Análisis de resultados globales de las pruebas Saber 11 de 2015 a 2019

Claudia Milena Arteaga Ceballos

Asesor

Julio Eduardo Mejía Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

Las bases de datos públicas compartidas en la plataforma “Datos Abiertos” dan la posibilidad de realizar ejercicios académicos con información real. En el ámbito académico, el acceso a este tipo de información fomenta el desarrollo de proyectos que buscan generar conocimiento a partir del análisis de datos reales, contribuyendo a la toma de decisiones basadas en evidencia. Al explorar las bases de datos académicas disponibles en la biblioteca de la universidad, encuentro que la revisión de los resultados de las pruebas Saber 11 se ha abordado desde muchas perspectivas, analizando aspectos como el desempeño académico, las condiciones socioeconómicas, la infraestructura educativa y otros elementos clave. El presente estudio se centra en el análisis de estos resultados con un enfoque basado en técnicas de ciencia de datos.

El conjunto de datos utilizado en este proyecto contiene múltiples variables demográficas, académicas y socioeconómicas. Estas variables permiten formular diversas preguntas de investigación, explorando la relación entre factores como el nivel educativo de los padres, el estrato socioeconómico, el tipo de colegio (público o privado), la ubicación geográfica y los puntajes obtenidos en las pruebas. Para ello, se trabajará con el set de datos de resultados únicos para las pruebas Saber 11 de los años 2010 al 2022. Sin embargo, el análisis se acotará a los años 2015 a 2019, 5 años en total (se excluyen los años 2020 en adelante para evitar variaciones por los efectos que pudo tener la pandemia en el desempeño de los estudiantes). La base de datos tiene 51 variables y, para el período de estudio (resultados 2015 a 2019) hay alrededor de 2.2 millones de registros. El proyecto abarcó diversas técnicas de ciencia de datos, incluyendo exploración y análisis de datos, limpieza y estandarización, análisis de correlaciones, y generación y evaluación de modelos predictivos. Entre los modelos a utilizar, se consideran regresión lineal simple y algoritmos de clasificación, lo que permitirá identificar patrones y

relaciones significativas entre las variables. Finalmente, se presentarán conclusiones y hallazgos relevantes sobre los factores que influyen en los puntajes de las pruebas Saber 11, contribuyendo al entendimiento del desempeño académico en Colombia.

Palabras claves: Educación secundaria, Pruebas Saber 11, ICFES, Ciencia de Datos, Aprendizaje Automático.

Abstract

Public databases shared on the “Datos Abiertos” platform enable academic exercises with real-world information. Access to such data fosters research projects that generate knowledge through data analysis, supporting evidence-based decision-making. A review of academic databases available at the university library reveals that Saber 11 test results have been analyzed from various perspectives, including academic performance, socioeconomic conditions, and educational infrastructure. This study focuses on analyzing these results using data science techniques.

The dataset used contains multiple demographic, academic, and socioeconomic variables, allowing for the exploration of relationships between parental education, socioeconomic status, school type (public or private), geographic location, and test scores. The study uses unique test result data from 2010 to 2022, but focuses on the 2015–2019 period to avoid distortions caused by the COVID-19 pandemic. The database includes 51 variables and approximately 2.9 million records for the selected period. The project applies various data science techniques, including data exploration, cleaning, standardization, correlation analysis, and predictive modeling. Methods such as simple linear regression and classification algorithms will be employed to identify patterns and significant relationships among variables. The study aims to provide insights into the factors influencing Saber 11 test scores, contributing to the understanding of academic performance in Colombia.

Keywords: Secondary education, Saber 11 Test, ICFES, Data Science, Machine Learning.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El sistema educativo en Colombia enfrenta desafíos significativos en materia de equidad y calidad, evidenciados en los resultados de las pruebas estandarizadas como las Saber 11. Estas evaluaciones permiten medir el conocimiento de los estudiantes, pero gracias a la encuesta sociodemográfica incluida en la prueba también se pueden evidenciar las desigualdades socioeconómicas, institucionales y regionales que persisten en el país. Factores como la naturaleza del colegio (público o privado), el calendario, el estrato de los estudiantes, incluso la ubicación geográfica influyen de manera determinante en el desempeño académico, perpetuando brechas que limitan las oportunidades de los jóvenes.

En este contexto, este proyecto aplicado tiene como propósito analizar los patrones y las relaciones entre las variables y los resultados de las pruebas Saber 11. Para ello, se implementó una metodología que integra análisis exploratorio de datos, estadísticas descriptivas y la construcción de modelos predictivos tanto de regresión como de clasificación, empleando exclusivamente variables categóricas provenientes de la encuesta mencionada anteriormente. Este enfoque no solo permite identificar las diferencias más relevantes en los puntajes, sino también evaluar hasta qué punto es posible anticipar el rendimiento académico de un estudiante a partir de las características de su entorno.

A lo largo del estudio se analizó el desempeño de distintos algoritmos tanto de regresión como de clasificación, aplicados a la predicción del puntaje global en las pruebas Saber 11. Para las tareas de regresión se utilizaron modelos como LinearRegression, BaggingRegressor, RandomForestRegressor y XGBRegressor, evaluados con dos esquemas de codificación de variables categóricas: Label Encoding y OneHot Encoding. En el caso de la clasificación binaria del puntaje global, se aplicaron algoritmos como XGBClassifier, RandomForestClassifier,

DecisionTreeClassifier, LogisticRegression y K-Nearest Neighbors, también bajo ambos tipos de codificación. El objetivo fue identificar las combinaciones más eficaces entre modelos y transformaciones para estimar el desempeño académico a partir de características exclusivamente categóricas.

Los resultados obtenidos muestran que, a pesar de las limitaciones inherentes al uso exclusivo de variables categóricas, es posible alcanzar niveles aceptables de predicción. En particular, los modelos basados en árboles de decisión, como XGBRegressor y XGBClassifier, mostraron una capacidad notable para modelar relaciones complejas entre las variables, logrando explicar hasta el 32.3% de la varianza del puntaje y alcanzando niveles de precisión cercanos al 68% en escenarios de clasificación binaria.

De esta manera, este estudio contribuye al entendimiento de los factores asociados al rendimiento académico en Colombia y propone una aplicación concreta de técnicas de ciencia de datos en el ámbito educativo, con potencial para orientar futuras investigaciones, políticas públicas y estrategias institucionales orientadas a reducir las desigualdades y mejorar los resultados de aprendizaje en contextos similares

Planteamiento del Problema

Hay una variedad de estudios realizados alrededor de los resultados de las pruebas Saber 11, se encuentra que hay muchos casos que se enfocan en unas pocas variables preseleccionadas para dar respuesta a un problema específico, a continuación, se listan unos cuantos.

Según Rodríguez et al. (2011) , el acceso limitado a las Tecnologías de Información y Comunicaciones (TIC) no solo incrementa la probabilidad de deserción escolar, sino que también reduce las oportunidades de los estudiantes para acceder a la educación superior. Esta relación parece intuitiva, ya que la disponibilidad de herramientas tecnológicas influye directamente en el proceso de aprendizaje. No obstante, es fundamental respaldar estas afirmaciones con evidencia.

En este sentido, en estudios como el de Mendoza-Lozano et al. (2021) se analiza la brecha digital entre estudiantes de secundaria en Colombia utilizando datos de las pruebas Saber 11 entre 2009 y 2018. Los autores examinan el acceso a computadoras e internet en los hogares, así como los determinantes socioeconómicos y geográficos que influyen en esta brecha. En este caso los autores resaltan que se observa una correlación positiva entre el acceso a TIC y los resultados en las pruebas Saber 11, aunque esta relación disminuyó ligeramente con el tiempo.

Así mismo, Rodríguez-Rosero et al. (2021) analizan los factores que influyen en el rendimiento académico de 14022 estudiantes de secundaria en el departamento de Nariño, Colombia, utilizando datos de las pruebas Saber 11 del año 2018. En este, concluyen que el rendimiento académico en Nariño está determinado por múltiples factores, donde el acceso a tecnología, el nivel educativo de los padres y las condiciones institucionales son fundamentales.

Por otro lado, el estudio de Bonilla-Mejía et al. (2024) muestra que los estudiantes de colegios rurales tienen menor desempeño que los estudiantes de colegios ubicados en zonas

rurales, en este se combinan datos educativos, geográficos y socioeconómicos con métodos espaciales para identificar el efecto causal del aislamiento en el aprendizaje. Aunque los datos son ricos en detalle, su confidencialidad limita la replicabilidad del estudio.

En Cabra-Hernández (2023) se revisa la relación de la posesión de bienes duraderos (lavadora, microondas, televisor, carro, computador, internet, consola de video juegos) en los resultados que obtienen los evaluados, en este sentido los bienes se clasifican en diferentes categorías donde, por ejemplo, tener consola de video juegos genera resultados más bajos versus a mejores resultados asociados a la tenencia de lavadora, computador e internet. El estudio utiliza datos de 364436 estudiantes de quinto y noveno grado en Colombia, obtenidos de la prueba nacional Saber del año 2017.

Por una línea muy similar, Barrios Aguirre et al. (2021) investiga cómo el acceso a computadores e internet en el hogar afecta el rendimiento académico de estudiantes colombianos en las pruebas Saber 11, este estudio tiene un alcance de 1578460 estudiantes entre 2017 y 2019. Como conclusión destaca que, aunque la tecnología puede ser una herramienta educativa valiosa, su impacto depende del contexto de uso y de variables socioeconómicas y culturales.

Los casos mencionados anteriormente abordan distintas preguntas y el uso de unas pocas variables para demostrar las hipótesis de los autores. En el presente ejercicio académico pretendo hacer una revisión un poco más exhaustiva de las 51 variables disponibles para determinar las más relevantes en los resultados de las pruebas de los estudiantes.

A la luz de lo mencionado anteriormente, la pregunta de investigación es ¿Cuáles son las variables más influyentes en el rendimiento de los estudiantes en las pruebas Saber 11 y cómo varía su impacto en el puntaje global?

Justificación

La educación es clave para impulsar el desarrollo social y económico de los países. En Colombia, la prueba Saber 11 se aplica a nivel nacional y es de carácter obligatorio para estudiantes cercanos a finalizar sus estudios de educación secundaria; esta es una evaluación estandarizada que mide a los estudiantes en diferentes competencias: lectura crítica, matemáticas, ciencias naturales, sociales y ciudadanas e inglés. Los resultados de esta prueba son decisivos para los estudiantes puesto que influyen directamente en sus oportunidades de acceder a la educación superior y al mercado laboral (Castro-Ávila & Ruiz-Linares, 2019). Sin embargo, así como lo menciona (Gómez, 2019), identificar las variables que realmente impactan los resultados de estas pruebas sigue siendo un desafío, debido a la complejidad y variedad de factores que pueden influir en el rendimiento académico.

Existen múltiples investigaciones que exploran los factores socioeconómicos, culturales y educativos relacionados con el rendimiento académico. Sin embargo, muchos de estos estudios no profundizan en el análisis específico de las variables dentro del contexto de las pruebas Saber 11, ni aplican enfoques metodológicos modernos como las técnicas de Machine Learning, que permiten identificar patrones complejos en los datos. Por otro lado, hay estudios que tienen un alcance acotado en variables específicas, regiones o ciudades de interés, el estudio de Rodríguez-Rosero et al. (2021) es un ejemplo de estudio focalizado, ya que realiza un análisis de los factores influyentes en el rendimiento de la educación media para el departamento de Nariño. Por la misma línea, está el estudio de Hernandez-Leal et al. (2021) cuyo objetivo es analizar los patrones educativos en instituciones de educación primaria y secundaria públicas en Colombia, pero a partir de datos para cuatro instituciones educativas de Norte de Santander.

Así mismo, otros estudios no exploran de manera específica las diferencias en el impacto

de las variables entre las distintas áreas evaluadas, lo que limita la capacidad para realizar intervenciones focalizadas y eficaces (Solano et al., 2022).

Dado el contexto de los estudios mencionados anteriormente, esta propuesta adopta un enfoque más global al considerar inicialmente las 51 variables disponibles. La única restricción aplicada es en el período de estudio, principalmente debido a limitaciones en la capacidad de cómputo, ya que se utilizarán plataformas gratuitas. Además, se busca mitigar posibles sesgos en los resultados derivados del impacto de la pandemia en el desempeño de los estudiantes.

El presente estudio busca determinar los factores que inciden en el desempeño en las pruebas Saber 11 mediante un enfoque analítico integral que combina técnicas de estadística descriptiva con modelos de aprendizaje automático. El análisis se centra en identificar no solo las variables con mayor influencia en los resultados generales, sino también las variaciones en su impacto según las distintas áreas evaluadas (inglés, matemáticas, lectura crítica, sociales y ciudadanas, ciencias naturales). Adicionalmente, se explorarán interacciones entre factores académicos y contextuales para aportar una comprensión más precisa de los determinantes del rendimiento educativo.

Adicionalmente, este proyecto aporta un análisis integral de las variables influyentes en el rendimiento académico en las pruebas Saber 11, también contribuye al campo de la educación al ofrecer un enfoque más completo y detallado sobre los factores que impactan el rendimiento que tienen los estudiantes en las pruebas Saber 11. Los resultados proporcionarán una base empírica para diseñar estrategias educativas más focalizadas y equitativas. La implementación de métodos avanzados como análisis de correlación y modelos de aprendizaje automático permitirá identificar patrones que podrían no aparecer con técnicas más básicas (Bravo et al., 2021).

Respecto a los beneficiarios de este estudio, sería posible incluir a los encargados de

generar políticas educativas, como autoridades gubernamentales y administradores de programas de educación, quienes eventualmente podrían usar los hallazgos para diseñar e implementar políticas que aborden las necesidades específicas de los estudiantes en distintas regiones y contextos para mejorar el rendimiento académico a nivel nacional. Por otro lado, también me considero una beneficiaria de este proyecto, ya que es un ejercicio académico que me permitirá fortalecer mis conocimientos en las técnicas de Machine Learning aprendidas a lo largo de la especialización y de mi futura línea de profundización, como profesional podré poner en práctica mis habilidades como especialista en Ciencia de Datos y Analítica.

Finalmente, este estudio se justifica porque es esencial para fortalecer la base de conocimientos sobre los elementos que afectan el rendimiento académico en las pruebas Saber 11, y también por el potencial impacto que lograría tener en la mejora de las políticas educativas y las estrategias de enseñanza. Los hallazgos permitirán obtener resultados más precisos y detallados de las situaciones que mejoran o perjudican el desempeño de los estudiantes, permitiendo apuntar a una situación hipotética de contribución al desarrollo de un sistema educativo más eficiente y equitativo en Colombia.

Objetivos

Objetivo General

Analizar las interacciones y efectos de las variables clave en los resultados de las pruebas Saber 11 (2015-2019), utilizando estadísticas descriptivas y técnicas de Machine Learning, para identificar patrones y tendencias en el rendimiento global.

Objetivos Específicos

Preprocesar y depurar los datos de las pruebas Saber 11, garantizando la calidad, consistencia e integridad de la información para su posterior análisis.

Realizar análisis exploratorio de datos utilizando herramientas estadísticas y visualizaciones gráficas que permitan identificar relaciones significativas entre las variables y el desempeño estudiantil.

Aplicar y comparar modelos de aprendizaje automático supervisado para evaluar su capacidad de detección de patrones complejos, utilizando métricas de desempeño para identificar los algoritmos más efectivos.

Marco Teórico

Antecedentes Prueba Saber 11

La prueba Saber 11 es un examen estandarizado que se realiza semestralmente por el ICFES, este es presentado en su mayoría por estudiantes de grado undécimo, los estudiantes son inscritos por medio de un establecimiento educativo (Ministerio de Educación Nacional, 2022); en una muy baja proporción también lo presentan individuos que no se inscriben a través de una institución educativa (ICFES, 2024), estos pueden ser bachilleres ya graduados o personas que hayan superado la prueba de validación del bachillerato. Los resultados de la prueba se utilizan como criterio de ingreso y/o admisión en instituciones de educación superior, adicionalmente permiten "medir la calidad de la educación de los colegios y producir información para la estimación del valor agregado de la educación superior" (ICFES, 2023). La prueba evalúa cinco áreas diferentes:

- Lectura crítica
- Ciencias sociales y ciudadanas
- Ciencias naturales
- Matemáticas
- Inglés

Los resultados del evaluado se dan tanto a nivel global como a nivel de cada área. Los resultados de cada área se miden en una escala de 0 a 100 puntos. El resultado global refleja el resultado obtenido en todas las áreas, esta escala va a 0 a 500 puntos, se hace un promedio ponderado donde el área de inglés tiene una ponderación de 1 punto y las demás áreas tienen un ponderado de 3 puntos (ICFES, 2022a).

Los resultados de las pruebas dan una lectura general del rendimiento académico del evaluado, de ahí que se tengan algunas categorías en los factores que pueden influir en dicho desempeño, entre las cuales se pueden nombrar:

- Demográficas: edad, género, nacionalidad.
- Socioeconómicas: estrato, el nivel educativo de los padres y tenencia de algunos recursos han sido identificados como predictores del rendimiento académico. Un estudio realizado por (Fundación ExE, 2024) muestra que los estudiantes de niveles socioeconómicos mayores obtienen mejores puntajes que los estudiantes en niveles más bajos.
- Institucionales: ubicación, calendario, jornada, carácter, género del colegio, influye en los resultados. Otro análisis de las pruebas Saber 11, muestra que las brechas de rendimiento entre instituciones públicas y privadas se han ampliado en los últimos años (Laboratorio de Economía de la Educación, 2024).

La base de datos abierta disponible en (ICFES, 2022b) consolida los resultados de las pruebas realizadas desde el año 2010 hasta el 2022, tiene un aproximado de siete millones de registros con 51 variables, donde se evidencia información demográfica, académica y socioeconómica. En el presente estudio solo se analizarán los datos de los años 2015 a 2019, que corresponden a un aproximado de 2.2 millones de registros.

Fundamentos Teóricos

El análisis de bases de datos complejas, como la que se usa en este estudio para las pruebas Saber 11, demanda metodologías rigurosas que garanticen validez y confiabilidad en los resultados. Como señalan Oreski et al. (2017), el uso de prácticas estandarizadas es fundamental para:

- Gestionar eficientemente grandes volúmenes de datos

- Minimizar errores en el procesamiento
- Asegurar que el análisis se pueda reproducir
- Facilitar la interpretación de resultados

Metodología CRISP-DM

Para este proyecto se utilizará CRISP-DM (Cross-Industry Standard Process for Data Mining), reconocida como el estándar más robusto en minería de datos (Martinez-Plumed et al., 2021). Esta metodología presenta varias ventajas clave (Wirth & Hipp, 2000):

- Flexibilidad: Permite iteraciones entre fases según necesidades del proyecto
- Completitud: Abarca todo el ciclo de vida del análisis de datos
- Aplicabilidad: Especialmente útil en contextos educativos por su adaptabilidad

Las 6 fases interconectadas de CRISP-DM son (Martinez-Plumed et al., 2021):

1. Comprensión del negocio: Definir los objetivos educativos (ej.: predecir bajo rendimiento en matemáticas).
2. Comprensión de los datos: Análisis exploratorio (EDA), visualización con gráficas como mapas de calor, diagramas de violín, etc.
3. Preparación de los datos (preprocesamiento): limpieza, normalización de las variables disponibles.
4. Modelado: selección y entrenamiento de algoritmos supervisados.
5. Evaluación: validación mediante métricas de desempeño (exactitud, precisión, F1-score).
6. Despliegue: generación de informes con hallazgos aplicables a políticas educativas.

Estadística Descriptiva

La estadística descriptiva es la parte de la estadística que abarca la recolección, análisis, interpretación, presentación y organización de datos numéricos con el fin de resumir sus características principales y descubrir patrones subyacentes (Dangeti, 2017). En el contexto de este proyecto, su aplicación permitirá:

1. Resumir y caracterizar los datos de las pruebas Saber 11 empleando medidas de resumen como la media, mediana y moda para la tendencia central, y la desviación estándar junto al rango intercuartílico para la variabilidad, lo que facilitará una comprensión inicial de la distribución de las variables.
2. Identificar relaciones preliminares entre variables mediante técnicas como:
   - Correlaciones para detectar asociaciones lineales o monotónicas entre variables socioeconómicas (ej.: estrato, acceso a internet) y rendimiento académico.
3. Visualización de datos mediante gráficos como:
   - Histogramas y diagrama de violín para evaluar distribuciones y detectar valores atípicos.
   - Mapas de calor para representar matrices de correlación.

Machine Learning

El Machine Learning (ML) es definido como “la ciencia y el arte de desarrollar algoritmos que permiten a los sistemas computacionales aprender patrones a partir de datos, sin ser programados explícitamente” (Géron, 2022). Los algoritmos de Machine Learning pueden ser supervisados o no supervisados, aunque hay otros tipos, éstos otros no se aplicarán en este proyecto. En los algoritmos supervisados se conocen con antelación las salidas o variables a

predecir, en los algoritmos no supervisados se tratan de formar grupos de datos sin predefinir una variable de salida (Dangeti, 2017).

Para este proyecto, se utilizará Machine Learning para analizar los datos de las pruebas Saber 11 (2015-2019), con el fin de:

1. Generar conocimiento mediante la identificación de patrones ocultos en los resultados académicos.
2. Entrenar modelos predictivos que permitan anticipar tendencias en el rendimiento estudiantil basándose en variables clave (socioeconómicas, institucionales, individuales).

En el proyecto se utilizarán las siguientes técnicas de Machine Learning:

- Regresión lineal: para predecir puntajes continuos en función de variables como estrato o tipo de colegio
- Regresión logística: para clasificación binaria/multiclase, en este caso se pueden definir categorías de acuerdo con el nivel del puntaje
- KNN: para clasificación de estudiantes y sus puntajes

Para medir el desempeño de los modelos se utilizarán las siguientes métricas:

Para casos de regresión:

- Error Cuadrático Medio (MSE)
- R^2
- Error Absoluto Medio (MAE)

Para casos de clasificación:

- Matriz de confusión
- Accuracy, Recall y F1-score

Metodología

A continuación, se listan las etapas que se llevaron a cabo para lograr los objetivos del presente proyecto.

Diseño de la Investigación

El presente estudio se enmarca en un enfoque cuantitativo, ya que se basa en el análisis de datos provenientes de la base de datos de los resultados de las pruebas Saber 11, disponible en el portal de datos abiertos de Colombia (ICFES, 2022b). La investigación es de tipo descriptivo y predictivo, ya que no solo se busca caracterizar y comprender las relaciones entre las variables, sino también identificar patrones y tendencias que permitan predecir el rendimiento académico en función de las variables clave.

Enfoque Metodológico

El estudio se divide en tres fases principales:

1. Fase de preprocesamiento y preparación de datos:

   - Se realizará una inspección detallada de la base de datos para identificar variables relevantes, valores faltantes, outliers y posibles errores en los datos.
   - Se aplicarán técnicas de limpieza y transformación de datos, como la imputación de valores faltantes, normalización de variables numéricas y codificación de variables categóricas.
   - Se seleccionarán las variables clave que afectan los resultados de las pruebas Saber 11, como el puntaje global, el tipo de institución educativa, el género, el estrato socioeconómico, entre otras.

2. Fase de análisis exploratorio de datos (EDA):

- Se utilizarán técnicas de estadística descriptiva para resumir y visualizar la distribución de los datos, así como las relaciones entre las variables.
- Se generarán gráficos como histogramas, diagramas de violín, matrices de correlación, entre otros para identificar tendencias, patrones y posibles correlaciones.
- Esta fase permitirá comprender mejor la estructura de los datos y formular hipótesis iniciales sobre los factores que influyen en el rendimiento académico.

3. Fase de modelado y análisis predictivo:

- Se aplicarán técnicas de Machine Learning para identificar patrones complejos en los datos y predecir el rendimiento académico en función de las variables seleccionadas.
- Se ejecutarán algoritmos de aprendizaje automático, como regresiones lineal y logística, clasificación, y otros.
- Se utilizarán métricas de evaluación según corresponda de acuerdo con el tipo de algoritmo, como el error cuadrático medio, precisión, recall y F1-score para comparar el desempeño de los modelos.
- El objetivo es identificar el modelo más efectivo para predecir el rendimiento y explicar la influencia de las variables clave.

Tipo de Estudio

El estudio es de tipo transversal, ya que se analizan datos correspondientes a un período específico (2015-2019). Además, es no experimental, dado que no se manipulan variables, sino que se analizan los datos existentes para identificar relaciones y patrones.

Fuente de Datos

La fuente de datos para este estudio es la base de datos “Resultados Únicos Saber 11”, disponible en el portal de datos abiertos de Colombia (ICFES, 2022b). Esta base de datos es

proporcionada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), entidad responsable de aplicar y evaluar las pruebas Saber 11 en Colombia.

Descripción de la Base de Datos

La base de datos original contiene información detallada sobre los resultados de las pruebas Saber 11 aplicadas a estudiantes de educación media en Colombia durante el período 2010-2022. Vale la pena aclarar que para este estudio, se seleccionaron específicamente los registros correspondientes al período comprendido entre 2015 y 2019. El conjunto de datos incluye variables relacionadas con el desempeño académico de los estudiantes, así como características sociodemográficas y contextuales. Algunas de las variables clave disponibles son:

- Puntajes por áreas: inglés, matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas.
- Puntaje global: resultado general de la prueba.
- Características del estudiante: género, residencia (departamento y municipio), fecha de nacimiento, etc.
- Información del colegio: nombre, ubicación (departamento y municipio), carácter, naturaleza, jornada, entre otras.
- Información familiar: tenencia de diferentes activos, nivel educativo de los padres, estrato socioeconómico, etc.

Estructura de la Base de Datos

La base de datos está organizada en formato tabular (filas y columnas), donde cada fila representa un estudiante y cada columna corresponde a una variable específica. El conjunto de datos del estudio incluye aproximadamente 2.9 millones y 41 variables, que la hacen una fuente robusta para el análisis.

