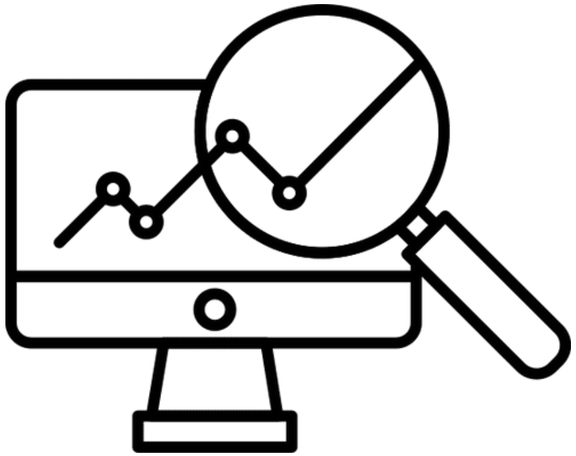
PROYECTO FINAL
PROCESAMIENTO DE DATOS A GRAN ESCALA

Juan Andres López Escalante

C O N T E N I D O S

C O N T E X T U A L I Z A C I Ó N

P R U E B A S S A B E R 1 1

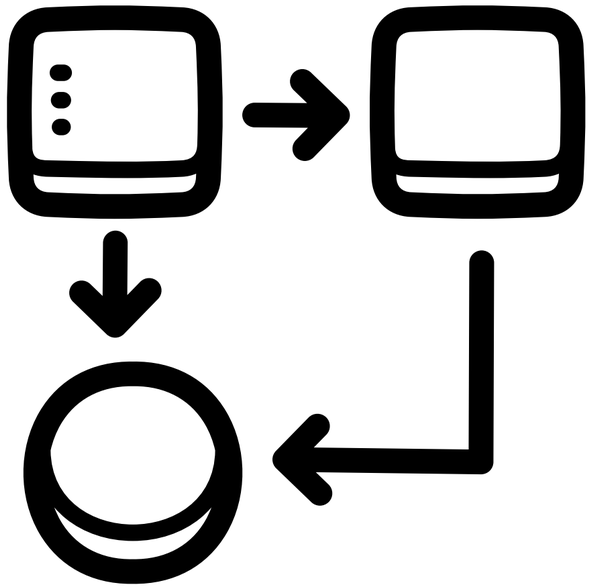


- Las pruebas Saber 11 son **exámenes estandarizados** que evalúan el nivel de competencias y conocimientos de los estudiantes que están por finalizar el grado undécimo de la educación media en **Colombia**.

Son administradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES) y son un requisito para obtener el título de bachiller.

PRESENTACIÓN DEL PROBLEMA

P R O B L E M A



- Colombia presenta disparidades tanto en el **rendimiento** del examen Saber 11 como en la cobertura de **internet fijo, el nivel académico de los padres**, además de **desigualdades** en diversas variables sociodemográficas.

Por ello, el Gobierno busca cuantificar cómo estas condiciones impactan los resultados educativos, para así **orientar** de manera más efectiva la asignación de recursos e inversiones.

SELECCIÓN DE LOS DATOS

SELECCIÓN DE DATOS

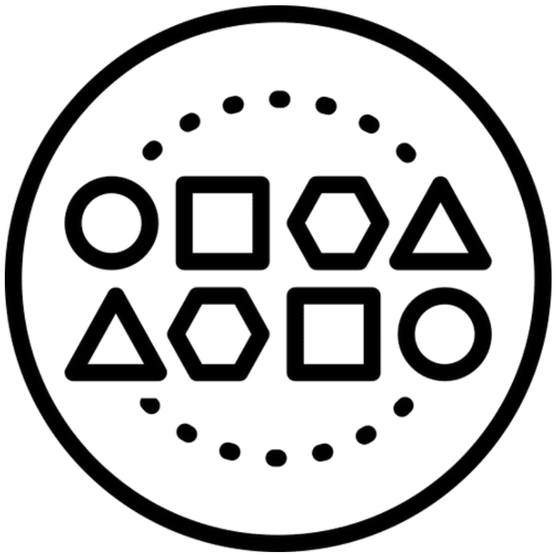
- Se identificó que las bases de datos del ICFES son de acceso público.

Por ello optamos por trabajar directamente con los archivos en crudo, ya que contienen la **información** detallada sobre cada **municipio** y la **conectividad** de los estudiantes.

Se opto por emplear las bases de datos de aquellos estudiantes de calendario A que presentaron la prueba entre los años 2017 - 2024.

