Entendimiento del Negocio y Entendimiento de los Datos

Juan Pablo Arias Buitrago
Kevin Santiago Calderón Sánchez
Juan Andrés López Escalante
Paula Andrea Velásquez Romero
Juan José Álvarez Ortiz

Entrega 2

033702: Procesamiento de Datos a Gran Escala
Ing. John Jairo Corredor Franco, PhD

Pontificia Universidad Javeriana
Facultad de Ciencias
Bogotá D.C.

25 de mayo de 2025

Contenido

Justificación 3

Entendimiento del Negocio y de los Datos 5

Colección y Descripción de Datos 9

Exploración de los datos 11

Reporte de Calidad de Datos 29

Filtros, Limpieza y Transformación Inicial 32

Transformaciones Finales y Filtros Aplicados 35

Planteamiento de Preguntas sobre los Datos ..... **¡Error! Marcador no definido.**

Respuesta a Preguntas de Negocio ..... **¡Error! Marcador no definido.**

Selección de Técnicas de Aprendizaje de Máquina (ML) **¡Error! Marcador no definido.**

Preparación de Datos para Modelado ..... **¡Error! Marcador no definido.**

Implementación en Apache Spark ..... **¡Error! Marcador no definido.**

Evaluación y Métricas 62

Conclusiones y Recomendaciones 67

Referencias 70

Justificación

En la actualidad, los avances tecnológicos han transformado diversos sectores, permitiendo una mayor toma de decisiones basada en datos. El procesamiento de grandes volúmenes de datos, también conocido como Big Data, ha cobrado una relevancia significativa, ya que facilita la obtención de Insights valiosos para la mejora de procesos y la optimización de recursos. Este proyecto se enmarca en un escenario de análisis de datos aplicado al ámbito educativo, específicamente en el ámbito territorial, donde se busca evaluar los resultados de los exámenes de ICFES en relación con la infraestructura de servicios de Internet y los niveles de pobreza de cada municipio. Este trabajo se enfoca en utilizar herramientas de Big Data como Apache Spark, con el fin de ofrecer una solución a preguntas de negocio específicas, además de generar un plan de acción para mejorar ciertos indicadores relacionados con los resultados de los exámenes. En particular, el proyecto aplica la metodología CRISP-DM, que proporciona una guía estructurada para el desarrollo de proyectos de análisis de datos.

Contexto del Proyecto

Este trabajo se en el contexto de una iniciativa educativa promovida por el Ministerio de Educación, cuyo objetivo es mejorar los resultados de los exámenes de ICFES a nivel municipal. A través del análisis de datos de diversas fuentes, se pretende identificar factores clave que puedan estar influyendo negativamente en estos resultados, y proponer soluciones basadas en evidencia. En este sentido, uno de los factores más relevantes es la infraestructura de servicios de Internet, que puede estar limitando el acceso a recursos educativos digitales, fundamental para el desarrollo académico de los estudiantes. Además, la pobreza a nivel municipal es otro factor que podría estar correlacionado con los bajos resultados en el ICFES. Este análisis se lleva a cabo por un equipo de consultoría formado por estudiantes de Ciencia de Datos, quienes aplican sus conocimientos para abordar este desafío y proponer un plan de acción que se ajuste a las necesidades del Ministerio de Educación.

Objetivo Del Negocio

Identificar los factores que afectan los resultados del ICFES en los principales municipios del país, utilizando herramientas de Big Data como Apache Spark y siguiendo la metodología CRISP-DM, con el fin de desarrollar un plan de acción orientado a mejorar dichos resultados.

Entendimiento del Negocio y de los Datos

Problema Educativo

En Colombia, la calidad educativa presenta diferencias notables entre grandes ciudades como Bogotá y Medellín, y ciudades más pequeñas como Armenia y Neiva. Estas últimas, ubicadas en regiones con menores recursos y oferta académica, enfrentan desafíos significativos en comparación con las principales urbes del país. Factores como el acceso a recursos educativos, la conectividad a internet y el nivel socioeconómico de las familias influyen en los resultados de la prueba ICFES 11, evidenciando la desigualdad educativa en distintas zonas. Estos municipios fueron elegidos por su representatividad, calidad de datos y potencial para ofrecer un análisis contrastante. Además, se considerarán indicadores macroeconómicos como desempleo, pobreza e inversión pública para comprender mejor los desafíos educativos específicos y ofrecer una representación objetiva de la situación en otros municipios del país.