Población y Muestra

La entrada inicial de información para este estudio tiene los resultados de las pruebas Saber 11 en Colombia durante el período 2010-2022. La muestra escogida incluye los datos que comprenden el período 2015-2019.

Esta delimitación temporal se estableció debido a que los años posteriores presentes en la base de datos (2020-2022) estuvieron marcados por las disrupciones educativas causadas por la época del COVID-19, lo que generó condiciones atípicas en la aplicación de las pruebas. Como señala (UNESCO, 2023) la pandemia modificó significativamente los sistemas educativos, transformando las modalidades de enseñanza y evaluación, así como introduciendo nuevos factores podrían sesgar el desempeño académico. En esa misma línea, (d'Orville, 2020) indica que la educación ya venía en crisis y que la pandemia fue un detonante, ya que alrededor de un 87% de la población estudiantil a nivel mundial se vio afectada por el cierre de los planteles educativos. Esta situación tan caótica y atípica supone afectaciones al proceso de aprendizaje y por supuesto genera consecuencias en los resultados de los evaluados.

La selección del quinquenio 2015-2019 permite trabajar con datos obtenidos bajo condiciones estables del sistema educativo, garantizando mayor consistencia en el análisis.

La muestra final incluye una selección representativa que considera diversidad geográfica (regiones), tipos de instituciones educativas (públicas/privadas), estratos socioeconómicos, entre muchos otros factores. Este enfoque busca capturar las dinámicas habituales del sistema educativo colombiano previo a la pandemia, ofreciendo así una base sólida para el análisis.

Acceso y Disponibilidad

La base de datos está disponible de manera gratuita en el portal de datos abiertos de Colombia, en formato CSV, lo que facilita su manipulación y análisis mediante herramientas como Python.

Calidad y Confiabilidad de los Datos

Los datos son recopilados y validados por el ICFES, lo que garantiza un alto nivel de confiabilidad y precisión. Sin embargo, como parte del proceso de investigación, se realizará una evaluación inicial de la calidad de los datos para identificar posibles problemas, como:

- Valores faltantes: datos incompletos en algunas variables.
- Outliers: valores atípicos que podrían distorsionar el análisis.
- Consistencia: verificación de que los datos cumplan con las expectativas lógicas (por ejemplo, que los puntajes estén dentro de los rangos permitido).

Procesamiento Inicial

Para garantizar la calidad de los datos, se realizará un procesamiento inicial que incluye:

- Limpieza de datos: depuración de registros duplicados, corrección de inconsistencias y tratamiento de valores faltantes.
- Selección de variables: identificación de las variables más relevantes para el análisis, descartando aquellas que no aporten información significativa.
- Transformación de datos: normalización de variables numéricas y codificación de variables categóricas para su uso en modelos de Machine Learning.

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) tiene como objetivo comprender la estructura de los datos, identificar patrones y formular hipótesis iniciales.

Estadísticas Descriptivas

- Cálculo de medidas de tendencia central (media, mediana, moda) y dispersión (desviación estándar, rango).
- Análisis de la distribución de los datos.

Visualización de Datos

- Generación de gráficos como histogramas, diagramas de violín, matrices de correlación y gráficos de dispersión.
- Identificación de relaciones entre variables y posibles correlaciones.

Técnicas de Machine Learning

En esta fase se aplican técnicas de Machine Learning para identificar patrones complejos y predecir el rendimiento académico.

Selección de Algoritmos

- Uso de algoritmos como regresión lineal, árboles de decisión y clasificación.
- Justificación de la selección de cada algoritmo en función de los objetivos del estudio.

Entrenamiento y Validación de Modelos

- División de los datos en conjuntos de entrenamiento y prueba.
- Aplicación de técnicas de validación cruzada (cross-validation) para evaluar el desempeño de los modelos.

Métricas de Evaluación

- Uso de métricas como el error cuadrático medio (MSE), precisión, recall y F1-score para comparar el desempeño de los modelos.

Herramientas y Software

Para el análisis de datos y la implementación de modelos de Machine Learning, se utilizarán las siguientes herramientas:

Lenguajes de Programación

- Python: con librerías como Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Geopandas, Polygon, shapely.

Entornos de Desarrollo

- Google Colab: para la documentación y ejecución del código.

Resultados

Selección de Datos y Variables

En esta fase se realizó la exploración de diferentes fuentes de información hasta llegar a un conjunto de datos que permitiera aplicar las herramientas y técnicas aplicadas en la especialización. Se optó por seleccionar una base de datos abierta de manera que se facilitara la autonomía del ejercicio y por las dificultades que representa trabajar con bases de datos de empresas.

Se seleccionó la base de datos del "Resultados Únicos Saber 11" publicada en (ICFES, 2022b). Esta base de datos es un archivo CSV que contiene el consolidado de resultados a nivel país desde el año 2010 al año 2022. Por la cantidad de datos, y por las razones ya enunciadas en otras secciones de este documento, se analizarán los datos comprendidos entre el año 2015 a 2019.

De este punto en adelante, todas las tablas, gráficas y resultados presentados se pueden evidenciar en el código que está compartido en el ***Apéndice A***.

Tal como se resume en la **Tabla 1**, la base de datos original cuenta con un total de 7109704 registros y 51 variables.

Tabla 1

Información Base de Datos de Estudio

Detalle base de datos "Resultados Únicos Saber 11"	
Cantidad de registros	7109704
Cantidad de variables (columnas)	51

Nota. Cantidad de información de la base de datos seleccionada.

Al considerar solamente los datos para los años 2015 a 2019 se tiene un total de 2267495 registros, a continuación, se muestra la **Tabla 2** con el desglose por período académico:

Tabla 2

Cantidad de Registros por Período

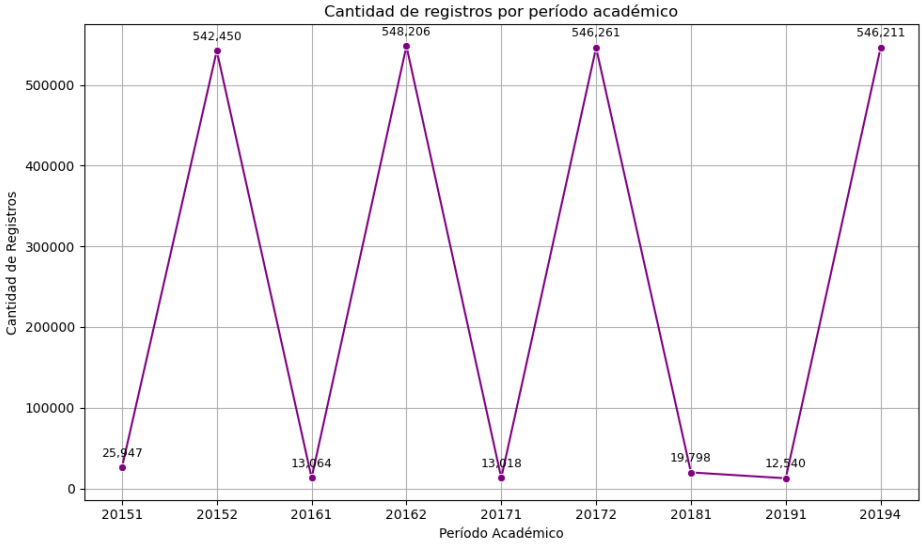
Período académico	Cantidad de registros
20151	25947
20152	542450
20161	13064
20162	548206
20171	13018
20172	546261
20181	19798
20191	12540
20194	546211

Nota. Cantidad de registros disponibles para los períodos académicos que serán parte del estudio.

Adicionalmente, la **Figura 1** muestra una gráfica de línea que permite observar los patrones de cantidad de evaluaciones por período. De esta vale la pena mencionar que los períodos 1 y 2 indican “primer” y “segundo” semestre del año respectivamente. En el primer semestre las pruebas son generalmente presentadas por estudiantes del Calendario B y en el segundo semestre se presentan las pruebas para el calendario A. Por razones desconocidas la base de datos del ICFES no presenta información para el período 20182 y en el caso del año 2019 el período se nombró 20194 y no 20192 como es esperado.

Figura 1

Cantidad de Evaluaciones por Período Académico



En cuanto a las variables, la base de datos inicial contaba con 51 en total de las cuales se descartaron algunas que no aportan información relevante para el estudio como el nombre del colegio, el consecutivo asignado a cada estudiante, el tipo de documento del estudiante, entre otras. A continuación, se listan las 41 variables que hacen parte del conjunto final para el análisis. Para facilidad en la presentación, las variables se dividieron en tres grupos.

En la **Tabla 3** se reúnen las características del plantel educativo al que pertenece el estudiante y sobre el cual se evalúan los resultados de la prueba, se indica el tipo de dato para cada característica, la descripción de la información que almacena, así como el número de registros que cuentan con información para la variable. Las variables del colegio ayudan a hacerse una idea del entorno institucional donde se llevan a cabo los procesos de formación, éstas pueden ayudar a explicar diferencias en los resultados académicos, especialmente en aspectos relacionados con la localización geográfica, su naturaleza (oficial o no oficial), su

jornada, el calendario escolar, entre otros que influyen en las condiciones educativas de los estudiantes.

Tabla 3

Características del Colegio

#	Nombre variable	Tipo	Descripción	Cantidad registros
1	COLE_AREA_UBICACION	Categórica nominal	Área de ubicación (Urbano/Rural)	2267495
2	COLE_BILINGUE	Categórica nominal	Si/No el colegio es bilingüe	2267495
3	COLE_CALENDARIO	Categórica nominal	Calendario del colegio (A, B, Otro)	1953219
4	COLE_CARACTER	Categórica nominal	Carácter del colegio (Académico, Técnico, etc.)	2267495
5	COLE_COD_DANE_ESTABLECIMIENTO	Categórica nominal	Código DANE asignado al colegio	2233668
6	COLE_COD_DANE_SEDE	Categórica nominal	Código DANE asignado a la sede del colegio	2267428
7	COLE_COD_DEPTO_UBICACION	Categórica nominal	Código del departamento donde está el colegio	2267495
8	COLE_COD_MCPIO_UBICACION	Categórica nominal	Código del municipio donde está el colegio	2267495
9	COLE_CODIGO_ICFES	Categórica nominal	Código ICFES del colegio	2267495
10	COLE_DEPTO_UBICACION	Categórica nominal	Nombre del departamento donde está el colegio	2267495
11	COLE_GENERO	Categórica nominal	Género del colegio (Femenino, Masculino, Mixto)	2267495
12	COLE_JORNADA	Categórica nominal	Jornada del colegio (Mañana, Tarde, Completa, etc.)	2267495
13	COLE_MCPIO_UBICACION	Categórica nominal	Nombre del municipio donde está el colegio	2267495
14	COLE_NATURALEZA	Categórica nominal	Naturaleza del colegio (Oficial, No oficial)	2267495
15	COLE_SEDE_PRINCIPAL	Categórica nominal	Si/No es la sede principal del colegio	2267495

Nota. Variables que muestran las características del colegio.

La **Tabla 4** presenta las características asociadas a los estudiantes que presentaron la prueba, igual que en la tabla anterior, se indica el tipo de dato para cada característica, la

descripción de la información que almacena, así como el número de registros que cuentan con información para la variable. Estas características incluyen aspectos como su lugar de residencia, nacionalidad, género y otras. Este tipo de información puede ayudar a identificar diferencias relacionadas con el lugar donde viven los estudiantes o su situación personal, que podrían afectar su desempeño académico.

Tabla 4

Características del Estudiante

#	Nombre variable	Tipo	Descripción	Cantidad registros
16	ESTU_COD_RESIDE_DEPTO	Categórica nominal	Código del departamento donde vive el estudiante	2266753
17	ESTU_COD_RESIDE_MCPIO	Categórica nominal	Código del municipio donde vive el estudiante	2266753
18	ESTU_DEPTO_RESIDE	Categórica nominal	Nombre del departamento donde vive el estudiante	2266753
19	ESTU_FECHANACIMIENTO	Categórica nominal	Fecha de nacimiento del estudiante	2267492
20	ESTU_GENERO	Categórica nominal	Género del estudiante	2264926
21	ESTU_MCPIO_RESIDE	Categórica nominal	Nombre del municipio donde vive el estudiante	2266753
22	ESTU_NACIONALIDAD	Categórica nominal	Nacionalidad del estudiante	2267495
23	ESTU_PAIS_RESIDE	Categórica nominal	País donde vive el estudiante	2267495
24	ESTU_PRIVADO_LIBERTAD	Categórica nominal	Si/No el estudiante está privado de la libertad	2267495

Nota. Variables que muestran las características del estudiante.

En la

Tabla 5 se muestran las características del entorno familiar de los estudiantes. Al igual que en los casos anteriores, se incluye el tipo de dato, una breve descripción del contenido de cada variable, y la cantidad de registros con información disponible. Estos datos permiten conocer aspectos generales respecto a la familia, como el nivel educativo de los padres, las condiciones materiales de la vivienda y el acceso a bienes y servicios. Esta información es útil

para analizar cómo las condiciones familiares pueden relacionarse con el rendimiento académico de los estudiantes.

Tabla 5

Características Familiares del Estudiante

#	Nombre variable	Tipo	Descripción	Cantidad registros
25	FAMI_CUARTOSHOGAR	Categórica ordinal	Cantidad de cuartos del hogar	2236779
26	FAMI_EDUCACIONMADRE	Categórica ordinal	Nivel educativo de la madre	2206893
27	FAMI_EDUCACIONPADRE	Categórica ordinal	Nivel educativo del padre	2206735
28	FAMI_ESTRATOVIVIENDA	Categórica ordinal	Estrato de la vivienda que habita el estudiante	2200939
29	FAMI_PERSONASHOGAR	Categórica ordinal	Cantidad de personas que habitan el hogar	2235089
30	FAMI_TIENEAUTOMOVIL	Categórica nominal	Si/No la familia tiene carro	2234094
31	FAMI_TIENECOMPUTADOR	Categórica nominal	Si/No la familia tiene computador	2236601
32	FAMI_TIENEINTERNET	Categórica nominal	Si/No la familia tiene internet	2205358
33	FAMI_TIENELAVADORA	Categórica nominal	Si/No la familia tiene lavadora	2237426

Nota. Variables que muestran las características de la familia del estudiante.

Finalmente, para cerrar las variables del estudio, la

Tabla 6 presenta las variables relacionadas a los puntajes de las Pruebas Saber 11. Conversando el mismo formato anterior, se indica el tipo de dato, una descripción de cada variable y la cantidad de registros disponibles. Estas variables tienen la información de los puntajes obtenidos por los estudiantes en las diferentes áreas evaluadas (matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas e inglés), así como el puntaje global. Estos datos pueden ayudar a observar posibles patrones de rendimiento y establecer comparaciones.

Tabla 6

Características para Resultados de las Pruebas

#	Nombre variable	Tipo	Descripción	Cantidad registros
34	PERIODO	Categórica nominal	Período académico en el que se presentó la prueba	2267495
35	DESEMP_INGLES	Categórica ordinal	Nivel de desempeño en inglés	2267495
36	PUNT_INGLES	Numérica	Puntaje de la prueba de inglés	2267476
37	PUNT_MATEMATICAS	Numérica	Puntaje de la prueba de matemáticas	2267495
38	PUNT_SOCIALES_CIUDADANAS	Numérica	Puntaje de la prueba de sociales y ciudadanas	2267495
39	PUNT_C_NATURALES	Numérica	Puntaje de la prueba de ciencias naturales	2267495
40	PUNT_LECTURA_CRITICA	Numérica	Puntaje de lectura crítica	2267495
41	PUNT_GLOBAL	Numérica	Puntaje global	2267495

Nota. Variables que muestran los resultados de las pruebas.

A nivel general, las tablas permiten notar que hay datos faltantes en varias variables. Esto puede ser normal en las variables socioeconómicas ya que éstas hacen parte de una encuesta que se incluye al final de la prueba. Sin embargo, en los puntajes se observa que, aunque el conjunto de datos seleccionado cuenta con un total de 2267495 registros, hay un conjunto de datos faltantes para el puntaje de inglés que solo cuenta con 2267476 registros.

Además, aunque algunas variables son claramente categóricas, fue necesario ajustar su tipo en Python, ya que originalmente estaban representadas por números utilizados para códigos o identificadores. Este fue el caso de las variables PERIODO, COLE_COD_DANE_ESTABLECIMIENTO, COLE_COD_DANE_SEDE, COLE_COD_DEPTO_UBICACION, COLE_COD_MCPIO_UBICACION, COLE_CODIGO_ICFES, ESTU_COD_RESIDE_DEPTO, ESTU_COD_RESIDE_MCPIO. De igual manera, las variables PUNT_INGLES y PUNT_MATEMATICAS no se estaban

reconociendo apropiadamente como números, por lo tanto, también se debió ajustar el tipo de dato; con esto las únicas variables numéricas del estudio son las relacionadas a los puntajes.

Análisis Descriptivo

Una vez seleccionadas las variables y con los ajustes respectivos de los tipos de datos, se procedió a la generación de las estadísticas descriptivas. En la **Tabla 7** presenta los resultados del análisis de estadística descriptiva de las diferentes áreas del conocimiento y el puntaje global:

Tabla 7

Estadísticas Descriptivas para las Variables Numéricas

Nombre variable	Cantidad	Media	Desviación Estándar	Mínimo	25%	50%	75%	Máximo
PUNT_INGLES	2267476	50.62	12.35	0	42	49	57	100
PUNT_MATEMATICAS	2267495	50.76	12.21	0	42	50	59	100
PUNT_SOCIALES_CIUDADANAS	2267495	49.55	11.67	0	41	49	58	100
PUNT_C_NATURALES	2267495	50.85	10.53	0	43	51	58	100
PUNT_LECTURA_CRITICA	2267495	52.18	10.08	0	45	52	59	100
PUNT_GLOBAL	2267495	254.09	50.21	0	217	251	288	494

Nota. Resumen de estadísticas descriptivas para las variables numéricas del ejercicio.

De las estadísticas anteriores se puede notar que del total de registros de la base de datos que son 2267495, todos los casos cuentan con información para los puntajes, excepto la prueba de inglés donde se evidencia la ausencia de información para 19 registros.

También es de resaltar que para lectura crítica el puntaje promedio es 52.18, el más alto de todas las áreas, lo que sugiere que los estudiantes tienen un mejor desempeño relativo en este caso. También presenta la menor desviación estándar, 10.08, indicando que los estudiantes tienden a obtener puntajes más uniformes. La mayoría de los estudiantes se encuentran entre los

puntajes de 45 (25%), 52 (50%) y 59 (75%), que muestra que los estudiantes tienen un desempeño más estable en esta área.

En cuanto al puntaje global la media es 254.09, que es un puntaje agregado de las diferentes áreas. La desviación estándar de 50.21 refleja una gran dispersión, ya que este puntaje es la combinación de varias áreas. Los estudiantes que tienen un puntaje muy bajo en algunas áreas y muy alto en otras contribuyen a esta variabilidad. Aunque gran parte de los estudiantes se encuentran por debajo de los 300 puntos, algunos logran puntajes cercanos al máximo.

De igual manera se generaron estadísticas para las variables categóricas con el fin de entender un poco el contexto y comenzar un entendimiento de alto nivel de los datos. A continuación, el detalle para cada grupo de variables.

La

Tabla *8* presenta un resumen estadístico de las variables relacionadas a las características de los colegios. Se incluyen el número total de registros disponibles por variable, la cantidad de valores únicos identificados, el valor más frecuente (moda: “Top”), su frecuencia absoluta y la frecuencia relativa correspondiente. Este análisis permite comprender la distribución y representatividad de las categorías más comunes en cada variable a nivel de colegio:

Tabla 8

Estadísticas para Variables del Colegio

Nombre variable	Cantidad	Valores únicos	Top	Frecuencia	Frecuencia relativa
COLE_AREA_UBICACION	2267495	2	URBANO	1933901	0.853
COLE_BILINGUE	1953219	2	N	1912245	0.979
COLE_CALENDARIO	2267495	3	A	2183882	0.963
COLE_CARACTER	2233668	4	ACADÉMICO	1228358	0.550
COLE_COD_DANE_ESTABLECIMIENTO	2267428	11104	105001000108	3783	0.002
COLE_COD_DANE_SEDE	2267495	12202	105001000108	3783	0.002
COLE_COD_DEPTO_UBICACION	2267495	33	11	370417	0.163
COLE_COD_MCPIO_UBICACION	2267495	1114	11001	370417	0.163
COLE_CODIGO_ICFES	2267495	17010	786	2902	0.001
COLE_DEPTO_UBICACION	2267495	34	ANTIOQUIA	298830	0.132
COLE_GENERO	2267495	3	MIXTO	2182459	0.962
COLE_JORNADA	2267495	6	MAÑANA	1127191	0.497
COLE_MCPIO_UBICACION	2267495	1409	BOGOTÁ D.C.	370417	0.163
COLE_NATURALEZA	2267495	2	OFICIAL	1645348	0.726
COLE_SEDE_PRINCIPAL	2267495	2	S	2200502	0.970

Nota. Resumen de estadísticas para variables asociadas al colegio.

A partir de los resultados presentados en la tabla anterior se observa que, a nivel de departamento, la mayoría de los colegios registran en Antioquía, con una predominancia del 13.2%. Pero a nivel de municipio, la mayoría de los colegios se ubica en Bogotá que registra un 16.3%, siendo la capital la que concentra la mayor cantidad de pruebas presentadas Este tipo de concentración puede representar la densidad poblacional de estas zonas e incluso mayor cobertura del sistema educativo.

Adicionalmente, la mayoría de los colegios son de calendario A (96.3%), mixtos (96.2%) y están ubicados en cascos urbanos (85.3%). En cuanto a la enseñanza de inglés, es contundente

que el 97.9% de los colegios no son bilingües lo cual puede ser un factor crucial para los resultados en los puntajes de dicha área.

A continuación, la **Tabla 9** presenta un resumen estadístico de las variables relacionadas con los estudiantes. Se incluyen aspectos demográficos como el lugar de residencia (departamento y municipio), la fecha de nacimiento, el género, entre otros. Para cada variable se reporta el número total de observaciones, la cantidad de valores únicos, el valor más frecuente (moda "Top"), su frecuencia absoluta y su frecuencia relativa. Este análisis permite identificar concentraciones poblacionales, posibles patrones de distribución geográfica y aspectos relevantes en la caracterización general de los evaluados:

Tabla 9

Estadísticas para Variables del Estudiante

Nombre variable	Cantidad	Valores únicos	Top	Frecuencia	Frecuencia relativa
ESTU_COD_RESIDE_DEPTO	2266753	34	11	375572	0.166
ESTU_COD_RESIDE_MCPIO	2266753	1120	11001	375572	0.166
ESTU_DEPTO_RESIDE	2266753	34	BOGOTÁ	375572	0.166
ESTU_FECHANACIMIENTO	2267492	18705	01/01/2000	1595	0.001
ESTU_GENERO	2264926	2	F	1233810	0.545
ESTU_MCPIO_RESIDE	2266753	1036	BOGOTÁ D.C.	375572	0.166
ESTU_NACIONALIDAD	2267495	84	COLOMBIA	2264768	0.999
ESTU_PAIS_RESIDE	2267495	84	COLOMBIA	2264768	0.999
ESTU_PRIVADO_LIBERTAD	2267495	2	N	2267177	1

Nota. Resumen de estadísticas para variables asociadas al estudiante.

En cuanto a las variables asociadas al estudiante, se observa una tendencia similar a la registrada en la ubicación de los colegios, donde Bogotá se destaca como la ciudad con mayor número de estudiantes, concentrando el 16.6% del total. Además, se evidencia una mayor

participación de mujeres en la prueba, representando el 54.5% de la población evaluada, esto eventualmente podría revelar algún patrón con respecto a los resultados obtenidos.

La **Tabla 10** presenta un resumen de las principales variables relacionadas con el entorno familiar de los estudiantes. Se incluyen indicadores como el número de cuartos en el hogar, el nivel educativo de los padres, entre otros, los cuales son relevantes para el análisis del contexto socioeconómico. Al igual que en las tablas anteriores, se reportan la cantidad total de registros, los valores únicos identificados, el valor más frecuente (moda "Top"), su frecuencia absoluta y su frecuencia relativa:

Tabla 10

Estadísticas para Variables de la Familia

Nombre variable	Cantidad	Valores únicos	Top	Frecuencia	Frecuencia relativa
FAMI_CUARTOSHOGAR	2236779	11	Tres	907545	0.406
FAMI_EDUCACIONMADRE	2206893	12	Secundaria (Bachillerato) completa	589324	0.267
FAMI_EDUCACIONPADRE	2206735	12	Secundaria (Bachillerato) completa	513182	0.233
FAMI_ESTRATOVIVIENDA	2200939	7	Estrato 1	830357	0.377
FAMI_PERSONASHOGAR	2235089	17	3 a 4	537509	0.240
FAMI_TIENEAUTOMOVIL	2234094	2	No	1716602	0.768
FAMI_TIENECOMPUTADOR	2236601	2	Si	1333056	0.596
FAMI_TIENEINTERNET	2205358	2	Si	1249814	0.567
FAMI_TIENELAVADORA	2237426	2	Si	1608343	0.719

Nota. Resumen de estadísticas para variables asociadas al entorno familiar del estudiante.

De la información anterior, respecto a la posesión de activos, se puede notar que predomina la presencia de lavadora, el 71.9% de los hogares tiene lavadora, lo que puede dar señal de que este bien se puede considerar esencial. A pesar de que la mayoría de los hogares pertenecen al estrato 1 (37.7%), la lavadora parece haber sido priorizada también en estos

hogares. El siguiente bien con mayor presencia en los hogares es el computador, presente en el 59.6% de los casos. Le sigue el acceso a internet, reportado para el 56.7% de los hogares, con apenas un punto porcentual de diferencia. Esta cercanía sugiere una posible asociación entre la tenencia de computador y la conectividad en los hogares. Estos dos recursos podrían facilitar un mayor acceso a materiales formativos y, tal vez, influir de positivamente en los resultados de las pruebas.

