Procesamiento de Datos a Gran Escala

Ing. John Jairo Corredor Franco, PhD

Proyecto Análisis Procesamiento de Datos a Gran Escala

Pontificia Universidad Javeriana

Facultad de Ciencias

Bogotá D.C.

25 de mayo de 2025

1. Justificación

La educación de calidad se constituye como un pilar fundamental e irrenunciable para el desarrollo económico y social sostenible de cualquier nación, al potenciar el capital humano, fomentar la innovación y promover la movilidad social. Sin embargo, el panorama colombiano se ve nublado por la persistencia de amplias y profundas desigualdades educativas. Estas diferencias no se distribuyen aleatoriamente, sino que afectan con especial crudeza a los municipios pequeños y a las áreas rurales, regiones históricamente marginadas que enfrentan múltiples desventajas acumulativas. Dentro de este complejo sostén de factores, la limitada y, en muchos casos, obsoleta infraestructura tecnológica emerge como una barrera crítica. Específicamente, la baja penetración de servicios de internet, sumada a la deficiente calidad y velocidad de las conexiones existentes, podría estar actuando como uno de los principales crecientes factores que aumentan esta inequidad educativa, creando una brecha digital que se superpone y agrava las brechas socioeconómicas ya existentes.

En este contexto, la realización de un análisis de datos exhaustivo y riguroso, que permita identificar y cuantificar con precisión cómo estos socioeconómicos influyen directamente en el desempeño académico de los estudiantes –medido a través de indicadores como resultados en pruebas estandarizadas, análisis descriptivo y predictivo– resulta crucial. Un diagnóstico certero no solo visibilizaría la magnitud del problema, sino que también proporcionaría al Ministerio de Educación Nacional y a otras entidades pertinentes la evidencia empírica necesaria para diseñar e implementar políticas públicas más precisas, efectivas y contextualizadas.

2. Entendimiento del Negocio

2.1 Contexto General

La educación es uno de los pilares fundamentales para el desarrollo social y económico de una nación. Sin embargo, en Colombia existen desigualdades marcadas en los resultados académicos obtenidos en las pruebas Saber 11 (ICFES), especialmente entre municipios de diferentes tamaños y niveles socioeconómicos. Este proyecto busca explorar estas desigualdades para identificar oportunidades de mejora, principalmente relacionadas con el acceso y calidad del servicio de internet.

2.2 Entendimiento detallado del negocio

En Colombia, los resultados de las pruebas Saber 11 reflejan importantes diferencias en el desempeño académico según la región y características socioeconómicas de los municipios. En general, los municipios grandes y con mejor infraestructura tecnológica suelen mostrar mejores resultados, mientras que municipios pequeños y rurales presentan desafíos importantes debido al acceso limitado a recursos educativos y conectividad.

El Ministerio de Educación ha reconocido que una posible causa de estas diferencias es el acceso desigual a internet, herramienta que se ha convertido en esencial para procesos de aprendizaje efectivos y acceso equitativo a recursos educativos digitales. Adicionalmente, factores como la pobreza multidimensional, el entorno familiar y la inversión municipal en educación también son determinantes claves que influyen en estos resultados.

Por lo tanto, se requiere un análisis exhaustivo y basado en datos que permita entender mejor la relación entre estos factores y los resultados académicos. Con esta información, el Ministerio podrá diseñar estrategias dirigidas específicamente a reducir brechas educativas, optimizar recursos invertidos y mejorar los indicadores territoriales de educación en Colombia.

3.1 Objetivo General

- Examinar la relación existente entre las condiciones socioeconómicas y el acceso a internet de los estudiantes con los puntajes globales obtenidos en las pruebas Saber 11, con el fin de generar recomendaciones basadas en evidencia para reducir brechas educativas.

3.2 Objetivos Específicos

- Realizar un procesamiento exhaustivo y riguroso de los conjuntos de datos disponibles para garantizar la calidad e integridad del análisis.
- Ejecutar un análisis descriptivo que permita comprender la estructura y comportamiento general de los datos disponibles, identificando tendencias y relaciones iniciales.
- Desarrollar modelos predictivos que permitan evaluar el impacto cuantitativo de factores socioeconómicos y de conectividad en el desempeño académico.
- Realizar una discusión detallada de los resultados obtenidos, identificando implicaciones prácticas y proporcionando recomendaciones concretas para el Ministerio de Educación.

4. Conjunto de Datos Seleccionados

Los datos utilizados en este proyecto fueron obtenidos de las bases de datos de acceso público del Instituto Colombiano para la Evaluación de la Educación (ICFES), las cuales ofrecen información detallada sobre los resultados del examen Saber 11. Esta fuente proporciona datos oficiales y estructurados que permiten realizar análisis robustos en el ámbito educativo colombiano.

Se seleccionaron los registros correspondientes a los años 2017 hasta 2024 con el objetivo de capturar la evolución de los factores socioeconómicos, tecnológicos y educativos a lo largo del tiempo. Esta serie temporal facilita la identificación de tendencias y posibles puntos de inflexión que afecten el rendimiento académico de los estudiantes.

El uso de datos multianuales permite además comparar distintos grupos de estudiantes bajo contextos políticos, sociales y económicos cambiantes. Así, se fortalece la posibilidad de comprender no solo las correlaciones contemporáneas, sino también las transformaciones estructurales que han tenido lugar en el país en términos de equidad educativa. Los datos utilizados en este proyecto fueron obtenidos de las bases de datos de acceso público del Instituto Colombiano para la Evaluación de la Educación (ICFES). Se seleccionaron los registros correspondientes a los años 2017 hasta 2024, lo que permite observar el comportamiento de los resultados educativos y su relación con variables socioeconómicas y tecnológicas a lo largo del tiempo. Esta selección temporal brinda una perspectiva más completa sobre la evolución de las condiciones estructurales y su influencia en el rendimiento académico de los estudiantes colombianos.

A continuación, se presenta el diccionario de datos:

Variable	Descripción	Justificación
periodo_examen	Código de año y semestre de presentación (e.g. 20172 = 2ª aplicación 2017)	Permite agrupar y comparar resultados por grupo temporal, seguir tendencias y controlar efectos de cada año/semestre.
calendario_colegio	Calendario académico del establecimiento (A, B, OTRO)	Filtra la segunda aplicación (calendario A) y controla posibles diferencias de grupos según calendario institucional.
cod_mcpio_colegio	Código DANE del municipio donde está ubicada la sede	Clave geográfica para unir con datos de penetración de internet, variables demográficas y construir el panel municipal.
puntaje_global	Puntaje total obtenido en Saber 11	Principal indicador de desempeño académico que se busca explicar y mejorar.
puntaje_lectura_critica	Puntaje en lectura crítica	Una de las áreas evaluadas; permite analizar impactos diferenciales según dominio de la competencia lectora.
puntaje_matematicas	Puntaje en matemáticas	Refleja el dominio de habilidades cuantitativas; útil para ver si conectividad afecta más a algunas áreas que otras.
puntaje_ciencias_naturales	Puntaje en ciencias naturales	Mide competencias científicas; sirve para desagregar el efecto de TIC en distintas áreas del conocimiento.
puntaje_sociales_ciudadanas	Puntaje en ciencias sociales y ciudadanas	Evalúa comprensión de contexto sociopolítico; relevante para entender si el acceso a información vía internet influye en el área.
puntaje_ingles	Puntaje en inglés	Indicador de competencia en lengua extranjera; proxy de exposición a contenidos digitales.
horas_internet_estudiante	Categoría ordinal de horas diarias de navegación no académica	Mide el uso individual de internet, contrastable con la cobertura del hogar y la penetración municipal.
indice_socioeconomico_estudiante	Índice Socioeconómico Individual (ISNI)	Control socioeconómico personal que explica parte de la

