada, análisis de residuos, validación segmentada por grado e interpretabilidad, que se presenta en el Capítulo 6.

6. EVALUAR EL DESEMPEÑO DE LOS MODELOS PREDICTIVOS DESARROLLADOS, CON EL PROPÓSITO DE SELECCIONAR AQUEL CON MAYOR CAPACIDAD EXPLICATIVA Y PREDICTIVA PARA EL CONTEXTO LOCAL

Este capítulo desarrolla el cuarto objetivo específico: evaluar el desempeño del modelo predictivo seleccionado y caracterizar sus fortalezas, limitaciones y aplicabilidad en el contexto educativo del Distrito de Barranquilla. La evaluación se organiza en cuatro componentes complementarios: un análisis de residuos que diagnostica sesgos sistemáticos y patrones de error de predicción (no supuestos del modelo, que XGBoost no requiere); una validación segmentada por subgrupos relevantes para la política educativa; una validación por nivel de desempeño oficial del ICFES; y un análisis de interpretabilidad mediante valores SHAP que cuantifica el aporte explicativo de cada variable a nivel global e individual. El modelo evaluado es el XGBoost optimizado del Capítulo 4, y todas las métricas se calculan sobre la cohorte de prueba de 2024 (16.854 estudiantes), que no participó en el entrenamiento ni en la selección de hiperparámetros.

6.1. Análisis de residuos del modelo XGBoost optimizado

Las métricas agregadas (R^2, MAE, RMSE) pueden ocultar patrones sistemáticos de error que solo se detectan al analizar los residuos. Al realizar este ejercicio se deben responder tres preguntas: ¿el modelo se equivoca de forma sesgada en algún rango del puntaje?, ¿la varianza del error es homogénea (homocedasticidad) o crece en los extremos (heterocedasticidad)?, ¿se distribuyen los residuos como una normal centrada en cero?

La Figura 12 resume el diagnóstico en cuatro paneles. Los hallazgos principales son los siguientes:

- Sesgo leve por deriva temporal. La media de los residuos es de +6,43 puntos y la mediana de +4,64 puntos, ambas positivas. El modelo subestima ligeramente a la cohorte 2024, lo cual es esperable bajo *split* temporal: el puntaje medio creció de 253,0 en el entrenamiento (2017–2022) a 260,8 en la prueba (2024), y el modelo, entrenado con cohortes anteriores, no anticipa por completo esa mejora en el puntaje global. La desviación estándar de los residuos es de 43,26 puntos, coherente con el RMSE de 43,73 puntos.

- Concentración en la zona central. La nube de predicciones se concentra a lo largo de la diagonal en el rango 200–320 puntos, donde el modelo funciona mejor. La mediana del error absoluto es de 29,9 puntos, pero en el percentil 95 alcanza 84,9 puntos, lo que evidencia mayor dificultad en los casos atípicos.

Distribución de residuos casi simétrica. El histograma muestra una forma centrada y prácticamente simétrica (asimetría = 0,17; curtosis = −0,08), y el *Q-Q plot* sigue la diagonal teórica en el rango central, con desviaciones esperables en las colas. Estos extremos provienen de casos

genuinamente difíciles de predecir, como estudiantes con puntajes excepcionales no anticipados por su contexto.

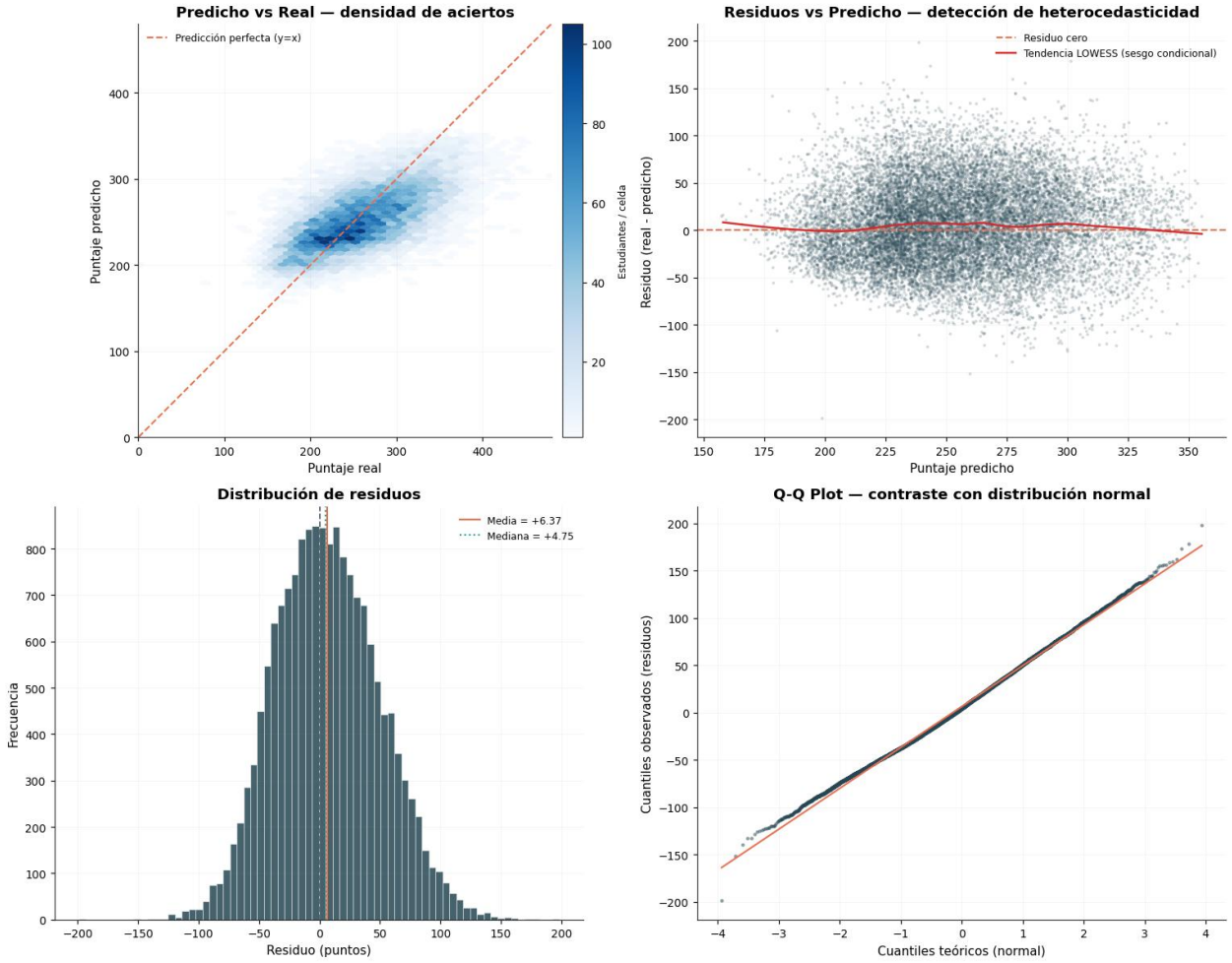


Figura 12. Diagnóstico de residuos del modelo XGBoost optimizado sobre el conjunto de prueba (cohorte 2024). Panel superior izquierdo: predicho vs. real (densidad *hexbin*); línea roja indica predicción perfecta. Panel superior derecho: residuos vs. predicho con curva LOWESS. Panel inferior izquierdo: histograma de residuos. Panel inferior derecho: *Q-Q* plot frente a una distribución normal.

6.1.1. Pruebas estadísticas sobre los residuos

Para verificar los patrones gráficos se aplicaron tres pruebas. La prueba de Shapiro-Wilk rechaza la normalidad de los residuos ($p \approx 2{,}66 \times 10^{-5}$), aunque su estadístico (0,9983, muy cercano a 1) confirma que la desviación es trivial: con muestras grandes la prueba detecta separaciones mínimas de la normal, por lo que el tamaño del efecto, no el p-valor, es lo relevante. La prueba de Breusch-Pagan rechaza la homocedasticidad ($p \approx 1{,}58 \times 10^{-16}$), lo que confirma una heterocedasticidad leve: el modelo es algo más impreciso al predecir puntajes extremos (muy altos o bajos). Finalmente, el estadístico de Durbin-Watson es 1,98, prácticamente igual a 2, lo que descarta autocorrelación de los residuos. En conjunto, las pruebas confirman un modelo con

residuos casi normales, sin autocorrelación y con heterocedasticidad moderada concentrada en los extremos.

6.2. Validación por subgrupos relevantes para la política educativa

Un modelo que funciona bien en promedio puede ocultar sesgos al desempeñarse de forma desigual entre subgrupos. Para que el modelo sea útil como insumo de política pública es indispensable verificar que su capacidad predictiva sea consistente entre los principales subgrupos del Distrito. Se evaluó el desempeño segmentado por grado educativo, por perfil socioeducativo (`cluster_id` del Capítulo 3) y por naturaleza del colegio. La Figura 13 compara R^2 y MAE de cada subgrupo contra el desempeño global (R^2 = 0,3795; MAE = 35,03).