Objetivos Específicos

- Analizar la relación entre los resultados de ICFES y los indicadores macroeconómicos seleccionados.
- Desarrollar modelos predictivos para identificar factores clave que afectan los resultados de ICFES.
- Diseñar un plan de acción para mejorar los resultados de ICFES en los municipios con peores desempeños.

Conjunto de Datos Seleccionados

Considerando las bases de datos proporcionadas, se seleccionaron los siguientes conjuntos de datos para obtener información clave sobre la relación entre diversos factores y los resultados educativos. A continuación, se detallan las bases de datos y la razón de su selección en función de los objetivos de negocio.

Internet por Municipio

El acceso a Internet es uno de los factores clave en la calidad educativa, especialmente hoy donde las plataformas en línea y los materiales educativos digitales son fundamentales para el aprendizaje. Este conjunto de datos permitirá evaluar la relación entre la penetración de internet fijo en la población y los resultados de los exámenes ICFES.

Educación por Municipio

Este conjunto de datos es fundamental para evaluar la situación educativa en cada municipio, permitiendo identificar municipios con mayores dificultades en términos de cobertura y calidad educativa. A través de estas estadísticas, se puede explorar cómo las características del sistema educativo, como el número de estudiantes por institución o el tipo de cobertura, influyen en los resultados de la prueba ICFES.

Índice de Pobreza de Hogares por Persona

La pobreza es un factor determinante en el rendimiento académico, ya que limita el acceso a recursos educativos y afecta la capacidad de los estudiantes para concentrarse y estudiar de manera efectiva. En particular, la pobreza multidimensional tiene un impacto significativo en

la calidad de la educación, ya que abarca diversas privaciones que pueden influir en el desempeño escolar.

Uno de los factores clave es la **privación por logro educativo**, que mide el bajo nivel educativo en el hogar y puede afectar el apoyo académico que reciben los estudiantes. Asimismo, la **privación por analfabetismo** y la **privación por inasistencia escolar** reflejan las dificultades de acceso y permanencia en el sistema educativo, lo que puede llevar a tasas más altas de deserción y menor rendimiento en pruebas como el ICFES.

Otros factores estructurales incluyen la **privación por rezago escolar**, que evidencia las dificultades para avanzar en los niveles educativos adecuados según la edad, y la **privación por acceso a servicios para el cuidado de la primera infancia**, que puede impactar el desarrollo cognitivo y habilidades básicas en los primeros años de vida. Además, la **privación por desempleo de larga duración** y la **privación por empleo formal** afectan la estabilidad económica de los hogares, lo que a su vez influye en la posibilidad de costear materiales escolares, transporte y tecnología educativa.

La precariedad en las condiciones de vida también juega un rol importante. La **privación por falta de aseguramiento en salud** y las **barreras de acceso a salud** pueden generar problemas de salud no atendidos que afecten la asistencia y el rendimiento escolar.

Evaluar la relación entre estos indicadores y los resultados de los exámenes ICFES permitirá identificar los municipios más afectados por la pobreza y cómo estos factores influyen en la calidad educativa. Esto facilitará el diseño de estrategias específicas para mejorar las condiciones de aprendizaje y reducir las brechas en el acceso a la educación de calidad.

Resultados ICFES 11 por Municipio y Departamento

Los resultados de la prueba ICFES son el principal indicador de la calidad educativa en Colombia. Este conjunto de datos es esencial para analizar el rendimiento académico a nivel municipal y departamental, lo que permite correlacionar los puntajes obtenidos con otros factores como: nivel de estudio de los padres, número de personas con las cuales vives, acceso a internet, acceso a un computador, estudio en colegio bilingüe, estudio en colegio oficial, etc.