		variabilidad en los puntajes y evita sesgos de confusión.
estrato_vivienda	Estrato socioeconómico de la vivienda según recibo de energía	Refleja la posición socioeconómica familiar; influye en acceso a recursos y apoyo educativo.
num_libros_hogar	Categoría ordinal de cantidad de libros en el hogar	Proxy de capital cultural familiar, correlaciona con rendimiento académico y ambiente de aprendizaje.
hogar_tiene_computador	Binaria: 1 si el hogar dispone de computador, 0 en caso contrario	Indicador de recursos TIC en el hogar, factor clave para actividades de estudio y acceso a contenidos digitales.
hogar_tiene_internet	Binaria: 1 si el hogar cuenta con conexión a internet, 0 en caso contrario	Mide cobertura doméstica directa; junto con la penetración municipal evalúa brechas de acceso a la red.
educacion_madre	Nivel educativo más alto alcanzado por la madre	Capital humano parental que influye en apoyo académico y aspiraciones; agrega contexto sociocultural.
educacion_padre	Nivel educativo más alto alcanzado por el padre	Complementa la información de capital humano familiar y sus efectos sobre el rendimiento estudiantil.
area_ubicacion_colegio	Área rural o urbano donde se ubica el colegio	Control geográfico esencial: diferencia infraestructuras, acceso a servicios y dinámicas urbanas vs. rurales.
caracter_colegio	Carácter del establecimiento (Académico, Técnico, Técnico/Académico, No Aplica)	Tipo de oferta educativa que puede modular el énfasis curricular y recursos disponibles.
naturaleza_colegio	Naturaleza del establecimiento (Oficial/Privado)	Control institucional clave: financiamiento, calidad de infraestructura y perfiles de población estudiantil.
jornada_colegio	Jornada en que opera la sede (Mañana, Tarde, Completa, etc.)	Representa el esquema horario y potencialmente la disponibilidad de servicios complementarios (tutorías, refuerzos).

colegio_bilingue	Indicador (S/N) de si el establecimiento es bilingüe	Factor de innovación y enfoque pedagógico que puede influir en resultados, especialmente en el área de inglés y uso de TIC.

5. Entendimiento de los Datos

Al unificar los datos de los distintos años disponibles, se conformó una base de datos con 4.761.554 registros y 86 columnas. Esta base contiene información de estudiantes de todo el país que presentaron el examen Saber 11 entre los años 2017 y 2024, lo cual constituye un insumo de alta representatividad para analizar tendencias educativas en el tiempo.

Las variables disponibles abarcan múltiples dimensiones: datos sociodemográficos del estudiante, características del entorno familiar, condiciones institucionales de los colegios, y puntajes obtenidos tanto globales como por área (lectura crítica, matemáticas, ciencias naturales, sociales y ciudadanía, e inglés). Esta riqueza de variables permite explorar cómo interactúan distintos factores en el rendimiento académico.

Un resumen estadístico de las variables numéricas revela que el puntaje global promedio fue de aproximadamente 250 puntos, con una desviación estándar de 50.83, mientras que los puntajes por área oscilaron entre 47 y 53 puntos en promedio, con distribuciones bastante simétricas. Respecto a las variables categóricas, se observa que la mayoría de los estudiantes se identifican con el género femenino (54.1%), pertenecen al estrato 2 (39.8%) y sus madres y padres tienen, en su mayoría, estudios de secundaria completa. Además, el 69.2% de los estudiantes reportaron tener acceso a internet en el hogar. Este conjunto de variables contextuales será clave en los análisis posteriores.

A partir del conjunto completo, se seleccionaron las variables más relevantes para el análisis descriptivo y predictivo, priorizando aquellas relacionadas con el contexto del estudiante, del

hogar, del colegio y los resultados del examen. Adicionalmente, se identificaron variables con proporciones importantes de datos faltantes, como estu_dedicacioninternet, fami_estratovivienda y fami_numlibros, lo que motivó una estrategia específica de tratamiento para mantener la calidad del análisis posterior.

Tipos de datos: Variables numéricas, categóricas, identificadores geográficos y temporales.

- Características personales del estudiante (género, etnia, NSE).
- Contexto familiar (estrato, educación de padres, acceso a internet y tecnologías).
- Características institucionales (naturaleza del colegio, ubicación, jornada, bilingüismo).
- Resultados del examen Saber 11 (puntajes y percentiles por área)

6. Reporte de Calidad de los Datos

A continuación, se presenta una tabla con el número de valores faltantes para cada variable clave del conjunto de datos.

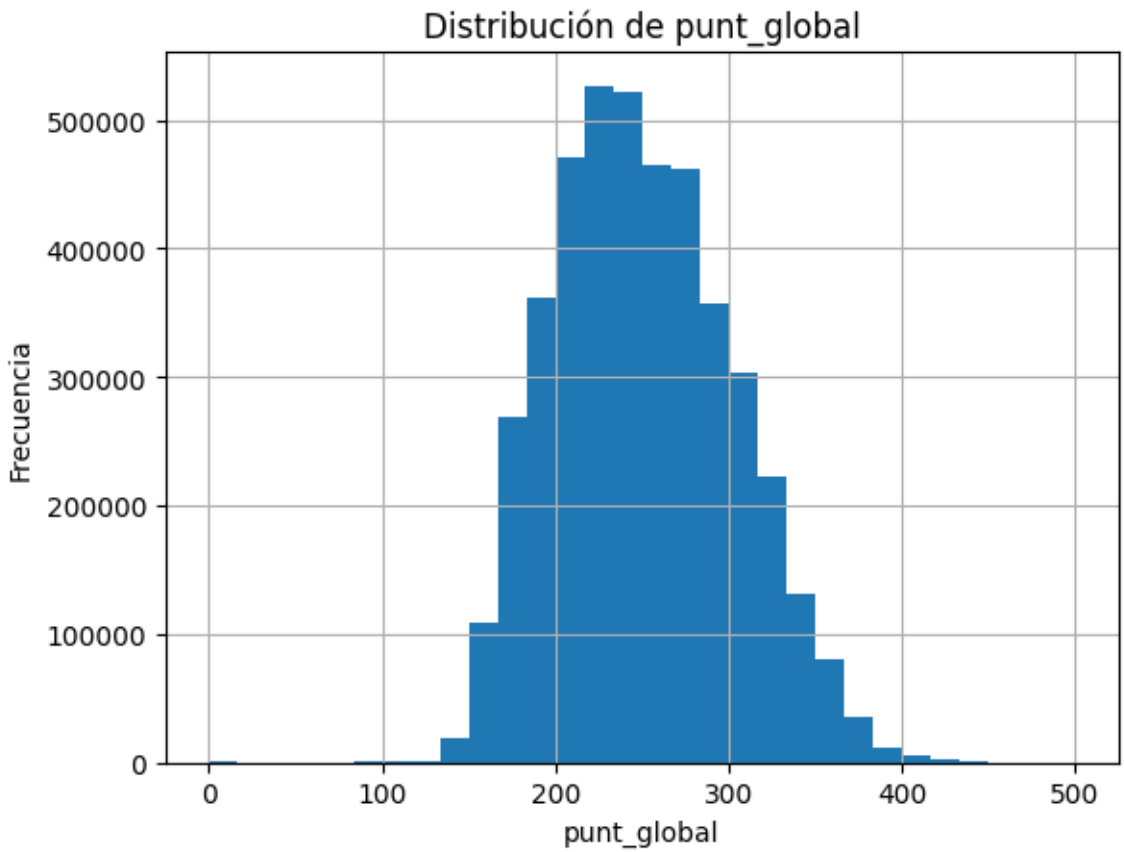
Variable	Valores Faltantes
periodo	0
cole_cod_mcpio_ubicacion	0
estu_genero	323
estu_fechanacimiento	1
estu_nse_individual	190782
estu_etnia	4154184
estu_discapacidad	0
estu_dedicacioninternet	292300
fami_estratovivienda	243359
fami_educacionmadre	272895
fami_educacionpadre	272002
fami_numlibros	354876
fami_personashogar	178680
fami_tieneinternet	276707
fami_tienecomputador	186331
fami_tieneconsolavideojuegos	191877
fami_cuartoshogar	183144
cole_naturaleza	0

cole_area_ubicacion	0
cole_calendario	0
cole_caracter	133762
cole_jornada	1911
cole_bilingue	734124
punt_global	0
percentil_global	32249
punt_lectura_critica	0
percentil_lectura_critica	0
punt_matematicas	0
percentil_matematicas	0
punt_c_naturales	0
percentil_c_naturales	0
punt_sociales_ciudadanas	0
percentil_sociales_ciudadanas	0
punt_ingles	30120
percentil_ingles	23692

Esta revisión permite identificar los campos que requieren imputación o tratamiento especial durante la fase de preparación. Destacan variables con grandes proporciones de datos nulos como estu_etnia, fami_numlibros, fami_estratovivienda, fami_educacionmadre, fami_educacionpadre, y fami_tieneinternet, lo cual puede afectar los análisis si no se gestionan adecuadamente. Esta información fue fundamental para implementar una estrategia de imputación tanto numérica como categórica durante el procesamiento.