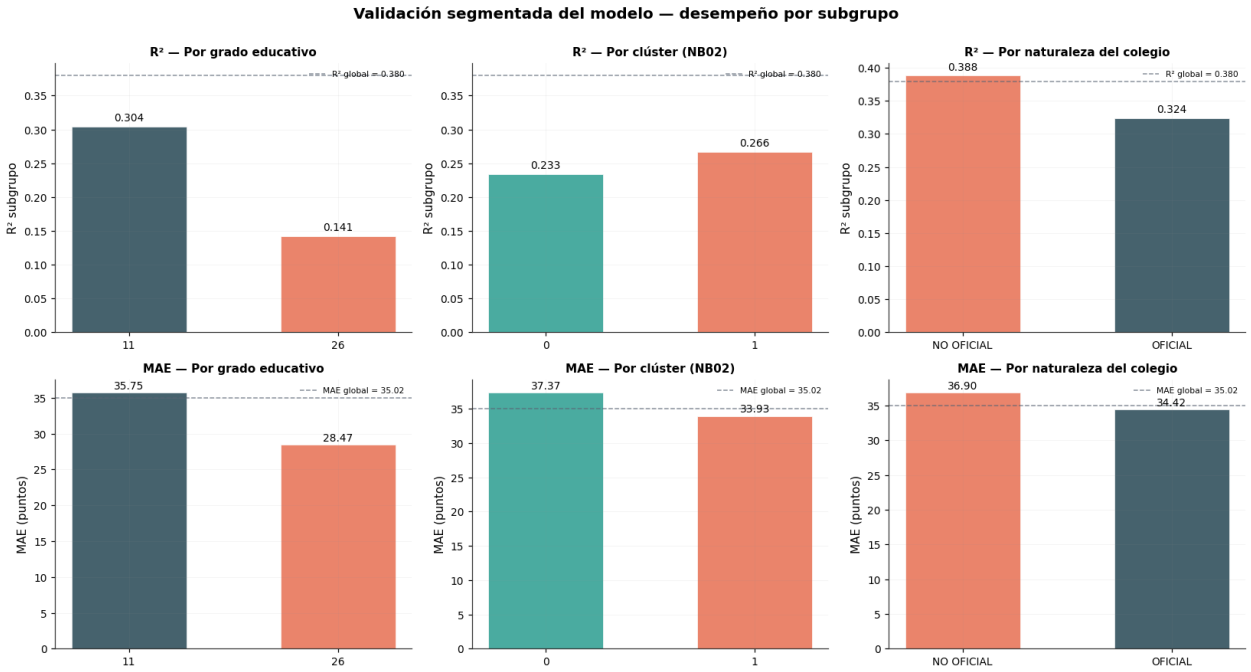


Figura 13. Comparación de R^2 y MAE del modelo entre subgrupos (grado, perfil socioeducativo y naturaleza del colegio). La línea punteada marca el desempeño global (R^2 = 0,380; MAE = 35,03).

6.2.1. Desempeño por grado educativo

La Tabla 9 reporta el desempeño desagregado por grado. En educación regular (grado 11), el modelo explica el 30,4% de la varianza con un MAE de 35,76 puntos y subestima el puntaje en 7,28 puntos en promedio. En educación de adultos/CLEI (grado 26), el R^2 cae a 0,1414, apenas el 14% de la varianza, pero el MAE es menor (28,47 puntos) y el sesgo casi nulo (−1,88 puntos). La paradoja se explica por la menor dispersión del puntaje en la población adulta, concentrada en torno a 203 puntos: el error en puntos absolutos es menor, aunque el modelo explique una fracción más pequeña de la varianza. La población de adultos tiene un desempeño mediado por

factores —trayectorias escolares interrumpidas, condiciones laborales— que las variables socioeducativas tradicionales no capturan bien.

Tabla 9. Desempeño del modelo XGBoost optimizado segmentado por grado educativo (cohorte 2024).

Grado	N	R^2	MAE (puntos)	Sesgo medio (puntos)	PG real medio
Regular (grado 11)	15.172	0,3043	35,76	+7,28	267,2
Adultos / CLEI (grado 26)	1.682	0,1414	28,47	−1,88	203,2

Fuente: elaboración propia.

6.2.2. Desempeño por perfil socioeducativo

La Tabla 10 segmenta por el perfil socioeducativo derivado en el Capítulo 3. El R^2 es similar en ambos perfiles (0,2332 en el clúster aventajado y 0,2664 en el vulnerable), y el MAE es algo menor en el clúster vulnerable (33,93 frente a 37,37 puntos), de nuevo por la menor dispersión relativa de su puntaje. Ambos perfiles presentan un sesgo positivo (subestimación) consistente con la deriva temporal global. La cercanía de los R^2 indica que el modelo no privilegia a un perfil sobre otro: su capacidad predictiva es comparable entre estudiantes aventajados y vulnerables.

Tabla 10. Desempeño del modelo XGBoost optimizado segmentado por perfil socioeducativo (`cluster_id`).

Perfil (clúster)	N	R^2	MAE (puntos)	Sesgo medio (puntos)	PG real medio
Clúster 0 (aventajado)	4.188	0,2332	37,37	+7,42	296,5
Clúster 1 (vulnerable)	12.666	0,2664	33,93	+6,11	249,0

Fuente: elaboración propia.

6.2.3. Desempeño por naturaleza del colegio

La Tabla 11 reporta el desempeño por naturaleza del colegio. El R^2 es más alto en colegios no oficiales (0,3883) que en oficiales (0,3240): en el sector privado las variables socioeducativas explican casi el 39% de la varianza del puntaje, frente al 32% en el oficial, una brecha de 0,066 en R^2. Paradójicamente, el MAE es menor en oficiales (34,44 frente a 36,85 puntos), porque sus puntajes están más comprimidos en torno a 253 puntos. La interpretación es relevante para la política pública: en el sector oficial intervienen factores institucionales no capturados por las variables socioeconómicas del estudiante —calidad del docente, recursos del aula, programas de intervención— con un peso explicativo mayor que en el sector privado. Cualquier despliegue del modelo en el sector oficial requeriría enriquecer los predictores con información institucional adicional.

Tabla 11. Desempeño del modelo XGBoost optimizado segmentado por naturaleza del colegio (cohorte 2024).

Naturaleza	N	R^2	MAE (puntos)	Sesgo medio (puntos)	PG real medio
No oficial	4.106	0,3883	36,90	+6,33	284,0
Oficial	12.748	0,3240	34,42	+6,47	253,3

Fuente: elaboración propia.

6.3. Validación por el nivel del desempeño

Para evaluar el comportamiento del modelo a lo largo de la escala de puntaje se segmentó la cohorte de prueba según cuatro niveles de desempeño del ICFES, cuyos cortes (190, 268 y 342 puntos) se derivan algebraicamente de la fórmula del puntaje global. En esta segmentación no se reporta el R^2: al definir los grupos mediante cortes del propio objetivo, la varianza interna de cada nivel se comprime y el R^2 diverge hacia valores negativos que no reflejan la calidad predictiva. Las métricas válidas son el MAE, el RMSE y el sesgo medio, reportados en la Tabla 12.

Tabla 12. Desempeño del modelo por nivel de desempeño oficial del ICFES (cohorte 2024). El R^2 se omite por no ser válido en segmentaciones por cortes del objetivo.

Nivel ICFES	N	PG real/pred	Sesgo (puntos)	MAE (puntos)	MAPE (%)
1 — Insuficiente	1.717	172/219	−46,91	47,00	27,77
2 — Mínimo	7.727	231/244	−12,63	24,98	11,10
3 — Satisfactorio	5.987	300/269	+31,21	36,21	11,91
4 — Avanzado	1.423	364/294	+70,09	70,12	19,16

Fuente: elaboración propia.

El patrón de signos del sesgo medio muestra un fenómeno de regresión hacia la media, comportamiento típico de los modelos predictivos cuando el objetivo tiene colas en ambos extremos. El modelo sobreestima a los estudiantes de bajo desempeño (Nivel 1: +47 puntos hacia arriba), predice con precisión la zona modal (Nivel 2: MAE de apenas 25 puntos) y subestima progresivamente a los de alto desempeño (Nivel 3: −31 puntos; Nivel 4: −70 puntos). La implicación de política pública es directa: el modelo es útil para evaluaciones diagnósticas y para clasificar estudiantes en rangos generales, pero no debe usarse para identificar talento excepcional (Nivel 4) ni casos extremos individuales en la cola inferior. La Figura 14 visualiza el MAE y el sesgo por nivel.

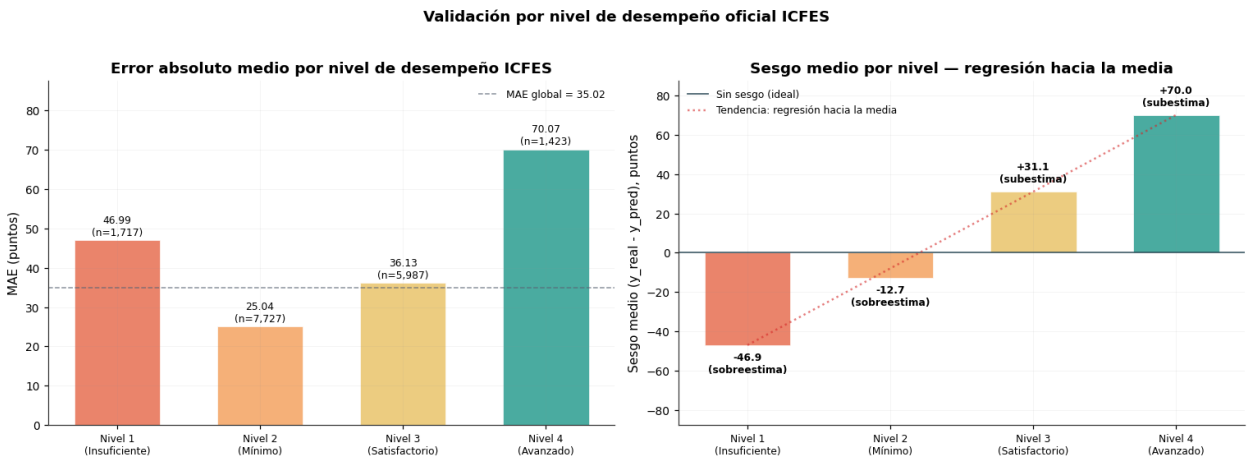


Figura 14. MAE y sesgo medio del modelo por nivel de desempeño oficial del ICFES.

6.4. Interpretabilidad mediante valores SHAP

Verificada la robustez estadística del modelo, el último paso consiste en abrir su caja negra y entender qué variables conducen las predicciones, en qué magnitud y en qué dirección. Para ello se utilizó la metodología SHAP (*SHapley Additive exPlanations*), que asigna a cada variable una contribución, positiva o negativa, al puntaje predicho de cada estudiante, con base en la teoría de los valores de Shapley de juegos cooperativos. SHAP ofrece dos niveles de interpretación: global (qué variables son más importantes en el conjunto completo) y local (qué variables empujaron la predicción de un estudiante concreto). Dado que el modelo final es XGBoost, se empleó shap.TreeExplainer, especializado y exacto para modelos basados en árboles, sobre una muestra aleatoria de 2.000 observaciones del conjunto de prueba.