S E L E C C I Ó N D E D A T O S

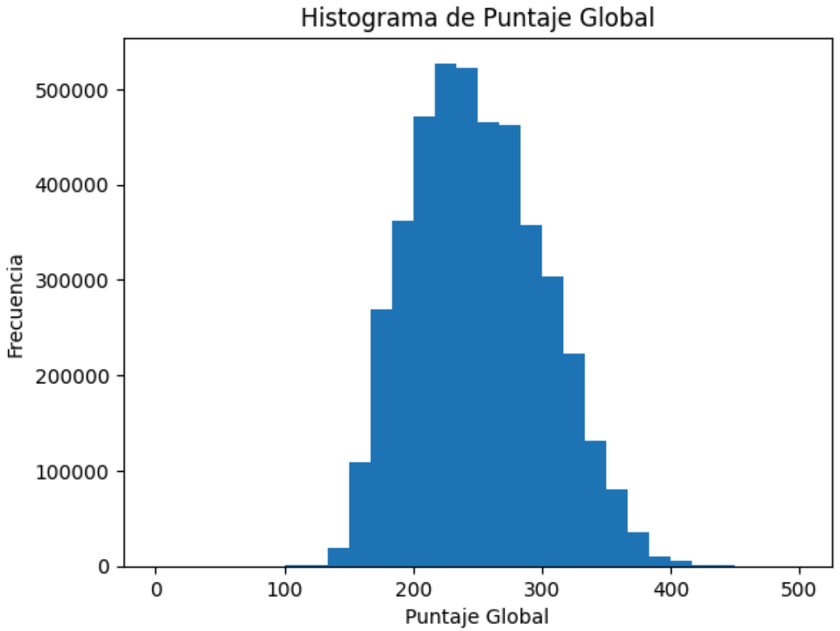
- Una vez importados los distintos archivos, se consolidaron en un solo conjunto de datos compuesto por **85** columnas y **4.761.554** registros.



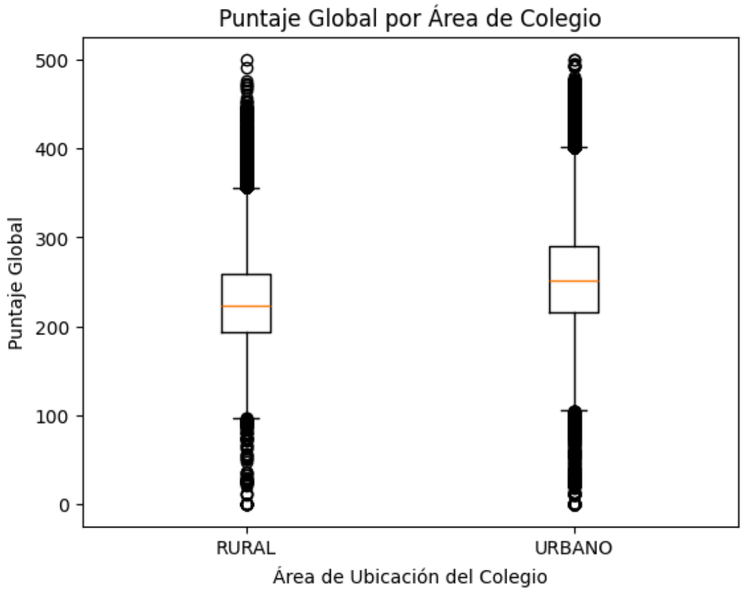
Variable	Descripción	Justificación
periodo_examen	Código de año y semestre de presentación (e.g. 20172 = 2ª aplicación 2017)	Permite agrupar y comparar resultados por cohorte temporal, seguir tendencias y controlar efectos de cada año/semestre.
calendario_colegio	Calendario académico del establecimiento (A, B, OTRO)	Filtra la segunda aplicación (calendario A) y controla posibles diferencias de cohortes 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 socio-político; 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.

EXPLORACIÓN DE LOS DATOS

A N Á L I S I S D E S C R I P T I V O

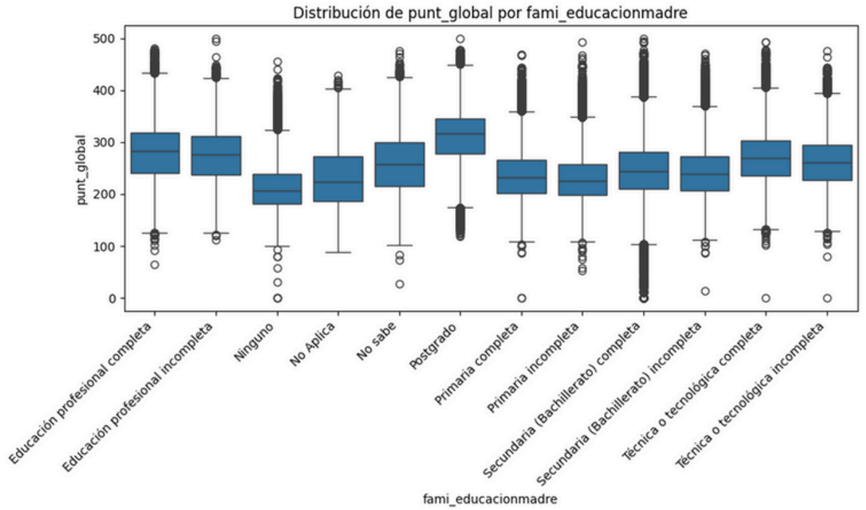


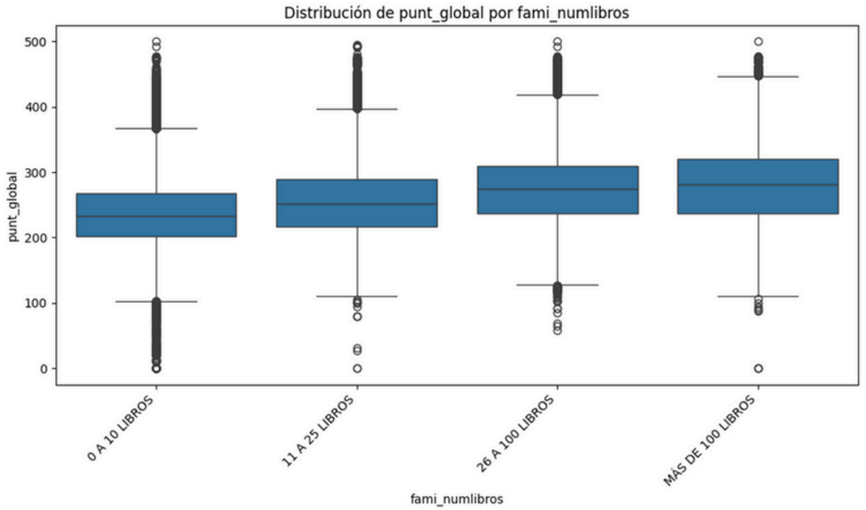
El puntaje global tiene una distribución **relativamente normal**; la cola superior sugiere un **subgrupo** destacado de **alumnos** de **muy alto desempeño**.