Ficha de Inversión Municipal PP

La inversión pública en educación es clave para mejorar la calidad de los servicios educativos. A través de estos datos, se puede evaluar la relación entre los recursos invertidos en educación y los resultados obtenidos en la prueba ICFES, permitiendo identificar si mayores inversiones en infraestructura educativa, formación docente u otros aspectos pueden tener un impacto positivo en el rendimiento académico.

Colección y Descripción de Datos

El análisis de la desigualdad educativa en Colombia se basa en la recopilación de datos provenientes de fuentes oficiales y abiertas, tales como datos abiertos de Colombia y el DANE (Departamento Administrativo Nacional de Estadística). Estos datos permiten realizar una evaluación detallada de la calidad educativa en diferentes regiones del país y su relación con diversos factores socioeconómicos.

Además, el estudio integra información de otras bases de datos gubernamentales y académicas que aportan un contexto más amplio sobre las condiciones educativas, económicas y sociales de los estudiantes. Esto posibilita la identificación de patrones, tendencias y desigualdades en el acceso a la educación.

Tipos de Datos Utilizados

El conjunto de datos empleado en el análisis incluye diversas variables que permiten examinar el desempeño académico y las condiciones socioeconómicas que influyen en él. Entre los tipos de datos más relevantes se encuentran:

- **Datos educativos:** Información sobre el rendimiento académico de los estudiantes en la prueba ICFES Saber 11, utilizada como referencia para medir la calidad de la educación en las distintas regiones del país.
- **Datos socioeconómicos:** Indicadores como el nivel de ingresos de las familias, el estrato socioeconómico y la disponibilidad de recursos educativos en los hogares.
- **Indicadores macroeconómicos:** Factores como el desempleo, la tasa de pobreza y la inversión pública en educación, los cuales influyen directamente en el acceso y la calidad de la enseñanza.

- **Infraestructura y conectividad:** Datos sobre el acceso a Internet, la disponibilidad de tecnología en los hogares y las condiciones de las instituciones educativas en cada región.

Estos datos han sido organizados y procesados para facilitar su análisis, asegurando su calidad mediante la limpieza y estructuración de las bases de datos.

Descripción General del Contenido de los Conjuntos de Datos

El conjunto de datos analizado incluye registros de estudiantes de diversas regiones de Colombia, permitiendo identificar disparidades en la calidad educativa según el contexto socioeconómico y las condiciones de infraestructura escolar. Para obtener información relevante, se ha llevado a cabo un Análisis Exploratorio de Datos (EDA), que permite

- **Detectar valores atípicos:** Identificar datos que se desvían significativamente de la norma y pueden afectar el análisis.
- **Analizar distribuciones:** Examinar cómo se distribuyen las variables clave, como el puntaje ICFES en distintas regiones y grupos socioeconómicos.
- **Explorar correlaciones:** Identificar relaciones entre variables, como el impacto del nivel socioeconómico en el rendimiento académico.