6.4.1. Importancia global de las variables

La Figura 15 presenta el ranking global de variables por su valor absoluto SHAP promedio, expresado en puntos del puntaje, y la Tabla 13 reporta los diez primeros lugares con su familia de pertenencia.

Tabla 13. Top 10 variables del modelo XGBoost optimizado por importancia global SHAP.

Rango	Variable	|SHAP| medio (puntos)	Familia
1	**NSE establecimiento**	**13,64**	**Colegio**
2	Dedicación a lectura	6,92	Estudiante
3	Edad de presentación	6,74	Estudiante
4	Educación de la madre	4,57	Familia
5	Educación del padre	4,13	Familia
6	Grado presentado	3,58	Estudiante
7	Número de libros en el hogar	3,34	Familia
8	Consumo de lácteos	2,82	Familia

Rango	Variable	|SHAP| medio (puntos)	Familia
9	Jornada del colegio: Mañana	2,38	Colegio
10	Dedicación a internet	2,11	Estudiante

Fuente: elaboración propia.

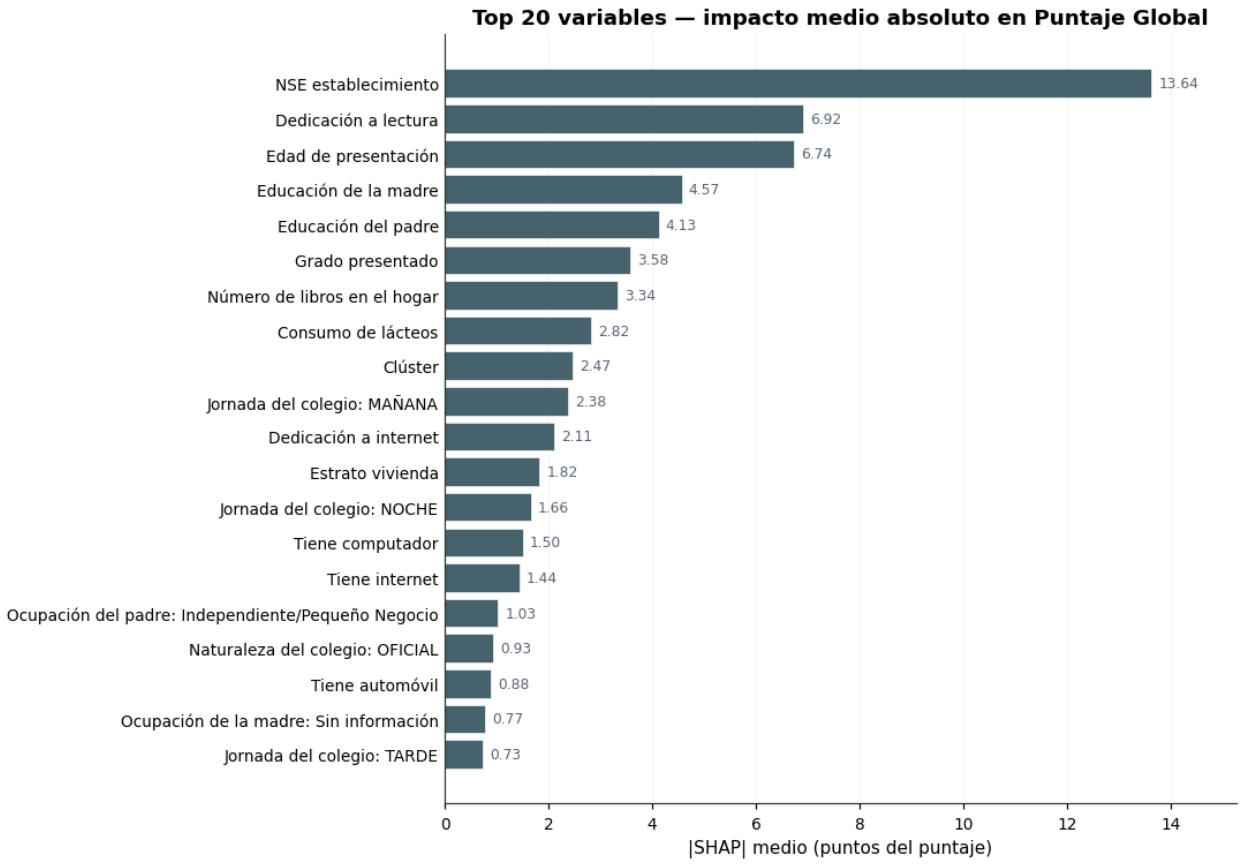


Figura 15. Top 20 variables ordenadas por impacto medio absoluto sobre el puntaje predicho (|SHAP| promedio en puntos).

Los hallazgos son los siguientes:

- NSE establecimiento (13,64 puntos) es la variable más influyente. El nivel socioeconómico del establecimiento, un índice contextual que sintetiza el entorno del colegio, casi duplica el impacto de la siguiente variable, lo que muestra que, en el modelo, el contexto institucional tiene mayor contribución predictiva que cualquier atributo individual del estudiante.

- Dedicación a lectura (6,92 puntos) constituye uno de los hallazgos más reveladores del proyecto: la dedicación a la lectura diaria es la segunda variable más influyente del modelo, solo por detrás del nivel socioeconómico del establecimiento y por encima de la educación de ambos padres, del número de libros en el hogar y de cualquier indicador de posesión

material. Su impacto supera al de tener internet o computador, lo que cuestiona las políticas centradas únicamente en dotar de tecnología, pues el acceso no garantiza el uso productivo. Este resultado matiza el énfasis de la literatura más cercana: mientras que Suaza-Medina y colaboradores [22] situaron el índice de nivel socioeconómico como la variable dominante del Saber 11°, sin destacar los hábitos de lectura, aquí un factor conductual y de capital cultural y, a diferencia del contexto socioeconómico, modificable mediante intervención educativa, rivaliza con los determinantes estructurales. La jerarquía es además coherente con Martínez-Abad y Chaparro [7], quienes concluyeron que los factores de orden personal pesan más que los de orden social o escolar.

- Edad de presentación (6,74 puntos) ocupa el tercer lugar: edades por encima de la esperada se asocian con rezago escolar, que correlaciona negativamente con el puntaje.

- La educación de la madre (4,57 puntos) supera a la del padre (4,13 puntos), hallazgo consistente con la investigación sobre capital cultural, donde el rol materno tiene mayor incidencia en el rendimiento escolar. La variable Grado presentado (3,58 puntos) captura el efecto negativo de la educación de adultos.

La Figura 16 complementa el ranking con la dirección del efecto mediante un diagrama *beeswarm*. En NSE establecimiento y en Dedicación a lectura, los valores altos (en rojo) se ubican a la derecha del cero, confirmando una relación monotónica creciente con el puntaje. En Edad de presentación, las edades altas se asocian con un efecto negativo. La estructura del diagrama para Jornada del colegio: Noche muestra que pertenecer a la jornada nocturna se asocia con un efecto negativo claro.

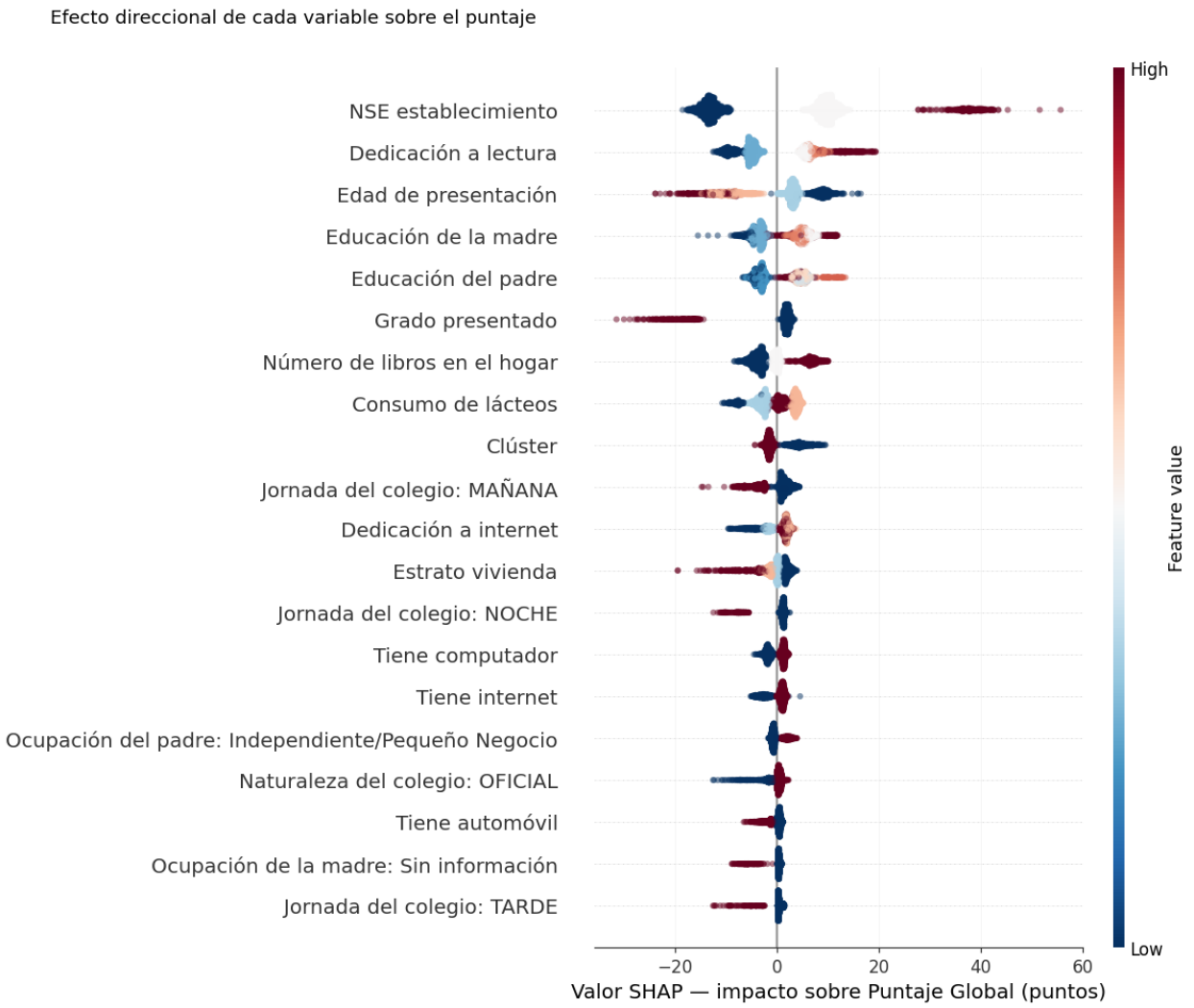


Figura 16. Diagrama *beeswarm* de los valores SHAP por variable. Cada punto es un estudiante; la posición horizontal indica magnitud y dirección del impacto sobre el puntaje, y el color codifica el valor de la variable (rojo = alto, azul = bajo).