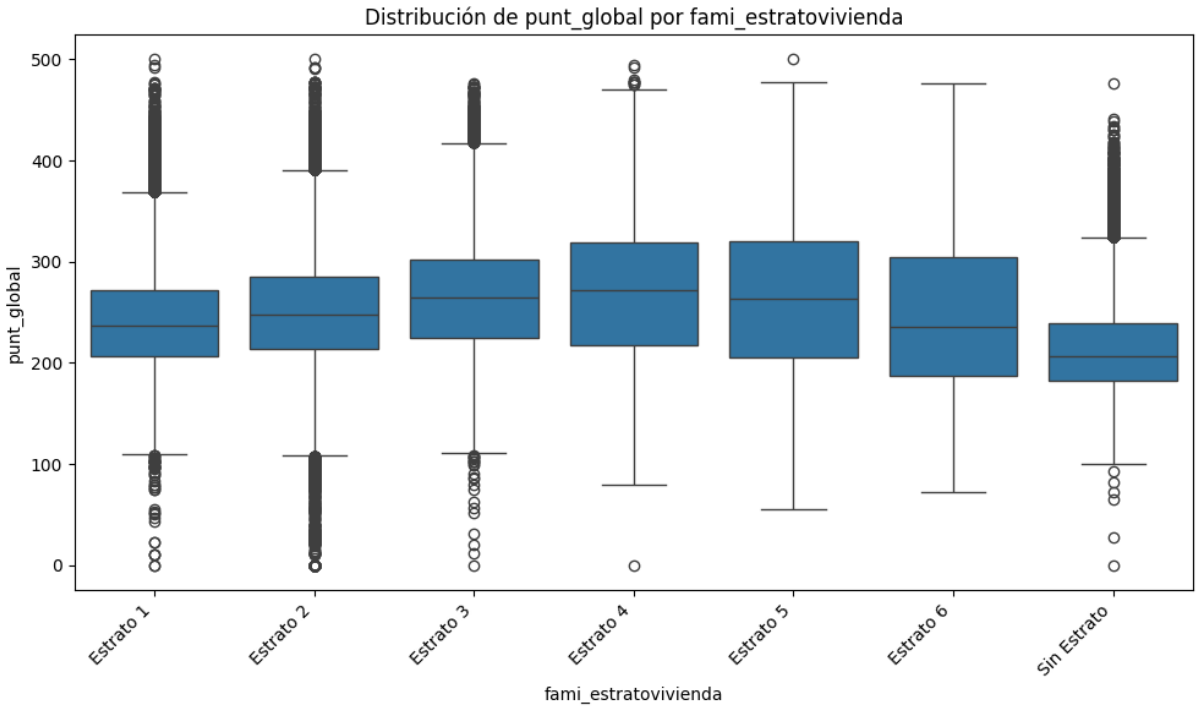
7. Análisis Descriptivo

La variable punt_global, que representa el puntaje total obtenido por los estudiantes en la prueba Saber 11, muestra una distribución aproximadamente normal. Tal como se observa en la Figura 1, la mayoría de los estudiantes se concentra en un rango de puntajes entre 200 y 300, con un pico de frecuencia en torno a los 250 puntos, que corresponde a la media general de la muestra.



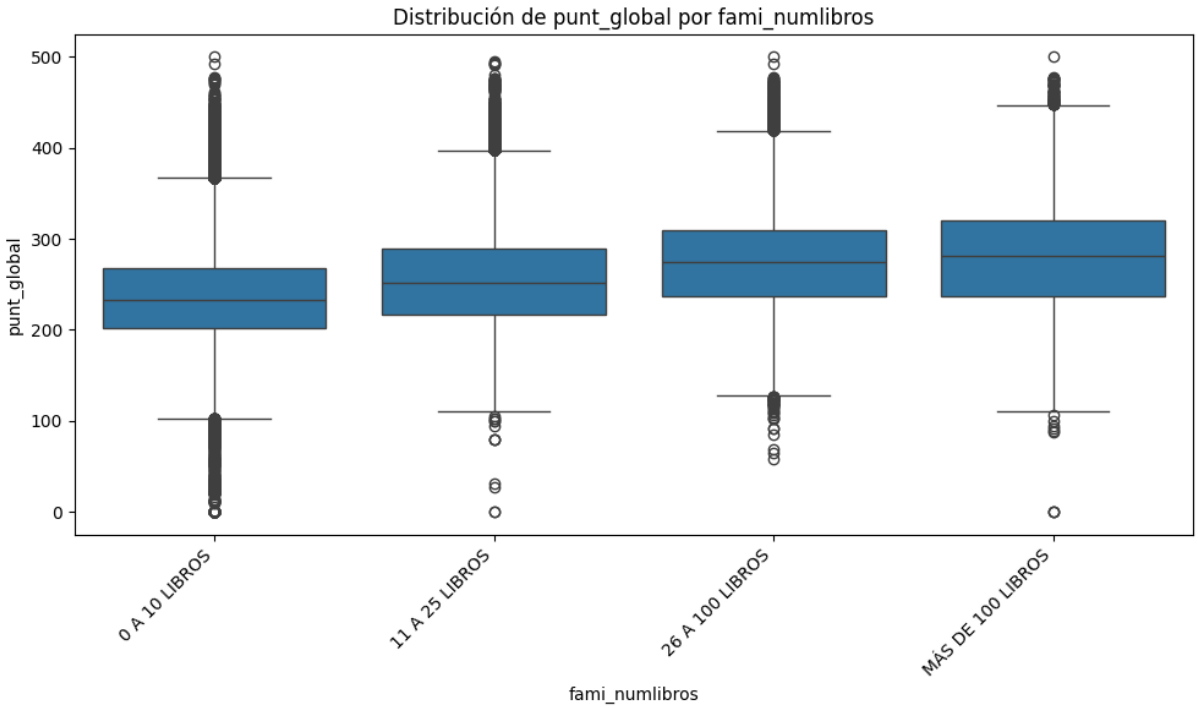
Esta forma de campana indica que los resultados del examen tienden a estar distribuidos de forma simétrica en torno a la media, con menor proporción de estudiantes en los extremos inferiores y superiores (menores de 150 o mayores de 400 puntos). Este comportamiento sugiere que el examen posee una capacidad diferenciadora estándar, siendo útil como variable dependiente en futuros análisis predictivos.

La Figura 2 muestra la distribución del puntaje global (punt_global) según el estrato socioeconómico reportado por los estudiantes en la variable fami_estratovivienda. Se observa una tendencia clara: a medida que aumenta el estrato, también tiende a incrementarse la mediana del puntaje, lo que sugiere una asociación positiva entre el nivel socioeconómico y el rendimiento académico.