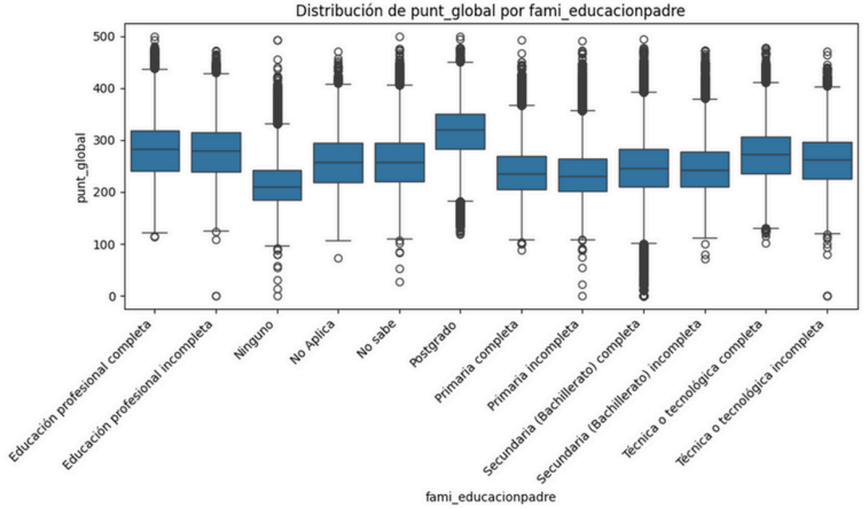


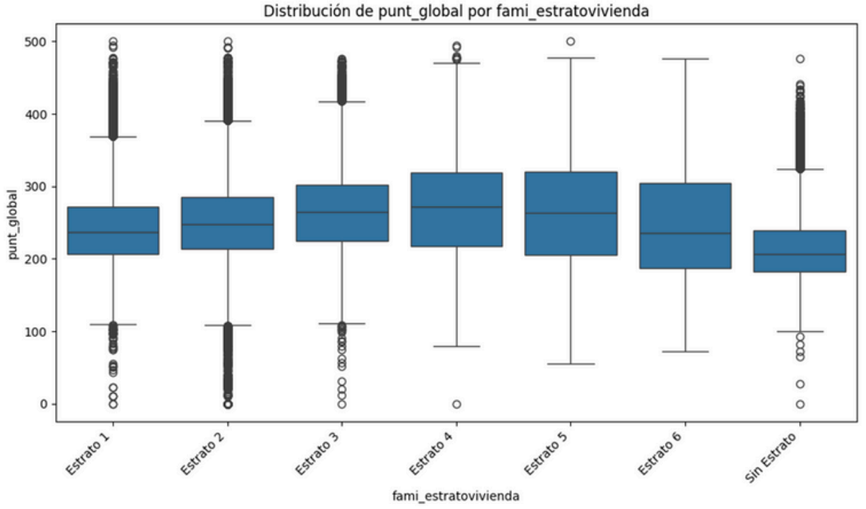
Existe una brecha **promedio** de 30 puntos a favor de los colegios **urbanos**; además, la variabilidad en contextos urbanos es mayor. Asi como una presencia de altos outliers en los datos.