6.4.2. Interpretabilidad local: explicaciones individuales

La interpretabilidad local descompone la predicción de un estudiante concreto en las contribuciones de cada variable, vista útil para que un rector o coordinador comprenda los factores detrás de un pronóstico y diseñe intervenciones focalizadas. Se seleccionaron dos casos ilustrativos a partir del valor base del modelo (253,03 puntos, el puntaje esperado promedio). La Figura 17 muestra el *waterfall plot* de un estudiante con predicción alta y la Figura 18 muestra el diagrama para un estudiante con una predicción baja.

El estudiante con predicción alta alcanza 355,49 puntos, 102,46 por encima del valor base: acumula ventajas socioeconómicas en múltiples dimensiones simultáneamente, alto nivel

socioeconómico del establecimiento, educación parental elevada, hábito de lectura y un perfil de colegio favorable. El estudiante con predicción baja obtiene 168,86 puntos, 84,17 por debajo del valor base, con el patrón inverso: bajo nivel socioeconómico del establecimiento, baja educación parental y escaso hábito de lectura. La descomposición confirma que el modelo no se apoya en una sola variable, sino en la acumulación coherente de factores contextuales y de hábito.

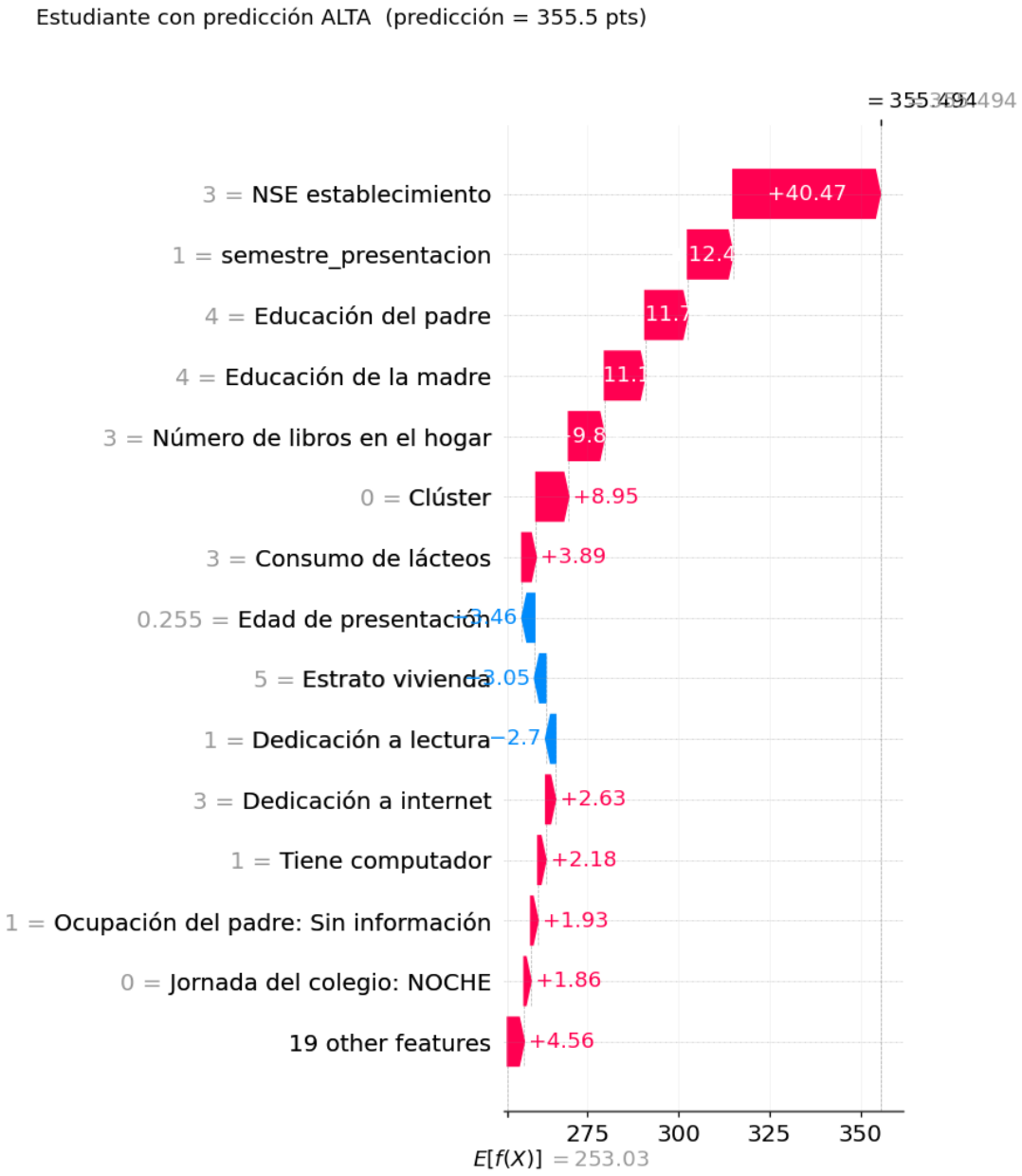


Figura 17. *Waterfall plot* del estudiante con predicción alta (355,5 puntos frente al valor base de 253,0 puntos; diferencia +102,5 puntos). Las barras rojas indican contribuciones positivas y las azules negativas, ordenadas por magnitud.

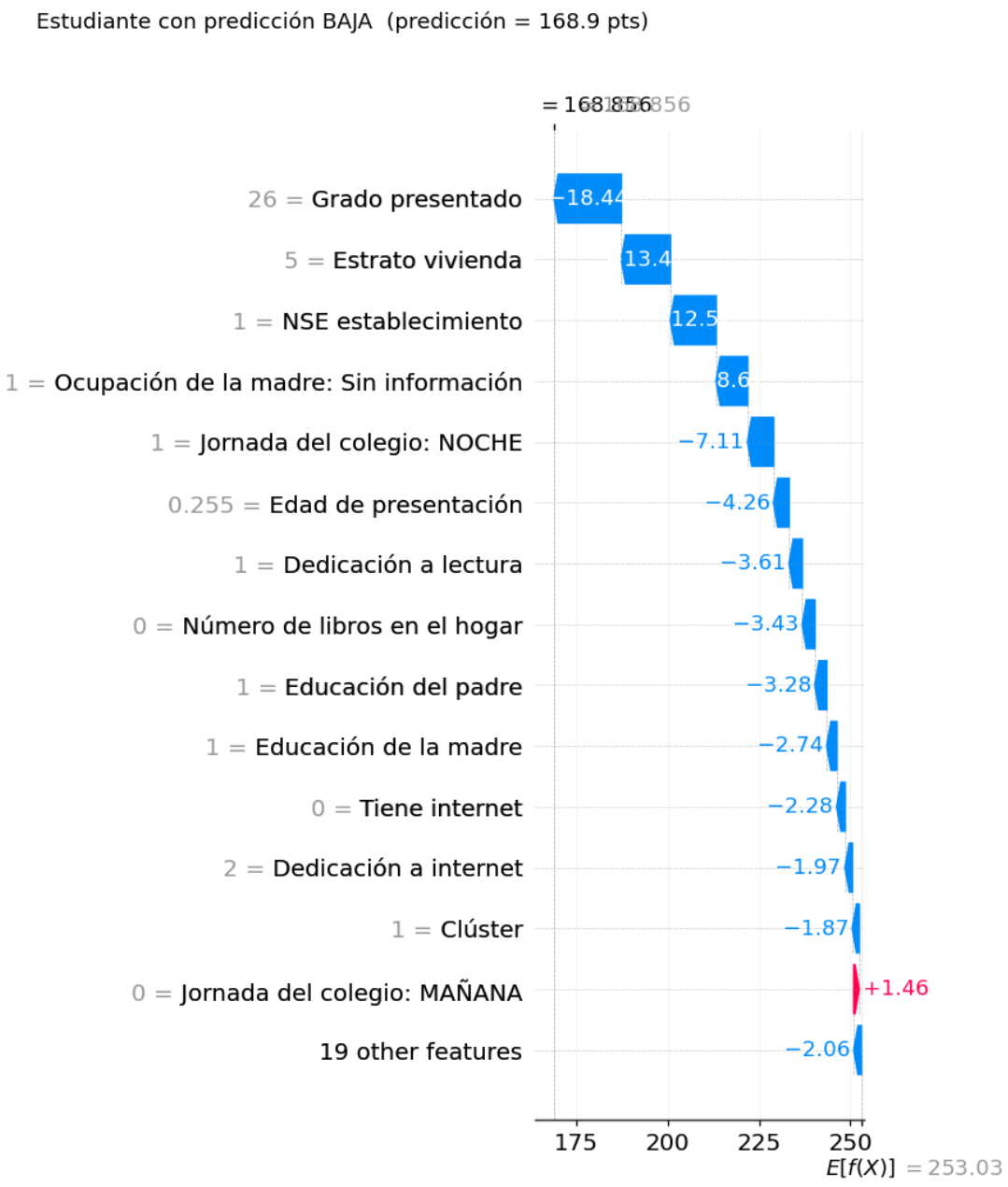


Figura 18. *Waterfall plot* del estudiante con predicción baja (168,9 puntos frente al valor base de 253,0 puntos; diferencia −84,2 puntos). El perfil de contribuciones es contrario al del caso anterior.