Finalmente, vale la pena mirar un poco más detenidamente el nivel educativo de los padres. Para ambos predomina la formación hasta el bachillerato, el 26.7% de las madres y el 23.3% de los padres alcanzaron este nivel educativo, lo que podría indicar que los padres pudieron enfrentar retos para acceder a la educación superior.

Visualización de Datos

Análisis de Correlación

Como se comentó en secciones anteriores, el conjunto de datos contaba originalmente con 51 variables, algunas de ellas no aportaban información útil para el análisis de este estudio, como el consecutivo del estudiante, su tipo de documento, entre otras que eran identificadores o datos redundantes. Por ello, se hizo una depuración para conservar únicamente las características que tuvieran valor analítico, así se redujo el total a 41. Si bien la reducción facilitó el manejo del conjunto de datos, tener 41 variables representaba un desafío importante para identificar posibles relaciones entre ellas. Por esta razón, se decidió realizar un análisis de correlación, con el objetivo de obtener medidas que ayudaran a depurar aquellas variables con bajas correlaciones y, de este modo, simplificar el análisis posterior.

El coeficiente que se utilizó en el análisis de correlación fue el coeficiente de correlación de Pearson. Como lo menciona (Schober & Schwarte, 2018) este coeficiente permite identificar

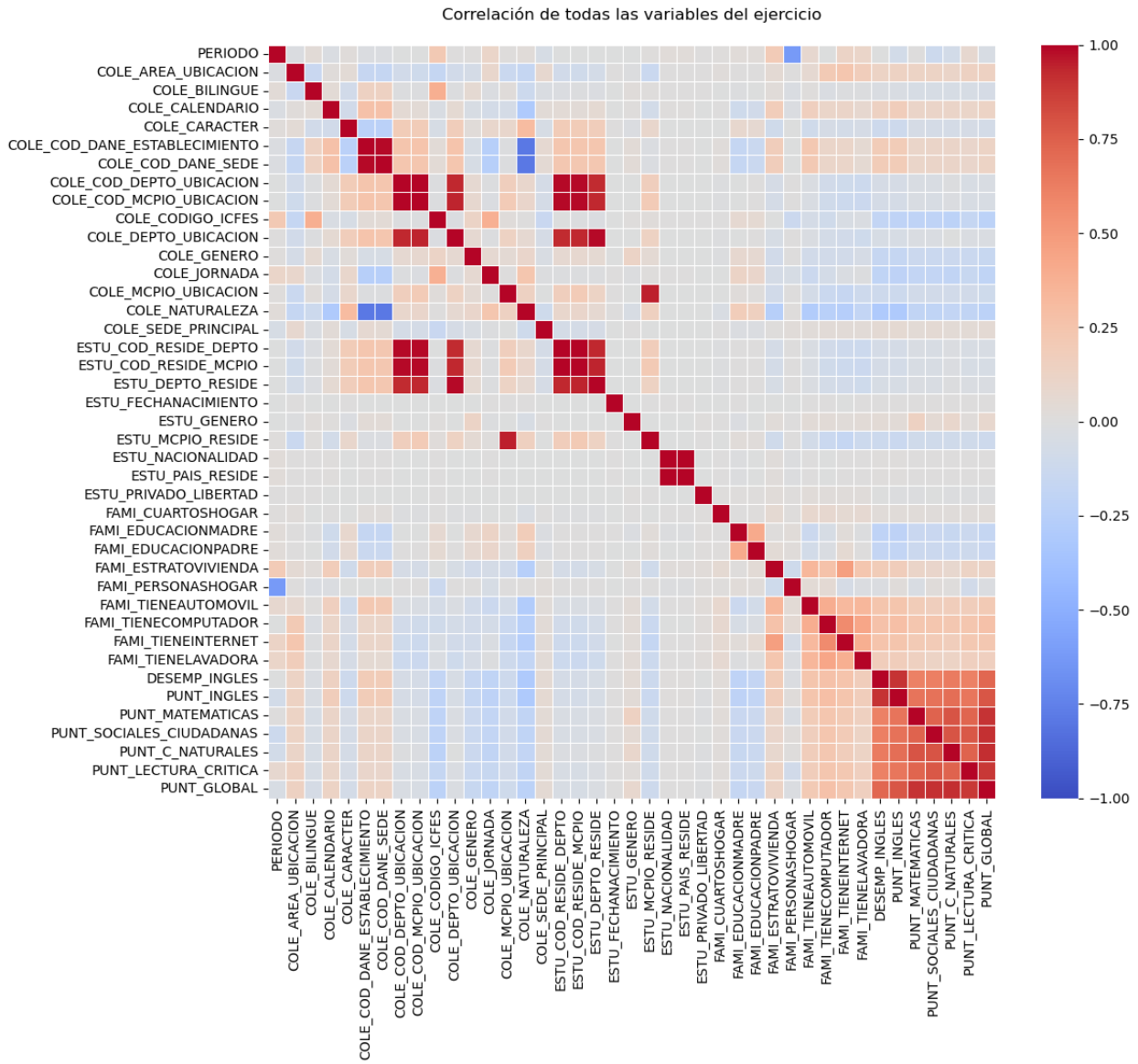
vínculos entre las variables, dando una idea de la fuerza y el sentido de su relación, proporcionando valores que van desde -1 (correlación negativa perfecta) hasta +1 (correlación positiva perfecta), siendo 0 el indicador de ausencia de correlación.

Dado que, a excepción de los puntajes, todas las variables son categóricas, se hizo la respectiva codificación para posibilitar el cálculo de correlaciones y su visualización a través de mapas de calor.

En la **Figura 2** se presenta el mapa de calor de la matriz de correlación entre todas las variables incluidas en el ejercicio, permitiendo identificar relaciones positivas o negativas entre ellas.

Figura 2

Matriz de Correlación de las 41 Variables Seleccionadas



En este caso la matriz generada es bastante grande ya que hay muchas variables en juego, a simple vista no es tan sencillo de interpretar Esta salida permite hacerse una idea de muy alto nivel sobre las variables que tienen mayor impacto en los puntajes que son el foco de este estudio.

Como se puede observar, la matriz de correlación es notablemente extensa debido al elevado número de variables incluidas en el análisis, lo que genera una mayor complejidad en la interpretación de las relaciones entre ellas. Aunque a simple vista puede parecer compleja de interpretar, esta representación permite obtener una visión general sobre la intensidad y dirección de las relaciones entre las distintas variables.

En la parte inferior derecha de la figura se observa un grupo de variables correspondientes a los puntajes por área (matemáticas, sociales y ciudadanas, ciencias naturales, lectura crítica e inglés), así como el puntaje global. Estas variables muestran una alta correlación positiva entre sí, lo cual se refleja en los tonos intensos. Este resultado es coherente con la metodología de cálculo del puntaje global, el cual se deriva directamente de los puntajes individuales por área.

Adicionalmente, se destacan correlaciones altas entre variables que hacen referencia a la ubicación geográfica de los estudiantes y de los colegios, como el departamento y el municipio de residencia o ubicación, lo cual es esperable dada la relación jerárquica entre estas divisiones territoriales. Como es habitual, la diagonal principal de la matriz presenta una correlación perfecta (valor de 1) entre cada variable y sí misma.

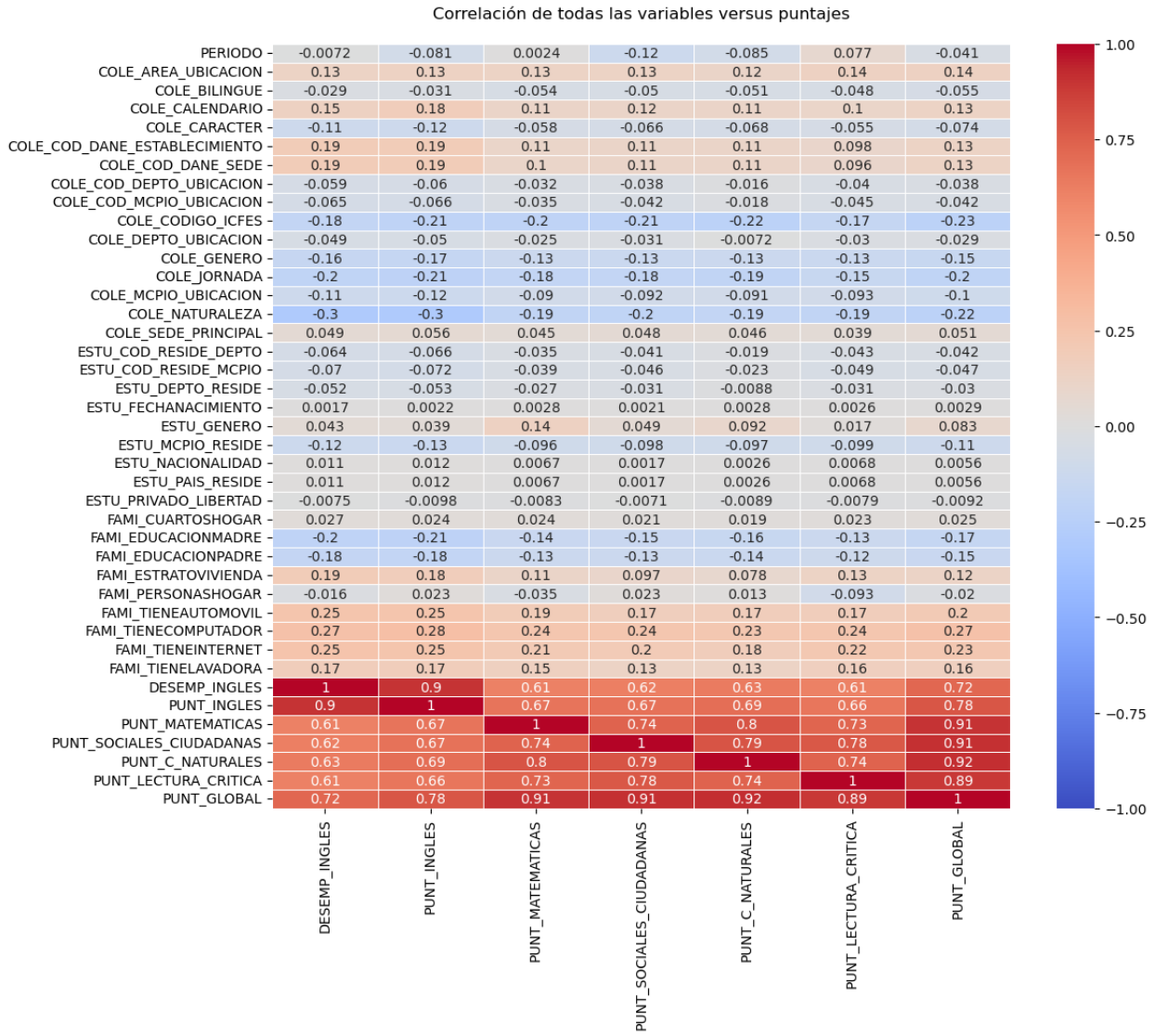
En conjunto, este análisis preliminar de correlaciones contribuyó a identificar asociaciones potencialmente relevantes para fases posteriores del estudio.

Con el fin de simplificar el análisis y facilitar su interpretación, se determinó revisar la matriz de correlación de manera reducida, enfocando la atención en la relación de todas las variables únicamente con los puntajes.

A continuación, en la **Figura 3** se muestra la matriz de correlación reducida. Como se mencionó previamente, no se incluyen las relaciones de las variables categóricas entre sí, solo se muestra la relación de las variables categóricas con los puntajes:

Figura 3

Matriz de Correlación de las 41 Variables Versus los Puntajes



Al ser una gráfica más pequeña, fue posible incluir el índice de correlación para ayudar un poco en caso de que la escala de colores no sea lo suficientemente clara. En este caso, es aún

más evidente observar que las correlaciones más fuertes se presentan entre los puntajes por áreas y el puntaje global. Como se comentaba anteriormente, este resultado era esperado, ya que el puntaje global se calcula a partir de los puntajes individuales por área, lo que evidencia esa relación preexistente. Puesto que esta relación es extremadamente fuerte y es clara, estas variables no se consideraron en los análisis posteriores. En su lugar, se enfocó el estudio en las variables categóricas, que aportan información sobre factores contextuales y externos al rendimiento académico.

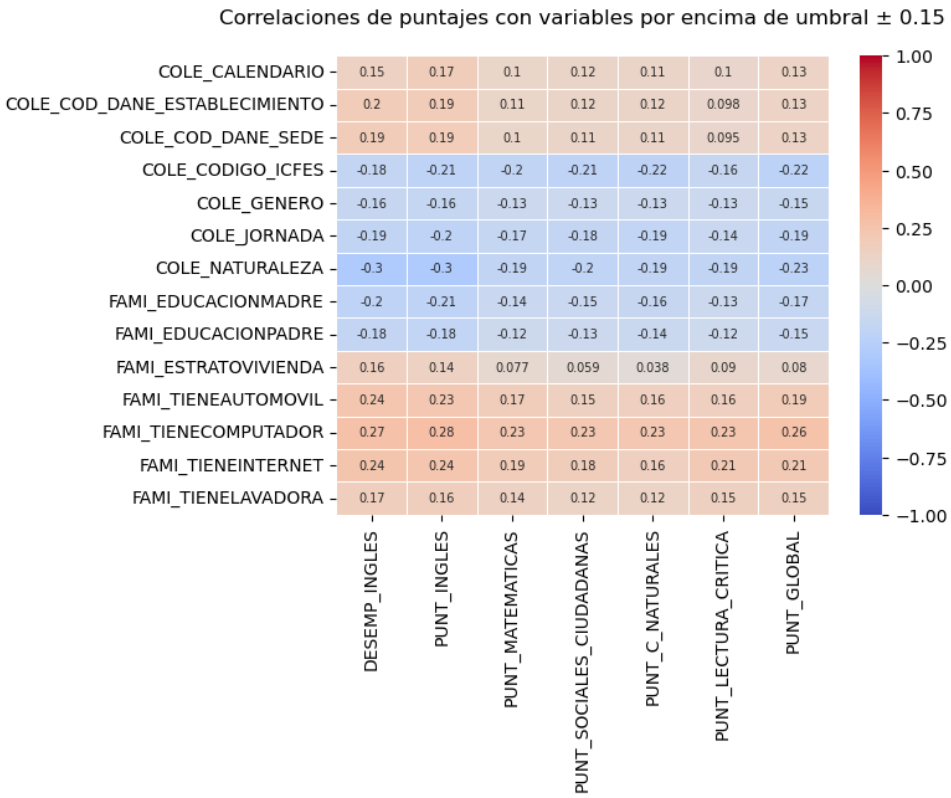
Con base en lo anterior, se procedió a establecer un criterio que facilitara la selección de las variables con las que se realizaría el análisis. Para ello, se tuvieron en cuenta los niveles definidos por la correlación de Pearson, donde un coeficiente entre 0 y 0.1 representa una correlación de muy baja intensidad, entre 0.1 y 0.3 es una correlación baja, entre 0.3 y 0.5 se considera correlación moderada y por encima de 0.7 se habla de altas correlaciones (Schober & Schwarte, 2018).

Con este criterio, inicialmente se pretendía tomar aquellas variables categóricas cuya correlación con alguno de los puntajes fuera como mínimo moderada (mayor o igual a 0.30), con el propósito de reducir la dimensionalidad del análisis y enfocar la atención en las variables con mejor capacidad para explicar el desempeño de los estudiantes.

Sin embargo, el proceso implicaba una reducción considerable en el número de variables, lo que resultaba en una cantidad muy limitada. Para lograr un equilibrio, se optó por filtrar las correlaciones iguales o superiores a ±0.15, que es el puntaje medio para correlaciones de baja intensidad. Como resultado, se obtuvo la matriz de correlación que se presenta en la **Figura 4**:

Figura 4

Matriz de Correlación de las Variables que Cumplen el Umbral



Esto permitió empezar a comprender los factores que influyen los posibles resultados obtenidos por un estudiante. El calendario del colegio, el colegio mismo, su género, jornada y naturaleza parecen determinar el rendimiento de los estudiantes. Así mismo, de las condiciones familiares, el nivel educativo de los padres y los activos familiares parecen ser relevante en los puntajes obtenidos.

Con la depuración realizada, gracias al análisis de correlación y el umbral definido, para etapas posteriores del desarrollo del proyecto se trabajó únicamente con las variables categóricas mostradas en el mapa de calor mostrado anteriormente. A continuación, se detalla el análisis para

algunas de estas variables agrupadas por características (colegio, activos familiares, nivel educativo de los padres).

Análisis de Puntajes

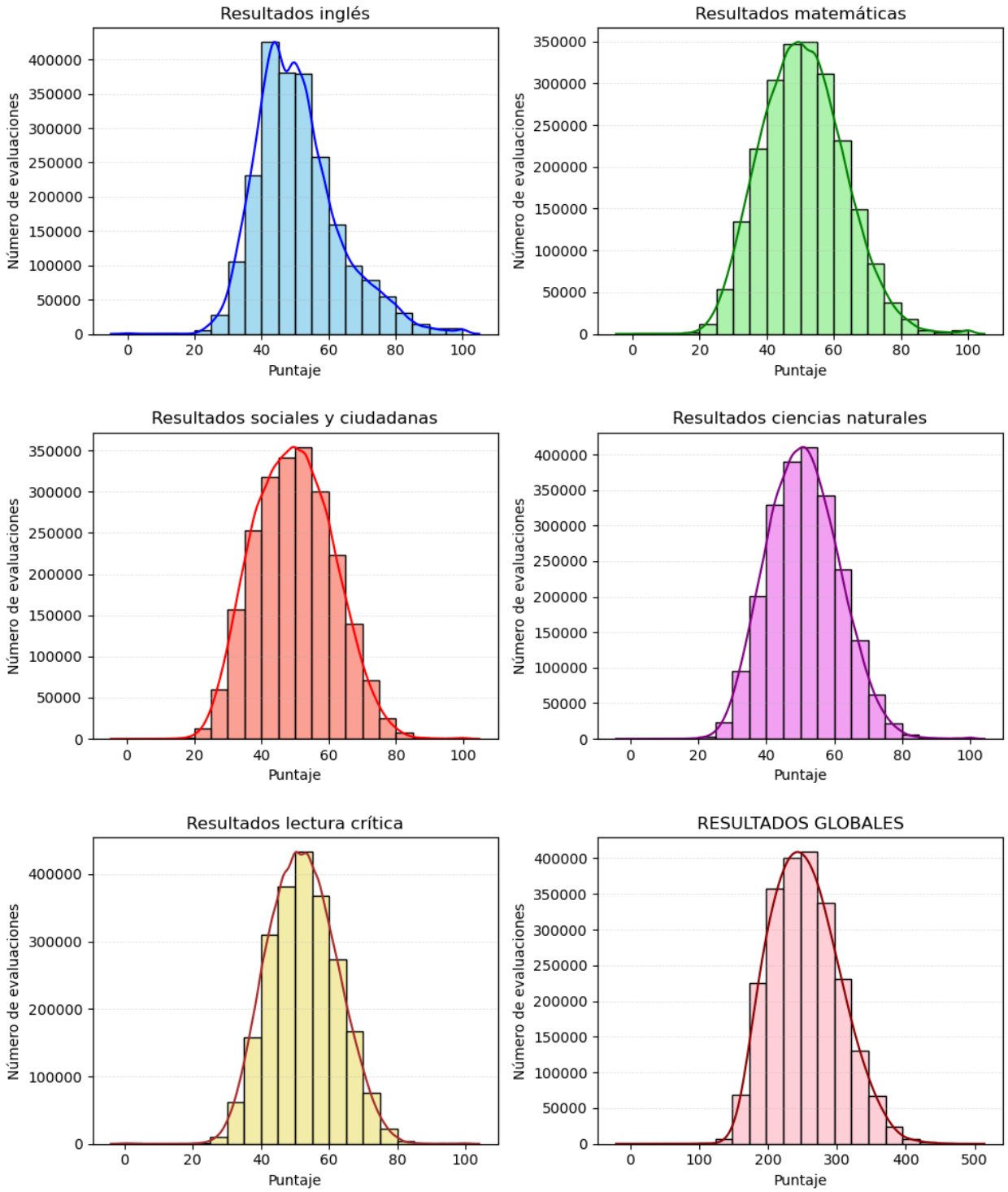
Una vez obtenidas las variables que cumplen con el umbral de correlación definido se procedió a realizar un mayor análisis de algunas de ellas.

Aunque los puntajes de las áreas no se consideraron para la aplicación de los modelos de Machine Learning, se hizo un análisis de su distribución como parte de la etapa exploratoria. En la

Figura *5*, se muestran los histogramas de los puntajes de todas las áreas:

Figura 5

Histogramas de los Puntajes



En general se observa que los puntajes siguen una distribución normal, indicando que la mayoría de los estudiantes obtuvieron puntajes alrededor de la media, los puntajes muy bajos o muy altos son menos.

Apoyando dichas visualizaciones con los datos estadísticos mostrados en secciones anteriores, se puede comentar que los promedios por área rondan los 50 puntos, con un rango que va desde los 49.55 puntos en sociales y ciudadanas hasta los 52.18 puntos en lectura crítica. Además, las desviaciones estándar de las áreas oscilan entre 10.08 (lectura crítica) y 12.35 (inglés), lo cual indica que los puntajes individuales presentan una dispersión moderada respecto a la media.

Es importante resaltar que los resultados de inglés muestran un comportamiento algo distinto a las demás áreas. El puntaje promedio es 50.62, lo que es comparable con otras áreas como matemáticas (50.76) y ciencias naturales (50.85), pero la desviación estándar en inglés es relativamente mayor (12.35 frente a 10.08 en lectura crítica). La distribución de inglés tiene una mayor concentración de puntajes en los rangos más bajos, haciendo que la distribución se sesgue hacia la izquierda. Los puntajes de 0 a 40 son más comunes en inglés, como lo indica el percentil 25 (42) y la mediana (49). En contraste, otras áreas tienen sus valores más bajos en un rango algo más alto (por ejemplo, el percentil 25 de lectura crítica es 45 y la mediana es 52). Asimismo, la gráfica permite distinguir dos picos, sugiriendo que es una variable bimodal, se observa un grupo de estudiantes con puntajes bajos (entre 30 y 40) y otro grupo de estudiantes con puntajes altos (entre 50 y 60). Esta bimodalidad puede deberse a diferentes niveles de contacto con el idioma, algunos parecieran tener sólidas bases mientras que otros tal vez han tenido formación básica o mínima. Esto puede tener relación con las diferencias entre estudiantes de colegios privados y públicos, urbanos vs. rurales, o entre regiones.

Respecto a los valores mínimos y máximos, todas las áreas reportan puntajes mínimos de cero, lo cual puede indicar la presencia de casos atípicos o errores de evaluación. Los puntajes máximos alcanzan el valor límite de la escala (100) en todas las áreas individuales, y 494 en el puntaje global. Las puntuaciones en los percentiles 25, 50 y 75 sugieren una distribución simétrica, con una mediana muy cercana a la media en cada caso.

Por otro lado, el puntaje global (PUNT_GLOBAL) tiene una media de 254.09 y una desviación estándar de 52.21, lo que es señal de que, aunque muchos de los estudiantes tienen entre 215-287 puntos, hay una dispersión considerable. El resultado global es el resultado de la combinación de los puntajes por área, donde la variabilidad vista en inglés podría estar contribuyendo a esta dispersión.

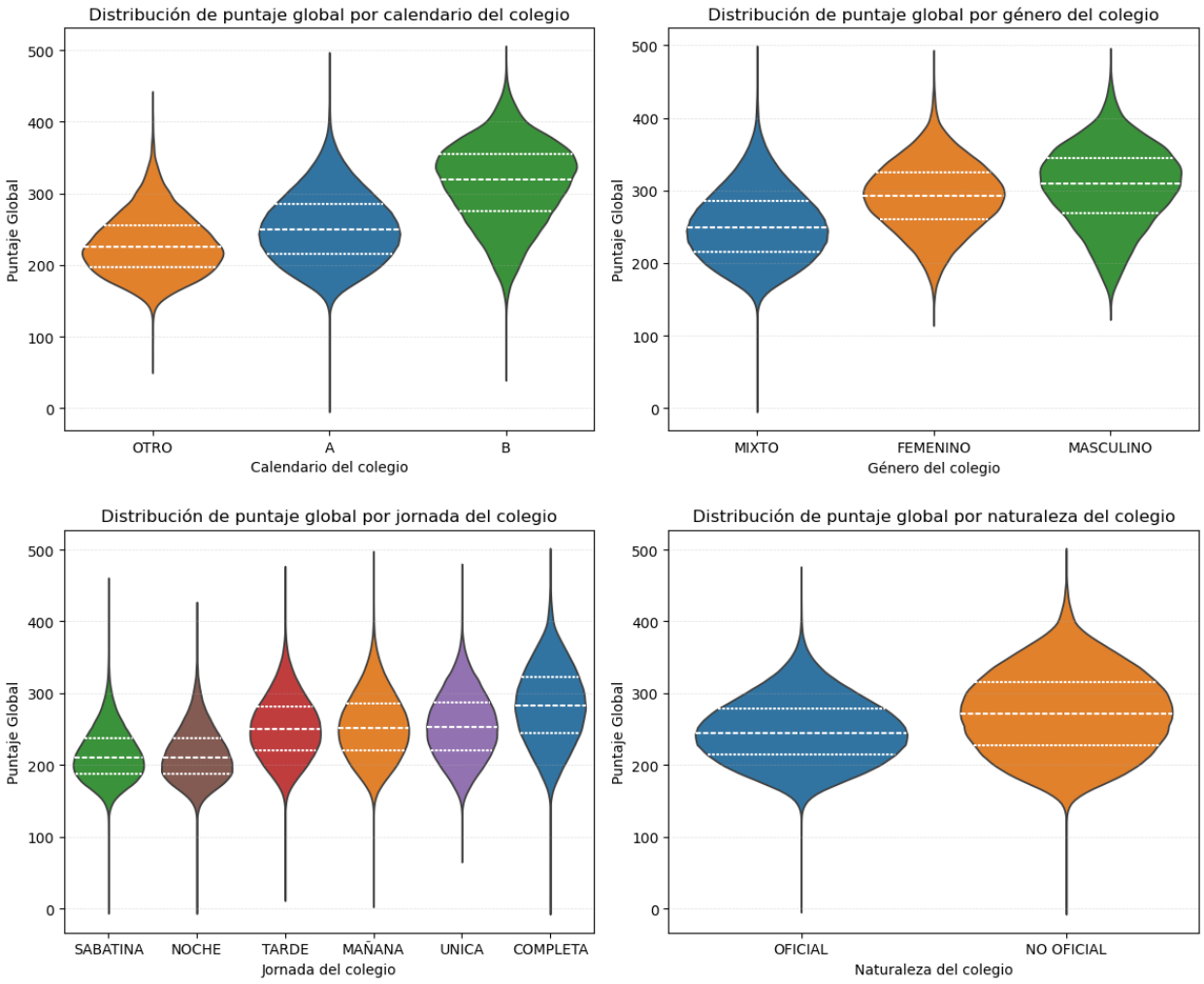
Estos resultados permiten concluir que, aunque existe una ligera variabilidad entre áreas, en general el comportamiento de los puntajes es consistente, ya que no se presentan distorsiones significativas ni desviaciones graves respecto a la normalidad.