Los estudiantes de estratos 4, 5 y 6 presentan no solo mayores medianas de puntaje, sino también una mayor dispersión hacia puntajes altos. En contraste, los estudiantes de estratos 1, 2 y especialmente aquellos que no reportan estrato (“Sin Estrato”) muestran puntajes más concentrados en la parte baja del rango, lo cual puede reflejar condiciones estructurales desfavorables para el aprendizaje.

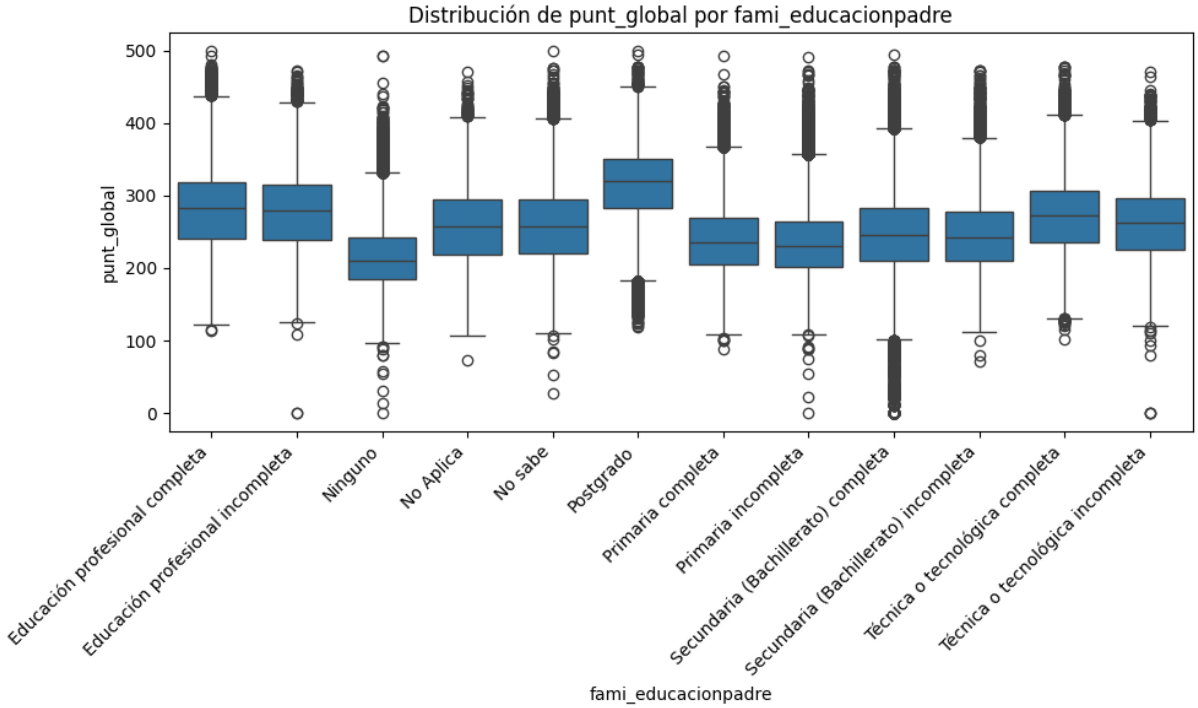
La Figura 3 presenta un análisis de la variable punt_global en función del número estimado de libros en el hogar, categorizado en cuatro niveles. Se observa una tendencia creciente: los estudiantes que reportan tener más libros en casa tienden a obtener puntajes globales más altos en la prueba Saber 11.

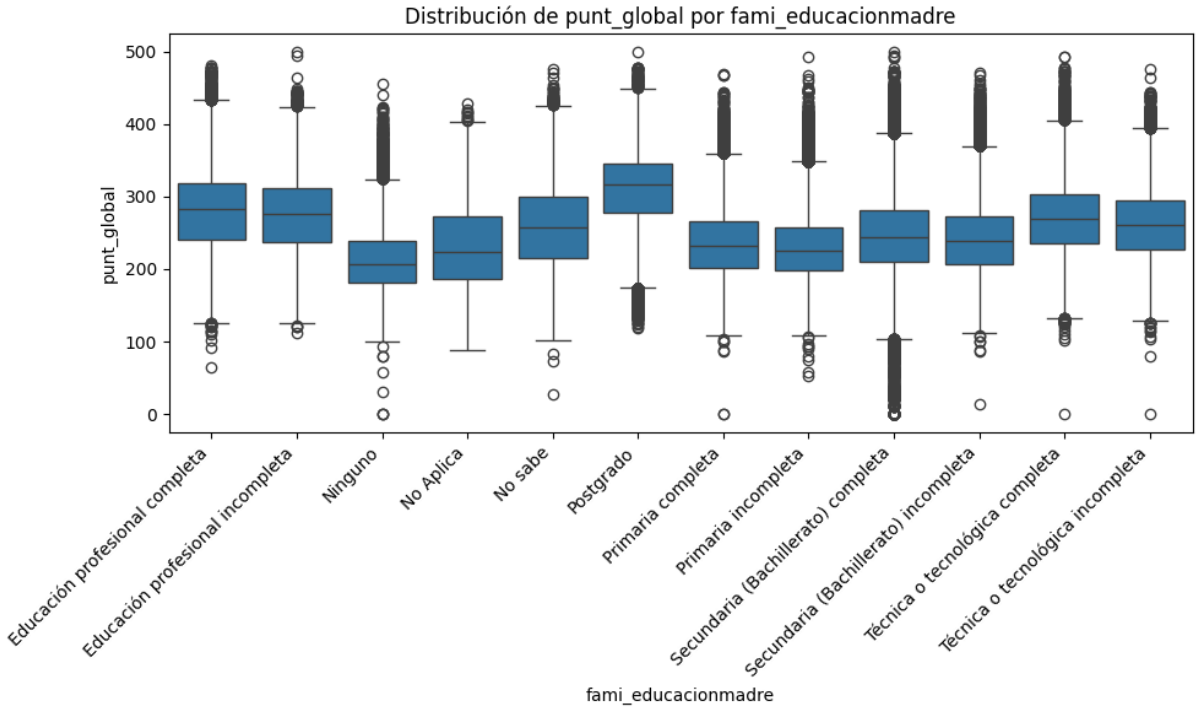


Los estudiantes con entre 0 y 10 libros tienen una mediana de puntaje más baja y mayor concentración de resultados en los rangos inferiores. En contraste, quienes indican tener más de 100 libros presentan una mediana de desempeño considerablemente superior, así como una mayor proporción de resultados altos, con menor densidad en los extremos inferiores.

Estos resultados sugieren que la disponibilidad de material de lectura en el hogar es un indicador relevante del capital cultural familiar, y que su presencia puede estar vinculada al estímulo académico y al desarrollo de habilidades cognitivas que se reflejan en el rendimiento educativo.

Las Figuras 4 y 5 presentan la distribución del puntaje global obtenido en la prueba Saber 11 en función del nivel educativo alcanzado por el padre y la madre del estudiante, respectivamente. En ambos casos, se observa una relación positiva entre el nivel educativo de los padres y el rendimiento académico del estudiante que presenta la prueba.





Los estudiantes cuyos padres o madres completaron estudios profesionales, técnicos, tecnológicos o de posgrado tienden a alcanzar puntajes significativamente más altos. Por el contrario, los estudiantes cuyos padres no tienen estudios formales, no saben su nivel educativo

o reportan niveles bajos (primaria incompleta o secundaria incompleta), concentran sus resultados en puntajes más bajos.

Con el objetivo de evaluar si existen diferencias estadísticamente significativas en el puntaje global (punt_global) entre los grupos definidos por diferentes variables categóricas, se aplicó una prueba ANOVA unidireccional para cada una de las siguientes variables: características del estudiante, contexto familiar y atributos del colegio.