A N Á L I S I S D E S C R I P T I V O









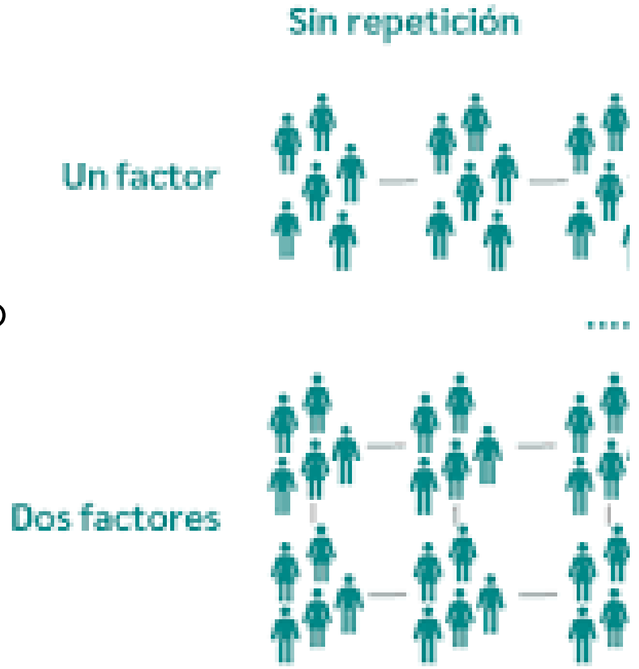
ANÁLISIS DESCRIPTIVO

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

PREPARACIÓN DE LOS DATOS

A N Á L I S I S I N F E R E N C I A L

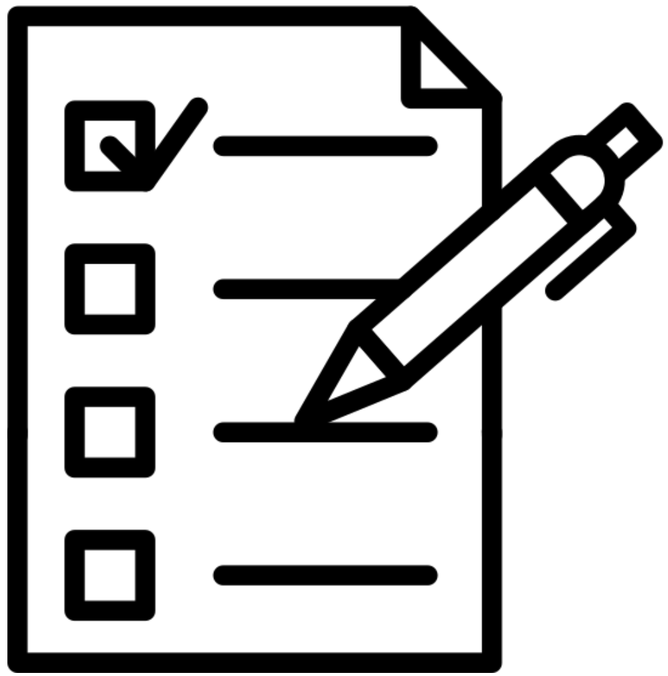
Con el objetivo de evaluar si existen diferencias estadísticamente significativas en el puntaje global entre los grupos definidos por variables categóricas, se aplicó una prueba **ANOVA** unidireccional sobre características del estudiante, el contexto familiar y atributos del colegio.



Los resultados indicaron que, en todos los casos evaluados (exceptuando cole_calendario), el valor p fue menor a 0.05, lo que permitió rechazar la hipótesis nula de igualdad de medias. Algunas variables mostraron una relación especialmente fuerte con el desempeño académico, como **fami_tienecomputador, fami_tieneinternet y fami_educacionmadre**, evidenciando una influencia directa del entorno familiar y tecnológico sobre los resultados de las pruebas saber 11.

RESULTADOS: HALLAZGOS PUROS

Filtrado por **calendario** A y **segunda aplicación** del año, se hizo la transformación de la variable fecha de nacimiento en **Edad** del estudiante.



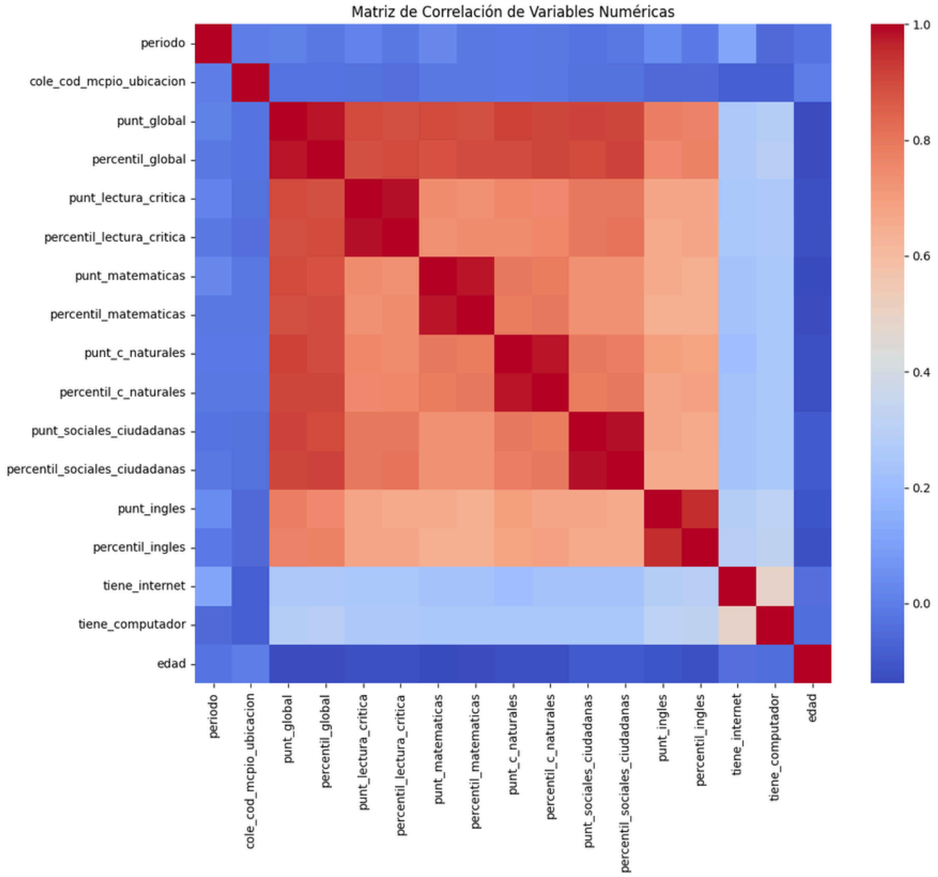
Se realizó un proceso de imputación donde los valores faltantes en variables **numéricas** fueron completados con la **media del municipio** y en su defecto, con la media global.

Para las variables **categóricas**, se utilizó la **moda agrupada por colegio**, preservando así la coherencia contextual.