Análisis de Puntaje Global versus Características del Colegio

En la **Figura 6** se muestran los resultados obtenidos al analizar el puntaje global con las variables asociadas al colegio: calendario, género, jornada y naturaleza. Es importante señalar que se tomaron únicamente las variables que cumplieron con el umbral definido anteriormente.

Figura 6

Distribución de Puntaje Global Versus Características del Colegio



La gráfica anterior muestra la distribución del puntaje global en la prueba Saber 11, según el calendario, el género, la jornada y la naturaleza del colegio. A partir de estos gráficos y los estadísticos descriptivos correspondientes, se identificaron patrones relevantes en el desempeño académico de los estudiantes.

En cuanto al calendario, los resultados evidencian diferencias marcadas según el tipo de calendario. Los colegios con calendario B, que usualmente son colegios privados, presentan los

puntajes más altos, con un promedio de 314.41 puntos y una mediana de 320, mientras que los de calendario A presentan una media de 254.38 y una mediana de 251, casi 60 puntos por debajo respecto al calendario B. En contraste, los colegios de “Otro” tipo de calendario, que corresponden a instituciones de educación acelerada u otras metodologías de enseñanza, obtienen los puntajes más bajos, con una media de 232.66 y mediana de 227, cayendo más o menos 22 puntos por debajo de los colegios de calendario A. Estos resultados indican una posible asociación entre el tipo de calendario y el rendimiento académico, siendo los colegios con calendario B, frecuentemente del sector privado, los que logran los puntajes más altos.

En relación con el género del colegio, se observa que los colegios masculinos tienen el promedio más alto con 305.91 puntos y los femeninos tienen un promedio de 292.35, cayendo casi 14 puntos. Por su parte, los colegios mixtos registran el promedio más bajo con 254.51 puntos, junto con una mayor dispersión en sus resultados. Aunque los colegios de un solo género obtienen puntajes más altos en promedio, la distribución sugiere que no hay diferencias estructurales significativas en la forma de los puntajes entre colegios masculinos y femeninos.

Asimismo, se identifican diferencias sustanciales en el rendimiento académico según la jornada del colegio. Las jornadas completa y única presentan los puntajes más altos, con promedios de 283.46 y 255.65 puntos, respectivamente. En contraste, las jornadas nocturna y sabatina muestran los puntajes más bajos, con medias de 219.04 y 218.28, respectivamente. Estas últimas también presentan una mayor concentración de puntajes en los niveles más bajos, lo cual podría reflejar factores contextuales que afectan el desempeño, como la necesidad de compatibilizar trabajo y estudio o limitaciones en la intensidad horaria.

Finalmente, se evidencian diferencias según la naturaleza del colegio. Los colegios no oficiales (privados) presentan una media de 274.88 puntos, mientras que los oficiales (públicos)

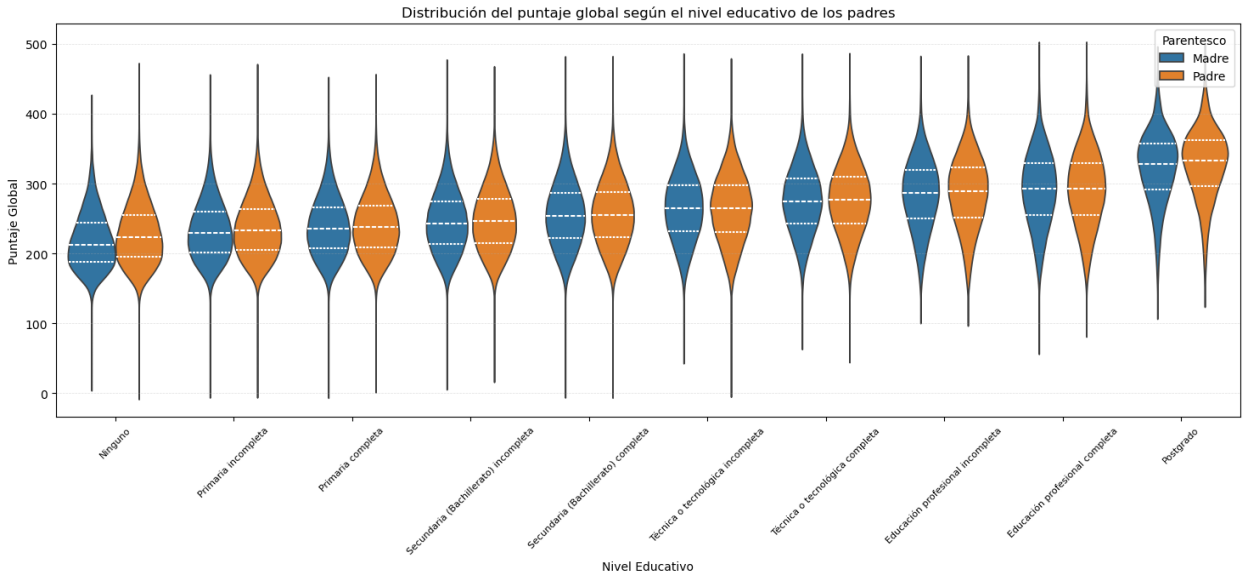
registran una media de 249.03 puntos. Aunque existe cierta superposición entre ambas distribuciones, los colegios privados tienden a concentrar una mayor proporción de puntajes en los niveles altos. Esta diferencia sugiere una posible influencia de los recursos institucionales y de los propios recursos familiares, ya que a menores posibilidades económicas la probabilidad de tener que asistir a un colegio público es mayor.

Análisis de Puntaje Global versus Nivel Educativo de los Padres

Posteriormente se procedió a hacer un análisis enfocado en la influencia del nivel de educativo de los padres, los resultados se muestran en la **Figura 7**:

Figura 7

Distribución de Puntaje Global Versus Educación de los Padres



La anterior figura muestra la distribución del puntaje global según el nivel educativo de los padres, diferenciando entre madre y padre. De manera general, se observa una clara tendencia donde a medida que el nivel educativo de los padres aumenta, también lo hace el puntaje global de los evaluados. Por ejemplo, los estudiantes cuyos padres tienen formación de posgrado

alcanzan promedios de 322.6 (madre) y 327.1 (padre) puntos, mientras que aquellos cuyos padres no tienen educación formal obtienen promedios significativamente más bajos entre 218.2 (madre) y 227.1 (padre).

Esta diferencia también se refleja en la dispersión de los resultados. En el caso de padres con nivel de posgrado, la mediana de puntaje global supera los 328 puntos, mientras que en padres sin estudios, la mediana no alcanza los 225. A pesar de estas diferencias, en todos los niveles educativos existe una amplia variabilidad de puntajes, incluso en los niveles más altos, se encuentran estudiantes con resultados bajos, y viceversa.

Respecto a la comparación entre madre y padre, las distribuciones son bastante similares en cada nivel educativo, sin una diferencia sustancial. En niveles como la educación profesional completa, tanto madres como padres presentan una mediana de 293 puntos, lo que sugiere que el impacto del nivel educativo no varía sustancialmente según el parentesco. Sin embargo, se observa una ligera ventaja a favor del nivel educativo del padre en algunos tramos, como en el nivel de posgrado, donde la mediana es 5 puntos mayor en comparación con la madre.

Por último, los niveles educativos más bajos, como primaria incompleta, presentan una mayor concentración de estudiantes con puntajes bajos: en madres con este nivel, el puntaje promedio es de 232.4 puntos, en contraste con los 291.5 puntos de aquellas con formación profesional completa.

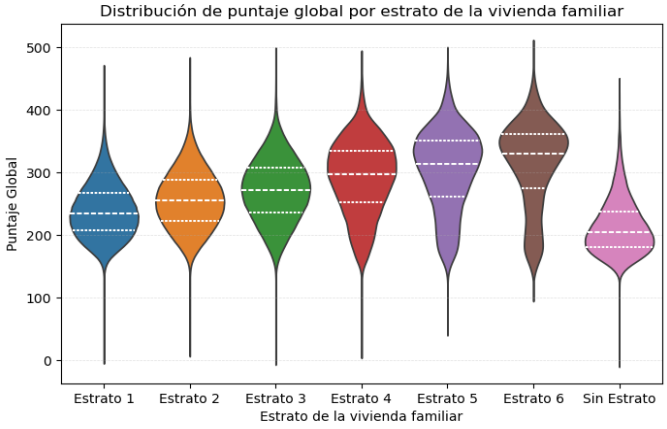
En resumen, los datos muestran que el nivel educativo de los padres está positivamente correlacionado con el desempeño de los estudiantes en la prueba. Este patrón sugiere que el entorno educativo y las oportunidades asociadas al nivel académico familiar tienen un rol importante en los resultados académicos.

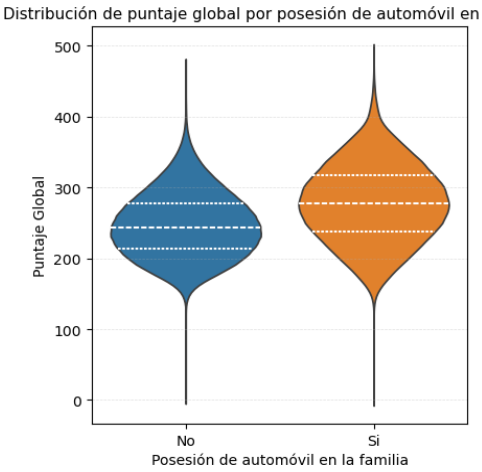
Análisis de Puntaje Global versus Condiciones Familiares

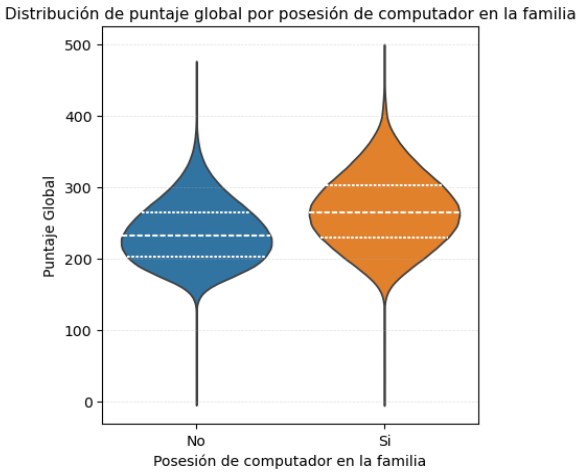
Después de hacer el análisis de puntajes globales a la luz del nivel educativo de los padres, se procedió a evaluar la relación de los puntajes globales con el resto de las variables familiares: estrato y posesión de diferentes activos. En la **Figura 8** se muestran los gráficos de distribución para dichas variables.

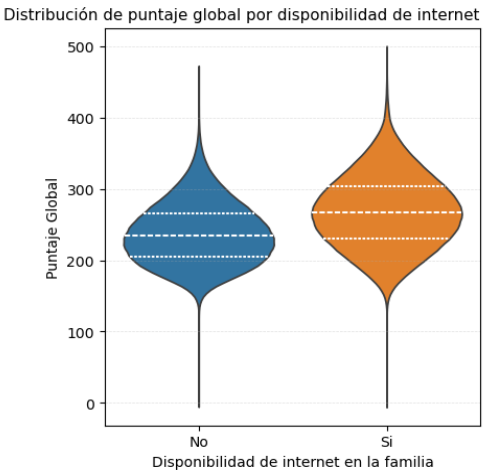
Figura 8

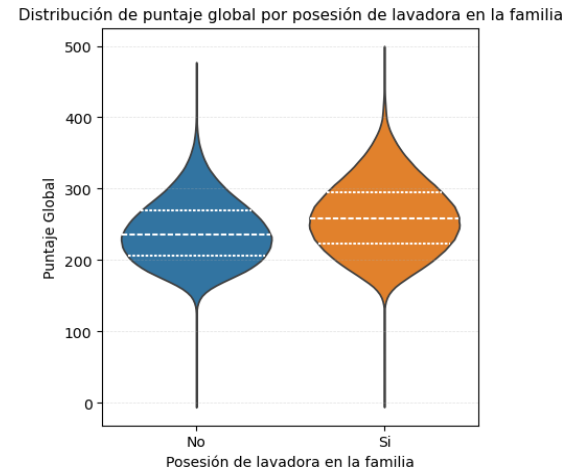
Distribución de Puntaje Global Versus Características Familiares











1. Distribución de puntaje global por estrato de la vivienda familiar

- Se observa que, en general, a medida que aumenta el estrato socioeconómico (de Estrato 1 a Estrato 6), la mediana del puntaje tiende a aumentar.
- Los estudiantes en estratos más altos (5 y 6) tienen distribuciones más altas en comparación con los de estratos más bajos.
- Los estudiantes sin estrato presentan una distribución con una media más baja, similar a los estratos bajos.
- Hay una relación entre el nivel socioeconómico y el puntaje global, donde los estudiantes de estratos más altos tienden a obtener mejores resultados.

2. Posesión de automóvil

- Los estudiantes con automóvil en la familia tienden a tener una distribución de puntajes más alta en comparación con los que no tienen.
- La mediana y el rango intercuartil son ligeramente mayores para quienes tienen automóvil.

3. Posesión de computador

- Se observa un patrón similar: quienes tienen computador en casa tienden a tener puntajes globales más altos.
- La distribución de puntajes es más amplia y más centrada en valores altos para los estudiantes que tienen un computador en casa.
- En contraste, los estudiantes que no tienen computador presentan una distribución más concentrada en los puntajes bajos y medios.

4. Disponibilidad de internet

- Tener acceso a internet en el hogar también parece estar asociado con un mejor desempeño en el puntaje global.
- La diferencia entre Si y No es notoria en la distribución.

5. Posesión de lavadora

- Aunque menos evidente que en los otros casos, también se nota una ligera tendencia en la que los estudiantes con lavadora en casa presentan una distribución de puntajes más elevada.

En general se puede decir que un mejor estrato socioeconómico, lo que tiene una relación con la posesión de bienes materiales y acceso a la tecnología en la familia, parece estar relacionado con un mejor desempeño en el examen. Esto puede estar asociado a un mejor acceso a recursos educativos y un entorno más favorable para el estudio.

Análisis de Puntaje Global por Departamento y Región

A pesar de que la variable COLE_COD_DEPTO_UBICACION no está dentro del umbral definido para el análisis, se decidió realizar una revisión extra que permitiera desglosar el comportamiento, pero consolidado a nivel de los departamentos de Colombia. A continuación, la **Tabla 11** consolida las estadísticas descriptivas de los puntajes globales obtenidos por los estudiantes, en cada departamento. La información incluye medidas de tendencia central, dispersión y distribución, lo que permite comparar el desempeño académico entre las diferentes entidades territoriales.

Tabla 11

Estadísticas de Puntajes a Nivel de Departamento

Departamento	Promedio	Desviación Estándar	Puntaje Mínimo	25%	50%	70%	Puntaje Máximo	Cantidad Pruebas	% Pruebas
BOGOTÁ, D.C.	273.83	49.30	49	238	272	307	494	370417	16.34
BOYACÁ	265.62	45.77	17	233	264	297	471	67373	2.97
SANTANDER	265.57	51.81	0	228	263	301	478	106654	4.70
CUNDINAMARCA	261.58	46.86	17	228	260	293	483	151575	6.68
NORTE DE SANTANDER	258.20	47.74	60	223	255	290	492	62155	2.74
RISARALDA	257.90	47.23	51	223	256	289	476	43438	1.92
VALLE DEL CAUCA	256.66	50.72	0	219	253	290	483	193867	8.55
QUINDIO	255.03	48.70	39	219	252	287	466	28526	1.26
META	254.85	45.40	57	222	253	285	472	48511	2.14
HUILA	253.51	47.56	21	218	250	285	469	56402	2.49
NARIÑO	253.31	50.92	6	216	251	288	471	69274	3.06
CALDAS	253.30	48.55	15	218	250	285	471	44107	1.95
CASANARE	252.46	45.03	31	220	250	282	452	23411	1.03
ANTIOQUIA	250.09	50.24	0	212	247	284	490	298830	13.18
ATLÁNTICO	248.88	51.56	12	210	244	283	483	127399	5.62
ARAUCA	248.08	44.97	0	215	245	278	424	12179	0.54
TOLIMA	246.90	45.57	50	213	244	277	460	68507	3.02
CESAR	243.78	45.84	53	210	239	273	465	51251	2.26
PUTUMAYO	242.67	44.79	107	209	239	272	456	17165	0.76
SAN ANDRÉS	242.44	49.12	3	205	237	276	409	2624	0.12
CÓRDOBA	242.28	46.37	0	208	237	272	471	77617	3.42
SUCRE	241.74	47.80	8	206	236	273	476	43562	1.92
CAQUETÁ	240.11	43.20	104	208	237	269	418	17375	0.77
GUAINÍA	238.75	46.78	137	202	237	271	403	941	0.04
CAUCA	237.10	46.67	0	202	232	266	462	60116	2.65
BOLÍVAR	235.54	49.04	0	198	228	266	475	101327	4.47
VICHADA	234.59	41.25	122	204	235	263	387	1905	0.08
GUAVIARE	233.17	41.14	99	203	229	260	407	3796	0.17
MAGDALENA	228.36	43.85	17	196	222	255	460	63392	2.80
LA GUAJIRA	227.51	45.44	11	193	222	256	459	31495	1.39
AMAZONAS	222.14	41.81	50	191	216	247	413	3345	0.15
VAUPÉS	217.90	38.54	138	190	212	239	376	1490	0.07
CHOCÓ	212.48	40.30	55	184	206	236	389	17469	0.77

Nota. Datos estadísticos para cada departamento, ordenado del más alto promedio al más bajo.

El análisis descriptivo de los puntajes globales de la prueba Saber 11, desagregado por departamento, revela una amplia variabilidad en el desempeño académico a nivel territorial en Colombia. Bogotá se posiciona como la entidad territorial con el puntaje promedio más alto (M = 273.83, DE = 49.30), superando significativamente la media nacional y concentrando el 16.34% de todas las pruebas aplicadas, lo que la convierte también en la región con mayor representatividad en la muestra.

Departamentos como Boyacá (M = 265.62), Santander (M = 265.57) y Cundinamarca (M = 261.58) también presentan puntajes promedio superiores al promedio nacional de 254.1 puntos, lo cual puede relacionarse con factores estructurales como cobertura educativa, calidad docente y condiciones socioeconómicas. En contraste, Chocó (M = 212.48, DE = 40.30), Vaupés (M = 217.90) y Amazonas (M = 222.14) se ubican en los últimos lugares del ranking, reflejando brechas persistentes en términos de equidad y acceso a una educación de calidad.

El rango de puntajes también presenta grandes diferencias entre territorios. Por ejemplo, Bogotá muestra un puntaje máximo de 494 y un mínimo de 49, mientras que regiones como Chocó y Vaupés registran valores máximos de 389 y 376 respectivamente, lo que sugiere un tope de desempeño más limitado. Departamentos como Bolívar y Valle del Cauca presentan amplias desviaciones estándar (DE = 49.04 y DE = 50.72, respectivamente), lo cual indica una alta dispersión de resultados, posiblemente asociada con desigualdades internas dentro de cada región.

En cuanto a la participación relativa, además de Bogotá, departamentos como Antioquia (13.18%), Valle del Cauca (8.55%) y Cundinamarca (6.68%) concentran un porcentaje elevado del total de pruebas, reflejando tanto su densidad poblacional como su peso en el sistema educativo colombiano. En cambio, departamentos como Vaupés (0.07%), Guainía (0.04%) y San

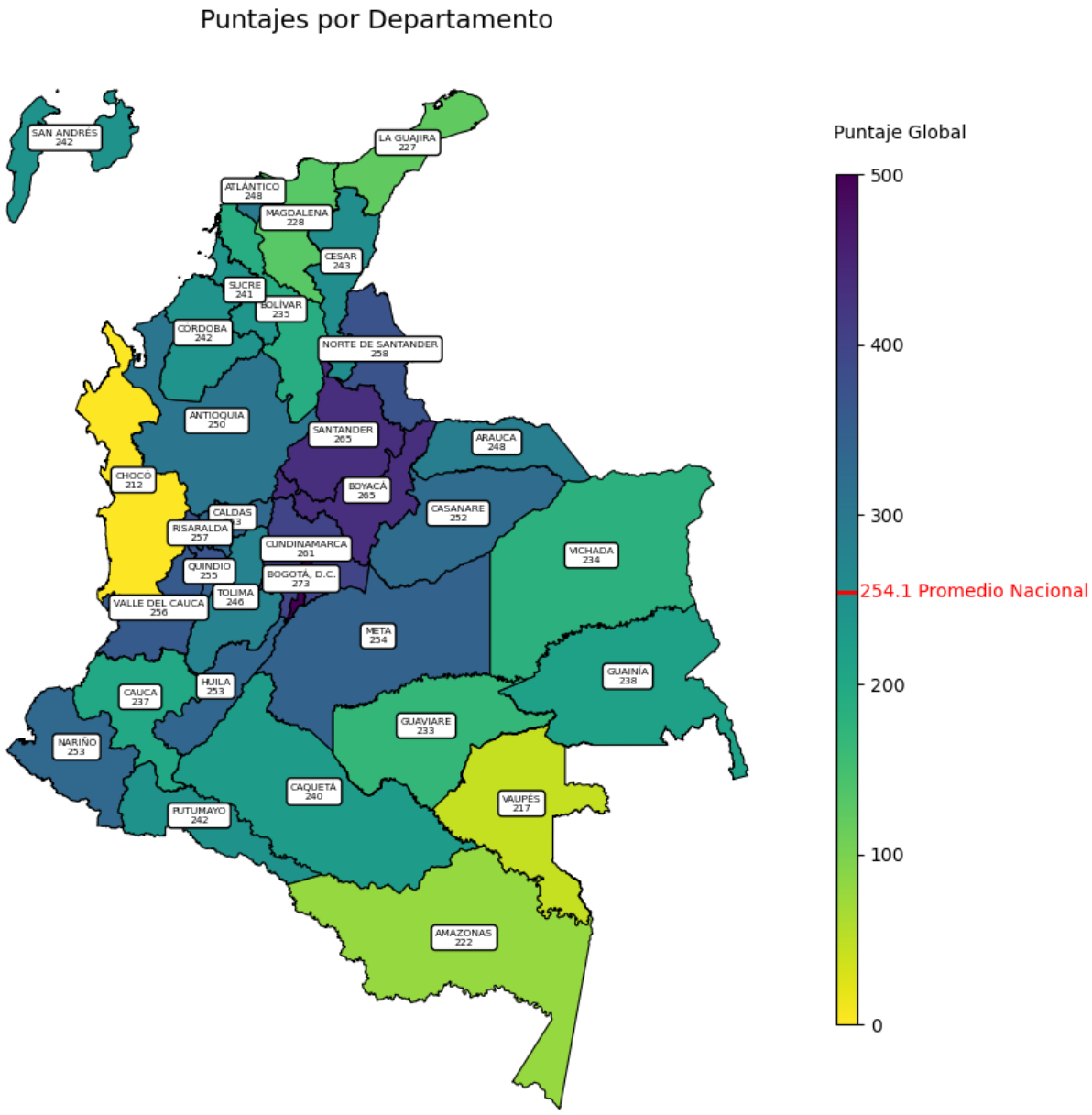
Andrés (0.12%) tienen una participación marginal, lo cual podría limitar el análisis estadístico detallado en estas zonas por el tamaño reducido de la muestra.

En resumen, los resultados evidencian un patrón desigual en el rendimiento académico regional, lo que enfatiza la necesidad de políticas públicas diferenciadas que atiendan tanto a los departamentos con bajo rendimiento como a aquellos con alta variabilidad interna.