Los resultados muestran que, en todos los casos (exceptuando cole_calendario), el valor p fue menor a 0.05, lo que indica evidencia suficiente para rechazar la hipótesis nula de igualdad de medias entre grupos. En particular, variables como fami_tienecomputador (F = 338,549.57), fami_tieneinternet (F = 259,218.44) y fami_educacionmadre (F = 57,617.07) presentaron estadísticos F muy elevados, lo cual sugiere una fuerte relación entre estas características y el desempeño académico. Entre las variables con mayor significancia destacan también: fami_estratovivienda, fami_numlibros, y estu_dedicacioninternet, relacionadas con el entorno de aprendizaje del estudiante. cole_naturaleza, cole_area_ubicacion, cole_jornada y cole_caracter, reflejando el impacto del contexto institucional. estu_genero, estu_etnia y estu_discapacidad, que evidencian diferencias que podrían estar asociadas a brechas estructurales.

En contraste, la variable cole_calendario no mostró diferencias significativas en los promedios de puntaje global entre sus categorías (valor p no significativo), posiblemente debido a que la mayoría de las instituciones operan bajo un mismo calendario académico o a que los tamaños muestrales por grupo fueron insuficientes.

Estos resultados ratifican la importancia de considerar el contexto sociodemográfico y educativo como determinante del rendimiento académico. En fases posteriores del análisis,

estas variables serán priorizadas en modelos predictivos para cuantificar su efecto individual y combinado sobre los resultados de las pruebas Saber 11.

8. Filtros y Limpieza de los Datos

Problemas Identificados:

- Valores faltantes en variables clave.
- Inconsistencias y duplicados iniciales.

El procesamiento de los datos comenzó con la estandarización de los nombres de las columnas, eliminando espacios innecesarios, convirtiéndolos a minúsculas y reemplazando espacios por guiones bajos, con el fin de garantizar uniformidad y evitar errores durante el análisis. A continuación, se eliminaron las columnas duplicadas para asegurar la integridad del esquema de datos.

Luego se identificaron las variables numéricas junto con la clave geográfica de agrupación por municipio (cole_cod_mcpio_ubicacion), lo cual permitió realizar una imputación específica para los valores faltantes. En este paso, los valores faltantes de cada variable numérica fueron reemplazados por la media correspondiente dentro del municipio. Si algún municipio no contaba con suficientes datos para calcular esta media (es decir, todos los valores eran nulos), se imputó con la media global de esa variable.

En cuanto a las variables categóricas, se imputaron sus valores faltantes utilizando la moda agrupada por colegio y por municipio de cada columna, lo que garantiza consistencia sin introducir sesgos geográficos específicos, la única excepción fue la variable **estu_etnia**, cuyo porcentaje de datos faltantes superaba el 87 %, y por tanto no se imputó mediante esta estrategia.

Posteriormente, todas las variables categóricas se convirtieron al tipo "***category***" para optimizar el almacenamiento y mejorar el rendimiento del modelado. Finalmente, se eliminaron todas las filas completamente duplicadas para asegurar un conjunto de datos limpio y no redundante, listo para los análisis posteriores.

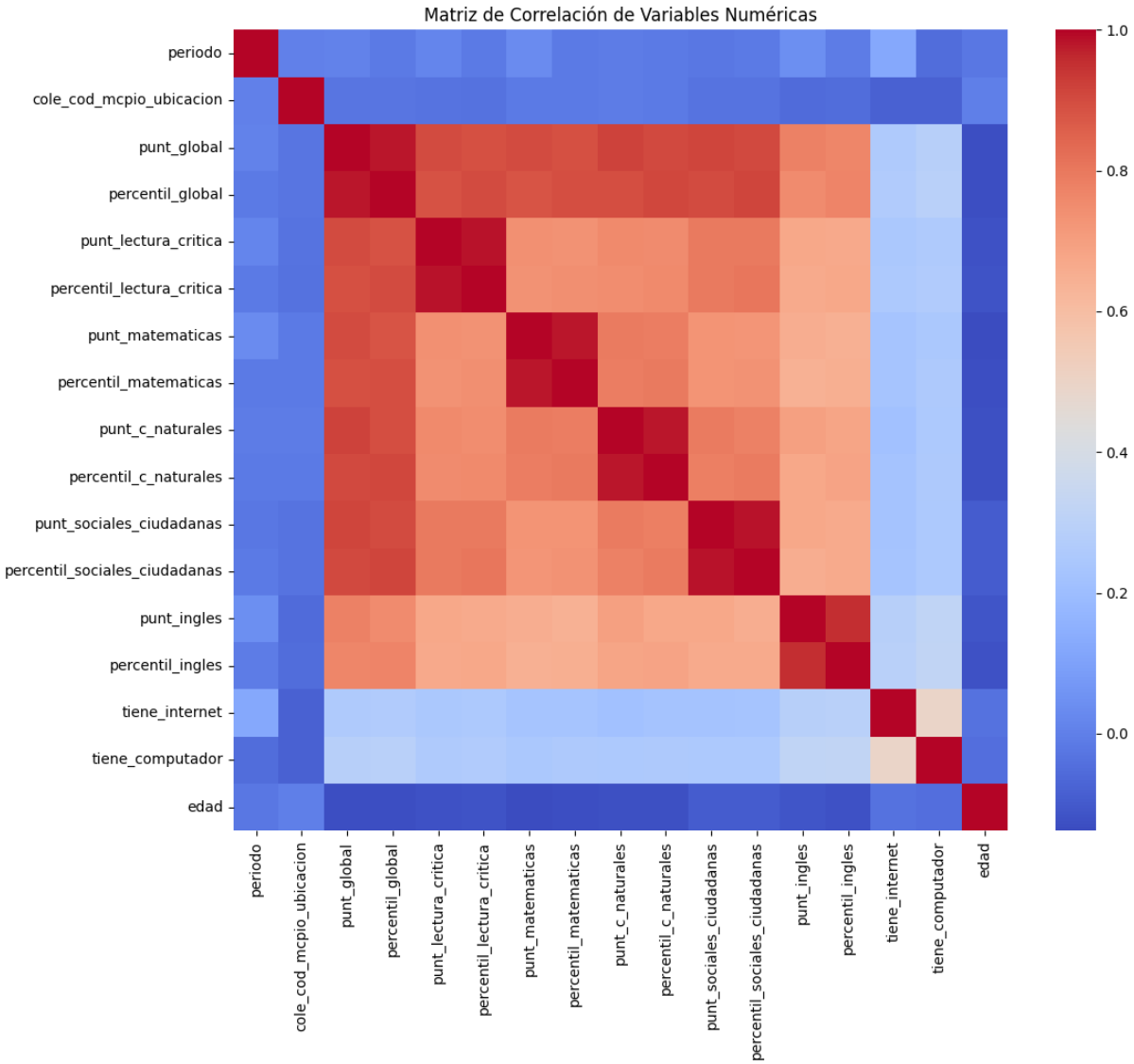
Como resultado de este pipeline de limpieza y transformación, se obtuvo un conjunto de datos final con 4.357.399 registros y 35 columnas, representando una versión depurada y optimizada de la base original para los fines del análisis descriptivo y predictivo del presente proyecto.

9. Modelado de Aprendizaje de Maquina

En el contexto del presente estudio, donde se busca predecir el puntaje global obtenido por los estudiantes en las pruebas Saber 11 a partir de una amplia gama de variables sociodemográficas, familiares e institucionales, la regresión regularizada surge como una herramienta estadística especialmente adecuada para abordar esta problemática.

Para construir el modelo predictivo que permita estimar el puntaje global obtenido por los estudiantes en las pruebas Saber 11, se seleccionaron variables representativas de tres dimensiones clave: características individuales del estudiante, contexto familiar y entorno institucional del colegio. Esta selección se basó tanto en criterios teóricos como en los resultados del análisis descriptivo y las pruebas estadísticas previas (como ANOVA), que demostraron la existencia de diferencias significativas en el rendimiento académico según dichas variables.