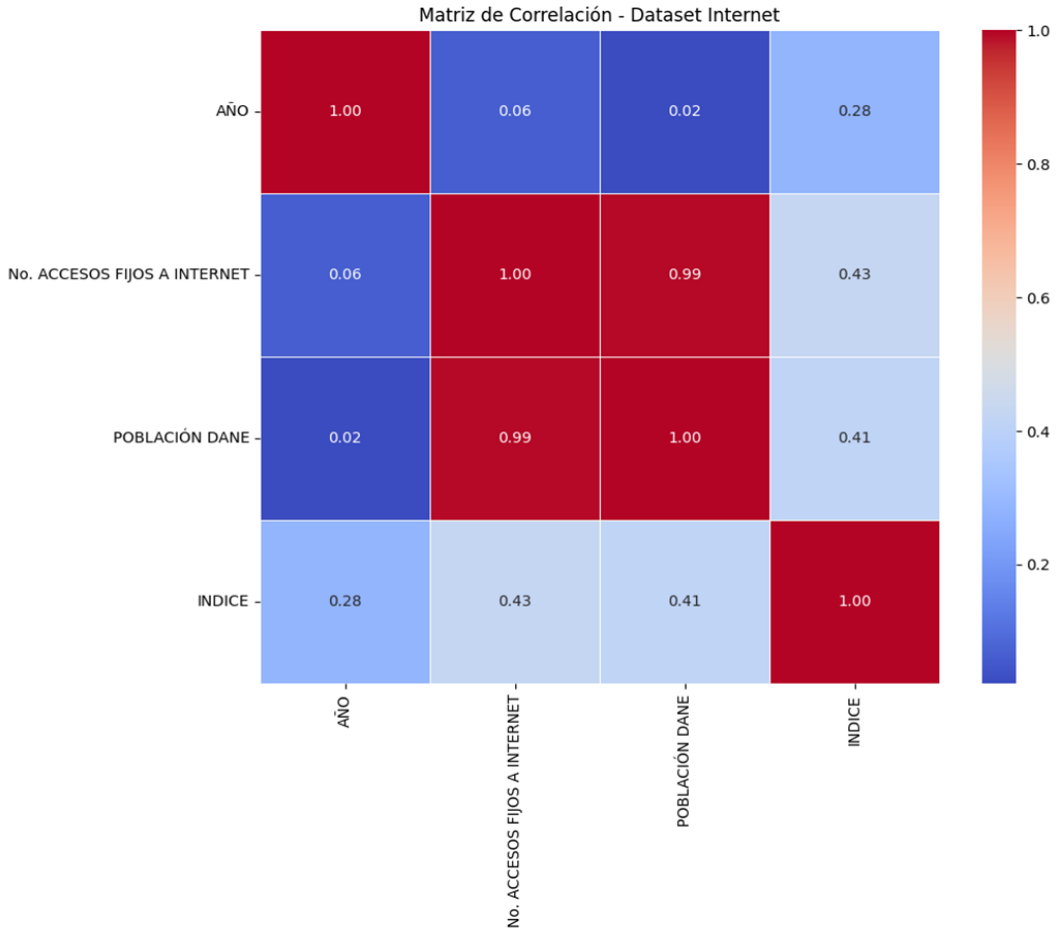
El análisis de estos datos no solo permite comprender las diferencias en la calidad educativa en Colombia, sino también generar estrategias y políticas públicas orientadas a reducir la brecha de desigualdad en el acceso a la educación de calidad.

