

Proyecto Final: Procesamiento de Datos

**CLASE:** Procesamiento de Datos

**PROFESOR:** John Corredor

**INTEGRANTES DEL EQUIPO:** Federico Quiroga, Ricardo Hurtado, Santiago Ramirez, Axel Caro

## Introducción

Los resultados del examen ICFES son un indicador clave de la calidad de la educación en el país y por ende es de gran utilidad saber las variables que influyen en los resultados del este, específicamente en el departamento de Cundinamarca. El objetivo de este proyecto definir estas variables más influyentes y proponer un plan de acción para influenciar positivamente en el ICFES de 2026. Se realizará esto con un enfoque analítico detallando todos los procesos que conlleva un análisis de este tipo, desde la limpieza hasta la creación de Modelos.

## Preparación del Conjunto de Datos para el Análisis del Puntaje ICFES

### Eliminación inicial de columnas

El conjunto de datos recolectado constaba de 83 variables, las cuales fueron depuradas y transformadas para garantizar la efectividad de los modelos predictivos aplicados. Los grupos de variables eliminadas fueron:

#### 1. Variables de identificación o administrativas

Ejemplos: ESTU\_TIPODOCUMENTO, ESTU\_CONSECUTIVO, ESTU\_FECHANACIMIENTO, PERIODO, COLE\_CODIGO\_ICFES, COLE\_COD\_DANE\_ESTABLECIMIENTO, COLE\_NOMBRE\_ESTABLECIMIENTO, COLE\_COD\_DANE\_SEDE, COLE\_NOMBRE\_SEDE

Razón de eliminación: estas variables corresponden a identificadores únicos o datos administrativos sin relación directa con el desempeño académico.

#### 2. Variables con alta granularidad geográfica

Ejemplos: ESTU\_PAIS\_RESIDE, ESTU\_COD\_RESIDE\_DEPTO, ESTU\_COD\_RESIDE\_MCPPIO, ESTU\_COD\_MCPPIO\_PRESENTACION, ESTU\_COD\_DEPTO\_PRESENTACION, COLE\_COD\_MCPPIO\_UBICACION, COLE\_COD\_DEPTO\_UBICACION

Razón de eliminación: se retuvieron únicamente los nombres del municipio y departamento de residencia para facilitar el análisis regional. Los códigos numéricos eran redundantes.

3. **Variables relacionadas con la presentación del examen**

Ejemplos: ESTU\_PRESENTACION SABADO,  
ESTU\_COD\_MCPIO\_PRESENTACION, ESTU\_MCPIO\_PRESENTACION,  
ESTU\_DEPTO\_PRESENTACION

Razón de eliminación: estas variables registran aspectos logísticos del examen sin impacto directo en el desempeño académico.

4. **Variables de desempeño en áreas individuales**

Ejemplos: PUNT\_LECTURA\_CRITICA, PUNT\_MATEMATICAS,  
PUNT\_C\_NATURALES, PUNT\_SOCIALES\_CIUDADANAS, PUNT\_INGLES y  
sus respectivos PERCENTIL y DESEMP

Razón de eliminación: se mantuvo únicamente PUNT\_GLOBAL para evitar multicolinealidad y centrar el análisis en el desempeño general.

Variables seleccionadas para el análisis

Después de eliminar las columnas mencionadas, quedaron las siguientes:

1. **Características del estudiante:**

ESTU\_GENERO, ESTU\_NACIONALIDAD, ESTU\_DEPTO\_RESIDE,  
ESTU\_MCPIO\_RESIDE

2. **Condiciones del hogar:**

FAMI\_ESTRATOVIVIENDA, FAMI\_EDUCACIONPADRE,  
FAMI\_EDUCACIONMADRE, FAMI\_TIENEINTERNET,  
FAMI\_TIENECOMPUTADOR, FAMI\_TIENESERVICIOTV,  
FAMI\_TIENECONSOLAVIDEOJUEGOS, FAMI\_NUMLIBROS,  
FAMI\_SITUACIONECONOMICA

3. **Hábitos de estudio:**

ESTU\_DEDICACIONLECTURADIARIA, ESTU\_DEDICACIONINTERNET,  
ESTU\_HORASSEMANA TRABAJA

4. **Características del colegio:**

COLE\_BILINGUE, COLE\_CALENDARIO, COLE\_AREA\_UBICACION

5. **Variable objetivo:**

PUNT\_GLOBAL

Imputación de valores nulos

Para asegurar la completitud del conjunto de datos antes del modelado, se llevaron a cabo procedimientos de imputación de valores nulos:

- Las variables categóricas fueron imputadas utilizando la moda, calculada dentro de los siguientes grupos definidos previamente:

```
columnas_grupo = [ 'ESTU_GENERO', 'ESTU_DEPTO_RESIDE',  
'ESTU_MCPPIO_RESIDE', 'COLE_DEPTO_UBICACION', 'COLE_MCPPIO_UBICACION',  
'COLE_NATURALEZA', 'COLE_CALEDARIO', 'ESTU_NACIONALIDAD' ]
```

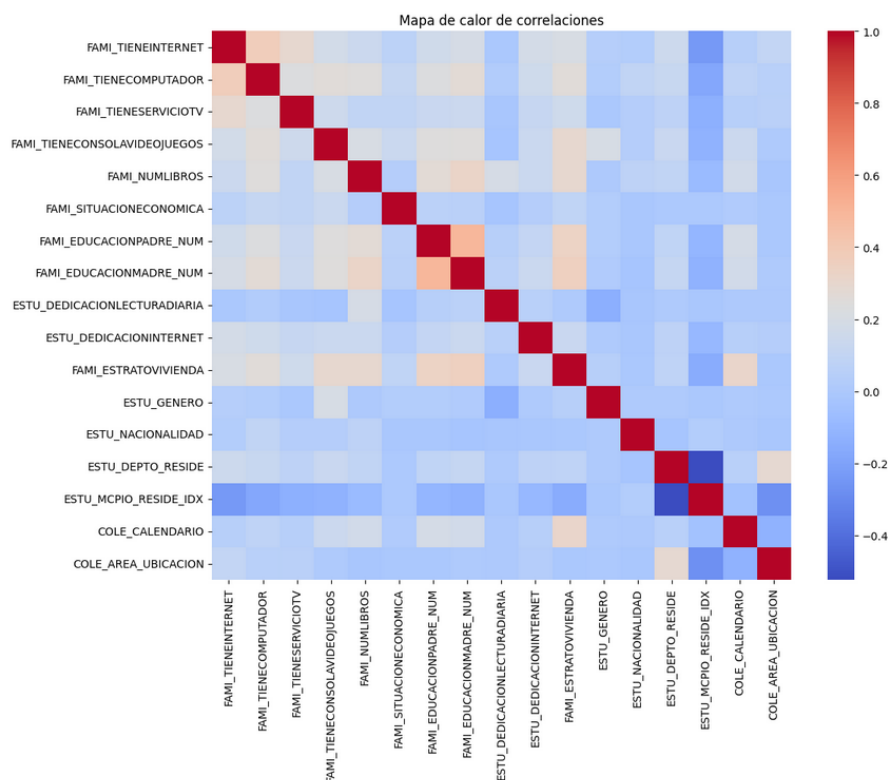
*Nota: esta imputación se realizó antes de eliminar algunas de las columnas mencionadas previamente.*

- Las variables numéricas con valores nulos fueron imputadas con su respectiva media.
- Los valores categóricos clasificados como "desconocido" fueron eliminados, ya que representaban un porcentaje muy bajo del total.