En la matriz de correlación presentada se observa una alta correlación entre el puntaje global (punt_global) y los puntajes por área, así como sus respectivos percentiles. Esto era esperado, ya que el puntaje global del examen Saber 11 es una suma ponderada de los puntajes individuales obtenidos en cada una de las áreas evaluadas (lectura crítica, matemáticas, ciencias naturales, sociales y ciudadanía, e inglés). Por esta razón, dichas variables están directamente incluidas en el cálculo del puntaje global.



Dado esto, incluir estas variables como predictoras en los modelos de aprendizaje automático generaría un sesgo artificial, ya que estaríamos utilizando componentes que determinan directamente la variable objetivo, lo que invalida el propósito del análisis predictivo. Por tanto, se decidió excluir estas variables del conjunto de predictores en los modelos para evitar problemas de sobreajuste y asegurar que las predicciones se basen únicamente en condiciones exógenas al examen (como características del estudiante, del hogar y del colegio).

Las **variables del estudiante** incluyen género, fecha de nacimiento (como proxy de edad), nivel socioeconómico individual, pertenencia étnica, condición de discapacidad y dedicación diaria

al uso de internet. Por parte del **contexto familiar**, se consideraron variables relacionadas con el nivel educativo de los padres, el estrato socioeconómico del hogar, el número de libros disponibles, el acceso a computador, internet y consolas de videojuegos, así como el número de personas y cuartos en el hogar. Finalmente, dentro de las **características del colegio**, se incluyeron el tipo de institución (oficial o privada), la ubicación geográfica (urbana o rural), el calendario académico, la jornada, el carácter del colegio (académico, técnico, etc.) y si ofrece educación bilingüe.

Dimensión	**Variables seleccionadas**	**Justificación**
Estudiante	estu_genero, estu_fechanacimiento, estu_nse_individual, estu_discapacidad, estu_dedicacioninternet	Factores personales y sociales que pueden influir directamente en las oportunidades educativas.
Contexto Familiar	fami_estratovivienda,fami_educacionmadre, fami_educacionpadre,fami_numlibros, fami_personashogar,fami_cuartoshogar,fami_tieneinternet, fami_tienecomputador, fami_tieneconsolavideojuegos	Reflejan el capital cultural, social y económico del hogar del estudiante.
Características Colegio	cole_naturaleza, cole_area_ubicacion,cole_caracter,cole_jornada, cole_bilingue	Variables institucionales que definen el entorno de aprendizaje formal del estudiante.

Antes de entrenar los modelos de aprendizaje automático, se realizó un paso de preprocesamiento sobre las variables predictoras para garantizar que estuvieran en un formato

adecuado para los algoritmos. En primer lugar, se identificaron todas las variables categóricas del conjunto de datos mediante el tipo de dato (object o category) y se convirtieron explícitamente a cadenas de texto (string). Esta conversión es necesaria para aplicar correctamente técnicas de codificación como **One-Hot Encoding**, que transforma variables categóricas en vectores binarios.

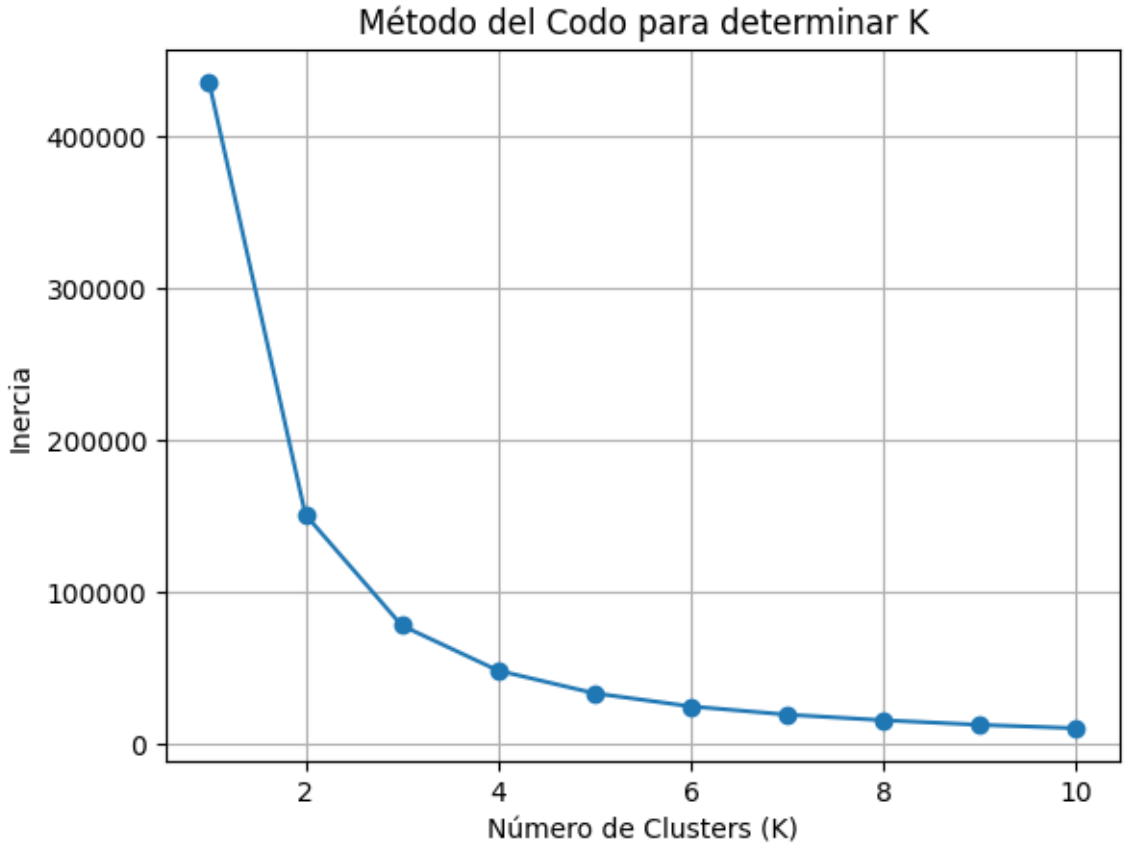
Posteriormente, se definió un **ColumnTransformer** que aplica **OneHotEncoder** a las variables categóricas, lo que permite representar cada categoría como una columna binaria. Esto convierte las variables cualitativas en una forma numérica interpretable por los modelos sin introducir supuestos de orden o distancia entre categorías. Aunque no se incluyó en esta etapa, cuando se usan variables numéricas también es común aplicar escalado mediante **StandardScaler**, especialmente en algoritmos sensibles a la magnitud de los datos (como regresiones o métodos basados en distancia). Este preprocesamiento es crucial para evitar sesgos y garantizar un entrenamiento adecuado del modelo.

En primer lugar, el modelo cuenta con una alta cantidad de variables explicativas, muchas de las cuales presentan correlaciones entre sí. Esta multicolinealidad puede afectar la estabilidad de los coeficientes en una regresión lineal ordinaria, generando resultados sensibles a pequeñas variaciones en los datos. La regresión regularizada, particularmente en sus variantes **Ridge** (penalización L2) y **Lasso** (penalización L1), permite mitigar este problema al introducir un término de penalización que reduce la magnitud de los coeficientes, controlando así la complejidad del modelo.

Además, la regularización permite **prevenir el sobreajuste**, un riesgo latente cuando se trabaja con grandes volúmenes de datos y múltiples variables. Al limitar la capacidad del modelo de ajustarse a las fluctuaciones aleatorias del conjunto de entrenamiento, se mejora su capacidad de generalización a nuevos datos, lo cual es esencial en contextos reales de aplicación educativa.