El análisis muestra una conclusión metodológica central: el desempeño en el Saber 11° en Barranquilla está fuertemente mediado por el contexto del establecimiento y por el capital cultural familiar —educación de los padres, libros en el hogar, hábitos de lectura— más que por

la riqueza material directa. Las variables de contexto institucional y de hábito superan en impacto SHAP a las variables de posesión material como el automóvil o los electrodomésticos.

6.5. Selección de modelo final y limitaciones

Con base en los análisis presentados, se selecciona como modelo final el XGBoost con hiperparámetros optimizados:

```
XGBRegressor(n_estimators=400, learning_rate=0.05, max_depth=4,
      subsample=0.8, colsample_bytree=1.0, reg_lambda=2.0)
```

Sus métricas finales sobre la cohorte de prueba (2024) son R^2=0,3795, MAE=35,03 puntos y RMSE=43,73 puntos. La selección no se basa únicamente en el mayor R^2 entre los candidatos, sino en el balance entre rendimiento predictivo, estabilidad bajo validación temporal, ausencia de patrones graves de error en el análisis de residuos e interpretabilidad mediante SHAP. El modelo es apto como herramienta de apoyo a la toma de decisiones en la Secretaría de Educación Distrital, con las siguientes condiciones y limitaciones explícitas:

- Techo predictivo. El R^2≈0,38 indica que las variables disponibles explican alrededor del 38% de la varianza del puntaje. El resto depende de factores no medidos: motivación, calidad docente específica, eventos individuales y factores institucionales del aula.
- Sesgo por deriva temporal. El modelo subestima en promedio 6,4 puntos a la cohorte 2024, reflejo del aumento del puntaje medio entre cohortes. Conviene reentrenarlo periódicamente con las cohortes más recientes.
- Regresión hacia la media. El modelo sobreestima a los estudiantes de bajo desempeño y subestima a los de alto desempeño (sesgo de −47 a +70 puntos entre niveles ICFES). Es útil para evaluación diagnóstica, pero no para identificar talento excepcional ni casos extremos individuales.
- Capacidad explicativa desigual por sector. El modelo explica mejor el sector no oficial (R^2=0,388) que el oficial (R^2=0,324). Su uso en política pública del sector oficial debe complementarse con variables institucionales.
- Menor capacidad en educación de adultos. El subgrupo de grado 26 tiene R^2=0,141; esta población heterogénea requiere análisis específico y, posiblemente, un modelo diferenciado.
- Imprecisión en los extremos. La heterocedasticidad leve evidenciada por Breusch-Pagan indica que el modelo es menos preciso al predecir puntajes muy bajos o altos.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

El desarrollo del proyecto demostró que es posible construir, con datos públicos y técnicas estándar de ciencia de datos, un modelo predictivo del desempeño académico contextualizado al Distrito de Barranquilla, interpretable y con valor para la toma de decisiones educativas. De este modo se da respuesta a la pregunta de investigación *¿cómo pueden los modelos de aprendizaje supervisado predecir el desempeño académico de los estudiantes del Distrito de Barranquilla e identificar los factores que expliquen dicha predicción?*: un modelo XGBoost optimizado, entrenado sobre 18 predictores, predice el puntaje global con un error medio de 35 puntos en una escala de 0 a 500 y explica cerca del 38% de su varianza sobre una cohorte futura no observada, al tiempo que identifica de forma explícita, mediante valores SHAP, los factores que sustentan cada predicción. La utilidad práctica del modelo no radica en la precisión punto a punto, sino en su capacidad para señalar de manera sistemática y consistente los factores que más inciden en el rendimiento, ofreciendo evidencia empírica para políticas educativas focalizadas.

Las preguntas de sistematización se responden de forma articulada con los resultados de cada fase:

- Sobre las variables relevantes: las fuentes oficiales contienen un conjunto reducido de variables con asociación significativa al desempeño. El análisis bivariado (ANOVA con η^2) y los métodos algorítmicos (Información Mutua y Random Forest) convergieron en 17 predictores, entre los que destacan el nivel socioeconómico del establecimiento, la educación de la madre y del padre, la jornada, la edad de presentación y los hábitos de lectura. Las variables de contexto institucional y capital cultural superan en poder explicativo a las de posesión material.

- Sobre los patrones que explican las diferencias de logro: el análisis no supervisado (FAMD + K-Means) reveló que la población se organiza en torno a un único eje socioeducativo dominante que la divide en dos perfiles, uno aventajado (≈31%) y otro vulnerable (≈69%), con una brecha de 48 puntos en el puntaje promedio. Esta estructura binaria confirma que las diferencias de logro en el Distrito están mediadas principalmente por el contexto socioeducativo agregado y no por una segmentación más fina.

- Sobre el modelo más adecuado: tras comparar cinco algoritmos bajo validación temporal, XGBoost resultó el de mejor balance entre capacidad predictiva y costo. Los modelos lineales mostraron una componente lineal dominante, mientras que Random Forest evidenció sobreajuste; el *boosting* secuencial capturó las interacciones no lineales sin memorizar el conjunto de entrenamiento. Aunque el modelo XGBoost optimizado presentó

el mejor desempeño, la magnitud de la mejora frente a los modelos lineales fue moderada, mostrando que la relación entre las variables socioeducativas y el puntaje global podría ser capturada mediante estructuras aproximadamente lineales.

- Sobre los criterios de selección y la variación del desempeño: la selección no se basó únicamente en el mayor R^2, sino en el balance entre rendimiento, estabilidad bajo validación temporal, ausencia de patrones graves de error en los residuos e interpretabilidad. La validación segmentada mostró que el desempeño varía de forma relevante entre subgrupos, mayor capacidad explicativa en el sector no oficial (R^2=0,390) que en el oficial (R^2=0,323), y menor en educación de adultos (R^2=0,140), lo que define las condiciones bajo las cuales el modelo es confiable. El análisis de interpretabilidad mediante valores SHAP identificó como variables de mayor importancia el nivel socioeconómico del establecimiento, la dedicación a la lectura diaria, la edad de presentación y la educación de los padres, lo que muestra que, dentro del modelo, el contexto institucional y el capital cultural familiar presentan mayor contribución predictiva que la riqueza material directa. Cabe subrayar que los valores SHAP explican la contribución de cada variable a las predicciones del modelo y no permiten inferir causalidad: una variable no causa un mayor puntaje, sino que se asocia con estimaciones más altas del modelo.

7.2. Trabajos futuros

El desarrollo del proyecto deja varias líneas de continuación que pueden fortalecer el alcance del modelo y su aplicabilidad. Se identifican cuatro grupos de extensiones prioritarias.

- Enriquecimiento con variables institucionales para el sector oficial. La menor capacidad explicativa del modelo en colegios oficiales (R^2=0,3240) sugiere que las variables socioeducativas del estudiante son insuficientes para capturar el desempeño en este sector. Una extensión natural consiste en incluir variables institucionales adicionales calidad y rotación docente, recursos del aula, programas de intervención pedagógica vigentes, indicadores de gestión escolar mediante acuerdos de intercambio de información con la Secretaría de Educación Distrital. Esta línea podría reducir la brecha de explicación entre sectores y mejorar el R^2 agregado del modelo.
- Análisis longitudinal de cohortes y estudios de panel. El presente proyecto adopta un enfoque transversal, evaluando cada cohorte como una observación independiente. Una extensión metodológicamente potente consiste en construir un panel longitudinal que siga a las mismas instituciones (no a los mismos estudiantes, dada la anonimización) a lo largo del período 2014–2024 e identifique trayectorias de mejora o deterioro institucional.

- Integración con datos de la Secretaría de Educación Distrital y datos de hogares del DANE. El cuestionario socioeconómico del ICFES tiene limitaciones inherentes (autoreporte, periodicidad fija). Una integración con registros administrativos de la Secretaría de Educación Distrital (matrícula, deserción, repitencia institucional) y con datos del DANE a nivel de barrio (índice multidimensional de pobreza, equipamiento urbano) permitiría enriquecer el contexto de cada estudiante y reducir el sesgo derivado del autoreporte. Esta línea exige acuerdos formales de transferencia de datos y un trabajo cuidadoso de georreferenciación.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] "¿Cómo está Barranquilla en educación? Retos y prioridades (2020-2023)", Fundación Empresarios por la Educación. [En línea]. Disponible en: https://www.fundacionexe.org.co/document/como-esta-barranquilla-en-educacion-retos-y-prioridades-2024-2027/

[2] "Informe Nacional de Resultados Saber 11° 2023", Instituto Colombiano para la Evaluación de la Educación – ICFES, Bogotá, Colombia, ago. 2025. [En línea]. Disponible en: https://www.icfes.gov.co/wp-content/uploads/2025/04/Informe_Saber11_2023.pdf

[3] C. Romero y S. Ventura, "Data mining in education", *WIREs Data Min. Knowl. Discov.*, vol. 3, núm. 1, pp. 12–27, ene. 2013, doi: 10.1002/widm.1075.