### Transformación de datos

- Todas las variables categóricas fueron transformadas a variables ordinales.
- Las columnas fueron convertidas al tipo `double` para asegurar su compatibilidad con los modelos predictivos.

### Normalización y revisión de correlación



Se elaboró una matriz de correlación para examinar relaciones entre variables. No se detectaron correlaciones fuertes, excepto una leve entre municipio y departamento:

	Feature1	Feature2	Correlation
130	ESTU_DEPTO_RESIDE	ESTU_MCPIO_RESIDE_IDX	-0.524174

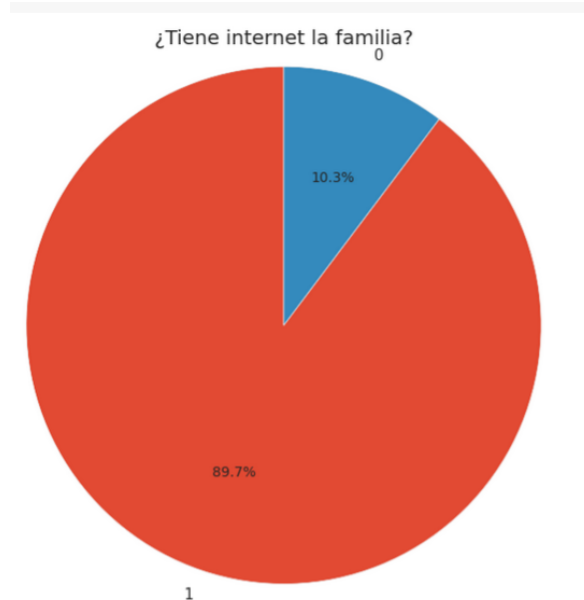
por lo cual se eliminó la variable de departamento.

Finalmente, se realizó una normalización de las variables numéricas usando `StandardScaler` de `pyspark` para unificar las escalas y facilitar la eficiencia del modelo.

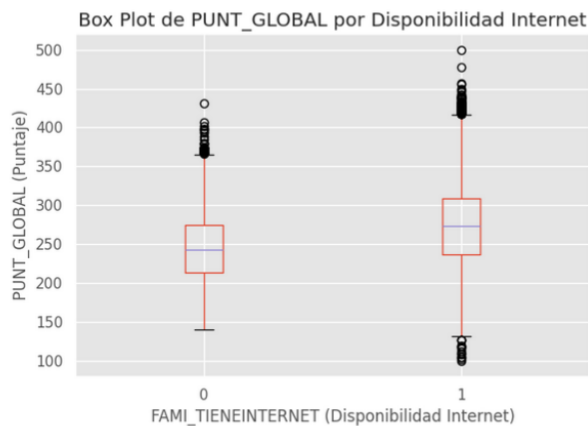
## Preguntas

*¿Existe una correlación entre el acceso a internet en el hogar y el puntaje global del ICFES?*

Se realizó un pie chart para identificar la proporción entre familias con y sin internet



89.7% de las personas tiene internet, se procede a revisar el efecto que tiene el internet en el puntaje con un boxplot:



Los estudiantes con acceso a internet (grupo 1) tienen: Mayor mediana del puntaje PUNT\_GLOBAL, Rango intercuartílico más alto (mayor dispersión positiva). Por otro lado, Los estudiantes sin acceso a internet (grupo 0) tienen: Una mediana claramente más baja. Distribución de puntajes más concentrada en rangos bajos. Por tanto SI, existe una correlación positiva entre el acceso a internet en el hogar y un mayor puntaje global en el ICFES.

*¿El tiempo dedicado al uso de internet influye en los resultados de las pruebas de matemáticas y ciencias naturales?*

El boxplot muestra la distribución del puntaje global obtenido por los estudiantes en relación con el tiempo que dedican diariamente al uso de internet. Se observa que los estudiantes que navegan en internet entre 1 y 3 horas, así como los que lo hacen por más de 3 horas, presentan medianas de puntaje más altas que aquellos que navegan por menos tiempo o que no lo hacen en absoluto. En contraste, quienes no utilizan internet o lo hacen por 30 minutos o menos tienen puntajes más bajos y distribuciones más concentradas en la parte inferior del gráfico. Además, hay una mayor cantidad de valores atípicos altos en los grupos con mayor uso de internet, lo que indica que algunos estudiantes con uso intensivo también alcanzan los puntajes más elevados.

Estos resultados sugieren una relación positiva entre el tiempo dedicado al uso de internet y el rendimiento académico global, al menos en términos de puntaje. Esta tendencia podría explicarse por el uso del internet como herramienta educativa, que permite el acceso a contenidos de apoyo, tutorías virtuales y plataformas de aprendizaje. Sin embargo, es importante señalar que esta relación no implica causalidad directa: factores como el propósito del uso del internet, el entorno socioeconómico y el acceso a

recursos educativos adicionales pueden estar influyendo también en los resultados.

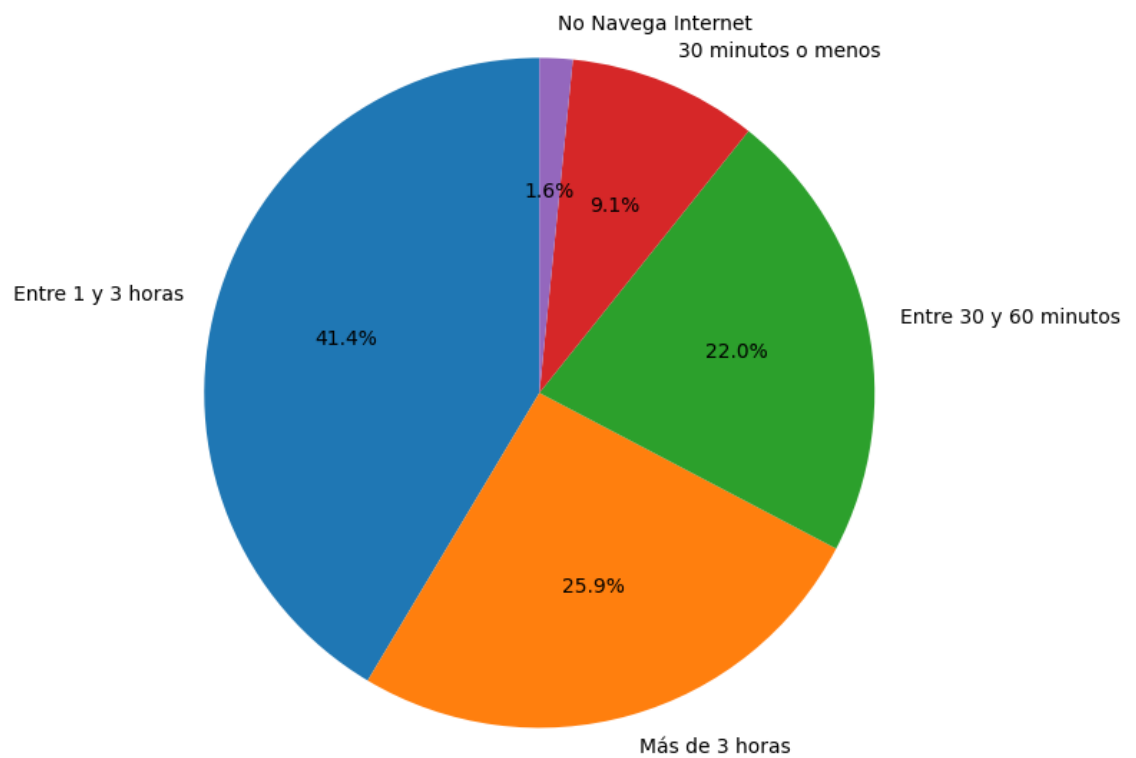
Distribución de la dedicación al uso de internet:

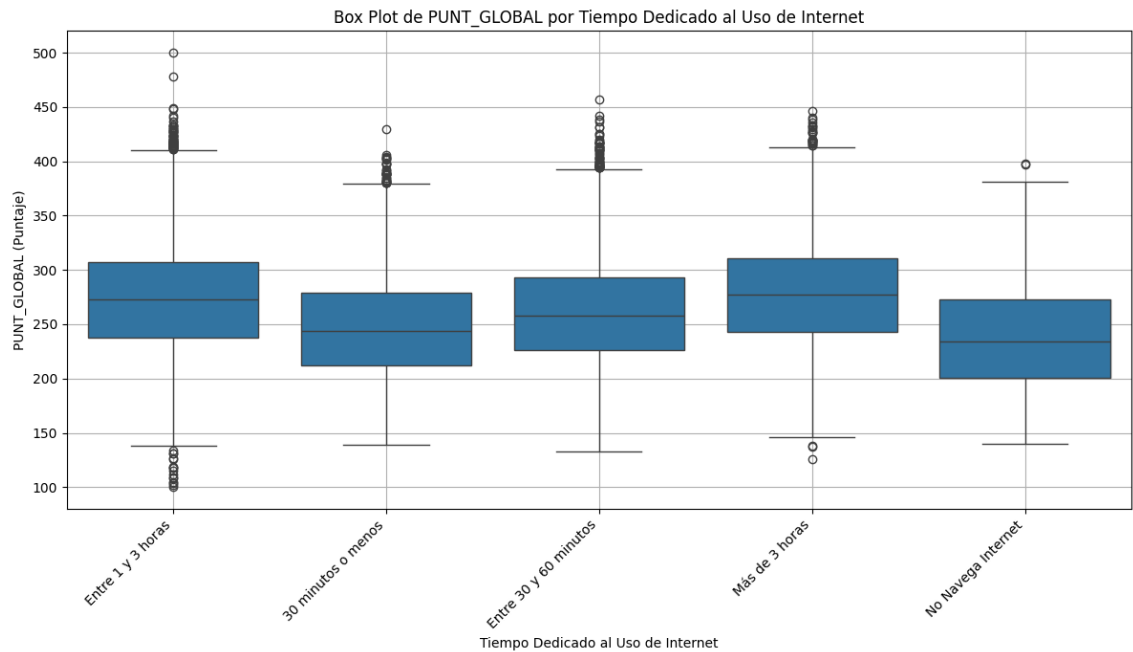
ESTU\_DEDICACIONINTERNET

Entre 1 y 3 horas	47565
Más de 3 horas	29686
Entre 30 y 60 minutos	25224
30 minutos o menos	10480
No Navega Internet	1827

Promedio de PUNT_GLOBAL por Tiempo Dedicado al Uso de Internet:	
ESTU_DEDICACIONINTERNET	
Más de 3 horas	277.179950
Entre 1 y 3 horas	273.180217
Entre 30 y 60 minutos	260.569180
30 minutos o menos	247.994943
No Navega Internet	239.059113

Distribución del Tiempo Dedicado al Uso de Internet





*¿Qué características socioeconómicas tienen en común los estudiantes con los puntajes más altos en el ICFES?*

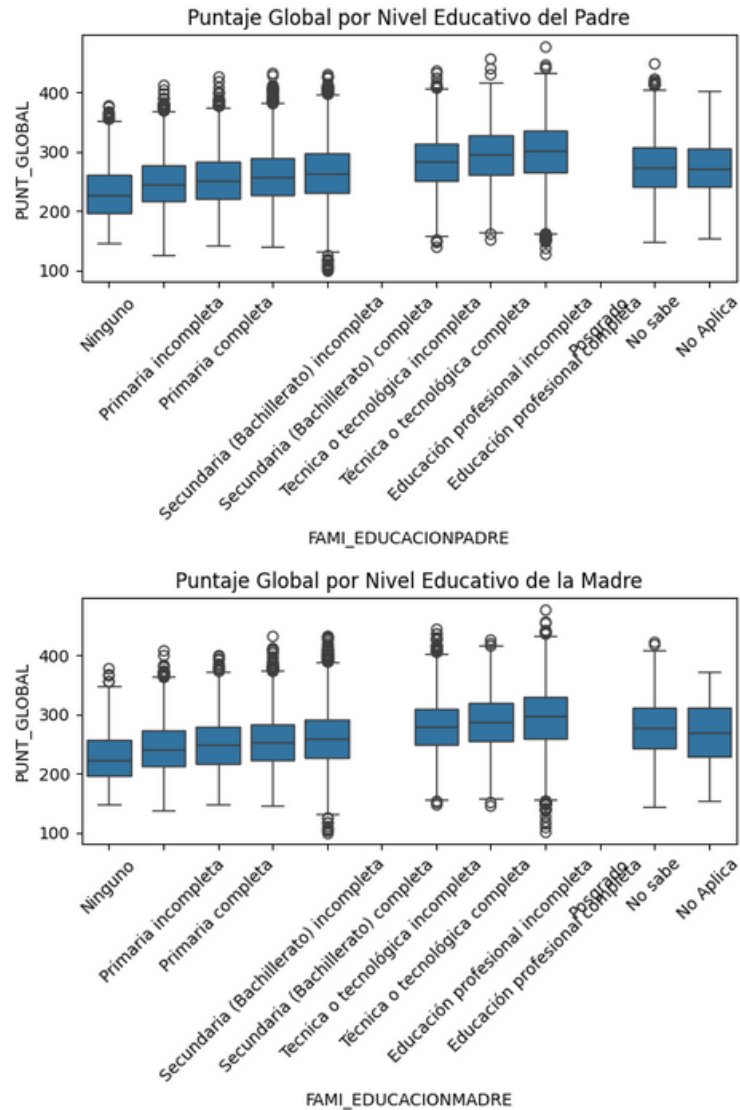
Los estudiantes con los puntajes más altos (top 10%) en el ICFES comparten varias características socioeconómicas y académicas. La mayoría son hombres (282 vs. 217 mujeres), viven en estratos altos (casi el 90% pertenece a los estratos 4, 5 o 6), y tienen padres con altos niveles educativos: más del 95% de los padres y madres tienen educación profesional o de postgrado. Todos cuentan con acceso a internet y computador en casa, y más del 90% reporta tener más de 25 libros. Además, la mayoría considera que su situación económica es igual o mejor que antes. Aunque más del 70% navega entre 1 y 3 horas al día, también es común que dediquen algo de tiempo a la lectura diaria. Casi ninguno trabaja y la totalidad proviene de colegios del calendario B, no bilingües, y mayoritariamente ubicados en zonas urbanas. Estos resultados sugieren una clara relación entre un entorno familiar con recursos, nivel educativo alto de los padres y mejores condiciones de estudio, con un mayor rendimiento académico.