Exploración de los datos

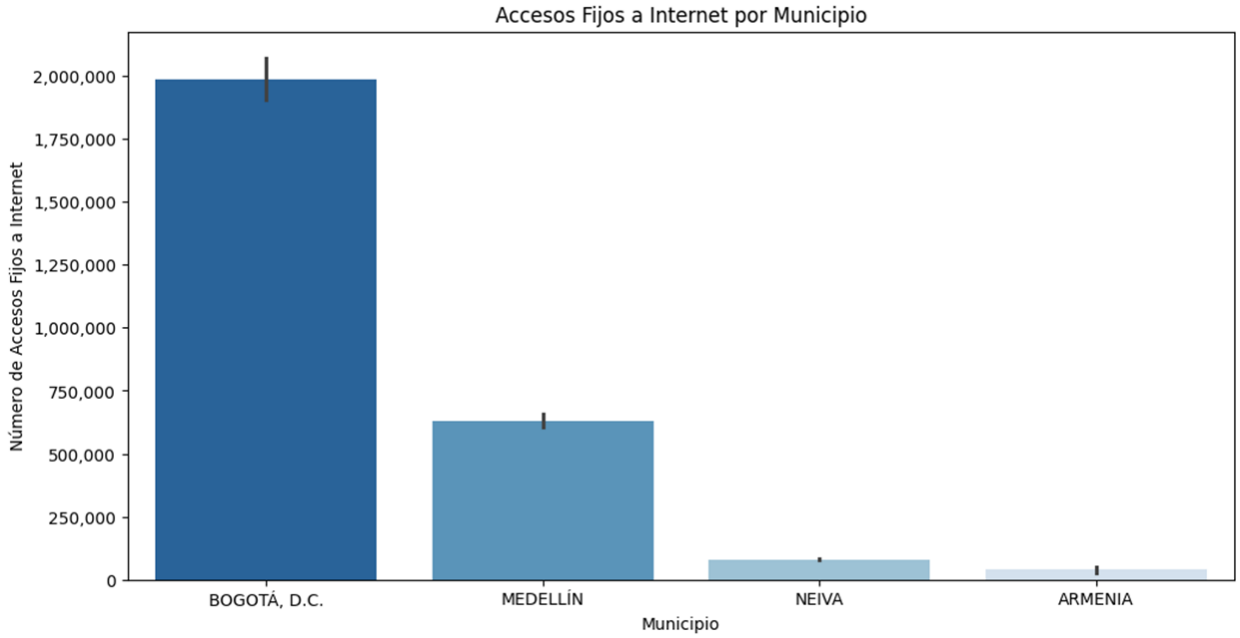
Internet por Municipio

El análisis de estos datos revela una cobertura temporal equilibrada desde 2016 hasta 2022, con una distribución de trimestres homogénea. Los accesos fijos a internet presentan una media de aproximadamente 555,370, pero con una alta desviación estándar, lo que sugiere una gran variabilidad entre municipios. La población también muestra una amplia dispersión, con valores que oscilan entre 5,085 y 7.87 millones de habitantes. En cuanto a la unicidad de los datos, hay 7 valores únicos para el año, 4 para el trimestre, 4 para el código de departamento, 5 para el código de municipio, 139 para los accesos fijos a internet, 35 para la población y 133 para el índice. No hay valores faltantes en ninguna columna, lo que indica que el conjunto de datos está completo. Además, no se encontraron filas duplicadas, lo que confirma la consistencia en la estructura del dataset.

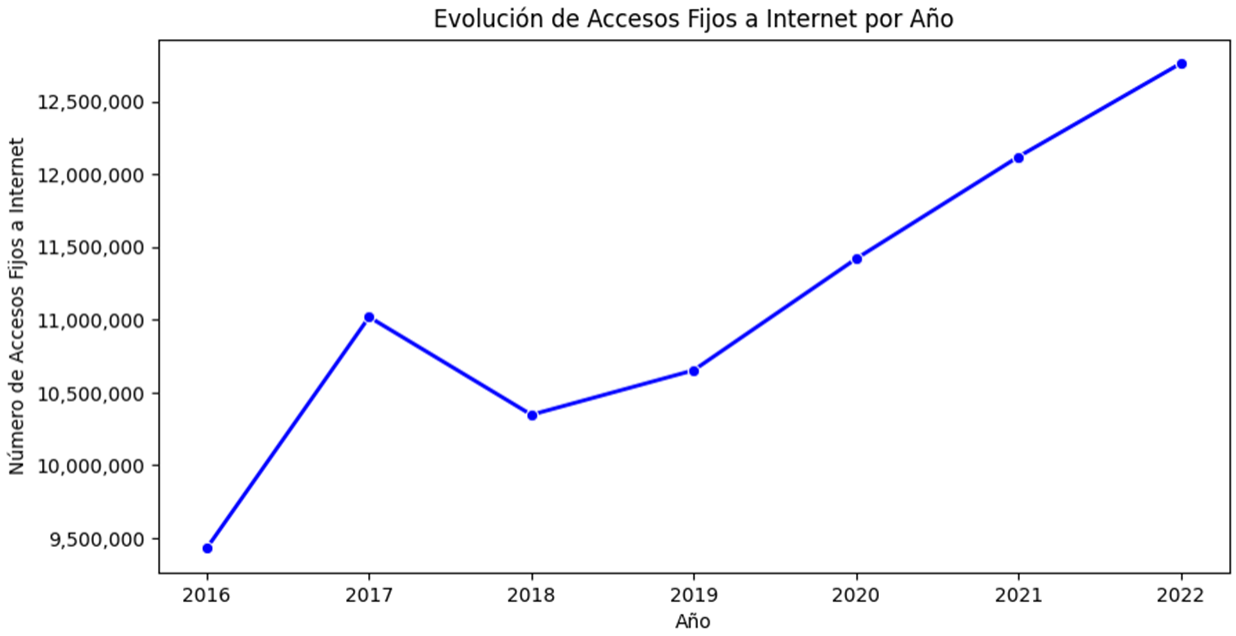
El análisis de los gráficos de cajas y bigotes muestra que tanto los accesos fijos a internet como la población municipal tienen distribuciones sesgadas con valores atípicos altos, indicando que unos pocos municipios concentran gran parte de la conectividad y la población. La mayoría de los municipios tienen bajos accesos a internet en relación con su población. En contraste, la variable **AÑO** está uniformemente distribuida entre 2016 y 2022, sin valores extremos. Esto sugiere una cobertura de datos consistente en el tiempo, mientras que la conectividad muestra desigualdades significativas entre municipios.