Para nuestro modelo de regresión Ridge alcanzó su mejor desempeño con un valor de penalización óptimo de $\alpha = 10$. Bajo este parámetro, se obtuvo un error cuadrático medio (RMSE) de 41.03 en el conjunto de entrenamiento y 42.09 en el conjunto de prueba, junto con un coeficiente de determinación (R^2) de 0.31 en ambos casos. Estos resultados indican una mala capacidad predictiva y generalización del modelo, con mínima diferencia entre el rendimiento en entrenamiento y prueba, lo que sugiere una adecuada regularización sin sobreajuste.

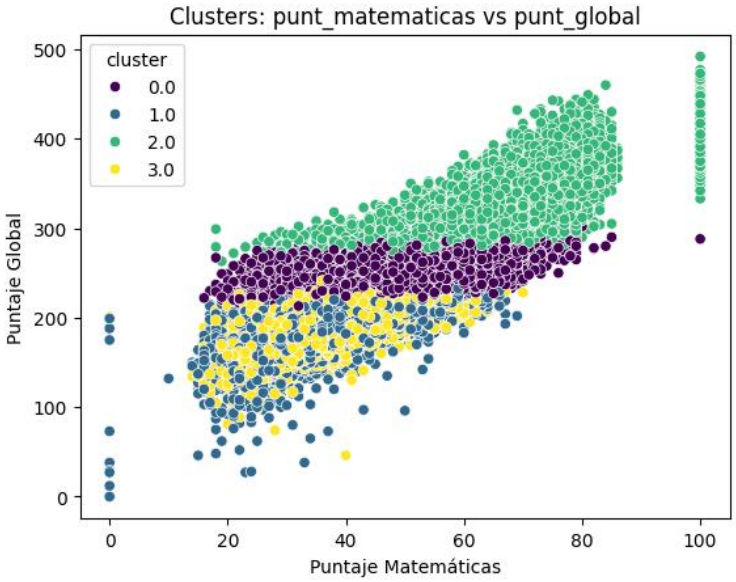
Con el fin de identificar patrones latentes en el rendimiento académico de los estudiantes, se implementó un modelo de clustering no supervisado mediante el algoritmo K-Means**,** tomando como base las variables numéricas estandarizadas relacionadas con el desempeño en las pruebas Saber 11. Esta técnica permitió agrupar a los estudiantes en subconjuntos homogéneos según su similitud en variables como punt_global.

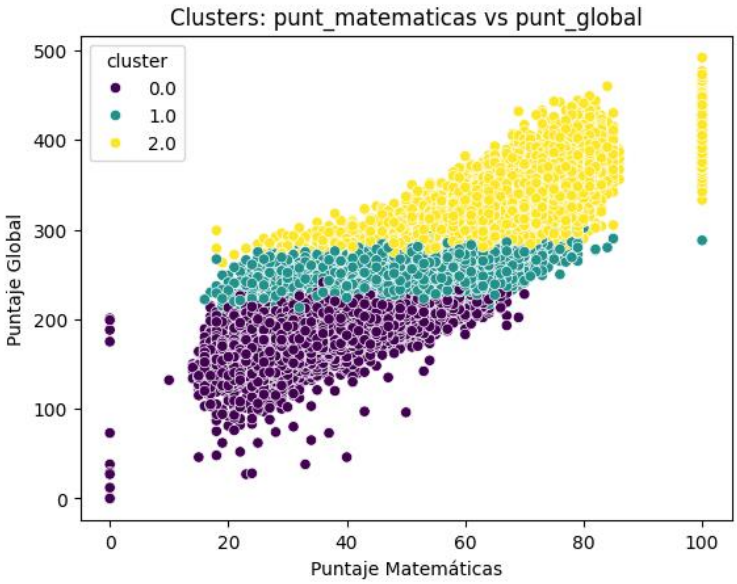


El modelo fue ajustado para generar tres clústeres dado el primer más un clúster de ajuste punto de inflexión de la prueba del codo (k = 3), valor determinado tras pruebas exploratorias que mostraron una segmentación significativa y balanceada de la población. Posteriormente, se visualizó la distribución de los clústeres en un plano bidimensional utilizando como referencia los puntajes de matemáticas (eje X) y el puntaje global (eje Y).

Con el objetivo de identificar perfiles estudiantiles diferenciados según condiciones sociodemográficas, se aplicó el algoritmo **K-Means**. Se exploraron dos configuraciones: una con **4 clústeres** y otra con **3 clústeres**, evaluando su capacidad para generar segmentos coherentes y útiles para la interpretación educativa.

Los resultados muestran que el algoritmo logró identificar perfiles académicos diferenciados:





Con la configuración de 4 clúster se permitió distinguir cuatro perfiles bien definidos, entre los que destacan:

- **Clúster 0**: Estudiantes adultos (edad media ~41 años) con rendimiento promedio-alto pese a condiciones socioeconómicas desfavorables. Representa un grupo resiliente en programas de educación para adultos.
- **Clúster 1**: Estudiantes jóvenes con **altas condiciones socioeconómicas y tecnológicas**, pero con **bajo rendimiento académico**, lo que sugiere un uso poco productivo de los recursos o posibles falencias motivacionales.
- **Clúster 2**: Jóvenes con **bajo nivel socioeconómico y acceso limitado a tecnología**, pero que logran el **mejor desempeño académico**. Refleja un segmento de alta resiliencia educativa.
- **Clúster 3**: Estudiantes con características intermedias en términos de edad, acceso, contexto familiar y rendimiento. Representa el grupo promedio del sistema educativo.

La segmentación en tres grupos resultó en una clasificación más simplificada:

- **Clúster 0**: Estudiantes con condiciones favorables (estratos altos, padres con mayor educación, acceso a computador e internet), con perfil similar al Clúster 1 del modelo de 4 clústeres. Dependiendo de su rendimiento, podría replicar la **paradoja del bajo desempeño con recursos**.
- **Clúster 1**: Perfil más vulnerable (estrato 1, baja escolaridad parental, acceso limitado a tecnología), con predominancia femenina y mayor edad promedio. Presenta similitudes con el Clúster 2 del modelo de 4 clústeres, que destacó por su **alto rendimiento en contextos adversos**.
- **Clúster 2**: Grupo con características socioeconómicas y educativas intermedias, similar al Clúster 3 de la segmentación anterior.

Esta segmentación aporta valor al permitir una visión más granular de los perfiles estudiantiles, facilitando el diseño de estrategias pedagógicas diferenciadas, priorización de recursos y focalización de intervenciones educativas según las características del grupo al que pertenece cada estudiante. Aunque esta configuración simplifica el panorama, diluye la visibilidad del grupo adulto (Clúster 0 del modelo de 4), y mezcla perfiles con dinámicas distintas.

10. Preguntas Planteadas con su repuesta

- ¿Existe una brecha de rendimiento académico entre estudiantes de diferentes niveles socioeconómicos?

A partir de los resultados observados en el boxplot que compara el puntaje global según el estrato de la vivienda, junto con la prueba ANOVA unidireccional, se concluye que existe una relación estadísticamente significativa entre estas variables. Sin embargo, esta relación no es estrictamente lineal, ya que, si bien el puntaje promedio tiende a aumentar con el estrato, a partir del estrato 4 se observa una leve disminución en la media. Esto sugiere que otros factores

podrían estar modulando el efecto del nivel socioeconómico sobre el rendimiento académico en los estratos más altos.

- ¿Qué factores del entorno educativo (características del colegio) influyen en el desempeño académico?

Con base en los análisis realizados, se identificaron varios factores del entorno educativo que presentan una relación significativa con el desempeño académico de los estudiantes. A través de la prueba ANOVA unidireccional, se encontró que variables como la **naturaleza del colegio** (oficial o privado), el **tipo de jornada académica**, el **carácter del establecimiento** (académico, técnico, etc.) y la **ubicación geográfica** (urbana o rural) tienen diferencias estadísticamente significativas en los puntajes globales obtenidos en la prueba Saber 11.