Este pipeline permitió obtener un dataset limpio y optimizado con **4.357.399 registros y 37 columnas**, listo para los análisis descriptivo y predictivo.

ANÁLISIS PREDICTIVO

SELECCION DE VARIABLES



- Las variables como los puntajes por área (matemáticas, lectura, etc.) son componentes directos del **puntaje global**. Incluirlas como predictoras en modelos de aprendizaje automático provocaría sesgo artificial y sobreajuste.

Excluir estas variables para garantizar que las predicciones se basen únicamente en condiciones externas al examen (hogar, estudiante, colegio).

S E L E C C I O N D E V A R I A B L E S

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.

M O D E L A D O

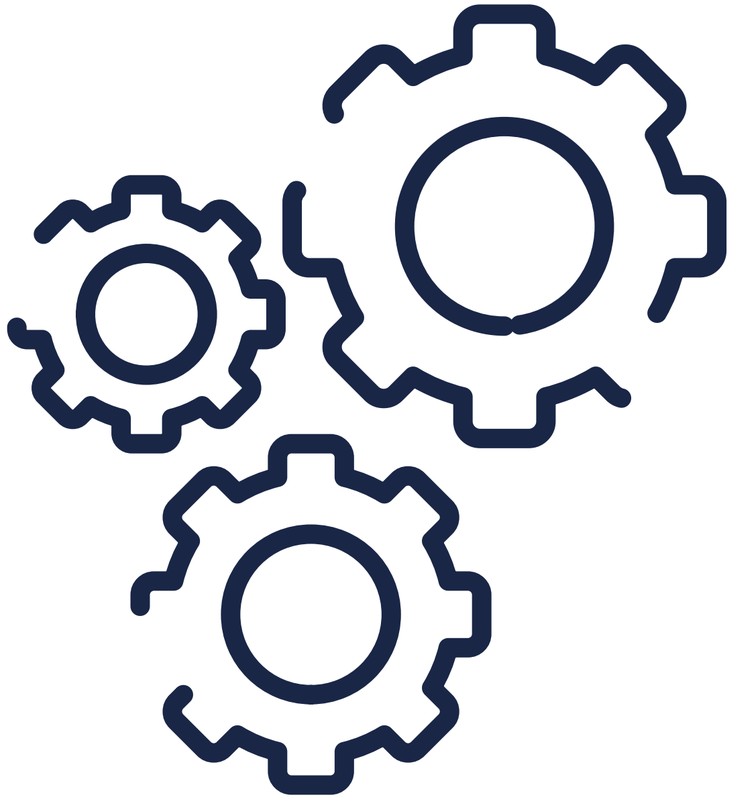
Se opto por utilizar los siguientes modelos:

Aprendizaje supervisado:

- **Regularized Regresor**

Aprendizaje no supervisado:

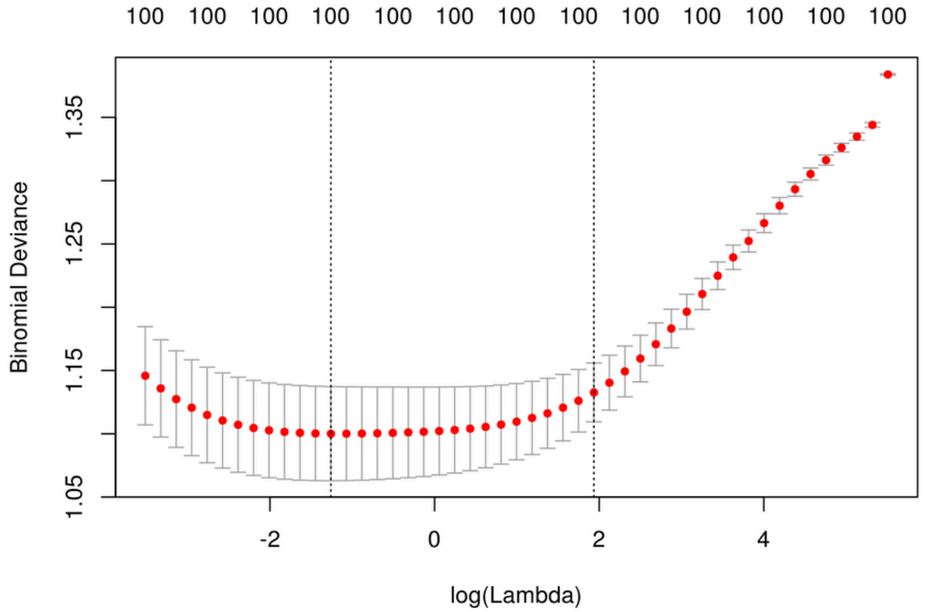
- **Kmeans Cluestering**



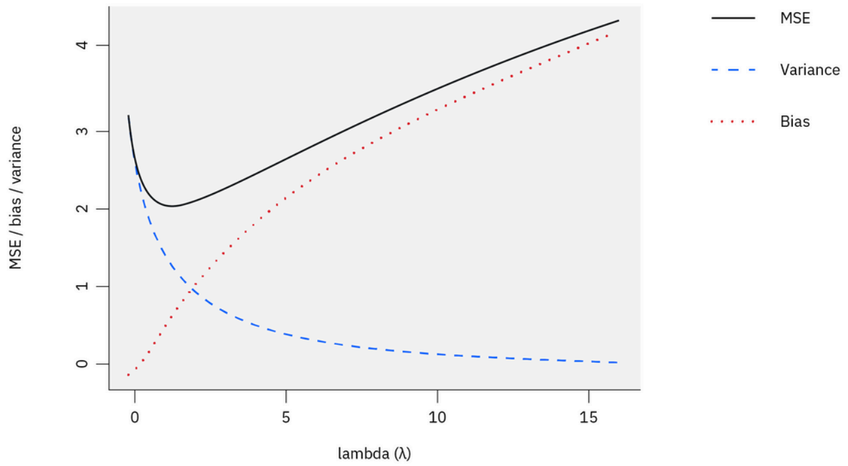
REGRESIÓN REGULARIZADA

La regresión regularizada, particularmente en sus variantes **Ridge** y **Lasso**, 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.