¿Cómo influye el nivel educativo de los padres en el desempeño de los estudiantes?

Se realiza un boxplot para realizar una inspección grafica



Visualmente se puede notar que a medida que aumente el nivel de educación aumenta el puntaje, se realiza una regresión lineal para corroborar:

Coeficientes del modelo: [2.2893406563082945, 3.827038162245616]  
Intercepto: 233.60546985245736  
 $R^2$ : 0.1446184326163864  
RMSE: 45.76175775517613  
MAE: 37.20024870653175

Al realizar la regresión lineal se ve que el nivel educativo de los padres muestra una relación positiva con el puntaje del ICFES. Específicamente, por cada aumento en una categoría del nivel educativo del padre, el puntaje promedio del ICFES incrementa en aproximadamente **2.28 puntos**. En el caso de la madre, un aumento en su nivel educativo se asocia con un incremento promedio de **3.38 puntos** en el puntaje del estudiante.

Si embargo, el coeficiente de determinación del modelo ( $R^2 = 0.144$ , es decir, 14.4%) indica que estas variables explican solo una **porción limitada de la variabilidad** total en los resultados del ICFES. El nivel educativo de los padres influye en el desempeño académico, pero hay otros factores relevantes que afectan el puntaje también.

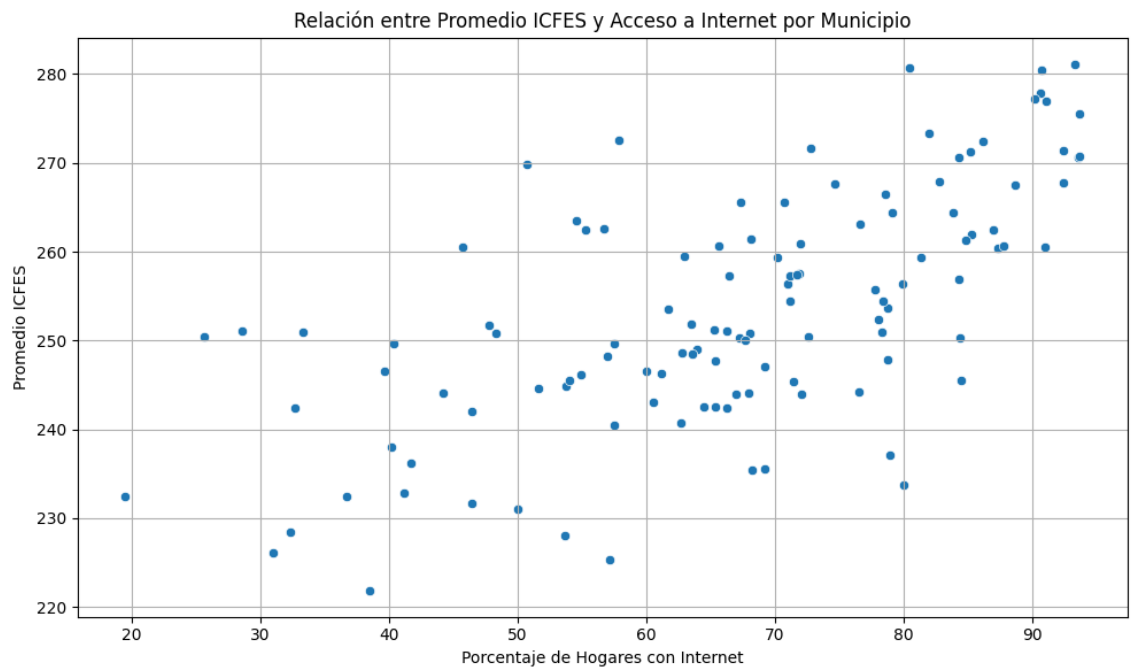
*¿Cuáles municipios presentan mejores y peores puntajes?*

Municipios con los promedios más bajos en el ICFES:	
ESTU_MCPIO_RESIDE	promedio_puntaje
BELTRÁN	221.85
ÚTICA	225.29
YACOPÍ	226.05
APULO	228.04
SAN CAYETANO	228.38
NARIÑO	230.97
JERUSALÉN	231.71
TOPAIPÍ	232.44
GUTIÉRREZ	232.49
EL PEÑÓN	232.84

Municipios con los promedios más altos en el ICFES:	
ESTU_MCPIO_RESIDE	promedio_puntaje
CAJICÁ	287.18
CHÍA	287.13
COTA	283.47
CHIPAQUE	280.75
SOPÓ	280.64
LA CALERA	278.69
ZIPAQUIRÁ	277.24
MOSQUERA	275.72
BOGOTÁ D.C.	273.66
FACATATIVÁ	272.59

Sí, existe una relación positiva clara entre el porcentaje de hogares con acceso a Internet y el desempeño promedio en el ICFES a nivel municipal. Los municipios más conectados presentan mejores resultados académicos en promedio.