[4] P. Amar Sepúlveda, "I Informe de Gestión 2025 – Secretaría Distrital de Educación", Secretaria de Educación del Distrito de Barranquilla, Barranquilla, Colombia, jun. 2025. [En línea]. Disponible en: https://barranquilla.gov.co/educacion/publicaciones/informes/informes-de-gestion-2/informes-de-gestion-2025-secretaria-distrital-de-educacion

[5] P. Amar Sepúlveda, "II Informe de Gestión 2025 – Secretaría Distrital de Educación", Secretaria de Educación del Distrito de Barranquilla, Barranquilla, Colombia, oct. 2025. [En línea]. Disponible en: https://barranquilla.gov.co/educacion/publicaciones/informes/informes-de-gestion-2/informes-de-gestion-2025-secretaria-distrital-de-educacion

[6] P. Amar Sepúlveda, "III Informe de Gestión 2025 – Secretaría Distrital de Educación", Secretaria de Educación del Distrito de Barranquilla, Barranquilla, Colombia, oct. 2025. [En línea]. Disponible en: https://barranquilla.gov.co/educacion/publicaciones/informes/informes-de-gestion-2/informes-de-gestion-2025-secretaria-distrital-de-educacion

[7] F. Martínez Abad y A. A. Chaparro Caso López, "Data-mining techniques in detecting factors linked to academic achievement", *Sch. Eff. Sch. Improv.*, vol. 28, núm. 1, pp. 39–55, ene. 2017, doi: 10.1080/09243453.2016.1235591.

[8] J. A. Solano, D. J. Lancheros Cuesta, S. F. Umaña Ibáñez, y J. R. Coronado-Hernández, "Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test", *Procedia Comput. Sci.*, vol. 198, pp. 512–517, 2022, doi: 10.1016/j.procs.2021.12.278.

[9] G. James, D. Witten, T. Hastie, y R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. en Springer Texts in Statistics, vol. 103. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-7138-7.

[10] Pagès, J., "Analyse factorielle de données mixtes", *Rev. Stat. Appliquée*, vol. 52, núm. 4, pp. 93–111, 2004.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[12] 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*, San Francisco California USA: ACM, ago. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[13] C. Bergmeir y J. M. Benítez, "On the use of cross-validation for time series predictor evaluation", *Inf. Sci.*, vol. 191, pp. 192–213, may 2012, doi: 10.1016/j.ins.2011.12.028.
[14] S. Lundberg y S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", el 25 de noviembre de 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.
[15] C. Márquez-Vera, A. Cano, C. Romero, y S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data", *Appl. Intell.*, vol. 38, núm. 3, pp. 315–330, abr. 2013, doi: 10.1007/s10489-012-0374-8.
[16] B. H. Quintero Londoño, J. A. Valero, y M. A. Vargas Peñaloza, "Informe Nacional de Resultados Saber 11° 2024", Instituto Colombiano para la Evaluación de la Educación – ICFES, Bogotá, Colombia, ago. 2025. [En línea]. Disponible en: https://www.icfes.gov.co/wp-content/uploads/2025/09/INFORME_NACIONAL_RESULTADOS_SABER_11_2024.pdf
[17] L. S. Rodrigues, M. Dos Santos, I. Costa, y M. A. L. Moreira, "Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review", *Procedia Comput. Sci.*, vol. 214, pp. 680–687, 2022, doi: 10.1016/j.procs.2022.11.229.
[18] T. M. Mitchell, *Machine learning*, Nachdr. en McGraw-Hill series in Computer Science. New York: McGraw-Hill, 2013.
[19] Instituto Colombiano para la Evaluación de la Educación (ICFES), "Guía metodológica para el acceso y uso de la información – Examen Saber 11°", Instituto Colombiano para la Evaluación de la Educación (ICFES), Bogotá, Colombia, ago. 2024.
[20] Congreso de la República de Colombia, *Ley 1324 de 2009 "Por la cual se fijan parámetros y criterios para organizar el sistema de evaluación de resultados de la calidad de la educación, se dictan normas para el fomento de una cultura de la evaluación, en procura de facilitar la inspección y vigilancia del Estado y se transforma el* ICFES*."*, vol. Diario Oficial No. 47.409. 2009.
[21] Colombia, *Decreto 869 de 2010: Por el cual se reglamenta el Examen de Estado de la Educación Media,* ICFES - *SABER 11°.*, vol. Diario Oficial No. 47.655. 2010.
[22] M. Suaza-Medina, R. Peñabaena-Niebles, y M. Jubiz-Diaz, "A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations", *Sci. Rep.*, vol. 14, núm. 1, p. 25306, oct. 2024, doi: 10.1038/s41598-024-76596-3.
[23] N. I. Mohd Talib, N. A. Abd Majid, y S. Sahran, "Identification of Student Behavioral Patterns in Higher Education Using K-Means *Clustering* and Support Vector Machine", *Appl. Sci.*, vol. 13, núm. 5, p. 3267, mar. 2023, doi: 10.3390/app13053267.
[24] A. Asselman, M. Khaldi, y S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm", *Interact. Learn. Environ.*, vol. 31, núm. 6, pp. 3360–3379, ago. 2023, doi: 10.1080/10494820.2021.1928235.
[25] Instituto Colombiano para la Evaluación de la Educación (ICFES), "Bases departamentales – Examen Saber 11°: Departamento del Atlántico (código DANE 8)". DataICFES. [Txt].

Disponible en: https://icfesgovco.sharepoint.com/sites/BasesDataIcfes (acceso con registro previo requerido)
[26] DataIcfes, “Diccionario Examen Saber 11°”. mayo de 2025.

ANEXO A. ANÁLISIS UNIVARIADO

Este anexo amplía el análisis univariado del conjunto de datos depurado, complementando la síntesis incluida en el cuerpo del documento. Sobre el *dataset* de modelado, conformado por la variable objetivo Puntaje Global (`punt_global`) y 42 predictores (41 categóricos y 1 numérico continuo), se examinó la distribución individual de cada variable con tres propósitos: verificar que la limpieza eliminó las anomalías detectadas en el diagnóstico estructural, caracterizar la forma, dispersión y asimetría de cada variable para anticipar el comportamiento de las pruebas bivariadas, e identificar categorías minoritarias o concentraciones extremas que pudieran condicionar la modelación.

A.1 VARIABLE OBJETIVO: PUNTAJE GLOBAL

La variable objetivo del enfoque de regresión es Puntaje Global (`punt_global`), el puntaje global del Examen Saber 11° en la escala teórica de 0 a 500 puntos. Como se resume en la Tabla A1, sobre los 133.985 registros válidos la media fue de 254,65 puntos y la mediana de 251, con una desviación estándar de 53,50. La distribución resultó ligeramente asimétrica a la derecha (asimetría = 0,30) y algo más plana que una distribución normal (curtosis = −0,46). Estos valores son coherentes con la referencia teórica del ICFES para 2014-II (media 250, desviación 50). El rango intercuartílico se extendió de 213 puntos (P25) a 293 puntos (P75), y los percentiles extremos P10 y P90 se ubicaron en 187 y 328 puntos, respectivamente.

Tabla A1. Estadísticos descriptivos de la variable objetivo Puntaje Global.

Estadístico	Valor
N (registros válidos)	133.985
Media	254,65
Mediana	251
Desviación estándar	53,50
Mínimo	0
Máximo	500
Asimetría (*skewness*)	0,30
Curtosis (exceso)	−0,46
Percentil 10 (P10)	187
Percentil 25 (P25)	213
Percentil 75 (P75)	293
Percentil 90 (P90)	328
Referencia teórica ICFES 2014-II (media / DE)	250 / 50

Fuente: elaboración propia con base en el dataset de modelado del Examen Saber 11° (Barranquilla).

La Figura A1 presenta la forma de la distribución mediante un histograma de densidad acompañado de su diagrama de caja y bigotes, donde se aprecian la posición de la media y la mediana frente a la referencia teórica de 250 puntos.

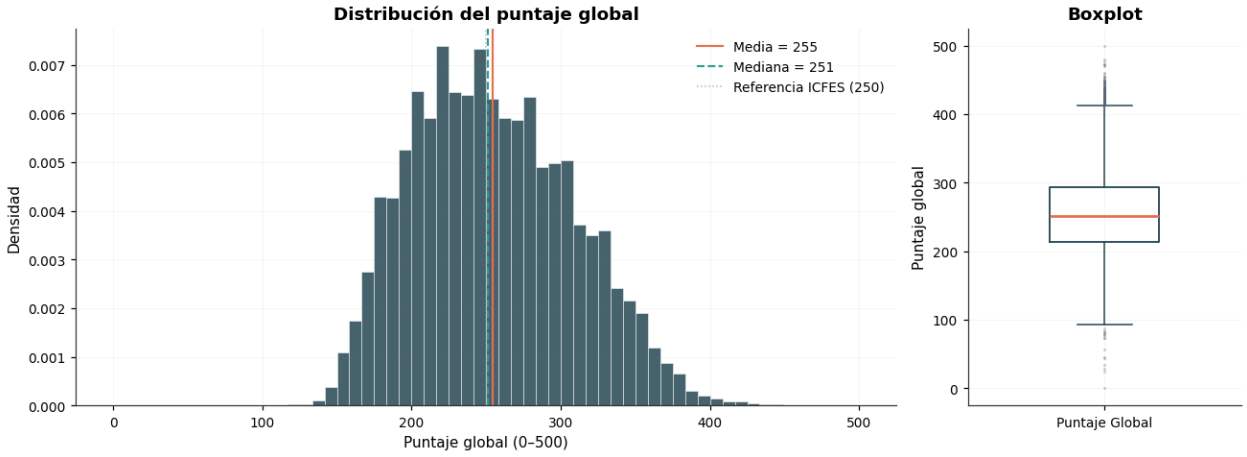


Figura A1. Distribución del puntaje global (punt_global): histograma de densidad y *boxplot*.