REGRESIÓN REGULARIZADA



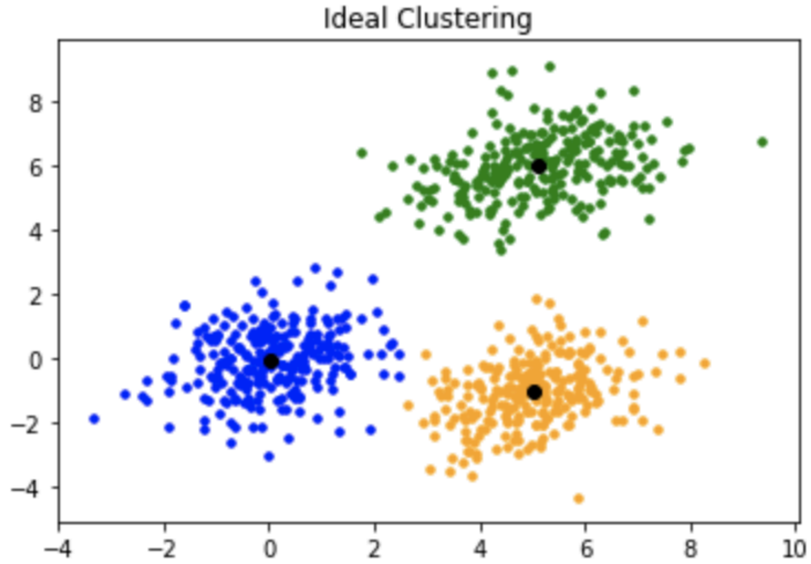
Ridge alcanzó su mejor desempeño con un valor de penalización óptimo de α = 0.05. Bajo este parámetro, se obtuvo un error cuadrático medio **(RMSE)** de 41.23 en el conjunto de entrenamiento y 42.09 en el conjunto de prueba. Junto con un **(R^2)** de 0.32 en **ambos casos.**

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.**

K M E A N S C L U S T E R I N G

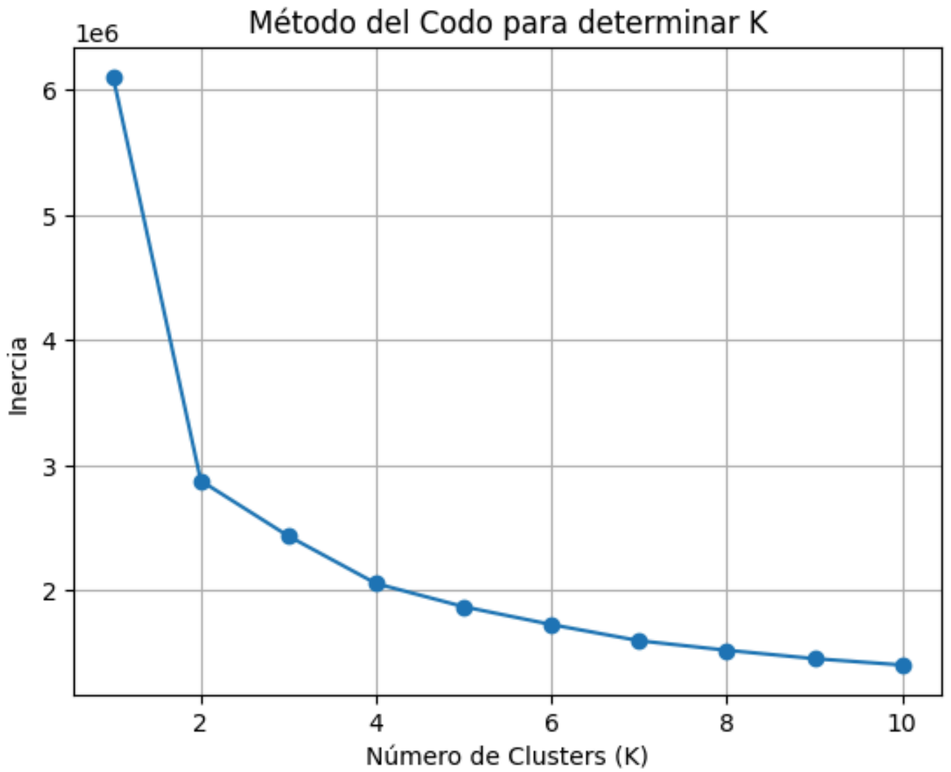
Método de aprendizaje no supervisado

Con el fin de **identificar patrones latente**s 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.