Con las estadísticas generadas para cada departamento, se realizó un procesamiento adicional de los datos, cruzando los departamentos con su respectiva información georreferencial. Se utilizó la librería geopandas para leer el shapefile “Versión MGN2023-Nivel Departamento” dispuesto en el geoportal del (DANE, 2023). Así mismo se usó como referencia el proyecto de GitHub en (Castañeda, 2023) para la generación del mapa de departamentos de Colombia que se muestra en la **Figura 9**. Una vez creado el mapa se incluyó información del promedio del puntaje global para cada departamento, así mismo, se muestra el promedio nacional de 254.1 como referencia para efectos comparativos:

Figura 9

Media del Puntaje Global por Departamento



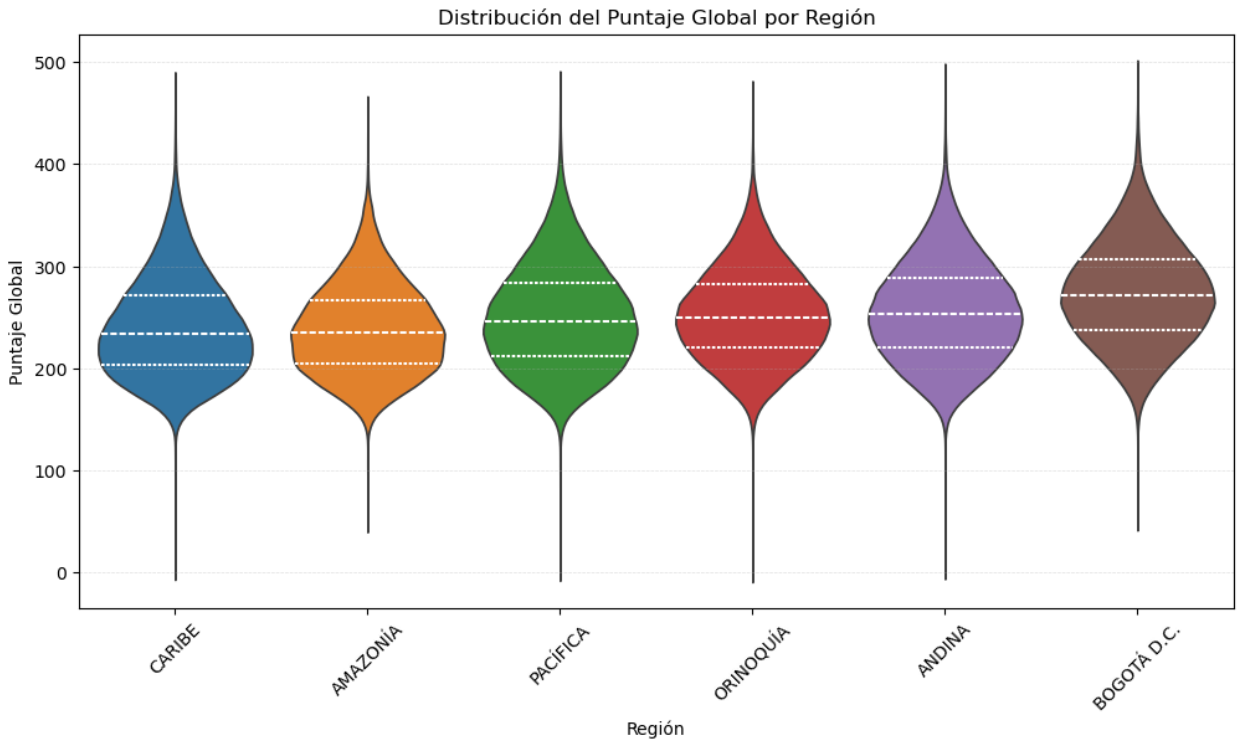
En la representación del mapa los colores más claros representan los menores puntajes. En este caso es muy claro el bajo desempeño académico de departamentos como Chocó, Vaupés, Amazonas, Guajira, Magdalena, Guaviare, Vichada, Bolívar, Guainía y Cauca, estos son los 10

departamentos con menor desempeño, muchos de estos siendo departamentos de la región Amazónica y del Caribe.

Posteriormente, se hizo un procesamiento extra en los datos, haciendo un mapeo de departamentos con su respectiva región y así se logró hacer un análisis de los puntajes por cada región. La **Figura 10** presenta la distribución del puntaje global por región:

Figura 10

Distribución de Puntaje Global a Nivel de Región



Se puede notar que la Amazonía y el Caribe tienen una mediana muy similar, en el caso de la Amazonía hay menos estudiantes con puntajes muy elevados y con puntajes muy bajos, el Caribe presenta más casos de puntajes elevados y puntajes bajos. La región Pacífica por su parte muestra una distribución de puntajes más homogénea que las dos regiones anteriores y presenta más concentración de puntajes por encima de la mediana. La mediana de la Orinoquía es muy

similar a la de la región pacífica, pero presenta más variabilidad. La región Andina permanece con una mediana cercana a la Orinoquía, pero tiene más estudiantes con puntajes altos. Finalmente, al llegar a Bogotá su distribución es más simétrica y presenta una concentración alta de puntajes en la parte media. Su mediana está en el rango 275-300 puntos. Sugiere que los estudiantes en esta región tienen un desempeño más uniforme y menos disperso.

En la **Tabla 12** se presentan en detalle los datos estadísticos reflejados en la **Figura 10**.

Tabla 12

Estadísticas de Puntajes a Nivel de Región

Región	Promedio	Desviación estándar	Puntaje mínimo	25%	50%	70%	Puntaje máximo	Cantidad pruebas	% pruebas
BOGOTÁ D.C.	273.83	49.30	49	238	272	307	494	370417	16.34
ANDINA	256.06	49.06	0	220	253	289	492	927567	40.91
ORINOQUÍA	252.79	45.29	0	220	250	283	472	86006	3.79
PACÍFICA	250.26	50.87	0	212	246	284	483	340726	15.03
CARIBE	240	48.57	0	203	234	271	483	498667	21.99
AMAZONÍA	238.37	43.99	50	205	235	267	456	44112	1.95

Nota. Datos estadísticos para cada región, ordenado del más alto promedio al más bajo.

El análisis regional de los puntajes globales en la prueba Saber 11, así como la revisión por departamentos, revelan desigualdades en el rendimiento académico. Bogotá como distrito capital se posiciona como la región con el promedio más alto (273.83), reflejando no solo una mayor inversión en educación, sino también mejores condiciones estructurales y de acceso. Esta región, además, presenta una desviación estándar de 49.30, lo que indica una dispersión moderada en los resultados, aunque menor que la observada en regiones como la Pacífica (50.87), donde la variabilidad en el desempeño sugiere diferencias significativas en las oportunidades educativas entre estudiantes del mismo territorio.

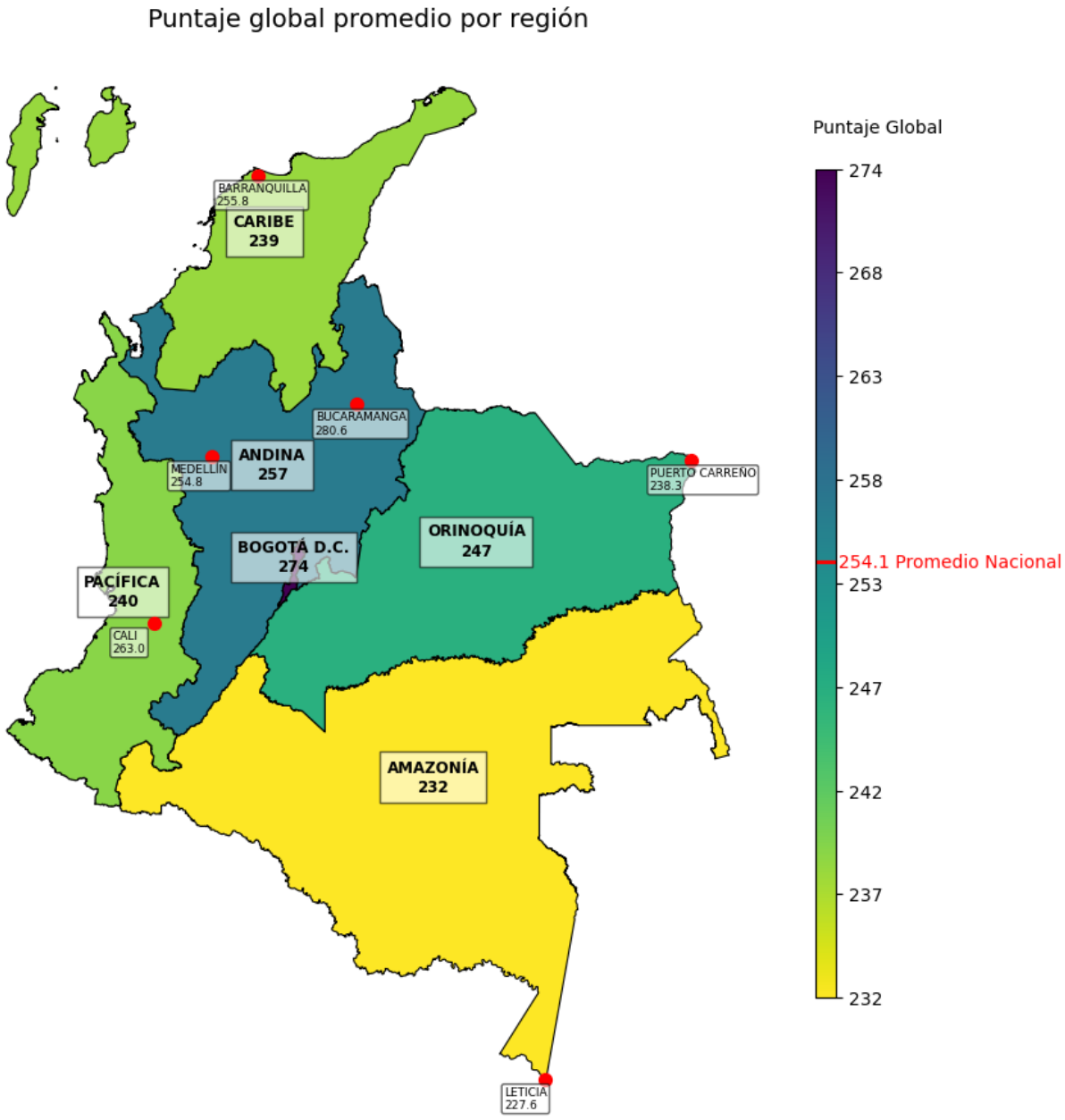
La región Andina, que concentra el mayor porcentaje de pruebas aplicadas (más del 40 %), tiene un promedio de 256.06 y una desviación estándar de 49.06, lo que apunta a una distribución del rendimiento similar a la de Bogotá, pero con una leve reducción en el promedio. Por otro lado, las regiones Caribe (240) y Amazonía (238.37) exhiben los promedios más bajos, acompañados de desviaciones estándar menores, 48.57 y 43.99 respectivamente, lo que puede interpretarse como una menor variabilidad, aunque en un contexto de bajo desempeño generalizado.

Estas diferencias reflejan no solo el impacto de los factores estructurales en la calidad educativa, sino también la persistencia de brechas regionales que limitan el acceso equitativo a oportunidades de aprendizaje de calidad. Reducir la dispersión y elevar los promedios en regiones históricamente rezagadas requiere fortalecer la inversión en formación docente, infraestructura escolar y acceso a tecnologías educativas, especialmente en zonas rurales y periféricas del país.

Así mismo, para generar la **Figura 11** se hizo uso de una representación georreferencial para diagramar los puntajes globales medios por región de acuerdo con lo que se mostró en la tabla anterior, adicionalmente se incluyó el promedio de la ciudad más importante para cada región:

Figura 11

Promedio de Puntaje Global a Nivel de Región



El mapa del puntaje global promedio por región, junto con los datos estadísticos, evidencia profundas brechas territoriales en el rendimiento académico de los estudiantes colombianos. Las regiones periféricas, caracterizadas por su geografía de difícil acceso y baja

densidad poblacional, como la Amazonía, la Pacífica, el Caribe y la Orinoquía, presentan los puntajes promedio más bajos del país. Es evidente que en estas regiones, muchas de ellas cubiertas por extensas zonas selváticas, las condiciones estructurales limitan el acceso a una educación de calidad.

La región Pacífica y la Caribe, por ejemplo, muestran promedios muy similares y consistentemente bajos, con 250.3 y 240 puntos respectivamente, lo que sugiere una tendencia compartida de atraso educativo. Por el contrario, los mejores desempeños se concentran en el centro del país, particularmente en la región Andina, con un promedio de 256.1, y en Bogotá, que alcanza los 273.8 puntos, ambos por encima del promedio nacional de 254.1. Esta concentración del rendimiento académico en el centro del país refleja, en gran medida, un modelo de desarrollo desigual, donde las oportunidades educativas están fuertemente centralizadas.

Más allá de los promedios regionales, al observar las principales capitales dentro de cada región, se hace aún más evidente la desigualdad interna. Ciudades como Bucaramanga, con un promedio de 280.6 puntos, superan con holgura la media de su región Andina, que se ubica en 256. De igual forma, Cali (263.0) y Medellín (254.8) también se sitúan por encima de sus respectivas regiones, lo cual indica que vivir en ciudades principales del país representa una ventaja sustancial para los estudiantes. Este patrón se repite en Barranquilla, donde el promedio de 255.8 contrasta fuertemente con el bajo desempeño general de la región Caribe, que apenas alcanza los 240 puntos.

Sin embargo, en regiones como la Amazonía y la Orinoquía, donde los promedios ya son bajos, ni siquiera las capitales logran acercarse al promedio nacional. Leticia (227.6) y Puerto Carreño (238.3) se ubican en el extremo inferior de la distribución, reflejando no solo las

carencias regionales sino también las dificultades específicas que enfrentan estas ciudades para garantizar una educación de calidad.

Claramente, el lugar de residencia influye significativamente en los resultados educativos. Las capitales, especialmente las del interior del país, actúan como focos de mejor desempeño dentro de contextos regionales más rezagados. Esto podría explicarse por una mayor inversión en infraestructura escolar, una mejor dotación docente y más oportunidades de acceso a servicios educativos complementarios. En cambio, las regiones periféricas enfrentan retos históricos asociados a la dispersión geográfica, baja conectividad y limitaciones presupuestarias, lo que se traduce en una menor equidad educativa.

En definitiva, los datos muestran que el sistema educativo colombiano no solo está fragmentado entre regiones, sino también dentro de ellas. Esta realidad plantea el desafío de diseñar políticas educativas con enfoque territorial que no solo eleven los promedios regionales, sino que también reduzcan las desigualdades estructurales que hoy condicionan el futuro de los estudiantes.

Preparación de Datos

Durante la fase de exploración y análisis de datos, se derivó la variable **región** para cada registro con el propósito de aplicar una librería de geolocalización, poner en práctica su uso y realizar una caracterización de los resultados a nivel regional. No obstante, debido a su baja correlación con la variable objetivo y con el fin de reducir la dimensionalidad del modelo, esta variable fue excluida del conjunto de características utilizadas en los modelos. Además, las pruebas realizadas evidenciaron que su inclusión incrementaba significativamente el tiempo de ejecución sin aportar mejoras en el desempeño de las métricas evaluadas.

A lo largo del desarrollo del proyecto, se realizaron numerosas iteraciones en el proceso de preparación de los datos, incluyendo codificaciones que optimizaran el desempeño de los algoritmos y la evaluación de múltiples combinaciones para obtener mejores resultados. En el caso de los modelos de clasificación, también se exploraron distintas formas de agrupar la variable objetivo, ajustando las categorías en busca de una mejor capacidad predictiva.

A continuación, en la **Tabla 13** se muestran las variables y la cantidad de registros donde había información faltante. Para mejorar los análisis, se eliminaron los registros. El total de registros antes de la limpieza de nulos era 2267495, total de registros luego de la limpieza de nulos: 2174129.

Tabla 13

Registros Faltantes por Variable

Variable	Registros faltantes
COLE_CALENDARIO	0
COLE_GENERO	0
COLE_JORNADA	0
COLE_NATURALEZA	0
FAMI_EDUCACIONMADRE	60602
FAMI_EDUCACIONPADRE	60760
FAMI_ESTRATOVIVIENDA	66556
FAMI_TIENEAUTOMOVIL	33401
FAMI_TIENECOMPUTADOR	30894
FAMI_TIENEINTERNET	62137
FAMI_TIENELAVADORA	30069

Nota. Detalle de la cantidad de registros con información faltante.

Asimismo, se realizó un tratamiento específico para algunas variables. Por ejemplo, en el caso del nivel educativo del padre y de la madre, se eliminaron los registros con los valores “No Aplica” y “No sabe”, dado que no aportan información útil para el análisis del puntaje global. En

la **Tabla 14** se puede ver la cantidad de datos que tenían este comportamiento. Luego de borrar los registros que tenían este tipo de información se pasó de 2174129 a 2039743 registros.

Tabla 14

Nivel Educativo de Padres sin Información

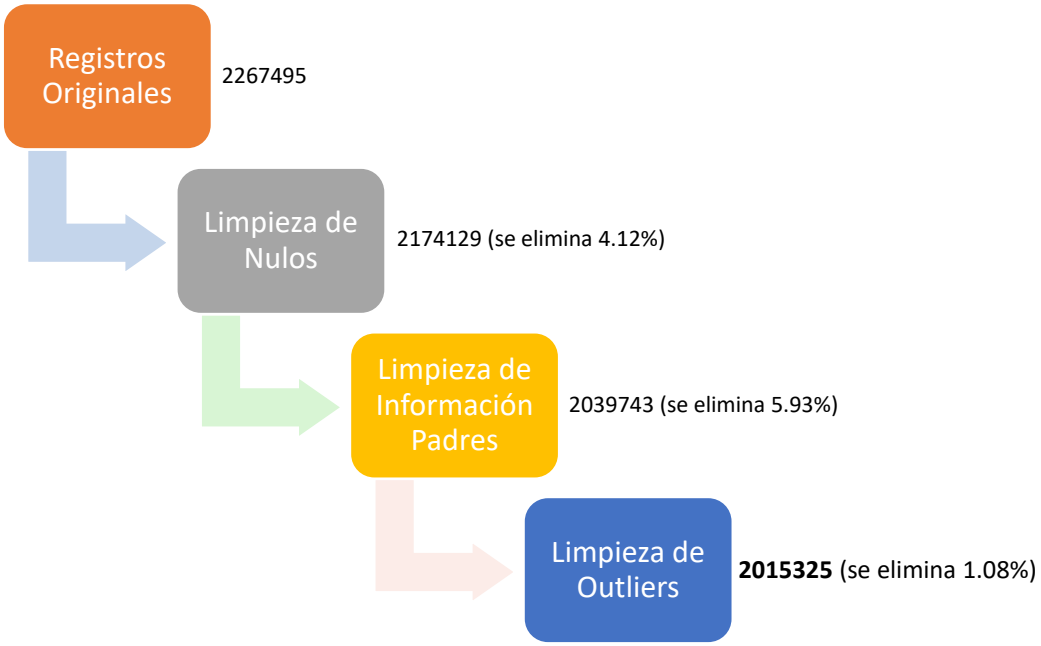
Variable	Valor	Conteo
FAMI_EDUCACIONMADRE	No Aplica	2032
FAMI_EDUCACIONMADRE	No sabe	34676
FAMI_EDUCACIONPADRE	No Aplica	11788
FAMI_EDUCACIONPADRE	No sabe	108903

Nota. Nivel educativo de padres irrelevante.

La siguiente estrategia de limpieza consistió en eliminar datos atípicos (outliers) de manera controlada, para ello, se acudió al cálculo de cuantiles. Se dividieron los datos en 1000 cuantiles de los cuales se eliminaron los datos del cuantil 1 y 999. Con esta limpieza se pasa de 2039743 registros a 2015325. A continuación, en la **Figura 12** se listan las etapas y la cantidad de registros que resultó luego de cada paso.

Figura 12

Etapas Preparación de Datos



Recapitulando la estrategia de preparación de datos, se aplicó una limpieza progresiva con el objetivo de asegurar la calidad, consistencia y confiabilidad de la información utilizada para el análisis. Esta limpieza se desarrolló en varias etapas, como se muestra en la gráfica anterior, comenzando con la depuración de registros con valores nulos en variables clave. En esta fase se eliminó aproximadamente el 4.12% de los datos originales, ya que la ausencia de información esencial no permitía una imputación precisa y habría generado sesgos en los resultados.

Posteriormente, se realizó una depuración centrada en la información relacionada con los padres. Esta fue la etapa más crítica y la que tuvo mayor impacto en la reducción del volumen de datos. Se detectaron numerosos registros con datos inválidos para el nivel educativo de los padres. Dado que esta información era vital para los objetivos del análisis, se optó por eliminar

aquellos registros que no cumplían con los criterios mínimos de integridad. Como resultado, en esta etapa se descartó un 5.93% adicional del total de registros.

Finalmente, se llevó a cabo la identificación y eliminación de outliers, es decir, valores extremos que no reflejaban patrones esperados y podían afectar negativamente las métricas estadísticas y el desempeño de los modelos. Esta última fase representó una reducción del 1.07% de los datos restantes.

En conjunto, el proceso de limpieza resultó en la eliminación de aproximadamente el 11% del total de registros originales. Esta depuración se considera justificada, ya que priorizó la calidad de los datos por encima del volumen, garantizando así que los análisis posteriores se realizaran sobre una base sólida, libre de errores críticos y suficientemente representativa.

Dado que todas las variables en el conjunto de datos son categóricas nominales, se emplearon dos técnicas de codificación para adaptarlas a los requisitos de los modelos utilizados. Se aplicó codificación por etiquetas (Label Encoding) en algunas variables, aunque estas no tienen un orden implícito, para simplificar su representación numérica. Además, se utilizó codificación one-hot (One Hot Encodeing) para evitar la introducción de relaciones inexistentes entre las categorías, garantizando que cada categoría se tratara de manera independiente. Estas transformaciones facilitaron la adecuada representación de los datos para el entrenamiento y evaluación de los modelos de clasificación y regresión empleados.

A continuación, en la **Figura 13** y **Figura 14** se muestran dos capturas de pantalla de cómo lucen los datos luego de las codificaciones, solo se muestran las variables relacionadas al calendario y género del colegio. Se muestra a manera de ejemplo, como con LabelEncoder son dos variables mientras que OneHotEncoder las descompone en seis columnas:

Figura 13

Variables Codificadas con Label Encoder

	COLE_CALENDARIO	COLE_GENERO
0	0	2
1	0	2
2	0	2
3	0	2
4	0	2
...	...	...
2836548	0	2
2836549	0	2
2836551	0	2
2836552	0	2
2836553	0	2

2487064 rows × 2 columns

Figura 14

Variables Codificadas con One Hot Encoder

	COLE_CALENDARIO_A	COLE_CALENDARIO_B	COLE_CALENDARIO_OTRO	COLE_GENERO_FEMENINO	COLE_GENERO_MASCULINO	COLE_GENERO_MIXTO
0	1.000	0.000	0.000	0.000	0.000	1.000
1	1.000	0.000	0.000	0.000	0.000	1.000
2	1.000	0.000	0.000	0.000	0.000	1.000
3	1.000	0.000	0.000	0.000	0.000	1.000
4	1.000	0.000	0.000	0.000	0.000	1.000
...	...	...	...	...	...	...
2836548	1.000	0.000	0.000	0.000	0.000	1.000
2836549	1.000	0.000	0.000	0.000	0.000	1.000
2836551	1.000	0.000	0.000	0.000	0.000	1.000
2836552	1.000	0.000	0.000	0.000	0.000	1.000
2836553	1.000	0.000	0.000	0.000	0.000	1.000

2487064 rows × 6 columns

Como es normal, para que los modelos de aprendizaje automático pudieran procesar adecuadamente las variables categóricas presentes en el conjunto de datos, fue necesario

transformarlas en representaciones numéricas. En este proceso se utilizaron dos técnicas muy usadas: Label Encoding y One Hot Encoding.

Label Encoding asigna un valor numérico a cada categoría dentro de una variable, lo que permite mantener la misma estructura original del conjunto de datos. Por ejemplo, si una variable como COLE_CALENDARIO tiene tres categorías distintas, estas se reemplazan por los valores 0, 1 y 2. Esta técnica tiene la ventaja de conservar la cantidad de columnas, lo cual puede resultar útil cuando se desea limitar la dimensionalidad del modelo.

Por otro lado, One Hot Encoding convierte cada categoría en una nueva columna binaria. De este modo, una variable como COLE_NATURALEZA, que distingue entre colegios "Oficiales" y "No Oficiales", genera dos columnas: una para cada categoría, marcando con 1 su presencia y con 0 su ausencia. Esto hace que el número de columnas aumente significativamente (en este caso de 11 a 44), lo que puede aportar mayor flexibilidad a los modelos, pero también incrementa la complejidad computacional. Para evitar el problema de colinealidad, una situación en la que una columna puede predecirse linealmente a partir de otra, se configuró el codificador para eliminar una categoría en las variables binarias. Por ejemplo, si una variable como FAMI_TIENE_INTERNET presenta solo dos categorías ("Sí" y "No"), se elimina una de las dos columnas generadas. Esta práctica, conocida como drop en la codificación one-hot, previene la redundancia sin pérdida de información, ya que la ausencia de una categoría implica la presencia de la otra.

Esta preparación de los datos permitió construir dos versiones distintas para el análisis: una utilizando Label Encoding y otra mediante One Hot Encoding. A partir de allí, fue posible evaluar diferentes modelos de aprendizaje automático con dos versiones de los datos, una para cada codificación.

Resultados de Modelos de Machine Learning

Una vez completada la preparación de los datos, se dio paso a la ejecución de múltiples iteraciones con distintos modelos de predicción, con el objetivo de estimar la variable PUNT_GLOBAL. En esta sección se detalla el proceso de aplicación de los diversos algoritmos de aprendizaje automático empleados durante el desarrollo del proyecto.

Regresión

El primer acercamiento para tratar de estimar la variable PUNT_GLOBAL fue utilizar técnicas de regresión, inicialmente se utilizó un mecanismo de selección de características para identificar las variables más influyentes en el puntaje global.

Selección de Características.