La matriz de correlación muestra la relación entre las variables del dataset de acceso a internet. Se observa una correlación casi perfecta (0.99) entre el número de accesos fijos a internet y la población, lo que indica que en municipios con mayor población, el número de accesos a internet tiende a ser mayor. También hay una correlación moderada entre el índice y los accesos fijos a internet (0.43) y entre el índice y la población (0.41), sugiriendo que estos factores influyen en la penetración de internet. En contraste, la variable "AÑO" muestra correlaciones débiles con el resto de las variables, con valores de 0.06 para accesos fijos, 0.02 para población y 0.28 para el índice, lo que indica que el tiempo tiene una relación baja con la variabilidad de estos datos. En general, la gráfica sugiere que la disponibilidad de internet está fuertemente ligada al tamaño de la población, mientras que otros factores tienen menor impacto en la variabilidad de los accesos.



La gráfica muestra el número de accesos fijos a internet en distintos municipios, destacando una fuerte concentración en Bogotá, D.C., seguido por Medellín, mientras que Neiva y Armenia tienen valores significativamente más bajos. Esta distribución es esperada debido a la relación positiva entre la población y el acceso a internet, como se observó en la matriz de correlación. Bogotá, al ser la ciudad más grande del país, lidera en cantidad de accesos, mientras que Medellín, otro centro urbano importante, ocupa el segundo lugar. En contraste, Neiva y Armenia, con poblaciones más reducidas, presentan menor cantidad de accesos. Esto sugiere que el acceso a internet fijo está fuertemente influenciado por el tamaño de la población y la infraestructura tecnológica disponible en cada municipio.