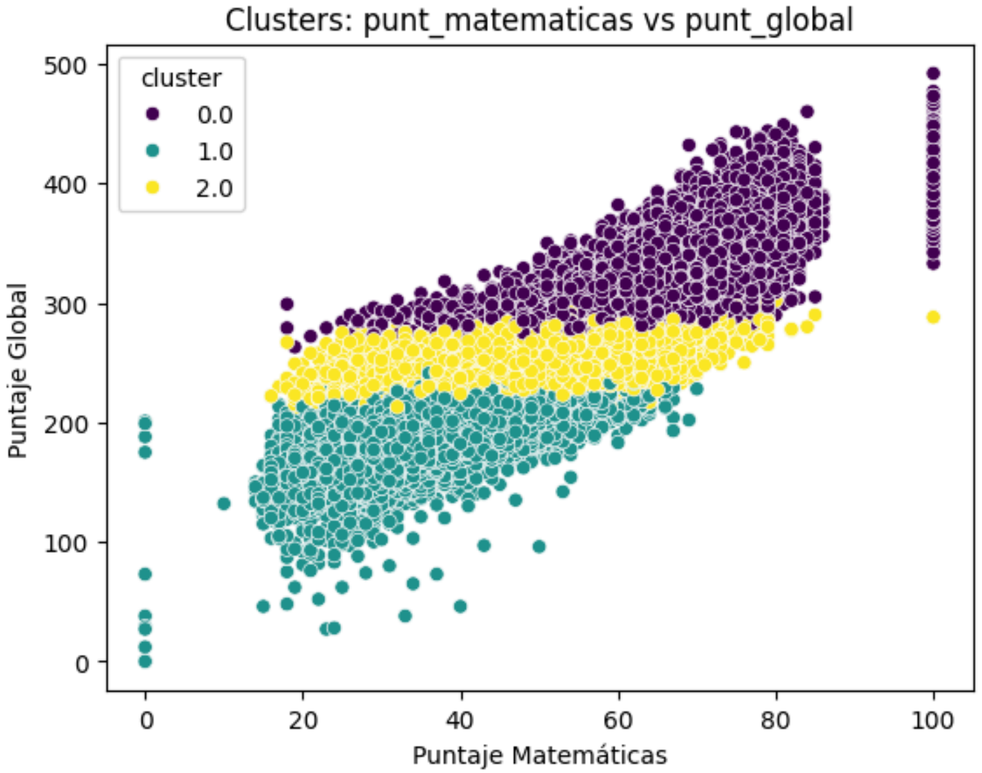


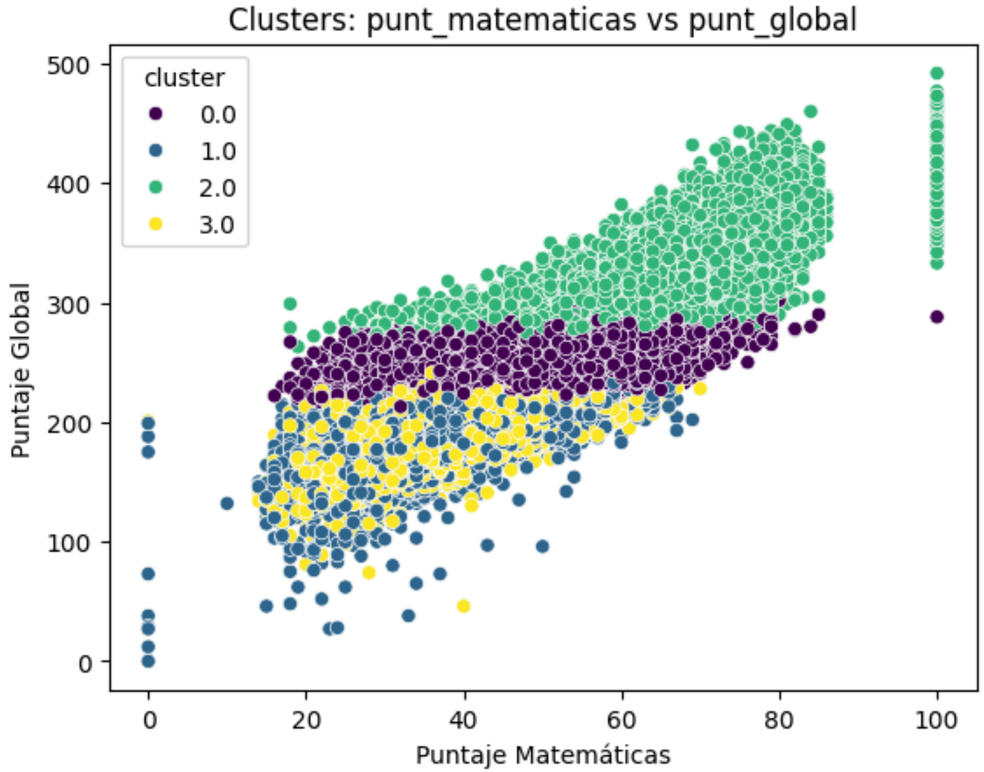
K M E A N S C L U S T E R I N G

Se aplicó el método del codo para determinar el **número** óptimo de clústeres (K) a utilizar en el algoritmo de agrupamiento, con el objetivo de identificar la cantidad adecuada de grupos que mejor representen la estructura subyacente de los datos.



K M E A N S C L U S T E R I N G





KMEANS CLUSTERING

3 CLUSTERS

- Clúster 0: NSE alto, buena conectividad, padres con educación alta.
- Clúster 1: NSE bajo, baja conectividad, más discapacidad reportada.
- Clúster 2: Perfil intermedio (estrato medio, acceso y rendimiento promedio).