Con los datos debidamente transformados y con el propósito de aplicar modelos de regresión, se hizo un análisis para determinar la influencia de variables independientes sobre el puntaje global. Para ello, se utilizó la función mutual_info_regression, que genera el nivel de relevancia entre cada característica y la variable objetivo, sin necesidad de que exista una relación lineal entre ellos (Scikit-learn, s. f.). Este análisis facilitó la identificación de los atributos con mayor poder explicativo, sirviendo como base para una selección de variables más enfocada y eficiente en el desarrollo de los modelos de regresión. La **Figura 15** presenta el resultado de dicho análisis para las características codificadas con LabelEncoder y la **Figura 16** muestra el caso con OneHotEncoder, en ambos casos se están considerando únicamente las 10 variables más relevantes.

Figura 15

Variables más Importantes Mutual_Info_Regression-Labelencoder

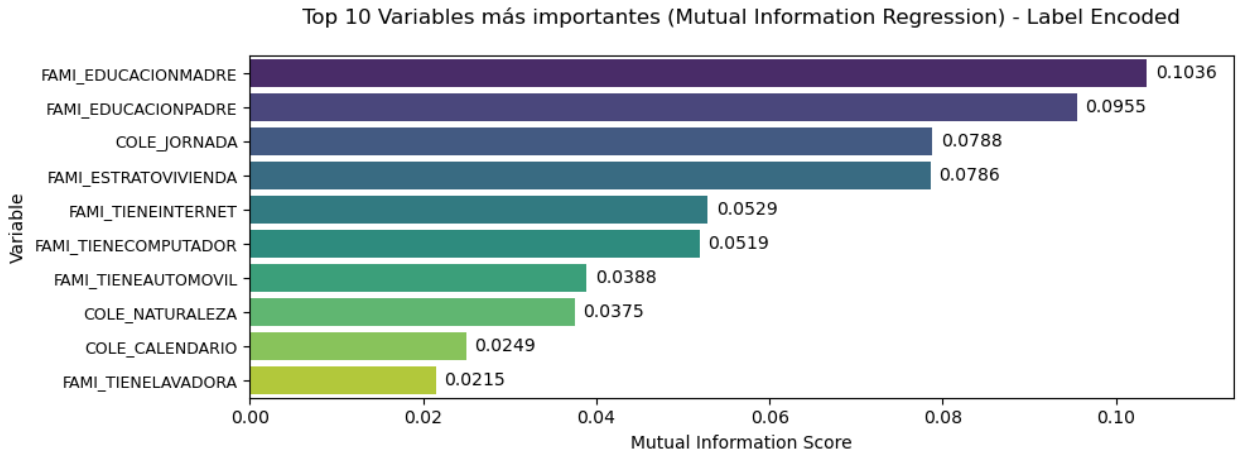
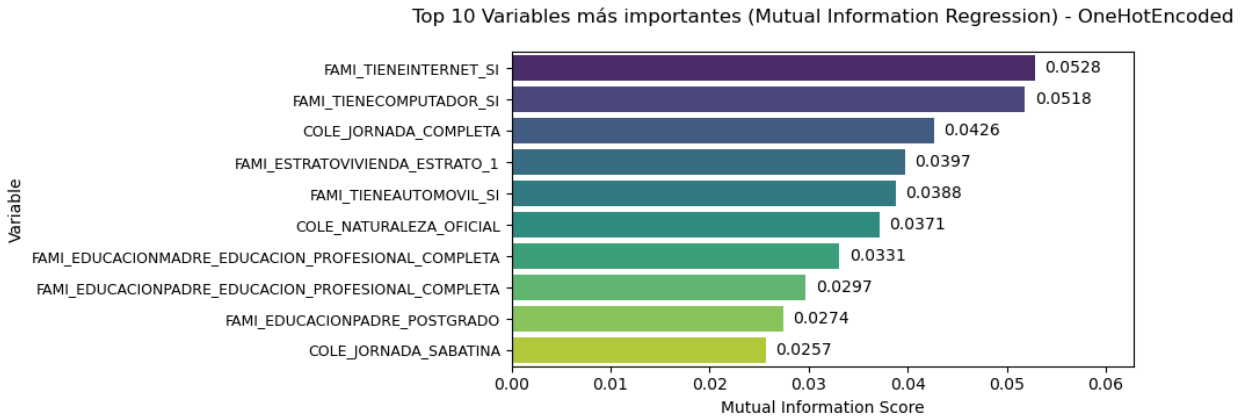


Figura 16

Variables más Importantes Mutual_Info_Regression-Onehotencoder



En los resultados del análisis se observa una diferencia notable. Con LabelEncoder, las variables más relevantes son el nivel educativo de los padres, la jornada escolar, el estrato y el acceso a internet. En cambio, al analizar las características codificadas con OneHotEncoder, las variables más influyentes incluyen el acceso a internet, la tenencia de computador, la jornada

completa, el estrato, la posesión de automóvil y la naturaleza del colegio. Hasta este punto, todas las variables parecen estar relacionadas con las capacidades económicas. Posteriormente, la educación profesional completa tanto en el padre como en la madre aparece como una variable significativa, lo que podría indicar una asociación con familias de mejores condiciones socioeconómicas.

El análisis de los resultados evidencia los cambios en los puntajes de acuerdo con el tipo de codificación, en LabelEncoder las dos primeras variables (educación de madre y padre) tienen un puntaje cercano a 0.1, mientras que al pasar a OneHotEncoder, las dos primeras variables (acceso a internet y tenencia de computador) tienen puntajes que inician en 0.05 que es alrededor de la mitad de lo que se genera en LabelEncoder. Es un comportamiento esperado en términos de puntajes, ya que OneHotEncoder aumenta la cantidad de variables, sin embargo, al tener mayor desglose podría pensarse que el puntaje puede ser más válido.

Las características mencionadas anteriormente se utilizarán para la ejecución de todos los modelos de regresión utilizados.

Resultados de Regresión.

Para predecir el puntaje global mediante modelos de regresión, se emplearon los algoritmos LinearRegression, BaggingRegressor, XGBRegressor y RandomForestRegressor. Cada uno de estos modelos fue entrenado utilizando tanto variables codificadas con LabelEncoder como con OneHotEncoder. Es importante señalar que, en cada tipo de codificación, las variables incluidas difieren, ya que su selección se basó en los niveles de importancia obtenidos a través de la función mutual_info_regression.

Para la implementación de los modelos de regresión, ya con los datos preparados, el proceso se llevó a cabo de la siguiente manera:

1. División de los Datos:

En todos los casos, se utilizó una estrategia de división de los datos en dos partes: el 80% de los datos fueron usados para entrenar los modelos y el 20% restante se utilizó para probar el desempeño.

2. Entrenamiento:

Para el entrenamiento, como se mencionó antes se utilizaron varios modelos: LinearRegression, BaggingRegressor, XGBRegressor y RandomForestRegressor. Cada modelo ajusta sus parámetros (coeficientes en el caso de la regresión lineal y árboles en los otros casos) para predecir la variable dependiente en función de las variables independientes.

3. Evaluación del Modelo:

Una vez entrenados los modelos, se evaluó su desempeño con los datos de prueba. Las métricas utilizadas para medir la calidad de los modelos fueron:

- R^2 (coeficiente de determinación): Indica qué tan bien los datos del conjunto de prueba son explicados por el modelo.
- MSE (error cuadrático medio): Mide la media de los errores al cuadrado, proporcionando una idea de qué tan lejos están las predicciones de los valores reales.
- RMSE (raíz del error cuadrático medio): Es la raíz cuadrada del MSE y proporciona la magnitud de los errores en las mismas unidades de la variable dependiente.
- MAE (error absoluto medio): Mide la media de los errores absolutos, lo que refleja la precisión del modelo sin tener en cuenta la dirección del error.

4. Resultados:

Los resultados obtenidos a partir de las métricas de desempeño permitieron evaluar el

rendimiento de los diferentes modelos utilizados y observar los mejores descriptores para el conjunto de datos del presente análisis.

Métricas Consolidadas de Regresión.

La **Tabla 15** muestra la consolidación de métricas obtenidas para cada modelo entrenado, ordenado por R2:

Tabla 15

Métricas para Modelos de Regresión

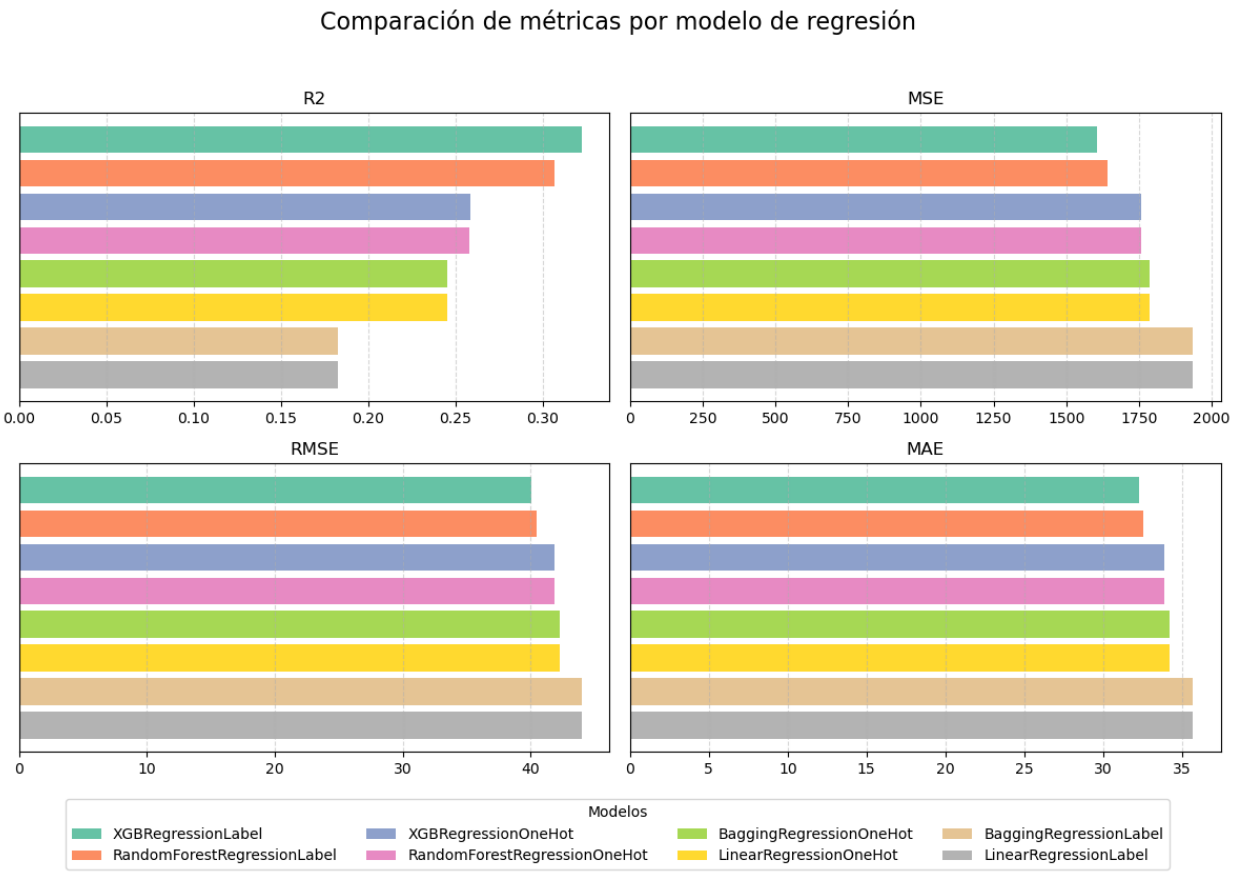
Modelo	R2	MSE	RMSE	MAE
XGBRegressionLabel	0.32222	1604.83571	40.06040	32.28062
RandomForestRegressionLabel	0.30672	1641.53155	40.51582	32.57450
XGBRegressionOneHot	0.25839	1755.98023	41.90442	33.88904
RandomForestRegressionOneHot	0.25821	1756.40756	41.90952	33.89325
BaggingRegressionOneHot	0.24552	1786.45028	42.26642	34.19669
LinearRegressionOneHot	0.24552	1786.45271	42.26645	34.19650
BaggingRegressionLabel	0.18300	1934.49386	43.98288	35.68133
LinearRegressionLabel	0.18300	1934.49518	43.98290	35.68128

Nota. Resultados obtenidos en los modelos de regresión, ordenados por R2.

En la **Figura 17** se presenta de manera visual la comparación de las métricas para cada modelo ejecutado:

Figura 17

Métricas de Regresión



De los resultados obtenidos, se resalta el modelo XGBRegressionLabel que obtuvo el mejor resultado en términos de R^2 (0.322), lo que indica que es el modelo que mejor explica la varianza del puntaje global entre los evaluados. Además, presentó el menor MSE, RMSE y MAE, lo cual refuerza su superioridad en cuanto a precisión y error promedio.

En el caso de XGBoost y Random Forest, los modelos entrenados con variables codificadas mediante LabelEncoder superaron a sus equivalentes con OneHotEncoder en todas las métricas evaluadas. Esto sugiere que, para estos algoritmos en particular, LabelEncoder ofreció una representación más eficiente de las variables categóricas. Sin embargo, este patrón

no se repite en la regresión lineal ni en Bagging, donde el rendimiento fue mejor con OneHotEncoder, lo que indica que el impacto del tipo de codificación depende del algoritmo utilizado.

Los modelos lineales y de bagging codificados con LabelEncoder presentaron los peores valores de R^2 (~0.183) y los mayores errores (MAE y RMSE), lo que indica que no lograron capturar adecuadamente la relación entre las variables y el puntaje global. La diferencia entre el mejor y el peor R^2 es considerable (de 0.322 a 0.183), lo que muestra que la elección del algoritmo y el tipo de codificación tienen un impacto significativo en la calidad de las predicciones.

Con base en los resultados obtenidos, se puede afirmar que, a nivel de regresión, el modelo XGBRegression con codificación LabelEncoder logra explicar aproximadamente el 32% de la varianza del puntaje global. Aunque este valor de R^2 no representa una predicción altamente precisa, sí indica que el modelo captura una parte significativa de la relación entre las variables independientes y la variable objetivo. Esto sugiere que, pese a la complejidad del fenómeno evaluado y la naturaleza categórica de los datos, es posible modelar el desempeño global de los estudiantes con un nivel razonable de ajuste, siendo XGBoost el algoritmo más eficaz dentro del conjunto evaluado.

Clasificación

A nivel de regresión, se identificaron múltiples desafíos para alcanzar un modelo con capacidad predictiva sólida. A pesar de probar diferentes algoritmos y estrategias de codificación, los valores obtenidos de R^2 y los errores asociados (MSE, RMSE y MAE) indicaron un rendimiento limitado en la explicación del puntaje global. Esto sugiere que la relación entre las variables independientes y la variable objetivo no es completamente lineal ni

fácilmente determinada por los modelos de regresión utilizados. Como resultado, se consideró necesario explorar enfoques alternativos, en particular modelos de clasificación, con el objetivo de reformular el problema de predicción en términos de categorías o niveles de desempeño, lo cual podría permitir una mayor capacidad de generalización y una mejor interpretación de los resultados.

A pesar de los avances en el preprocesamiento, los ejercicios de clasificación presentaron varios desafíos. Inicialmente, se definieron cuatro categorías para el puntaje global, establecidas de forma empírica mediante un proceso de prueba y error que incluyó el uso de cuartiles y otros criterios. Con estas categorías se realizaron múltiples iteraciones empleando regresión logística, en un intento por optimizar el rendimiento del modelo. Posteriormente, se incorporaron algoritmos como Random Forest y XGBoost para explorar mejoras en la capacidad predictiva. Sin embargo, los resultados obtenidos no fueron del todo satisfactorios: las clases correspondientes a los extremos —la más baja y la más alta— presentaban métricas de desempeño muy limitadas e, incluso, en algunos casos, la categoría superior no era predicha en absoluto. Si bien algunos ajustes en la estrategia de categorización permitieron cierta mejora en la detección de estas clases, el rendimiento global seguía siendo bajo, con valores de exactitud generalmente entre el 55 % y el 58 %.

En un intento adicional por mejorar el rendimiento de los modelos, se optó por una nueva categorización del puntaje global en tres niveles de desempeño. No obstante, esta estrategia tampoco condujo a mejoras significativas en los resultados. Se realizaron múltiples iteraciones, ajustando de forma reiterada los rangos de puntaje para definir las tres categorías, pero las métricas obtenidas continuaron siendo insatisfactorias.

Finalmente, se decidió simplificar la clasificación a dos categorías: una para identificar los puntajes por debajo del promedio nacional y otra para los puntajes por encima de dicho valor. Este umbral se estableció en 254.1, correspondiente al promedio nacional previamente señalado en el mapa que representa los puntajes por región en Colombia.

Las siguientes secciones muestran los pasos más importantes para la elaboración y pruebas de los modelos de clasificación.

Creación de Categorías para la Etiqueta.

En esta sección se muestra el detalle de las categorías finales con las que se realizaron pruebas de los algoritmos. Como se mencionó anteriormente, se hicieron pruebas con múltiples rangos, sin embargo, para dar cierre a las pruebas se llegó a los valores que se muestran en las siguientes tablas. Para la sección de los resultados de clasificación, se mostrarán unos casos puntuales para las cuatro y tres categorías, sin embargo, el énfasis del análisis fue en la agrupación de dos categorías.

En las siguientes tablas se pueden observar la versión final de las categorías utilizadas en el ejercicio, junto con detalles de la cantidad de registros para cada categoría y los puntajes mínimos y máximos que componen el rango. La **Tabla 16** muestra el detalle para cuatro categorías, la **Tabla 17** expone la situación con tres categorías y finalmente la **Tabla 18** expone el escenario para dos categorías. Nótese que como se mencionó en la sección de preparación de datos, hubo una limpieza de outliers, así que esa es la razón por la que en los tres casos la categoría más baja inicia en el puntaje 159 y la categoría más alta cierra en el puntaje 422.

Tabla 16

Puntaje Global - Cuatro Categorías

Categoria_4_Num	Categoria_4	Cantidad	Mínimo	Máximo
0	Baja	268197	162	200
1	Media	1008901	201	270
2	Alta	254623	271	290
3	Muy Alta	483604	291	424

Nota. Categorización del puntaje global en cuatro niveles.

Tabla 17

Puntaje Global - Tres Categorías

Categoria_3_Num	Categoria_3	Cantidad	Mínimo	Máximo
0	Baja	672965	162	230
1	Media	673139	231	275
2	Alta	669221	276	424

Nota. Categorización del puntaje global en tres niveles.

Tabla 18

Puntaje Global - Dos Categorías

Categoria_2_Num	Categoria_2	Cantidad	Mínimo	Máximo
0	Debajo	1027602	162	253
1	Encima	987723	254	424

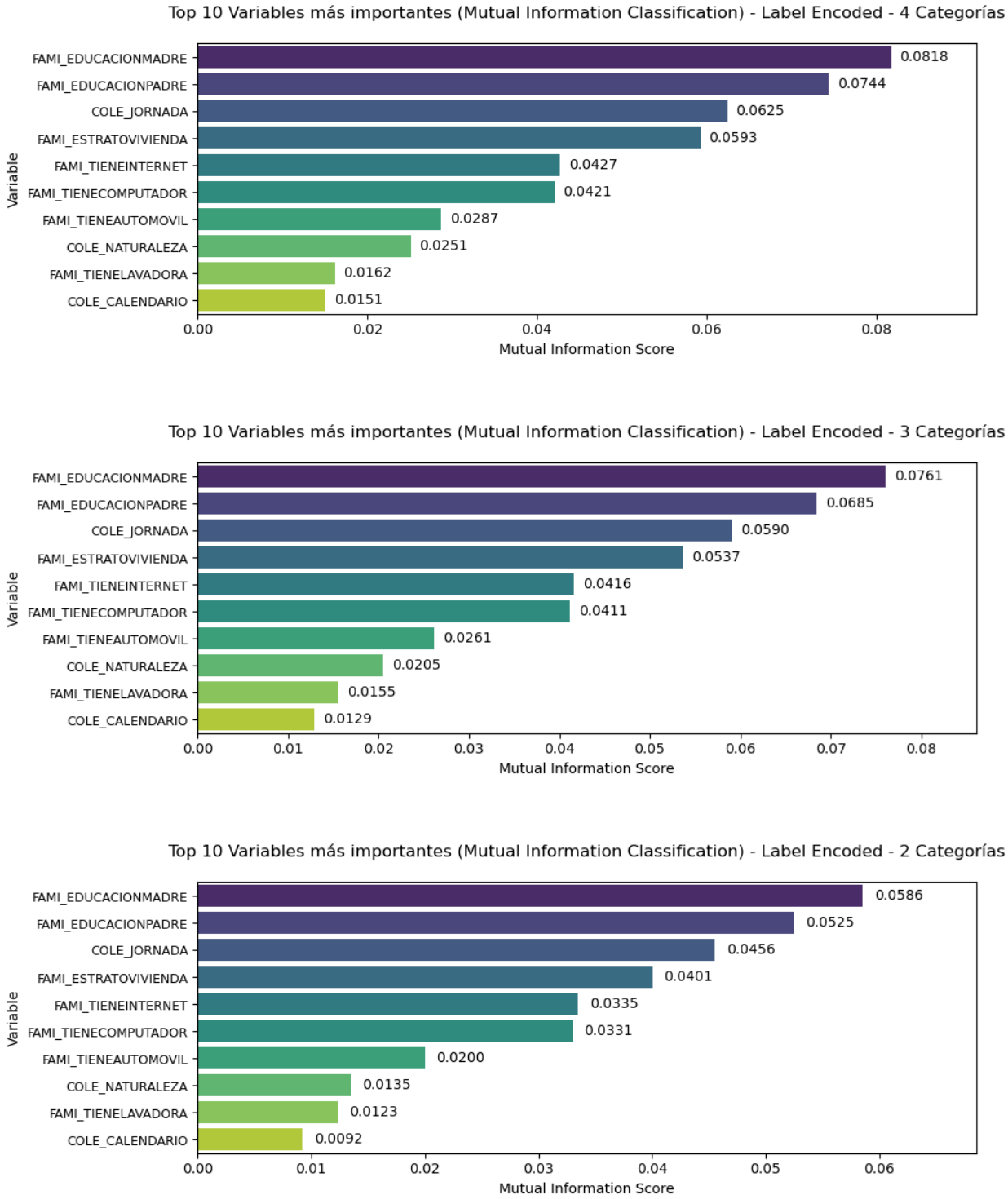
Nota. Categorización del puntaje global en dos niveles: por encima y por debajo del promedio nacional.

Selección de Características.

Una vez definidas las categorías del puntaje global, se realizó un análisis para identificar la influencia de las variables independientes, siguiendo un enfoque similar al aplicado en los modelos de regresión. Dado que se trata de un problema de clasificación, se empleó la función mutual_info_classif, la cual permitió determinar las variables más relevantes. Estas variables fueron luego utilizadas en los modelos de clasificación. A continuación, se presentan los resultados del análisis utilizando codificación con LabelEncoder y OneHotEncoder, mostrando únicamente las 10 variables más importantes en cada caso. Además, dado que se generaron tres versiones categorizadas del puntaje (con cuatro, tres y dos niveles), se incluye el análisis correspondiente para cada una. La **Figura 18** presenta el detalle de variables codificadas con LabelEncoder para cuatro, tres y dos categorías.

Figura 18

Variables más Importantes Mutual_Info_Classif-Labelencoder

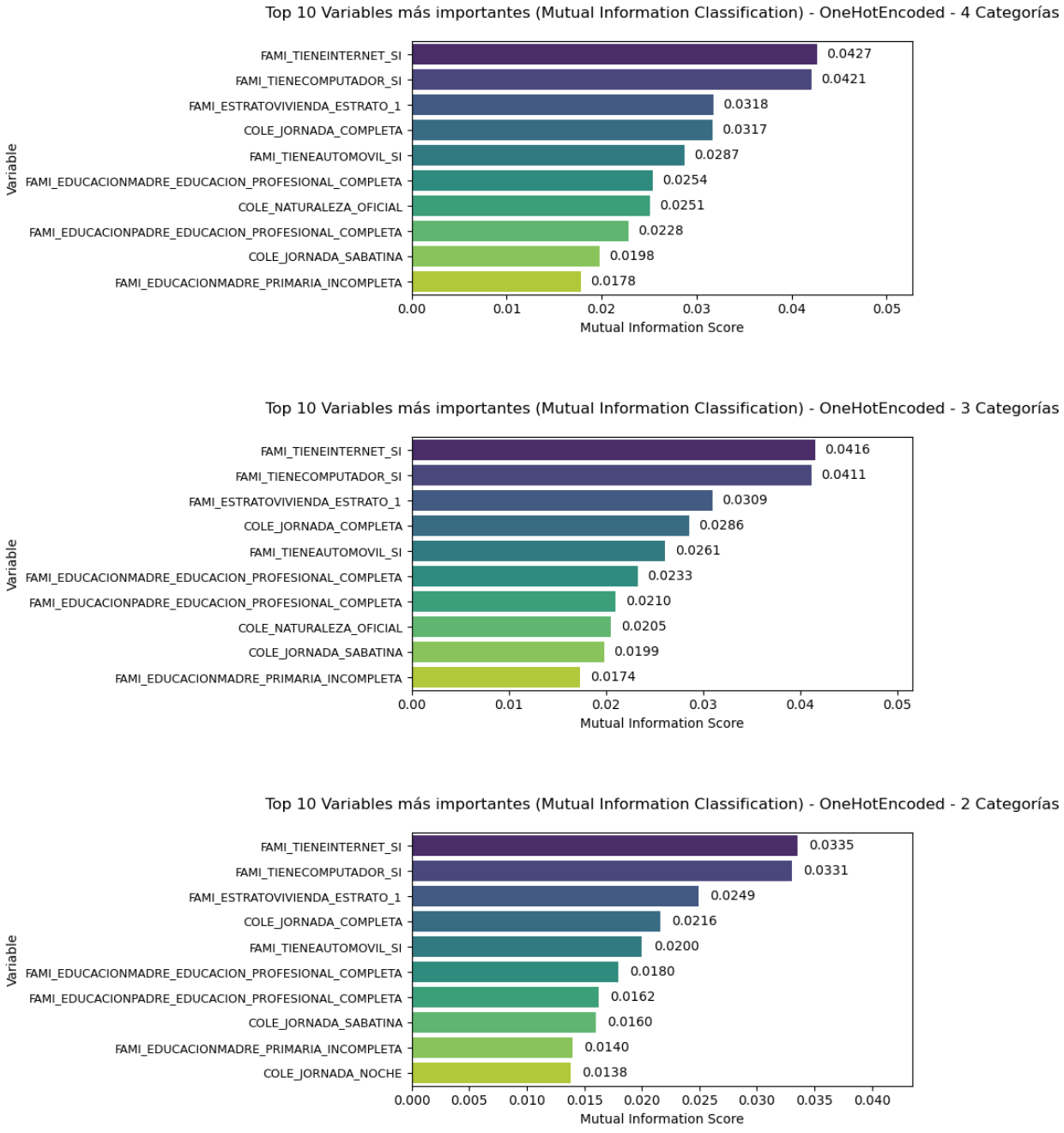


En el caso del análisis de características con LabelEncoder como la cantidad total de variables es 11 y se pretendía escoger 10, no se esperaba una diferencia muy notoria entre los pesos de las variables para cada categorización. Los pesos por variable cambian un poco entre las diferentes categorizaciones, pero el orden de influencia se mantiene en todos los casos. De igual manera, la variable que quedó por fuera fue el género del colegio, lo que indica que esta variable no aporta información suficiente a la determinación del nivel del rendimiento académico (categoría para el puntaje global).