A.2 PREDICTOR CUANTITATIVO CONTINUO: EDAD DE PRESENTACIÓN

Tras el tipado semántico de la Fase 2, el único predictor cuantitativo continuo es `la Edad de presentación`. Como se observa en la Tabla A2, la edad de presentación se concentró fuertemente en torno a los 17 años (mediana = 17; media = 17,38; desviación estándar = 2,18 años). La distribución resultó marcadamente leptocúrtica (curtosis = 146,29) y muy asimétrica a la derecha (asimetría = 9,84): el rango intercuartílico fue de apenas un año (Q1 = 17, Q3 = 18) y la cola superior se extendió de forma atenuada pero persistente hasta los 72 años. Esa cola corresponde principalmente a la población de educación para adultos identificada en el diagnóstico estructural, cuya presencia justifica conservar de manera simultánea `Edad de presentación` y `Grado presentado` como predictores diferenciados.

Tabla A2. Estadísticos descriptivos del predictor cuantitativo continuo Edad de presentación.

Estadístico	Valor
N (registros válidos)	133.985
Media	17,38
Mediana	17,00
Desviación estándar	2,18
Mínimo	12
Primer cuartil (Q1)	17
Tercer cuartil (Q3)	18

Estadístico	Valor
Máximo	72
Asimetría (*skewness*)	9,84
Curtosis (exceso)	146,29

Fuente: elaboración propia con base en el dataset de modelado del Examen Saber 11°.

La Figura A2 ilustra esta concentración y la cola derecha extendida mediante el histograma y el *boxplot* de la variable.

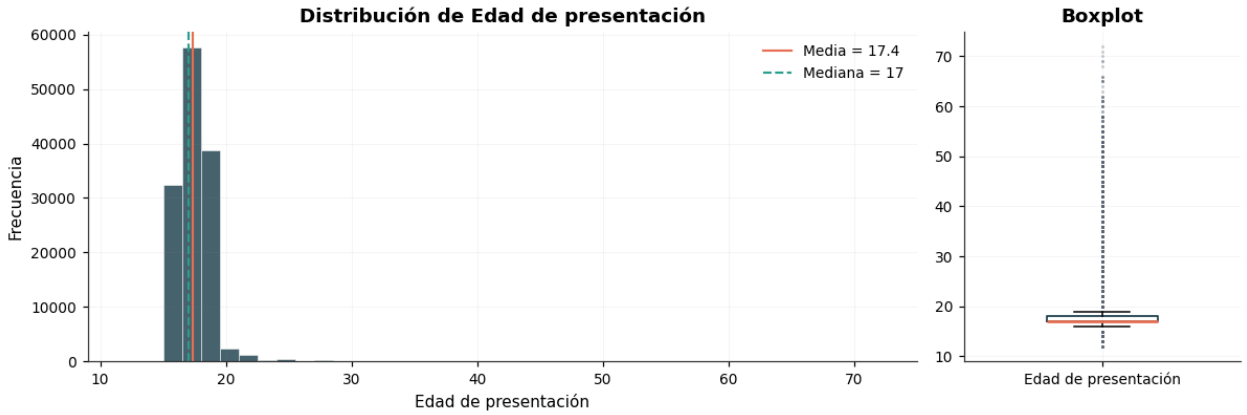


Figura A2. Distribución de la edad de presentación (edad_presentacion): histograma y *boxplot*.

A.3 PREDICTORES CATEGÓRICOS

El inventario de cardinalidad sobre los 41 predictores categóricos arrojó la composición que resume la Tabla A3: 18 variables binarias (2 categorías), 15 de cardinalidad baja (3 a 5 categorías) y 7 de cardinalidad media (6 a 10 categorías). El bloque binario agrupa los indicadores de tenencia de bienes y servicios del hogar (por ejemplo, `tenencia de computador` o `internet`) y características institucionales como `la naturaleza o carácter del colegio`. El bloque de cardinalidad media reúne las variables con mayor número de niveles, encabezadas por los rangos de trabajo de los padres y los niveles educativos parentales.

Tabla A3. Inventario de cardinalidad de los 41 predictores categóricos del *dataset* depurado.

Banda de cardinalidad	N.º de categorías	N.º de variables	Ejemplos representativos
Binarias	2	18	Indicadores de tenencia (fami_tiene*), cole_naturaleza, cole_calendario, cole_bilingue, estu_grado
Baja cardinalidad	3 – 5	15	cole_caracter, fami_situacioneconomica, fami_numlibros, estu_dedicacioninternet

Banda de cardinalidad	N.º de categorías	N.º de variables	Ejemplos representativos
Media cardinalidad	6 – 10	7	cole_jornada, fami_educacionmadre/padre, fami_estratovivienda, fami_trabajolabor*, anio_presentacion

Fuente: elaboración propia con base en el dataset de modelado del Examen Saber 11°.

La Tabla A4 detalla, para los 20 predictores categóricos de mayor cardinalidad, el número de categorías, la categoría modal y la proporción de registros que esta concentra.

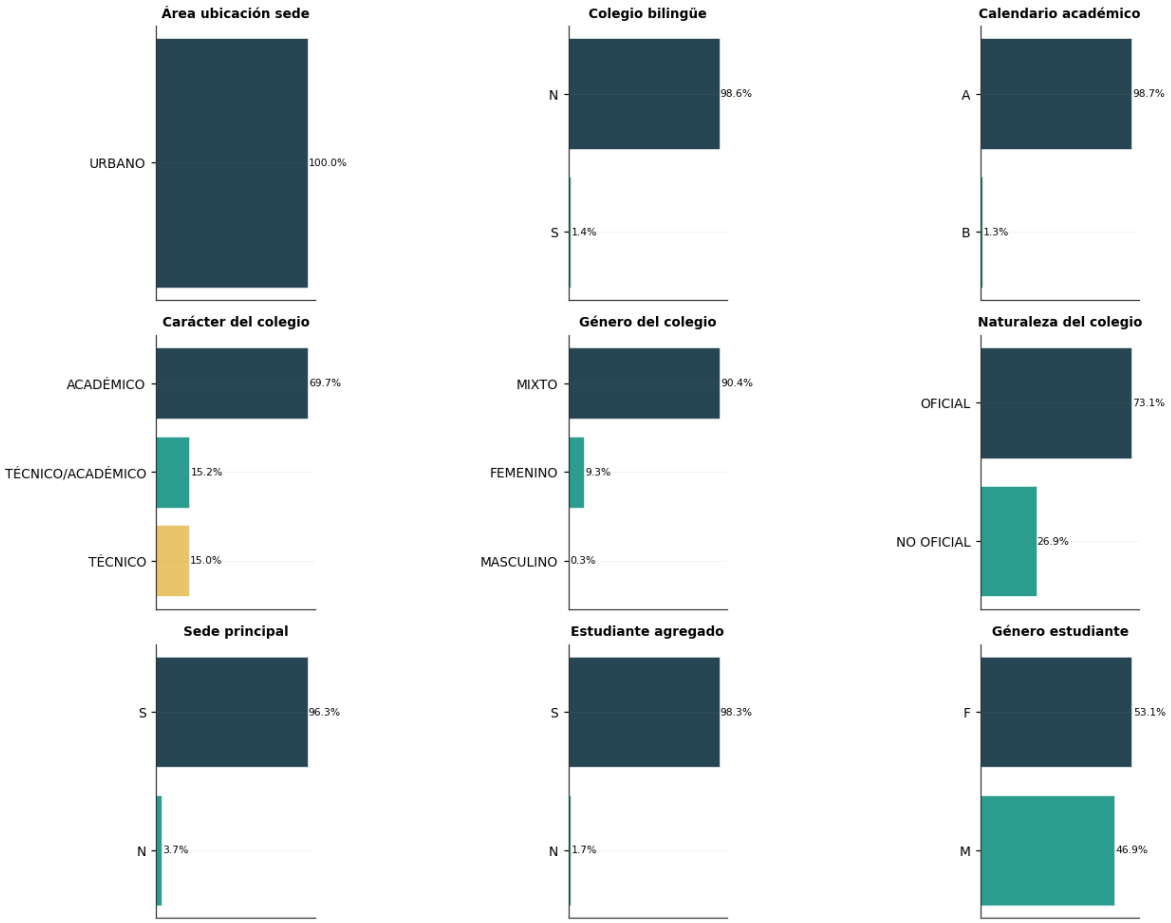
Tabla A4. Inventario detallado de los 20 predictores categóricos de mayor cardinalidad: categoría y porcentaje modales.

Etiqueta	N.º cat.	Categoría modal	% modal
Año de presentación	8	2018	13,3
Trabajo de la madre	7	Hogar/No trabaja	39,3
Trabajo del padre	7	Operativo/Servicios	31,0
Educación del padre	6	Secundaria	43,6
Educación de la madre	6	Secundaria	42,8
Estrato de la vivienda	6	Estrato 1	36,9
Jornada del colegio	6	UNICA	36,9
Personas en el hogar	5	3 a 4	44,2
Dedicación a internet	5	Entre 1 y 3 horas	33,3
Dedicación a lectura diaria	5	30 minutos o menos	37,5
Horas de trabajo semanal	5	0	77,1
NSE del establecimiento	4	2	47,2
Consumo carne/pescado/huevo	4	Todos o casi todos los días	47,8
Número de libros en el hogar	4	0 A 10 LIBROS	46,8
Consumo cereal/frutas/legumbres	4	1 o 2 veces por semana	45,0
Consumo leche/derivados	4	Todos o casi todos los días	33,6
Carácter del colegio	3	ACADÉMICO	69,7
Situación económica	3	Igual	59,4
Pertenencia étnica	3	No	74,3
Es repitente	3	No	98,5