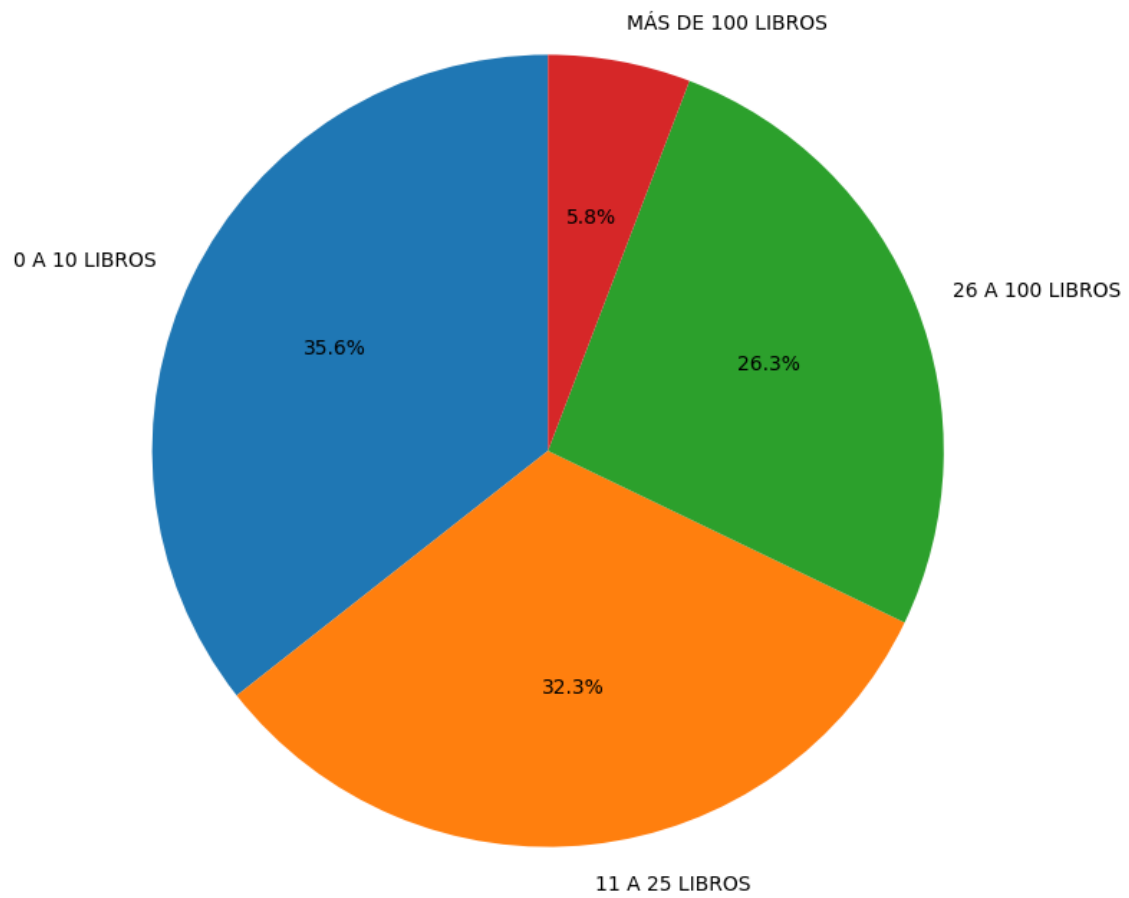


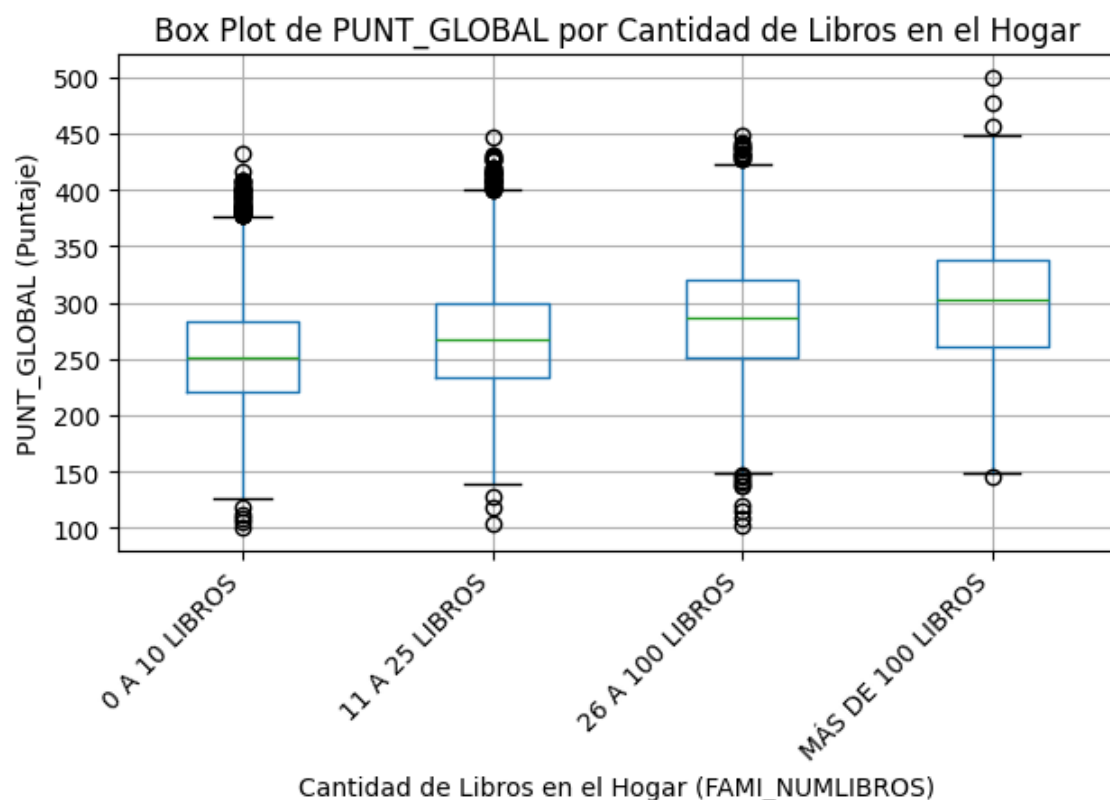
- Hay una clara correlación positiva: a mayor acceso a internet en el hogar, mayor es el puntaje promedio del ICFES en el municipio.
- Municipios con más del 80% de hogares con internet tienden a tener puntajes por encima de 260, incluso llegando a 280+.
- Variabilidad en municipios con bajo acceso:
- En municipios con menos del 50% de hogares conectados, los puntajes son más dispersos y generalmente más bajos (220–260).

### *¿Influye la cantidad de libros en casa en la lectura crítica?*

Sí, existe una relación clara y positiva entre la cantidad de libros en el hogar y el puntaje global en el ICFES (incluyendo lectura crítica). A medida que el número de libros aumenta, también lo hace el rendimiento académico del estudiante.

Distribución de Cantidad de Libros en el Hogar (FAMI\_NUMLIBROS)





Se uso un pie chart para segmentar la cantidad de libros en las familias de los estudiantes, esto ayudo a tener un mejor analisis y comparar con los promedios de los resultados del ICFES, se presenta los siguientes resultados:

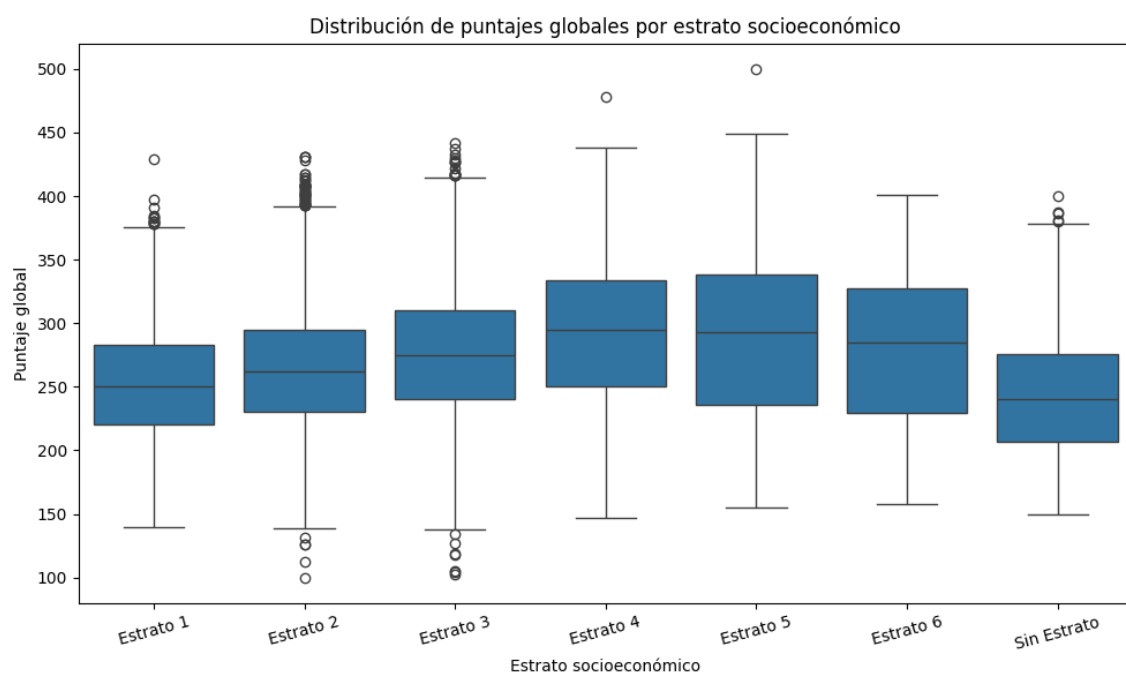
- Incremento claro y progresivo en la mediana del puntaje global del ICFES a medida que aumenta la cantidad de libros.
- La categoría de “Más de 100 libros” tiene la mediana más alta y los valores más consistentes en el rango alto.
- La categoría de “0 a 10 libros” tiene la mediana más baja y distribución más dispersa hacia puntajes bajos.
- Existe menor dispersión y más concentración de puntajes altos en los hogares con más libros.

*¿El estrato socioeconómico del estudiante tiene un impacto significativo en sus resultados académicos?*

Sí. Los datos muestran una relación clara entre el estrato socioeconómico y el puntaje global en las pruebas Saber 11. En promedio, los estudiantes de estrato 1 obtienen 252 puntos, mientras que los de estrato 4 alcanzan casi 291, una diferencia de cerca de 40 puntos. Esta tendencia es consistente: a mayor estrato, mayor puntaje promedio.

estrato	n_estudiantes	promedio_puntaje	desviacion_std
Estrato 1	12032	252.3031914893617	43.72174196468081
Estrato 2	47836	263.05351618028266	45.08957972320104
Estrato 3	42415	274.92985971943887	48.258004674376956
Estrato 4	8800	290.46488636363637	54.9380243137684
Estrato 5	1724	288.3897911832947	60.9972267541051
Estrato 6	568	276.34859154929575	61.1953961172675
Sin Estrato	1407	241.6453447050462	47.07558785896673

Esta grafica muestra los promedios de puntaje global por estrato, permite ver esta tendencia ascendente de forma clara.



Además, La grafica de boxplot refuerza el hallazgo: no solo la mediana aumenta progresivamente, sino que la dispersión del puntaje también crece en los estratos más altos, lo que sugiere una mayor diversidad de resultados dentro de esos grupos.

Por último, una prueba ANOVA confirmó que estas diferencias no son producto del azar. El resultado fue estadísticamente significativo, con un **estadístico F de 88.88** y un **p-valor de  $4.27 \times 10^{-109}$** , lo que indica que las diferencias en los puntajes entre estratos son altamente significativas.

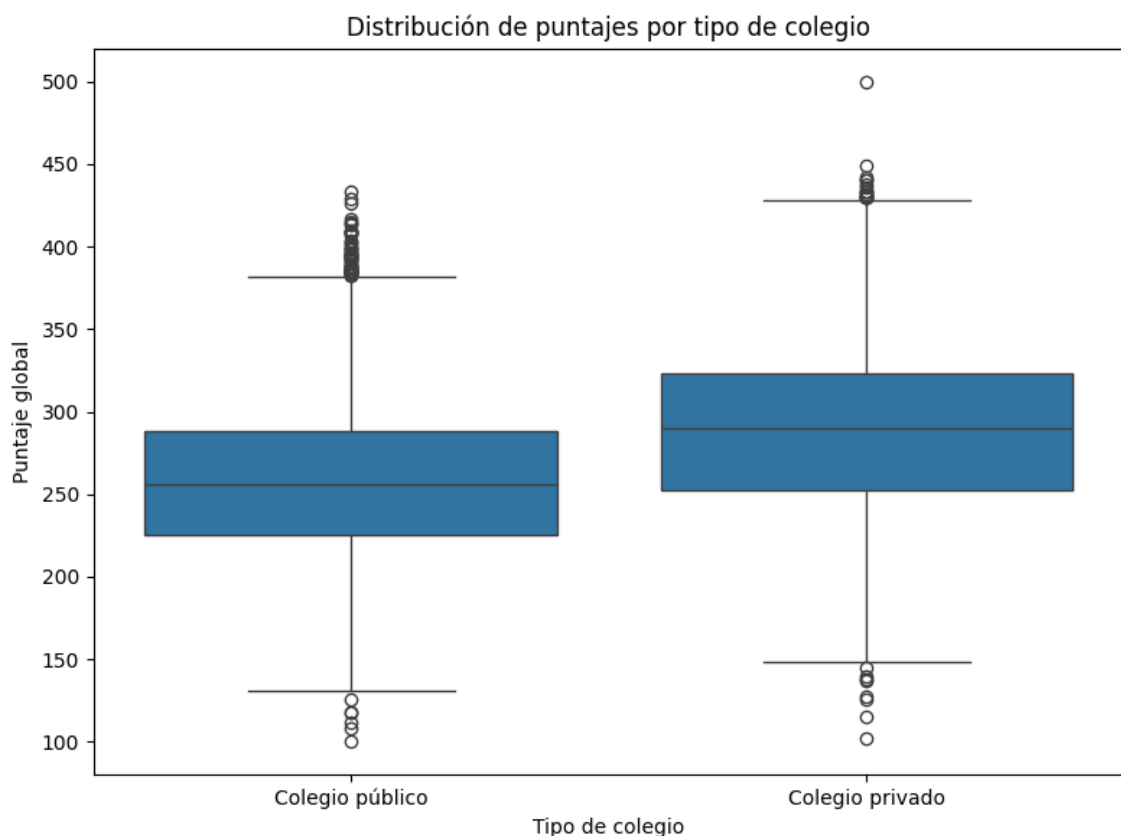
En resumen, los estudiantes de estratos más altos tienden a obtener mejores resultados en las pruebas Saber 11, lo que refleja una posible influencia directa de las condiciones socioeconómicas sobre el desempeño académico.