A continuación, la **Figura 18** presenta el resultado de variables codificadas con OneHotEncoder para cuatro, tres y dos categorías.

Figura 19

Variables más Importantes Mutual_Info_Classif-Onehotencoder



En el caso del análisis de características con OneHotEncoder, el número de características iniciales era 44 y también se requería escoger 10 esto para facilitar la ejecución de los modelos,

reducir capacidad computacional y ejecutar solo con lo necesario, las demás variables representan ruido y un sobre esfuerzo en los modelos. En este caso, se nota que para los tres escenarios las 10 variables más importantes son las mismas, pero el orden de importancia varía un poco.

Ya con las salidas de estos análisis se realizaron los ajustes respectivos en el conjunto de datos de entradas para la aplicación de los modelos de clasificación.

Resultados de Clasificación.

Inicialmente se implementaron los modelos LogisticRegression, XGBClassifier, RandomForestClassifier y KNeighborsClassifier para predecir el puntaje global en cuatro categorías. Se realizaron diversas pruebas ajustando los rangos de clasificación, pero las métricas obtenidas no fueron satisfactorias. Posteriormente, se aplicó un enfoque similar utilizando tres categorías, lo cual mostró mejoras leves en el desempeño. Ante estos resultados, y con el objetivo de obtener conclusiones más claras, se optó por simplificar el problema a una clasificación binaria, distinguiendo entre estudiantes que se encuentran Encima o Debajo del promedio nacional del puntaje global, establecido en 254.1 puntos.

Para la implementación de los modelos de clasificación, ya con los datos preparados y con la variable de salida categorizada, el proceso se llevó a cabo de la siguiente manera:

1. División de los Datos: En todos los casos, se utilizó una estrategia de división de los datos en dos partes: el 80% de los datos fueron usados para entrenar los modelos y el 20% restante se utilizó para probar el desempeño.

2. Entrenamiento: Como se mencionó antes, se utilizaron varios modelos: LogisticRegression, RandomForestClassifier, XGBClassifier, KNeighborsClassifier y DecisionTreeClassifier. Cada uno de estos modelos fue ajustado para aprender a predecir la

variable categórica objetivo a partir de las variables independientes seleccionadas, aplicando sus respectivos algoritmos de aprendizaje.

3. Evaluación del Modelo: Una vez entrenados, los modelos fueron evaluados utilizando el conjunto de prueba. Para medir su rendimiento en tareas de clasificación, se utilizaron las siguientes métricas:

- Accuracy: Proporción de predicciones correctas respecto al total de casos.
- Precision (ponderada): Mide la exactitud de las predicciones positivas, considerando el balance entre clases.
- Recall (ponderado): Mide la capacidad del modelo para identificar correctamente todas las instancias de cada clase.
- F1 Score (ponderado): Promedio armónico entre precisión y recall, útil cuando existe un desbalance entre clases.
- Matriz de confusión: Permite visualizar los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada clase.

4. Resultados: Los resultados obtenidos a partir de las métricas de desempeño permitieron evaluar el rendimiento de los diferentes modelos utilizados y observar los mejores descriptores para el conjunto de datos del presente análisis. Más adelante se presentará un resumen de los resultados de todos los modelos. Sin embargo, se muestra primero el análisis de regresión logística, luego, se muestra el caso de KNN que tuvo un tiempo de ejecución muy alto respecto al resto de casos, su comportamiento se salió de lo considerado normal ya que la ejecución tardó alrededor de una hora versus los de más casos que segundos o minutos. Posteriormente, se mostrarán los árboles de decisión generados para los modelos de este tipo de algoritmo.

Caso Inicial Regresión Logística.

A continuación, la **Tabla 19** presenta a modo de ejemplo, los resultados obtenidos al aplicar el modelo de regresión logística para clasificar el puntaje global en cuatro, tres y dos categorías. Si bien se realizaron pruebas similares con otros algoritmos de clasificación, se optó por mostrar únicamente los resultados de la regresión logística ya que los hallazgos fueron similares en los demás tipos de algoritmos. Los resultados evidenciaron que, en este conjunto de datos, reducir el número de categorías contribuye a simplificar el problema y mejora el desempeño predictivo del modelo.

Tabla 19

Evaluación de Regresión Logística: Número de Categorías vs. Desempeño

Modelo	Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)
LogisticRegressionOneHot_C2	0.66521	0.66766	0.66521	0.66455
LogisticRegressionLabel_C2	0.63109	0.63105	0.63109	0.63106
LogisticRegressionOneHot_C3	0.48259	0.47970	0.48259	0.48011
LogisticRegressionLabel_C3	0.46660	0.45580	0.46660	0.45640
LogisticRegressionOneHot_C4	0.31260	0.44996	0.31260	0.28844
LogisticRegressionLabel_C4	0.30890	0.43878	0.30890	0.29544

Nota. Impacto que tienen las técnicas de codificación y la cantidad de clases objetivo en el desempeño de la regresión logística.

A partir del análisis realizado, se observó que los modelos presentan un mejor rendimiento cuando la variable objetivo se define como binaria. Esto se demuestra al comparar los resultados, donde la regresión logística con dos categorías y codificación OneHotEncoder obtuvo el accuracy más alto (0.665), seguida por la misma configuración pero con LabelEncoder (0.631). Sin embargo, al aumentar a tres categorías, el accuracy disminuye aproximadamente

0.18 puntos, reduciendo significativamente la capacidad predictiva de los modelos. Esta tendencia se acentúa al utilizar cuatro categorías, donde el accuracy cae a menos de la mitad del obtenido en el caso binario.

Por ello, se optó por continuar el estudio utilizando la versión binaria de la variable objetivo. Aunque OneHotEncoder mostró métricas ligeramente superiores en la regresión logística, se decidió trabajar con ambos métodos de codificación en distintos algoritmos para analizar su impacto en el desempeño de los modelos. Esta decisión se basó en la posibilidad de que existieran variaciones no observadas inicialmente, como diferencias en el tiempo de procesamiento, la interpretabilidad de los resultados o el comportamiento en modelos. Así, se garantiza un análisis más robusto y generalizable.

Caso Especial KNN.

Durante la fase de pruebas surgieron nuevos desafíos relacionados con el rendimiento computacional de los modelos. A pesar de haber limitado el análisis a las 10 variables más relevantes, se observó que algunos algoritmos, como K-Nearest Neighbors (KNN), presentan una alta sensibilidad al tipo de codificación empleada. Específicamente, la implementación de OneHotEncoder generó un aumento considerable en el tiempo de procesamiento, alcanzando cerca de 59 minutos, en comparación con los aproximadamente 5 minutos requeridos al utilizar LabelEncoder sobre el mismo conjunto de datos. La **Figura 20** y la **Figura 21** ilustran esta diferencia en los tiempos de ejecución según el tipo de codificación utilizada.

Figura 20

Tiempo de Ejecución KNN con LabelEncoder

```
[47]  5 min
#KNN, con LabelEncoder, 5 mins,  3 neighbors, accuracy 0.63

#Si las relaciones son no lineales, la regresión logística tal vez no es la mejor opción
#KNN PUNT_GLOBAL_CATEGORIA_2_NUM

#Predecir PUNT_GLOBAL_CATEGORIA_2_NUM a partir de las siguientes variables

#Definir datos de entrenamiento y pruebas para el modelo
X = X_LabelEncoded[top_features_label_cl_2['Feature']].copy()
y = data_scope['PUNT_GLOBAL_CATEGORIA_2_NUM']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#Entrenar el modelo
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

#Predecir
y_pred = knn.predict(X_test)

#Métricas del modelo
print("\n:::::::CLASSIFICATION REPORT FOR KNN LABEL ENCODED 2 CATEGORIES:::::::\n")
print(classification_report(y_test, y_pred))

# Guardar métricas generales en variables
accuracy_knn_label_cl_2 = accuracy_score(y_test, y_pred)
precision_knn_label_cl_2 = precision_score(y_test, y_pred, average='weighted', zero_division=0)
recall_knn_label_cl_2 = recall_score(y_test, y_pred, average='weighted')
f1_knn_label_cl_2 = f1_score(y_test, y_pred, average='weighted')

# Imprimir detalles
print("\n:::::::METRICS FOR KNN LABEL ENCODED 2 CATEGORIES::::::\n")
print(f"Accuracy: {accuracy_knn_label_cl_2:.3f}")
print(f"Precision (weighted): {precision_knn_label_cl_2:.3f}")
print(f"Recall (weighted): {recall_knn_label_cl_2:.3f}")
print(f"F1 Score (weighted): {f1_knn_label_cl_2:.3f}")

# Matriz de confusión
plt.figure(figsize=(8,6))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt="d", cmap="Blues")
plt.xlabel("Predicción")
plt.ylabel("Verdadero")
plt.title("\nMatriz de Confusión - KNN - LabelEncoded - C2\n")
plt.show()
```

Figura 21

Tiempo de Ejecución KNN con OneHotEncoder

```
59 min

#KNN, con OneHotEncoder,

#Si las relaciones son no lineales, la regresión logística tal vez no es la mejor opción
#XGB PUNT_GLOBAL_CATEGORIA_2_NUM

#Predecir PUNT_GLOBAL_CATEGORIA_2_NUM a partir de las siguientes variables

#Definir datos de entrenamiento y pruebas para el modelo
X = X_OneHotEncoded[top_features_one_hot_cl_2['Feature']].copy()
y = data_scope['PUNT_GLOBAL_CATEGORIA_2_NUM']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#Entrenar el modelo
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

#Predecir
y_pred = knn.predict(X_test)

#Métricas del modelo
print("\n:::::::CLASSIFICATION REPORT FOR ONE HOT LABEL ENCODED 2 CATEGORIES:::::::\n")
print(classification_report(y_test, y_pred))

# Guardar métricas generales en variables
accuracy_knn_one_hot_cl_2 = accuracy_score(y_test, y_pred)
precision_knn_one_hot_cl_2 = precision_score(y_test, y_pred, average='weighted', zero_division=0)
recall_knn_one_hot_cl_2 = recall_score(y_test, y_pred, average='weighted')
f1_knn_one_hot_cl_2 = f1_score(y_test, y_pred, average='weighted')

# Imprimir detalles
print("\n:::::::METRICS FOR KNN ONE HOT ENCODED 2 CATEGORIES:::::::\n")
print(f"Accuracy: {accuracy_knn_one_hot_cl_2:.3f}")
print(f"Precision (weighted): {precision_knn_one_hot_cl_2:.3f}")
print(f"Recall (weighted): {recall_knn_one_hot_cl_2:.3f}")
print(f"F1 Score (weighted): {f1_knn_one_hot_cl_2:.3f}")

# Matriz de confusión
plt.figure(figsize=(8,6))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt="d", cmap="Blues")
plt.xlabel("Predicción")
plt.ylabel("Verdadero")
plt.title("\nMatriz de Confusión - KNN - OneHotEncoded - C2\n")
plt.show()
```

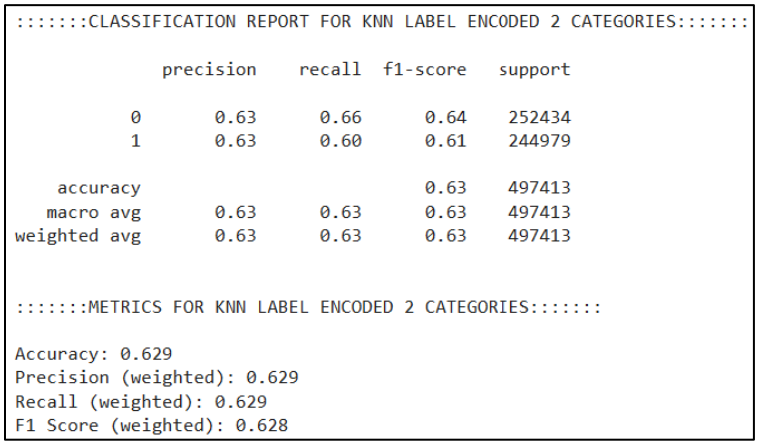
Aunque se esperaba que la codificación con OneHotEncoder ofreciera un mejor rendimiento del modelo debido a su representación más detallada de las variables categóricas, los resultados obtenidos no justificaron el alto costo computacional. A pesar de requerir casi 59 minutos de procesamiento, en contraste con los 5 minutos necesarios al emplear LabelEncoder, las métricas de desempeño fueron incluso ligeramente inferiores.

La comparación de los reportes de clasificación, que se presentan en la **Figura 22** y **Figura 23**, muestra que ambos enfoques produjeron resultados prácticamente equivalentes. Para

el caso de LabelEncoder el accuracy fue de 0.63, mientras que para OneHotEncoder fue 0.62, así mismos el f1-score para ambas clares es igual, 0.64 para la clase 0 y 0.61 para la clase 1. Esto sugiere que, para este conjunto de datos y modelo, el uso de OneHotEncoder no representa una ventaja significativa y solo introduce una expansión innecesaria del espacio de características, con un impacto negativo en la eficiencia del procesamiento.

Figura 22

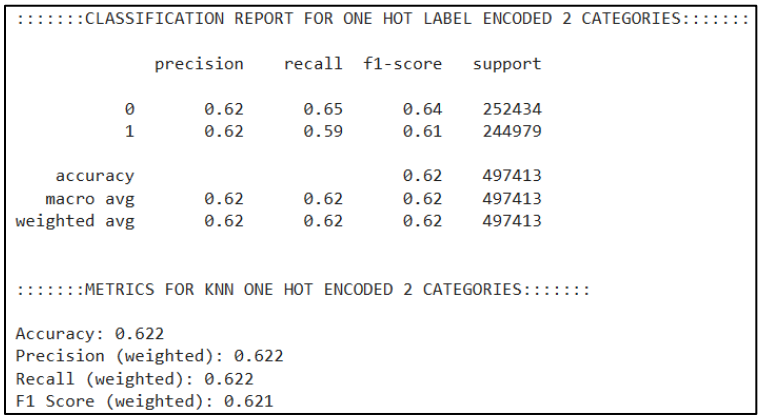
Métricas KNN LabelEncoder



```
:::::::CLASSIFICATION REPORT FOR KNN LABEL ENCODED 2 CATEGORIES:::::::

              precision    recall  f1-score   support

           0       0.63      0.66      0.64    252434
           1       0.63      0.60      0.61    244979

    accuracy                           0.63    497413
   macro avg       0.63      0.63      0.63    497413
weighted avg       0.63      0.63      0.63    497413


:::::::METRICS FOR KNN LABEL ENCODED 2 CATEGORIES:::::::

Accuracy: 0.629
Precision (weighted): 0.629
Recall (weighted): 0.629
F1 Score (weighted): 0.628
```

Figura 23

Métricas KNN OneHotEncoder



```
:::::::CLASSIFICATION REPORT FOR ONE HOT LABEL ENCODED 2 CATEGORIES:::::::

              precision    recall  f1-score   support

           0       0.62      0.65      0.64    252434
           1       0.62      0.59      0.61    244979

    accuracy                           0.62    497413
   macro avg       0.62      0.62      0.62    497413
weighted avg       0.62      0.62      0.62    497413


:::::::METRICS FOR KNN ONE HOT ENCODED 2 CATEGORIES:::::::

Accuracy: 0.622
Precision (weighted): 0.622
Recall (weighted): 0.622
F1 Score (weighted): 0.621
```

Gráficas para Árboles de Decisión.

En el caso de los árboles de decisión, se mantuvo la misma estrategia que con los otros algoritmos: se utilizaron solo 10 variables de entrada, tanto con LabelEncoder como con OneHotEncoder. En la **Figura 24** y **Figura 25** se presentan los árboles de decisión obtenidos en cada caso, que por facilidades de visualización se generaron solo hasta el cuarto nivel.

Figura 24

Árbol de Decisión LabelEncoder

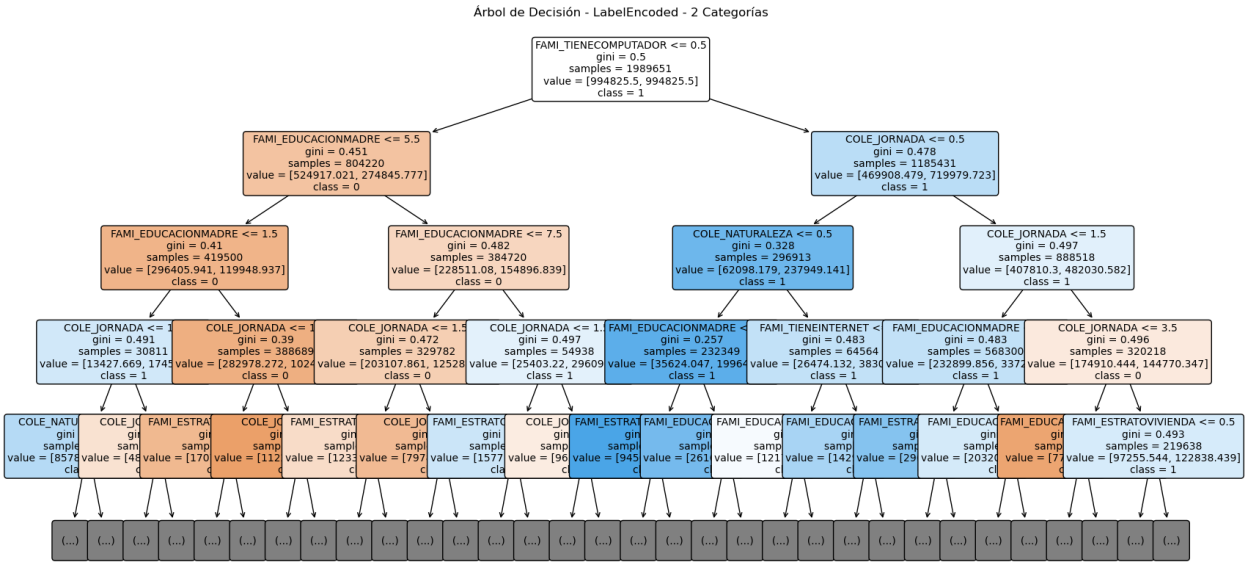
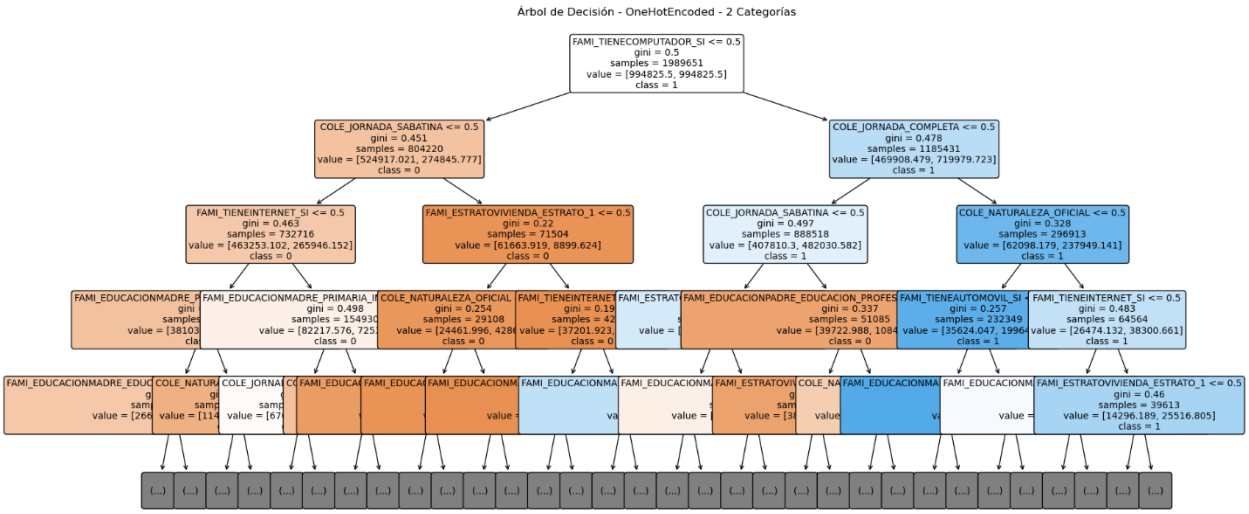


Figura 25

Árbol de Decisión OneHotEncoder



Al analizar los árboles, se observa que ambos comienzan con la misma variable, FAMI_TIENE_COMPUTADOR para LabelEncoder y FAMI_TIENE_COMPUTADOR_SI para OneHotEncoder.

Sin embargo, al revisar el segundo nivel de ambos árboles, se evidencian las diferencias derivadas de las codificaciones utilizadas. En el caso de LabelEncoder, la educación de la madre adquiere relevancia desde este nivel. En cambio, para OneHotEncoder, la educación de la madre solo aparece a partir del cuarto nivel. En este último caso, las tres jornadas escolares del colegio parecen aportar más información al modelo.

Es importante mencionar que en el caso del árbol con LabelEncoder se puede ver que hay nodos como “FAMI_EDUCACIONMADRE” <= 5.5”, “FAMI_EDUCACIONMADRE” <= 1.5” o FAMI_EDUCACIONMADRE” <= 7.5”, lo cual hace evidente el hecho de que se trata a la variable como ordinal. Para asegurar una mejor captura de la información, se podrían hacer codificaciones personalizadas para asegurar que se tenga el orden correcto. De igual manera,

para variables que no son ordinales como es el caso de COLE_JORNADA, FAMI_TIENE_COMPUTADOR se podría tener una codificación con OneHotEncoder. Se podría hacer una mezcla de ambas codificaciones y validar si el modelo genera mejores patrones de aprendizaje.

En la siguiente sección se muestran los resultados consolidados para todos los modelos de clasificación entrenados.

Métricas Consolidadas de Clasificación.

La **Tabla 20** muestra la consolidación de métricas obtenidas para cada modelo entrenado para la clasificación de dos categorías de puntaje global por encima o por debajo del promedio nacional, ordenado por Accuracy:

Tabla 20

Métricas para Modelos de Clasificación

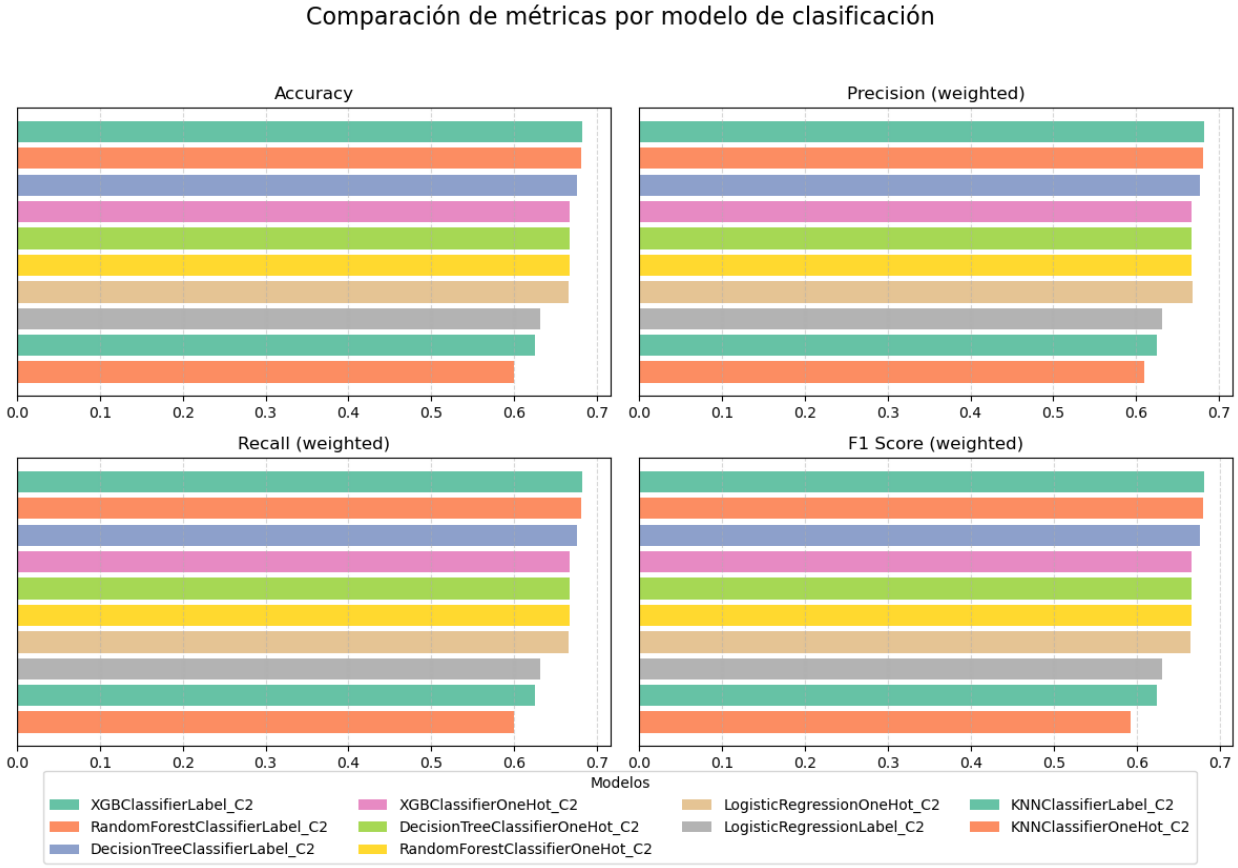
Modelo	Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)
XGBClassifierLabel_C2	0.68224	0.68262	0.68224	0.68179
RandomForestClassifierLabel_C2	0.68030	0.68027	0.68030	0.68026
DecisionTreeClassifierLabel_C2	0.67622	0.67653	0.67622	0.67623
XGBClassifierOneHot_C2	0.66683	0.66738	0.66683	0.66616
DecisionTreeClassifierOneHot_C2	0.66681	0.66679	0.66681	0.66668
RandomForestClassifierOneHot_C2	0.66681	0.66679	0.66681	0.66668
LogisticRegressionOneHot_C2	0.66521	0.66766	0.66521	0.66455
LogisticRegressionLabel_C2	0.63109	0.63105	0.63109	0.63106
KNNClassifierLabel_C2	0.62480	0.62481	0.62480	0.62440
KNNClassifierOneHot_C2	0.59961	0.60970	0.59961	0.59270

Nota. Resultados obtenidos en los modelos de clasificación, ordenados por Accuracy.