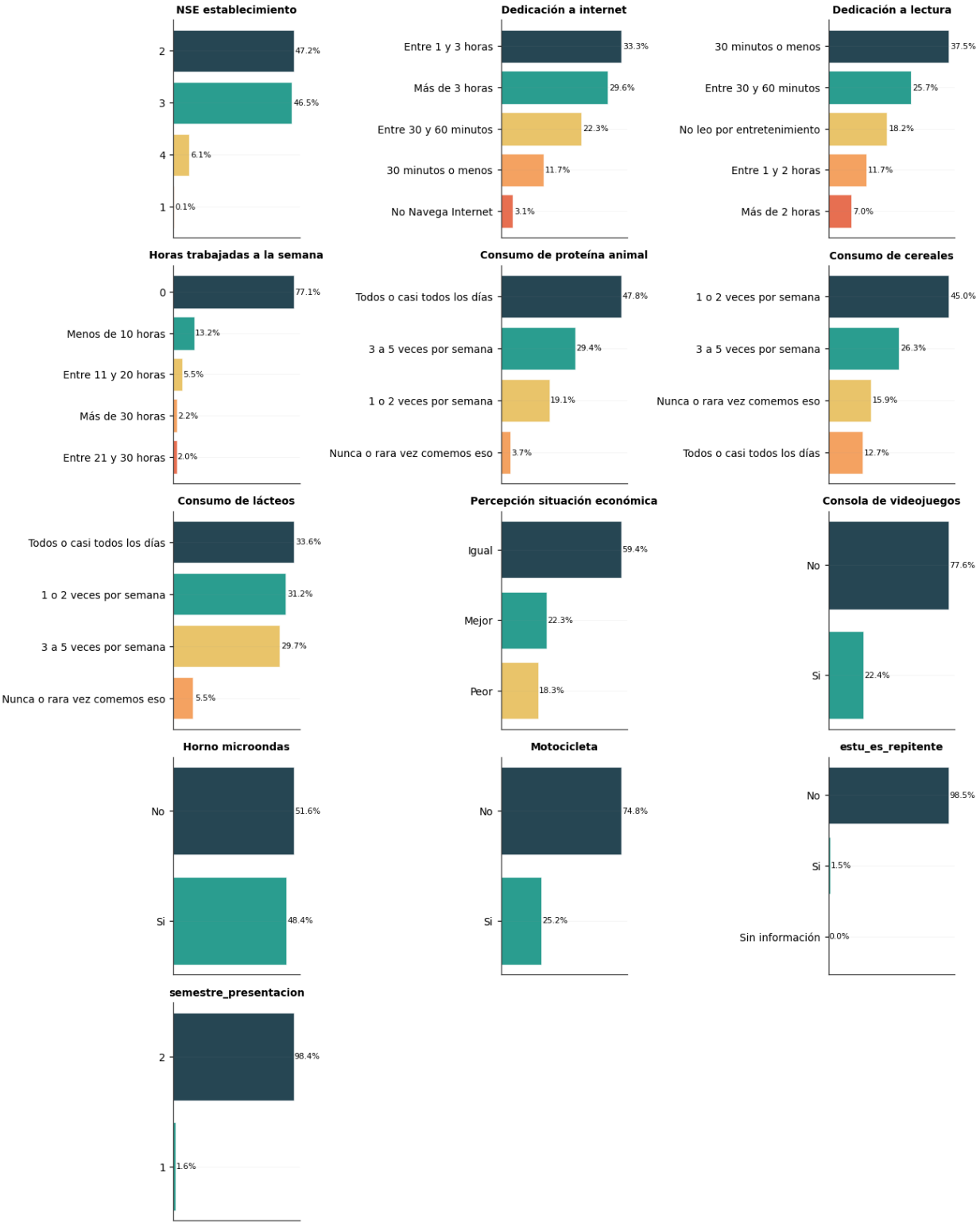
Fuente: elaboración propia con base en el dataset de modelado del Examen Saber 11°. Variables ordenadas de forma descendente por número de categorías.

Las Figuras A3 y A4 presentan la distribución de frecuencias de cada predictor categórico, agrupados por nivel de cardinalidad para una lectura ordenada. La Figura A3 reúne los predictores de baja cardinalidad (5 categorías o menos), incluyendo el bloque de variables binarias.

Figura A3. Distribución de los predictores categóricos de baja cardinalidad (≤ 5 categorías).



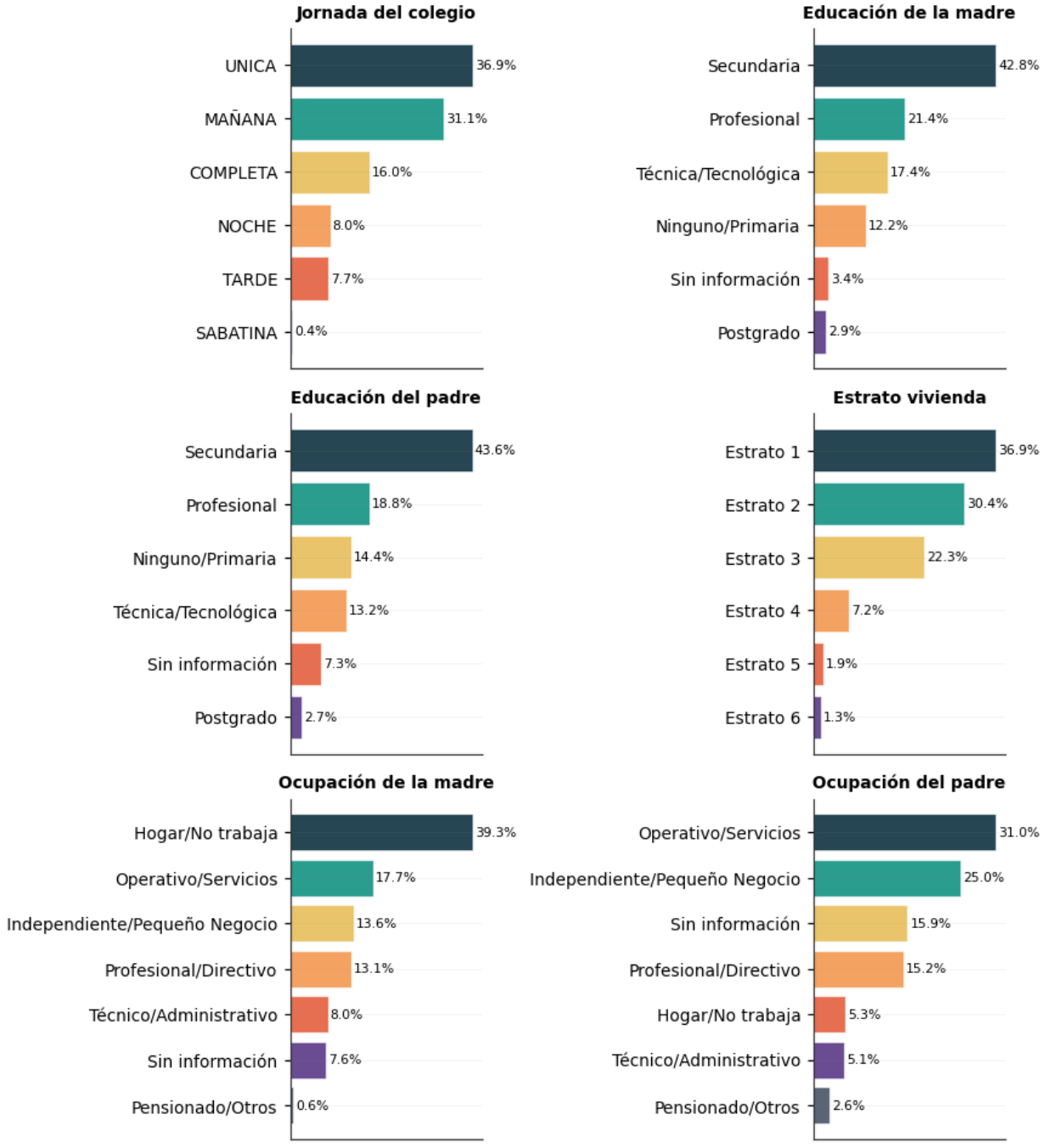

Sede principal
S
96.3%
N
3.7%
Estudiante agregado
S
98.3%
N
1.7%
Género estudiante
F
53.1%
M
46.9%
Grado presentado
11
89.1%
26
10.9%
Privado de libertad
N
99.9%
S
0.1%
Número de cuartos del hogar
3 a 4
51.4%
1 a 2
43.0%
5 a 6
5.6%
Personas en el hogar
3 a 4
44.2%
5 a 6
35.6%
7 a 8
10.1%
9 o más
5.2%
1 a 2
4.9%
Tiene automóvil
No
74.6%
Si
25.4%
Tiene computador
Si
56.5%
No
43.5%
Tiene internet
Si
72.8%
No
27.2%
Tiene lavadora
Si
83.4%
No
16.6%
Servicio de TV
Si
75.9%
No
24.1%
Número de libros en el hogar
0 A 10 LIBROS
46.8%
11 A 25 LIBROS
30.6%
26 A 100 LIBROS
17.8%
MÁS DE 100 LIBROS
4.7%
Tiene etnia
No
74.3%
Sin información
24.7%
Si
1.0%
Discapacidad
N
99.0%
S
1.0%

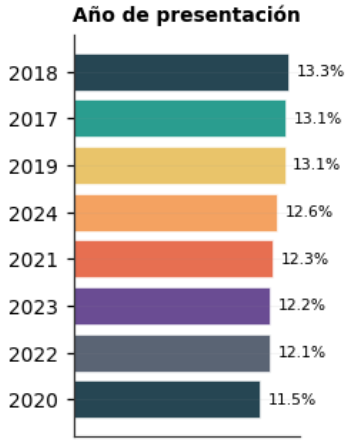


Fuente: elaboración propia con base en el dataset de modelado del Examen Saber 11°.

La Figura A4 reúne los siete predictores de cardinalidad media (entre 6 y 10 categorías), que concentran las variables socioeconómicas y educativas con mayor número de niveles.

Figura A4. Distribución de los predictores categóricos de media cardinalidad (6 a 10 categorías).





Fuente: elaboración propia con base en el dataset de modelado del Examen Saber 11°.

En conjunto, el análisis univariado confirmó que la depuración produjo distribuciones consistentes con la naturaleza de cada variable y con la referencia teórica del examen, y permitió anticipar dos rasgos relevantes para la modelación: la fuerte concentración de la edad en torno a los 17 años con una cola adulta minoritaria, y la presencia de predictores categóricos altamente concentrados en una sola categoría, cuyo aporte informativo se valoró posteriormente en el análisis bivariado y en la selección de características.