*¿Los estudiantes de colegios privados obtienen mejores puntajes en comparación con los de colegios públicos?*

Sí. Los resultados muestran que los estudiantes de colegios privados (identificados como “NO OFICIAL”) obtienen **puntajes globales significativamente más altos** que aquellos de colegios públicos (“OFICIAL”). En promedio, los estudiantes de colegios privados alcanzan 286.9 puntos, mientras que los de colegios públicos obtienen 256.8, una diferencia de aproximadamente **30 puntos**.

naturaleza	n_estudiantes	promedio_puntaje	desviacion_std
NO OFICIAL	44294	286.8781324784395	50.02500089720878
OFICIAL	70488	257.1153387810691	43.69847401979227

Este patrón se observa claramente en la **tabla**, que presenta los promedios de puntaje por tipo de colegio. La diferencia es consistente y evidente.



Además, la gráfica **boxplot** ilustra cómo los estudiantes de colegios privados no solo tienen una mediana más alta, sino también una mayor dispersión de puntajes, lo cual sugiere una diversidad más amplia de rendimiento dentro de ese grupo.

Para validar esta diferencia de manera estadística, se aplicó una prueba t de dos muestras independientes (Welch's t-test), obteniendo un **estadístico t de -77.15** y un **p-**



**valor de 0.0** (redondeado). Esto indica que la diferencia en los puntajes es **estadísticamente significativa**, es decir, no se debe al azar.

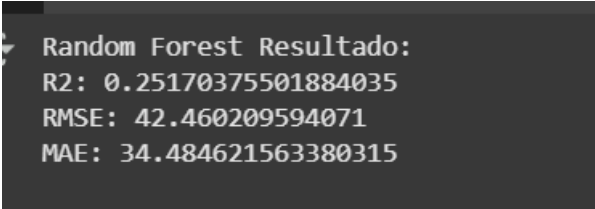
En conclusión, **los estudiantes de colegios privados obtienen, en promedio, mejores puntajes en las pruebas Saber 11 que los estudiantes de colegios públicos**. Esta diferencia puede estar relacionada con factores como mayores recursos institucionales, mejores condiciones de estudio y características socioeconómicas del entorno educativo.

### Creación modelo Predictivo

Se seleccionaron la regresión lineal múltiple y el algoritmo de *k-means* como modelos predictivos para llevar a cabo el análisis. Además, el proceso se desarrolló en dos entornos distintos: Google Colab, que presenta limitaciones para el manejo de grandes volúmenes de datos, y el clúster creado específicamente para esta asignatura, el cual está optimizado para el procesamiento eficiente de datos a gran escala

### Modelo Regresión Lineal

Se creó un modelo ajustado que buscaba mejorar los resultados ajustando los parámetros y optimizando el proceso, a continuación, se presentan resultados y el cómo se hizo:



```
Random Forest Resultado:  
R2: 0.25170375501884035  
RMSE: 42.460209594071  
MAE: 34.484621563380315
```

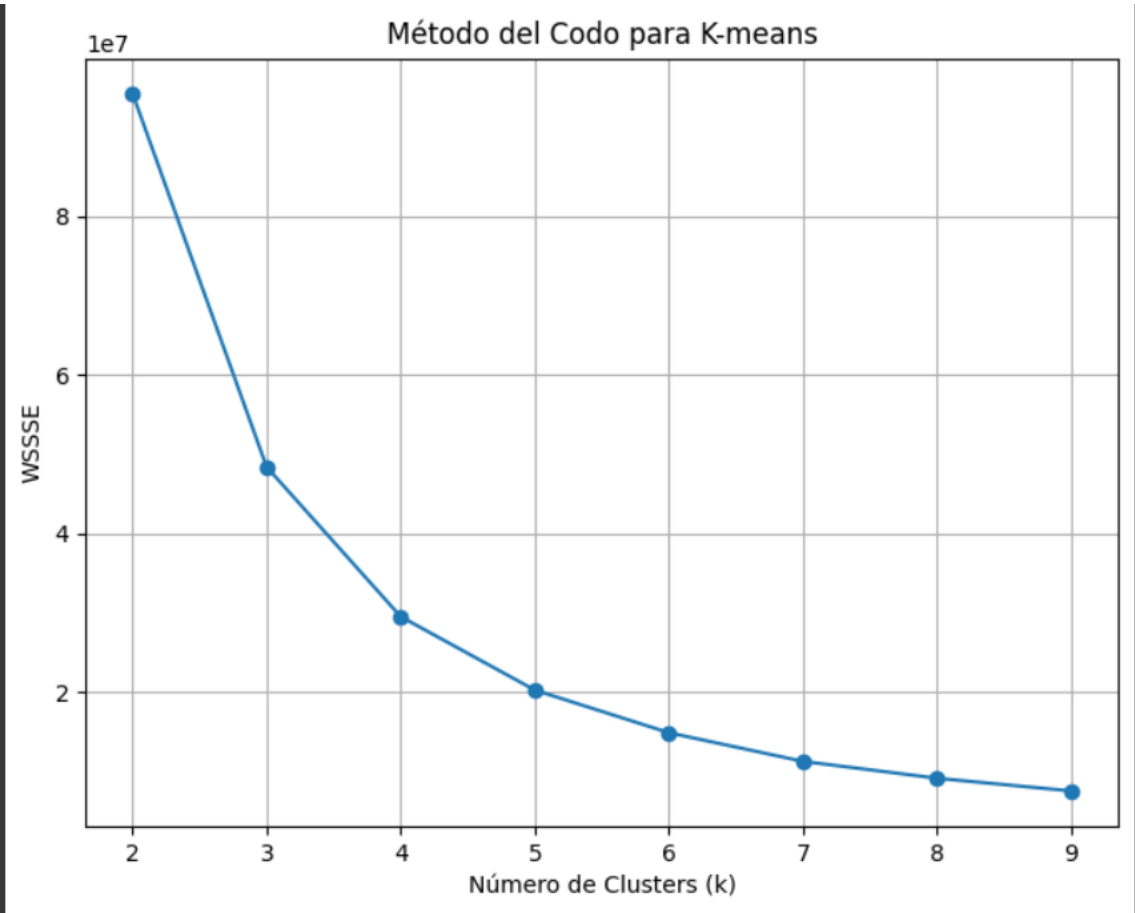
Comparación con el Modelo Normal (no ajustado):

**R<sup>2</sup>:** El modelo ajustado (0.2517) tiene un R<sup>2</sup> ligeramente superior al modelo normal (0.2399). Esto indica que el ajuste (las modificaciones o afinaciones que realizaste) logró capturar una pequeña fracción adicional de la variabilidad en los datos.

**RMSE:** El modelo ajustado (42.46) tiene un RMSE ligeramente menor que el modelo normal (43.13). Un RMSE más bajo es deseable, ya que indica que las predicciones del modelo ajustado son, en promedio, un poco más cercanas a los valores reales.

**MAE:** De manera similar al RMSE, el modelo ajustado (34.48) tiene un MAE ligeramente menor que el modelo normal (34.96). Esto refuerza la idea de que las predicciones del modelo ajustado tienen un error absoluto promedio menor.

Se concluye de la comparación que, el modelo Random Forest ajustado presenta métricas de rendimiento marginalmente mejores que el modelo normal. El aumento en el  $R^2$  y la disminución en el RMSE y MAE sugieren que los ajustes realizados tuvieron un impacto positivo en la capacidad predictiva del modelo, aunque la mejora es modesta.