En particular, los estudiantes que asisten a colegios **privados**, ubicados en zonas **urbanas**, con jornadas **diurnas** (mañana o tarde) y con un carácter **académico** tienden a obtener mejores resultados. Estos hallazgos sugieren que las condiciones institucionales del colegio tienen un impacto directo en el aprendizaje, posiblemente asociado a diferencias en recursos pedagógicos, infraestructura, acceso a tecnología y capital académico del entorno escolar.

Por tanto, el contexto educativo no solo influye en el acceso a oportunidades de aprendizaje, sino que también puede contribuir a ampliar o reducir las brechas de rendimiento entre estudiantes de distintos territorios y niveles socioeconómicos. Estos factores deben ser considerados al formular políticas públicas orientadas a mejorar la equidad y calidad en la educación.

- ¿Qué tan bien puede predecirse el puntaje global del estudiante a partir de sus condiciones?

El modelo predictivo implementado mediante regresión Ridge, utilizando las características sociodemográficas, familiares y escolares de los estudiantes como variables explicativas, logró un **error cuadrático medio (RMSE)** de **41.03** en el conjunto de entrenamiento y **42.09** en el conjunto de prueba. El **coeficiente de determinación (R^2)** obtenido fue de **0.32** en ambos conjuntos, lo cual indica que aproximadamente el **32% de la variabilidad en el puntaje global** puede ser explicada por las variables incluidas en el modelo.

Estos resultados sugieren que, si bien las condiciones del estudiante tienen un impacto relevante en su desempeño académico, una parte importante del rendimiento sigue estando influenciada por otros factores no observados en los datos disponibles, como la motivación individual, la calidad del cuerpo docente, o incluso eventos contextuales. En consecuencia, aunque el modelo ofrece información útil para la segmentación y priorización de intervenciones, su capacidad predictiva es **mala** y debe ser complementada con análisis cualitativos o variables adicionales para lograr una comprensión más completa del fenómeno educativo.

- ¿Existe una diferencia significativa y cuantificable en el desempeño de los estudiantes que reportan tener acceso a internet en el hogar frente a los que no?

Sí, el análisis demostró una diferencia estadísticamente significativa. La prueba ANOVA unidireccional para la variable fami_tieneinternet arrojó un estadístico F considerablemente alto ($F = 259{,}218.44$) con un valor $p < 0.05$. Esto indica que los estudiantes que reportan tener acceso a internet en sus hogares tienden a obtener, en promedio, puntajes globales significativamente distintos (y descriptivamente más altos, como se infiere del contexto) en las pruebas Saber 11 en comparación con aquellos que no disponen de este recurso. Este hallazgo subraya la relevancia del acceso a internet como un factor asociado al rendimiento académico.

- ¿Qué impacto específico tienen el nivel educativo de los padres sobre el puntaje global obtenido de los estudiantes?

El análisis confirma un fuerte impacto positivo de ambos factores. Las pruebas ANOVA para fami_educacionmadre ($F = 57{,}617.07$, $p < 0.05$) y fami_educacionpadre (significativo, aunque F no se cita explícitamente en ese resumen, se indica que es una de las de mayor significancia) muestran que, a mayor nivel educativo de los padres, los estudiantes tienden a obtener mejores puntajes.

- ¿Es posible identificar agrupaciones de estudiantes con características socioeconómicas que podrían beneficiarse de intervenciones gubernamentales?

Sí, el análisis de clustering K-Means, particularmente con la configuración de 4 clústeres basada en condiciones sociodemográficas, logró identificar perfiles estudiantiles distintos y relevantes para la política educativa. Por ejemplo, se identificó un "Clúster 2" de jóvenes con bajo nivel socioeconómico y acceso limitado a tecnología que, sin embargo, logran el mejor desempeño académico, indicando alta resiliencia. En contraste, el "Clúster 1" agrupó a estudiantes jóvenes con altas condiciones socioeconómicas y tecnológicas, pero con bajo rendimiento, sugiriendo posibles falencias motivacionales o un uso no productivo de los recursos. Estos perfiles permiten al Ministerio considerar estrategias focalizadas que atiendan las necesidades y potencien las fortalezas de cada grupo.

- ¿Se puede cuantificar todas las variables que influyen en el puntaje global de los estudiantes obtenidos en las pruebas?

No es factible cuantificar e incorporar la totalidad de las variables que impactan el desempeño académico en un modelo predictivo. Muchos factores cruciales son inherentemente difíciles de medir a gran escala, como la motivación intrínseca del

estudiante, la calidad específica de la pedagogía docente en el aula, o variables de bienestar personal como las horas de estudio efectivo y de sueño. Estos elementos, junto con otros constructos psicológicos y variables contextuales no siempre disponibles, constituyen una parte significativa de la variabilidad en los resultados que los modelos basados en datos observables limitados no pueden capturar.

Por consiguiente, la porción de la varianza en los puntajes que el modelo no explica (un porcentaje relevante en este caso, dado el R^2 de 0.32) se atribuye en gran medida a la influencia de estas variables latentes o no medidas. Esto subraya que, si bien los modelos predictivos son valiosos para identificar factores socioeconómicos y contextuales relevantes, siempre existirán limitaciones inherentes a su capacidad para predecir con exactitud el rendimiento individual debido a la complejidad y multidimensionalidad del fenómeno educativo.

- ¿Cómo afecta la disponibilidad de libros en el hogar en los puntajes globales de los estudiantes?

Se revela una asociación positiva y estadísticamente significativa entre la cantidad de libros reportados en el hogar y los puntajes globales obtenidos por los estudiantes. Como se observa en la Figura 3, que ilustra la distribución del puntaje global según el número estimado de libros en casa, existe una tendencia creciente clara: a medida que aumenta la categoría de libros disponibles, también lo hace la mediana del puntaje global.

Específicamente, los estudiantes que reportan tener una menor cantidad de libros (entre 0 y 10) tienden a presentar una mediana de puntaje más baja y una mayor concentración de sus resultados en los rangos inferiores de la distribución

11. Conclusión

El presente proyecto permitió identificar y analizar, mediante técnicas de estadística descriptiva, inferencial y aprendizaje de máquina, los principales factores asociados al desempeño académico de los estudiantes en las pruebas Saber 11. A partir del procesamiento de más de cuatro millones de registros provenientes del ICFES entre 2017 y 2024, se evidenció que las condiciones socioeconómicas del hogar, el acceso a recursos tecnológicos, y las características institucionales del colegio tienen un impacto significativo en el puntaje global obtenido por los estudiantes.

El análisis ANOVA confirmó que variables como el estrato socioeconómico, el nivel educativo de los padres, la tenencia de computador o internet, y el tipo de colegio presentan diferencias estadísticamente significativas en los resultados académicos. Asimismo, los modelos predictivos implementados, aunque con capacidad mala ($R^2 \approx 0.32$), demostraron que una parte importante del rendimiento puede explicarse con base en estas condiciones estructurales. Complementariamente, el uso de técnicas de clustering permitió segmentar a los estudiantes en grupos con perfiles académicos diferenciados, facilitando una comprensión más profunda de las desigualdades educativas.

En conjunto, estos hallazgos proporcionan una base sólida para que el Ministerio de Educación pueda diseñar intervenciones focalizadas, mejorar la asignación de recursos y desarrollar políticas públicas más equitativas. Se concluye que el análisis de datos masivos no solo permite describir el estado actual del sistema educativo, sino también anticipar escenarios y orientar decisiones estratégicas para reducir brechas y promover una educación de mayor calidad y equidad en Colombia.

Referencias

- Instituto Colombiano para la Evaluación de la Educación – ICFES. (2017–2024). *Bases de datos del examen Saber 11*. Recuperado de https://www.icfes.gov.co/resultados/bases-de-datos
- Romero, C., & Ventura, S. (2020). *Educational Data Mining and Learning Analytics: An updated survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355
- Baker, R. S. (2019). *Challenges for the future of educational data mining: The Baker Learning Analytics Prizes*. Journal of Educational Data Mining, 11(1), 1–17.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.