También se presentan los resultados en la **Figura 26** para una mejor interpretación visual, posteriormente se mencionan las observaciones más relevantes de esta sección:

Figura 26

Métricas de Clasificación



El análisis comparativo de los modelos de clasificación utilizados para predecir el puntaje global categorizado en dos niveles (por encima y por debajo del promedio de 254.1) revela diferencias importantes en el desempeño según el tipo de codificación de variables y el algoritmo empleado. La metodología incluyó la selección de diez variables por conjunto a partir del criterio de información mutua y se evaluaron distintos modelos bajo esquemas de codificación LabelEncoder y OneHotEncoder.

De manera consistente, los modelos construidos con variables codificadas mediante LabelEncoder superan a sus equivalentes entrenados con OneHotEncoder en todas las métricas evaluadas. Esto se evidencia especialmente en los algoritmos más robustos como XGBoost y Random Forest, donde las versiones con LabelEncoder muestran mejores niveles de exactitud y equilibrio entre precisión y sensibilidad. El modelo XGBClassifierLabel_C2 se destaca como el mejor clasificador, alcanzando una precisión del 68.26%, un valor idéntico para el recall, y un F1 score de 68.18%. Estas métricas no solo indican un buen rendimiento general sino también una adecuada capacidad para balancear los errores de clasificación entre ambas clases, lo cual es relevante dado que se trabaja con una variable objetivo binaria construida a partir del promedio nacional.

Muy cerca en desempeño se encuentra el Random Forest con LabelEncoder, con una precisión y recall apenas inferiores, ambos en torno al 68.03%, y un F1 score de 68.02%. El algoritmo de árboles de decisión también muestra un buen comportamiento con LabelEncoder, aunque ligeramente por debajo de los dos anteriores. En contraste, cuando los mismos algoritmos se entrenan con variables codificadas mediante OneHotEncoder, sus métricas tienden a disminuir de manera leve pero sistemática. Tanto XGBoost como Decision Tree y Random Forest con OneHotEncoder presentan valores muy similares entre sí, todos alrededor del 66.68% de exactitud y F1, lo cual refleja una pérdida de rendimiento frente al uso de codificación por etiquetas, posiblemente debido a la alta dimensionalidad generada por este tipo de transformación.

Otros algoritmos como la regresión logística y K-Nearest Neighbors muestran resultados más modestos, especialmente en sus versiones con LabelEncoder. La regresión logística con codificación por etiquetas apenas alcanza un 63.11% de precisión, mientras que KNN muestra

los peores resultados, particularmente con OneHotEncoder, donde la precisión cae hasta el 59.96% y el F1 score se reduce aún más, evidenciando una pobre capacidad de generalización para este problema.

Estos hallazgos complementan el análisis realizado mediante regresión, donde se había determinado que el modelo XGBRegressor con LabelEncoder lograba explicar el 32% de la varianza del puntaje global continuo, lo cual, aunque modesto, ya anticipaba que este tipo de algoritmo podría adaptarse bien al patrón de datos. En conjunto, tanto en tareas de regresión como de clasificación, XGBoost con variables codificadas por etiquetas se posiciona como la mejor alternativa para predecir el desempeño académico global, ofreciendo el mejor balance entre exactitud, consistencia y capacidad explicativa en el contexto del conjunto de datos y las transformaciones aplicadas. Esto sugiere que la estructura no lineal del modelo y su capacidad para manejar interacciones complejas entre variables categóricas lo convierten en la opción más adecuada para abordar este problema predictivo.

Conclusiones

El análisis exploratorio, estadísticas descriptivas, junto con técnicas de correlación y visualización de datos, permitieron identificar patrones que muestran diferencias en los resultados de las pruebas Saber 11. Se evidenció que las características socioeconómicas, familiares y del colegio condicionan el desempeño de los estudiantes en la prueba, lo que motiva a cuestionar las desigualdades del sistema educativo de nuestro país. A continuación, se detallan los hallazgos más relevantes.

El análisis de los puntajes y las condiciones de los colegios reveló algunas relaciones significativas. En primer lugar, se identificó una diferencia importante según la naturaleza del colegio: los colegios privados obtienen en promedio 274.88 puntos, mientras que los públicos alcanzan un promedio de 249.03 puntos, una brecha de casi 26 puntos. Esta diferencia sugiere que los recursos institucionales y las condiciones socioeconómicas influyen considerablemente en los resultados académicos, dado que los estudiantes de sectores con menos recursos suelen asistir con mayor frecuencia a colegios públicos, donde enfrentan barreras como el acceso limitado a materiales educativos, dificultades de conectividad, entornos poco favorables para el estudio y mayores responsabilidades familiares o laborales. Estas condiciones pueden afectar su desempeño académico y limitar sus oportunidades de aprendizaje y preparación para las pruebas.

Por otro lado, también se encontraron diferencias marcadas según el tipo de calendario escolar. Los colegios con calendario B (principalmente privados) presentan los puntajes más altos, con un promedio de 314.41 puntos, lo que representa una diferencia de más de 60 puntos frente a los colegios con calendario A con un promedio de 254.38 puntos. Los colegios con “Otro” tipo de calendario, que suelen corresponder a modalidades de educación no convencional

o acelerada, muestran el desempeño más bajo, con un promedio de 232.66 puntos, lo que refuerza la idea de que las condiciones institucionales impactan directamente en el rendimiento.

Siguiendo con las características internas de los colegios, se observaron diferencias por género y jornada. Los colegios masculinos registraron el promedio más alto (305.91 puntos), seguidos por los femeninos (292.35 puntos), mientras que los mixtos tuvieron el promedio más bajo (254.51 puntos). En cuanto a la jornada escolar, los mejores resultados se concentraron en las jornadas completa (283.46 puntos) y única (255.65 puntos), en contraste con los promedios más bajos observados en las jornadas nocturna (219.04) y sabatina (218.28). Estos datos refuerzan la relación entre las condiciones de oferta educativa y las oportunidades de desempeño académico, destacando la necesidad de políticas que aborden las brechas estructurales en el sistema educativo.

En cuanto a los puntajes, la mayoría de los estudiantes tienen alrededor de 50 puntos en cada área, con lectura crítica un poco por encima de las demás con un promedio de 52.18, mientras que las demás áreas muestran promedios más bajos, ubicándose entre 49.55 y 50.85. Sociales y ciudadanas tiene el promedio más bajo que es 48.97, lo que sugiere que esta área podría requerir más apoyo. Inglés, aunque no es el área con el promedio más bajo, muestra una dispersión considerable de 12.35 puntos, lo que podría indicar diferencias en el acceso a la enseñanza del idioma. Los puntajes se mueven en un rango entre 0 y 100 puntos para todas las áreas. La variabilidad de los puntajes es moderada, con una desviación estándar de 11.36 puntos en promedio lo que indica que la mayoría de los estudiantes se concentra en un rango de puntajes intermedio sin extremos pronunciados. Una proporción considerable de estudiantes, cercana al 50%, se encuentra entre 40 y 60 puntos que es la franja media, esto sugiere que la mayoría de los estudiantes logran adquirir conocimientos básicos, pero se evidencia una capacidad reducida para

desarrollar desempeños excepcionales, sugiriendo posibles falencias en estrategias de profundización curricular, currículos académicos y calidad en general.

Se evidencian marcadas desigualdades en el desempeño académico de los estudiantes en Colombia a nivel de departamento; por ejemplo, mientras Bogotá (visto no propiamente como departamento si no distrito especial) presenta un puntaje promedio de 273.83 que es el más alto. Por su parte, departamentos como Chocó y Vaupés reflejan brechas significativas, con promedios de 212.48 y 217.90 cada uno, estos son indicios de que requieren atención diferenciada en las políticas educativas.

La alta concentración de pruebas en Bogotá (16.34%), Antioquia (13.18%) y Valle del Cauca (8.55%) sugiere que estos territorios no solo lideran en volumen, sino también ejercen mayor influencia en las estadísticas nacionales, lo que resalta la importancia de fortalecer la equidad en el acceso y la calidad educativa en geografías con menor participación. Así como se notó a nivel de departamento, los resultados evidencian marcadas diferencias en el desempeño académico de los estudiantes según la región. Ya se estableció anteriormente que Bogotá (también diferenciada como región) presenta el promedio más alto de 273.83 puntos y una desviación estándar de 49.30**,** con una distribución de puntajes que muestra mayor concentración en niveles altos, reflejada en sus percentiles 25, 50 y 75 de 238, 272 y 307 respectivamente. Esto sugiere condiciones educativas más favorables en comparación con otras regiones del país. En contraste, las regiones Caribe y Amazonía presentan promedios más bajos de 240 y 238.37, con percentiles que se ubican en rangos significativamente inferiores de 205, 235 y 267, lo que podría reflejar limitaciones estructurales en el acceso y la calidad de la educación. Estas diferencias regionales señalan la necesidad urgente de políticas educativas

diferenciadas que respondan a los contextos territoriales, con el objetivo de avanzar hacia una mayor equidad y calidad en el sistema educativo nacional.

Una vez descubiertas las relaciones más importantes y revelados los patrones de afectación de las variables sobre el rendimiento académico, se usaron algoritmos de Machine Learning para generar predicciones del puntaje global. A continuación, se detallan los hallazgos más importantes al respecto.

En particular, el análisis hecho para algoritmos de regresión permitió evaluar el desempeño de distintos modelos para la predicción del puntaje global en las pruebas Saber 11, empleando exclusivamente variables categóricas. Se utilizaron cuatro algoritmos principales: LinearRegression, BaggingRegressor, XGBRegressor y RandomForestRegressor, en combinación con dos enfoques de codificación de variables: Label Encoding y OneHot Encoding.

Los resultados evidenciaron que el modelo XGBRegressor con variables codificadas mediante Label Encoding obtuvo el mejor desempeño general, alcanzando un coeficiente de determinación (R^2) de 0.323. Este valor, aunque no representa una capacidad predictiva alta, indica que el modelo fue capaz de explicar aproximadamente el 32 % de la varianza del puntaje global, superando al resto de los algoritmos evaluados en todas las métricas consideradas (MSE, RMSE y MAE).

Asimismo, se observó que los modelos basados en árboles de decisión (XGBoost y Random Forest) obtuvieron mejores resultados con Label Encoding, mientras que los modelos lineales y de Bagging se desempeñaron de manera más favorable con OneHot Encoding. Este hallazgo sugiere que la elección del tipo de codificación debe considerarse en función del algoritmo empleado, ya que impacta significativamente en la calidad de las predicciones.

Por otra parte, los modelos LinearRegression y BaggingRegressor con Label Encoding presentaron los resultados más bajos en términos de R^2 (~0.178) y los errores más altos, lo que indica una capacidad limitada para capturar la relación entre las variables independientes y el puntaje global.

Para cerrar lo relativo a regresión, los hallazgos de este estudio muestran que, a pesar de trabajar exclusivamente con variables categóricas, es posible alcanzar un nivel aceptable de predicción del puntaje global en las pruebas Saber 11. El modelo XGBRegressor con Label Encoding se destacó como la opción más eficaz (explicando alrededor del 32% de la varianza), lo cual demuestra el potencial de este enfoque para estudios educativos similares. No obstante, la varianza explicada por los modelos sugiere que existen otros factores no considerados en el presente análisis que podrían tener un papel importante en el desempeño académico de los estudiantes.

Con el fin de llegar a mejores predicciones, se continuó con la generación de modelos de clasificación, esta vez transformando la variable continua del puntaje global en una variable categórica binaria que indicara si el estudiante se encontraba por encima o por debajo del promedio nacional (254.1 puntos). Esta nueva formulación permitió evaluar la capacidad de distintos algoritmos para identificar patrones de desempeño académico de forma más robusta. Al igual que en la fase de regresión, se utilizó únicamente información proveniente de variables categóricas, lo que supuso un reto adicional en términos de representación y generalización de los datos. Sin embargo, los resultados obtenidos fueron alentadores. El modelo XGBClassifier con LabelEncoder alcanzó una precisión de 68.26%, un recall del mismo valor y un F1 score de 68.18%, lo cual representa un desempeño significativamente superior frente a otros clasificadores evaluados. Esta consistencia entre precisión y sensibilidad sugiere que el modelo

no solo acierta con frecuencia, sino que también mantiene un adecuado equilibrio en la clasificación de estudiantes en ambas categorías.

A nivel comparativo, modelos como Random Forest con LabelEncoder se aproximaron al rendimiento del XGBClassifier, logrando una precisión de 68.03% y un F1 score de 68.02%, lo que refuerza la superioridad de los algoritmos basados en árboles de decisión para este tipo de problema. En contraste, la versión del XGBClassifier entrenada con OneHotEncoder disminuyó su rendimiento a una precisión de 66.68%, evidenciando que la elección del tipo de codificación tiene un impacto concreto, posiblemente asociado a la mayor dimensionalidad y dispersión que introduce OneHotEncoder en conjuntos de datos con muchas variables categóricas. Algoritmos como Logistic Regression o K-Nearest Neighbors, aunque conceptualmente distintos, mostraron desempeños menos competitivos, el modelo de regresión logística con LabelEncoder se quedó en un 63.11% de precisión y KNN con OneHotEncoder descendió hasta un 59.96%, mostrando además un F1 score inferior al 59%, lo cual refleja su limitada capacidad de adaptación a la estructura del problema.

Estas observaciones no solo confirman los hallazgos obtenidos durante la fase de regresión, donde XGBRegressor con LabelEncoder logró explicar el 32.3% de la varianza del puntaje global, sino que refuerzan la idea de que los modelos de tipo boosting, en combinación con una codificación por etiquetas, son especialmente efectivos en contextos donde las variables categóricas dominan la representación de los datos. En conjunto, tanto para tareas de predicción continua como de clasificación binaria, XGBoost se consolidó como la alternativa más sólida, ofreciendo no solo el mejor desempeño cuantitativo, sino también una notable estabilidad entre diferentes configuraciones. Esto sugiere que su arquitectura, basada en el ensamblaje de árboles optimizados en etapas sucesivas, es capaz de capturar las relaciones complejas y no lineales entre

las características educativas de los estudiantes, lo cual lo convierte en una herramienta valiosa para apoyar la toma de decisiones en el ámbito educativo colombiano.

Recomendaciones

Durante el desarrollo del presente proyecto, enfocado en el análisis de los puntajes de la prueba Saber 11 mediante modelos de regresión y clasificación, se identificaron diversos retos metodológicos y analíticos. Estos desafíos permitieron reflexionar sobre aspectos susceptibles de mejora que podrían optimizar tanto la calidad de los modelos predictivos como la comprensión del fenómeno educativo analizado.

En primer lugar, se sugiere ampliar el conjunto de variables utilizadas. Aunque en este estudio se seleccionaron las diez características más relevantes según medidas de importancia, la inclusión de variables adicionales relacionadas con aspectos socioeconómicos, demográficos o contextuales podría enriquecer los modelos y mejorar su capacidad predictiva. Este tipo de información podría capturar factores latentes que actualmente no están representados en el análisis.

En segundo lugar, se recomienda explorar técnicas de preparación de datos distintas a las empleadas. Si bien se utilizaron métodos tradicionales como Label Encoding y OneHot Encoding, sería pertinente evaluar otras alternativas, como la codificación de frecuencia o por media de la variable objetivo, que podrían representar de manera más eficiente las variables categóricas y evitar el aumento excesivo de dimensionalidad observado con algunas técnicas. Este tipo de codificaciones puede ser especialmente útil en modelos sensibles al volumen de variables, como *k*-NN, donde se evidenció un alto tiempo de procesamiento.

Otra oportunidad de mejora está relacionada con la optimización de los hiperparámetros de los modelos. Aunque se realizaron pruebas preliminares con herramientas como GridSearchCV y RandomizedSearchCV, los resultados obtenidos no variaron significativamente respecto a los valores iniciales y, además, estas técnicas resultaron costosas en términos

computacionales. No obstante, en un escenario con mayor disponibilidad de recursos, la optimización sistemática de los hiperparámetros podría mejorar el ajuste de los modelos y su rendimiento.

Asimismo, es aconsejable considerar modelos más avanzados. A pesar de que algoritmos como XGBoost y Random Forest mostraron un desempeño adecuado, técnicas como redes neuronales profundas o modelos híbridos podrían ser capaces de identificar relaciones no lineales más complejas entre las variables. La exploración de estos enfoques podría ofrecer una mejora significativa en la capacidad de predicción.

La implementación de validación cruzada también es una recomendación clave para garantizar la robustez del análisis. En lugar de utilizar una única partición entre datos de entrenamiento y prueba, la validación cruzada permite estimar el desempeño del modelo de forma más estable, reduciendo la posibilidad de que los resultados estén sesgados por una división aleatoria específica. Esto es especialmente relevante para estudios donde el tamaño de la muestra puede limitar la generalización de los hallazgos.

Otro aspecto relevante es el establecimiento de un proceso de monitoreo y actualización periódica de los modelos. Dado que los resultados de las pruebas Saber 11 se actualizan con el tiempo, es fundamental reentrenar los modelos regularmente e incorporar nuevas observaciones a medida que estén disponibles. Esta práctica no solo mejora la precisión de las predicciones, sino que también garantiza que los modelos sigan siendo pertinentes en contextos educativos cambiantes.

Por último, se destaca la importancia de una documentación exhaustiva del proceso analítico. Registrar detalladamente cada etapa del análisis, desde el preprocesamiento hasta la evaluación final, permite garantizar la replicabilidad del estudio y facilita la comprensión de las

decisiones tomadas. Una documentación clara contribuye además a la transparencia del trabajo y a su utilidad en investigaciones futuras.

En conjunto, al abordar los desafíos y recomendaciones señaladas, el análisis de los puntajes de la prueba Saber 11 podrá beneficiarse de mejoras en la precisión, interpretabilidad y robustez de los modelos predictivos. Continuar experimentando con distintas técnicas y configuraciones optimizará los resultados, permitiendo una mejor comprensión de los factores que influyen en el desempeño de los estudiantes en la prueba.

Referencias Bibliográficas

Barrios Aguirre, F., Forero, D. A., Castellanos Saavedra, M. P., & Mora Malagón, S. Y. (2021). The Impact of Computer and Internet at Home on Academic Results of the Saber 11 National Exam in Colombia. *SAGE Open*, *11*(3). https://doi.org/10.1177/21582440211040810/SUPPL_FILE/SJ-PDF-1-SGO-10.1177_21582440211040810.PDF

Bonilla-Mejía, L., Londoño-Ortega, E., & Henao, M. F. (2024). Geographic isolation and learning: Evidence from rural schools in Colombia. *Economics of Education Review*, *99*, 102522. https://doi.org/10.1016/J.ECONEDUREV.2024.102522

Bravo, L. E. C., López, H. J. F., & Trujilllo, E. R. (2021). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. *Revista Boletín Redipe*, *10*(13), 171-190. https://doi.org/10.36260/RBR.V10I13.1737

Cabra-Hernández, H. W. (2023). Three approaches to modeling the relationship among durable goods, academic achievement, and school attendance in Colombia. *Heliyon*, *9*(12), e22732. https://doi.org/10.1016/J.HELIYON.2023.E22732

Castañeda, J. A. (2023, agosto 8). *TopoJSON of Colombia's departments and towns*. https://github.com/jacasta2/colombian_map/blob/main/from_shapefiles/create_from_shapefile.ipynb

Castro-Ávila, M., & Ruiz-Linares, J. (2019). La educación secundaria y superior en Colombia vista desde las pruebas Saber. *Praxis & Saber*, *10*(24), 341-366. https://doi.org/10.19053/22160159.v10.n25.2019.9465

DANE. (2023). *Geoportal DANE- Página de descarga datos geoestadísticos - Versión MGN2023-Nivel Departamento*. https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/datos-geoestadisticos/

Dangeti, P. (2017). *Statistics for Machine Learning : Build Supervised, Unsupervised, and Reinforcement Learning Models Using Both Python and R*. Packt Publishing.

d'Orville, H. (2020). COVID-19 causes unprecedented educational disruption: Is there a road towards a new normal? *Prospects*, *49*(1-2), 11-15. https://doi.org/10.1007/S11125-020-09475-0/METRICS

Fundación ExE. (2024, julio). *Pruebas saber 11° 2023. Análisis de resultados*. https://prod.fundacionexe.org.co/en/document/pruebas-saber-11-analisis-de-resultados/

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (Inc. O'Reilly Media, Ed.; 2.a ed.). https://books.google.com.co/books?id=HHetDwAAQBAJ&printsec=frontcover&redir_esc=y&hl=es&pli=1#v=onepage&q&f=false

Gómez, N. A. C. (2019). Factores de la calidad educativa desde una perspectiva multidimensional: Análisis en siete regiones de Colombia. *Plumilla Educativa*, *23*(1), 121-139. https://doi.org/10.30554/plumillaedu.1.3350.2019

Hernandez-Leal, E., Duque-Mendez, N. D., & Cechinel, C. (2021). Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions. *Heliyon*, *7*(9), e08017. https://doi.org/10.1016/J.HELIYON.2021.E08017

ICFES. (2022a). *Acerca del examen Saber 11*. http://www.icfes.gov.co/evaluaciones-icfes/acerca-del-examen-saber-11/

ICFES. (2022b, junio 7). *Resultados únicos Saber 11 | Datos Abiertos Colombia*. https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-11/kgxf-xxbe/about_data

ICFES. (2023). *Data ICFES*. http://www.icfes.gov.co/data-icfes/

ICFES. (2024, noviembre 18). *Glosario*. http://www.icfes.gov.co/atencion-y-servicios-a-la-ciudadania/glosario-icfes/

Laboratorio de Economía de la Educación. (2024). *Pruebas Saber 11: una década de análisis*. Informe No. 92. https://lee.javeriana.edu.co/-/lee-informe-92

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048-3061. https://doi.org/10.1109/TKDE.2019.2962680

Mendoza-Lozano, F. A., Quintero-Peña, J. W., & García-Rodríguez, J. F. (2021). The digital divide between high school students in Colombia. *Telecommunications Policy*, *45*(10), 102226. https://doi.org/10.1016/J.TELPOL.2021.102226

Ministerio de Educación Nacional. (2022, septiembre 2). *Pruebas Saber*. https://www.mineducacion.gov.co/portal/micrositios-preescolar-basica-y-media/Evaluacion/Evaluacion-de-estudiantes/397384:Pruebas-saber

Oreski, D., Pihir, I., & Konecki, M. (2017). CRISP-DM process model in educational setting. *Economic and Social Development: Book of Proceedings*, 19-28. https://www.zbw.eu/econis-archiv/bitstream/11159/645/1/Book_of_Proceedings_esdPrague_2017_Online.pdf#page=29

Rodríguez, C., Sanchez Torres, F., & Zúñiga, J. M. (2011). Impacto del Programa Computadores para Educar en la deserción estudiantil, el logro escolar y el ingreso a la educación superior. *Documentos CEDE*, *8744*. https://EconPapers.repec.org/RePEc:col:000089:008744

Rodríguez-Rosero, D. D., Ordoñez Ortega, R. E., & Hidalgo-Villota, M. E. (2021). Determinantes del rendimiento académico de la educación media en el departamento de Nariño, Colombia. *Lecturas de Economía*, *94*, 87-126. https://doi.org/10.17533/udea.le.n94a341834

Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Scikit-learn. (s. f.). *mutual_info_regression — scikit-learn 1.6.1 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

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

UNESCO. (2023, abril 20). *One year into COVID-19 education disruption: Where do we stand?* https://www.unesco.org/en/articles/one-year-covid-19-education-disruption-where-do-we-stand

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, *1*, 29-39. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Apéndices

Apéndice A

Código Python en Google Colab URL

https://colab.research.google.com/drive/14cHtoED68MpFkh-go6Dz7pio7tUAT8g6?usp=sharing

Apéndice B

Video de Socialización

VideoProyectoAplicadoClaudiaArteaga.mp4