Análisis del modelo ajustado:

Método del Codo: Se aplicó correctamente el Método del Codo calculando la Suma de Cuadrados Dentro del Conjunto (WSSSE) para un rango de números de clústeres (de 2 a 9). El WSSSE mide la compacidad de los clústeres. Graficar el WSSSE frente al número de clústeres ayuda a identificar un punto de "codo", donde la tasa de disminución del WSSSE se ralentiza. Este codo sugiere un número razonable de clústeres que equilibra la varianza dentro del clúster con el número de clústeres.

Clustering: Luego se ajustó el modelo K-Means a los datos (df\_kmeans) utilizando un k elegido (el número óptimo de clústeres que determinas a partir del gráfico del codo, en este caso tres clústeres fue lo óptimo).

Centroides: Se imprimen los centroides de los clústeres. Los centroides representan los valores medios de cada característica para los puntos de datos dentro de ese clúster.

Valores medios por clúster: Se calcula e imprime los valores medios de las características originales dentro de cada clúster. Al comparar los valores medios entre clústeres, se puede entender cómo los clústeres difieren en términos de las variables originales (por ejemplo, las familias en cluster 0 tienen niveles educativos más altos y más acceso a internet que las familias en los otros clúster).

Comparación con un modelo sin ajuste (escalado):

Si se realiza el clustering K-Means con las características sin escalar, los resultados podrían ser significativamente diferentes.

En resumen, el escalado fue crucial para obtener clústeres que reflejan la estructura subyacente de los datos considerando la influencia conjunta de múltiples características. El análisis de las medias por clúster en la escala ajustada proporciona una base sólida para describir y entender las diferencias entre los grupos de estudiantes identificados.

## Entorno de ejecución:

Para evidenciar la eficiencia del procesamiento distribuido, se ejecutó el mismo proceso de carga y análisis de datos en dos entornos distintos:

Entorno	Tiempo de Ejecución
Google Colab	10 minutos y 12 segundos
Clúster con Spark	1 minuto y 57 segundos

Este resultado refleja cómo el uso de Apache Spark en un entorno distribuido mejora significativamente los tiempos de procesamiento, especialmente en tareas con grandes volúmenes de datos como la lectura y análisis de archivos. Siendo el Clúster.

## Bono: Red Neuronal

Se construyó una red neuronal secuencial compuesta por tres capas ocultas densamente conectadas con 128, 64 y 32 neuronas respectivamente, todas con activación ReLU. Se incorporaron capas de **Batch Normalization** para estabilizar el entrenamiento y **Dropout** para reducir el sobreajuste. La capa de salida tiene una sola neurona para realizar predicción de valores continuos. El modelo contiene un total de 12.289 parámetros, de los cuales 11.905 son entrenables.

Model: "sequential_1"		
Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 128)	1,152
batch_normalization (BatchNormalization)	(None, 128)	512
dropout (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 64)	8,256
batch_normalization_1 (BatchNormalization)	(None, 64)	256
dropout_1 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 32)	2,080
dense_6 (Dense)	(None, 1)	33
Total params: 12,289 (48.00 KB)		
Trainable params: 11,905 (46.50 KB)		
Non-trainable params: 384 (1.50 KB)		

Tras entrenarse con Early Stopping y evaluarse con múltiples métricas, el modelo obtuvo un error absoluto medio (MAE) de 34.61 puntos y un  $R^2$  de 0.2093, lo que indica que, si bien logró aprender ciertos patrones relevantes, aún explica solo una parte limitada de la variabilidad en los resultados académicos.

Se concluye que, aunque el enfoque es prometedor, se requiere incorporar más variables predictoras o probar modelos alternativos para lograr una predicción más precisa y robusta del desempeño estudiantil.

### Conclusiones / recomendaciones

El análisis realizado, utilizando técnicas de **regresión lineal**, **Random Forest** y **clustering con K-means**, permitió identificar que el **contexto familiar y educativo** de los estudiantes que presentan el ICFES es un factor determinante en su rendimiento académico.

La regresión lineal mostró que variables como el **nivel educativo del padre y la madre**, el **estrato socioeconómico** y la **cantidad de libros en el hogar** tienen una correlación positiva con el puntaje global del ICFES. Aunque el modelo explicó una fracción moderada de la variabilidad ( $R^2 \approx 0.25$ ), los resultados fueron

consistentes con la teoría: a mejores condiciones familiares, mayor desempeño académico.

Además, el modelo Random Forest identificó las variables más influyentes en la predicción del puntaje:

- **Nivel educativo de la madre:** 39% de importancia.
- **Cantidad de libros en casa:** 18%.
- **Nivel educativo del padre:** 16%.
- **Acceso a computador y estrato socioeconómico:** también relevantes.

Las **condiciones del entorno familiar** son mucho más influyentes que factores escolares o individuales aislados.

**Implicación para el plan de acción:** Es fundamental diseñar estrategias que fortalezcan el entorno educativo en el hogar. Por ejemplo:

- Programas de alfabetización para padres.
- Campañas de donación de libros.
- Subsidios para acceso a tecnologías (computadores, internet).

Al aplicar K-means y reducir las dimensiones con PCA, se identificaron **tres grupos de estudiantes claramente diferenciados**:

- **Clúster 0:** Estudiantes con mejores condiciones familiares (más libros, padres con mayor nivel educativo, acceso a computador) y puntajes más altos ( $\approx 332$ ).
- **Clúster 1:** Estudiantes con menores condiciones y puntajes significativamente bajos ( $\approx 213$ ).
- **Clúster 2:** Grupo intermedio, con condiciones mixtas y puntajes medios ( $\approx 271$ ).

Estos clústeres permiten visualizar cómo los factores sociales y económicos se agrupan y reflejan directamente en los resultados.

**Implicación para políticas públicas:** Los estudiantes del Clúster 1 deben ser **priorizados en las intervenciones**, puesto que se caracterizan por tener:

- Bajos niveles educativos en los padres.
- Poca disponibilidad de libros en el hogar.
- Bajo estrato socioeconómico.
- Menor acceso a computador e internet.
- Mayor cantidad de horas dedicadas al trabajo semanal.

Estas condiciones reflejan una situación de vulnerabilidad que limita su desempeño académico. Las políticas deben enfocarse en:

- Garantizar conectividad y recursos básicos.
- Identificar municipios críticos con bajo puntaje promedio.
- Implementar tutorías comunitarias y acompañamiento educativo.
- Crear espacios extracurriculares con apoyo académico
- Diseñar programas de formación para padres que fortalezcan el hábito de lectura y el valor de la educación en casa.

### Conclusión Final

El análisis sugiere que para mejorar los resultados del ICFES en Cundinamarca no basta con intervenir en las instituciones educativas. Se requiere un enfoque integral que **involucre el hogar, los hábitos de estudio y el acceso a recursos educativos básicos**. Las variables con mayor impacto no son necesariamente las más visibles, pero sí las más transformadoras si se intervienen adecuadamente.