4 CLUSTERS

- Clúster 0: Adultos mayores a 25 años con desempeño promedio-alto, bajos recursos.
- Clúster 1: Jóvenes con altos recursos pero bajo rendimiento ("paradoja").
- Clúster 2: Jóvenes en condiciones adversas, con mejor rendimiento.
- Clúster 3: Perfil promedio en contexto y desempeño.

PREGUNTAS DE NEGOCIO

¿Existe una brecha de rendimiento según nivel socioeconómico?

- Sí, la ANOVA y el boxplot muestran diferencias significativas en punt_global según estrato.
- El puntaje promedio aumenta con el estrato, pero disminuye levemente a partir del estrato 4.
- La relación no es lineal: podrían influir otros factores no observados.

¿Qué factores del entorno educativo influyen en el rendimiento?

- Variables como naturaleza del colegio, jornada, carácter y ubicación muestran diferencias significativas.
- Mejor desempeño en estudiantes de colegios privados, urbanos, diurnos y de carácter académico.
- Estas condiciones reflejan diferencias en recursos, infraestructura y entorno institucional.

PREGUNTAS DE NEGOCIO

¿Qué tan bien puede predecirse el puntaje global?

- Modelo Ridge: RMSE ≈ 41–42, R^2 = 0.32 (32% de la varianza explicada).
- El modelo tiene capacidad predictiva mala.
- Las condiciones del estudiante son relevantes, pero hay factores externos no observados que también influyen.

¿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?

- ANOVA: F = 259,218.44, $p < 0.05$ (significativo).
- Sí hay diferencias en puntaje según acceso a internet.
- Estudiantes con internet tienden a obtener mejores resultados

P R E G U N T A S D E N E G O C I O

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

- ANOVA: Educación madre $F = 57{,}617.07$, $p < 0.05$.
- Mayor educación de padres se ve representado como mejores puntajes en los estudiantes.

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

- Sí. Clustering (K=4) permitió segmentar perfiles clave.
- Ej: Clúster 2 = alta resiliencia; Clúster 1 = bajos resultados con altos recursos.
- Estos grupos permiten diseñar intervenciones focalizadas.

PREGUNTAS DE NEGOCIO

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

- No. $R^2 = 0.32$ → muchos factores clave no están en los datos.
- Faltan variables como motivación, hábitos, calidad docente.
- Los modelos ayudan, pero tienen limitaciones estructurales.

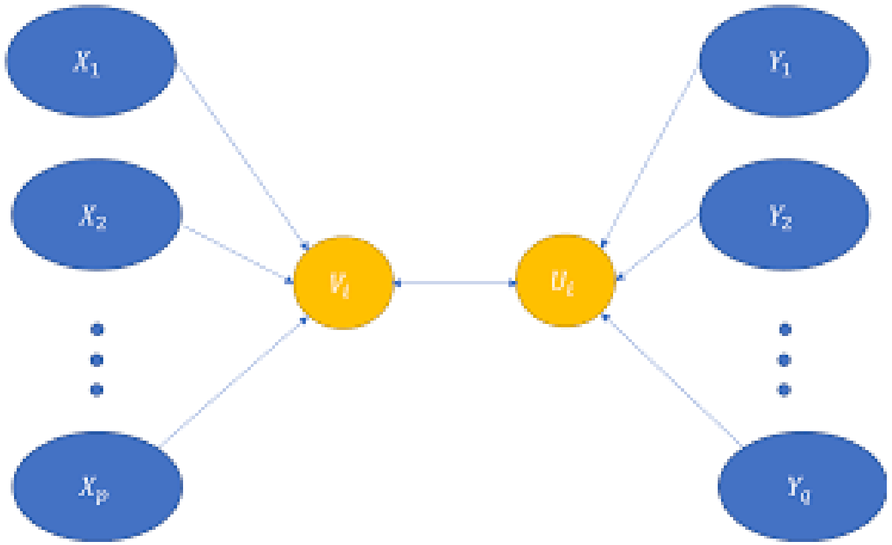
¿Cómo afecta la disponibilidad de libros en el hogar en los puntajes globales de los estudiantes?

- Sí. Tendencia creciente clara en mediana del puntaje.
- Más libros tienen una tendencia a un mejor desempeño global en la prueba.
- Refleja el valor del capital cultural del hogar.

R E C O M E N D A C I O N E S

- Dado que una porción significativa de la varianza en el puntaje_global permanece sin explicar, lo que sugiere la influencia de variables latentes o no medidas (como la **calidad docente**, la **motivación estudiantil** o **hábitos de estudio detallados**), se recomienda para futuras investigaciones el uso de técnicas multivariadas más avanzadas como el Análisis de Correlación Canónica (ACC).

Este análisis permitiría examinar la relación, no solo con el puntaje global, sino entre un conjunto de variables predictoras (socioeconómicas, familiares, de contexto escolar) y un conjunto de variables de resultado (los puntajes desagregados por área: matemáticas, lectura crítica, ciencias, sociales, inglés).



C O N C L U S I O N E S

- Las condiciones socioeconómicas, el acceso a tecnología y las características del colegio influyen significativamente en el rendimiento académico.

- El modelo predictivo explica cerca del 32% de la variabilidad en el puntaje global.

- El clustering permitió identificar perfiles estudiantiles diferenciados.

- A raíz de estos hallazgos se puede **focalizar** a los estudiantes con mas **opciones** de **mejora** a través de **políticas educativas**

R E F E R E N C I A S

- 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.

G R A C I A S