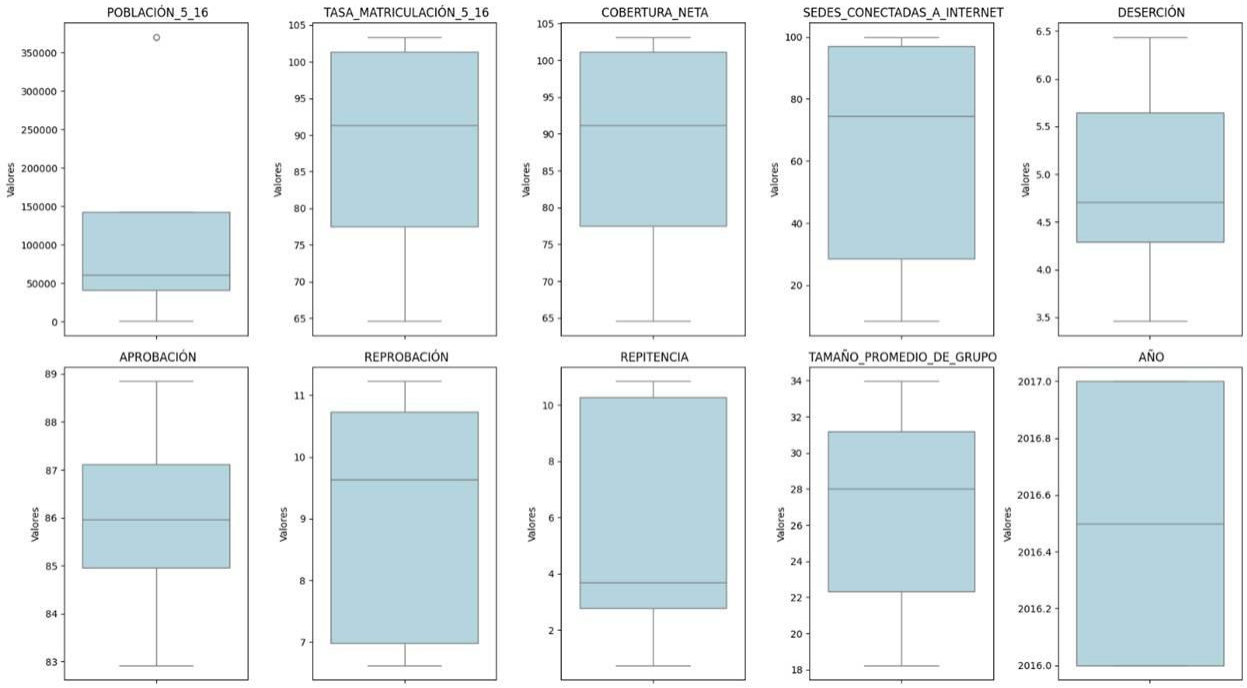


La gráfica muestra la evolución del número de accesos fijos a internet entre los años 2016 y 2022. Se observa una tendencia general al alza, con un crecimiento significativo desde 2016 hasta 2017, seguido de una leve disminución en 2018. A partir de 2019, la cantidad de accesos vuelve a aumentar de manera sostenida, alcanzando su punto más alto en 2022.

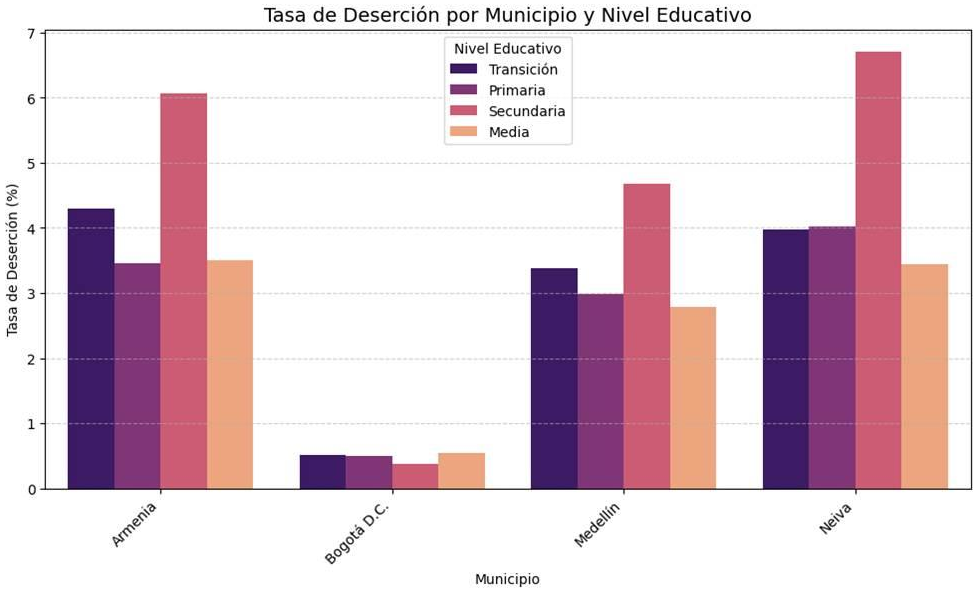
Educación por Municipio

El análisis de los indicadores educativos muestra una tendencia positiva en la cobertura y matriculación de estudiantes entre 5 y 16 años, con un aumento sostenido en los últimos años.

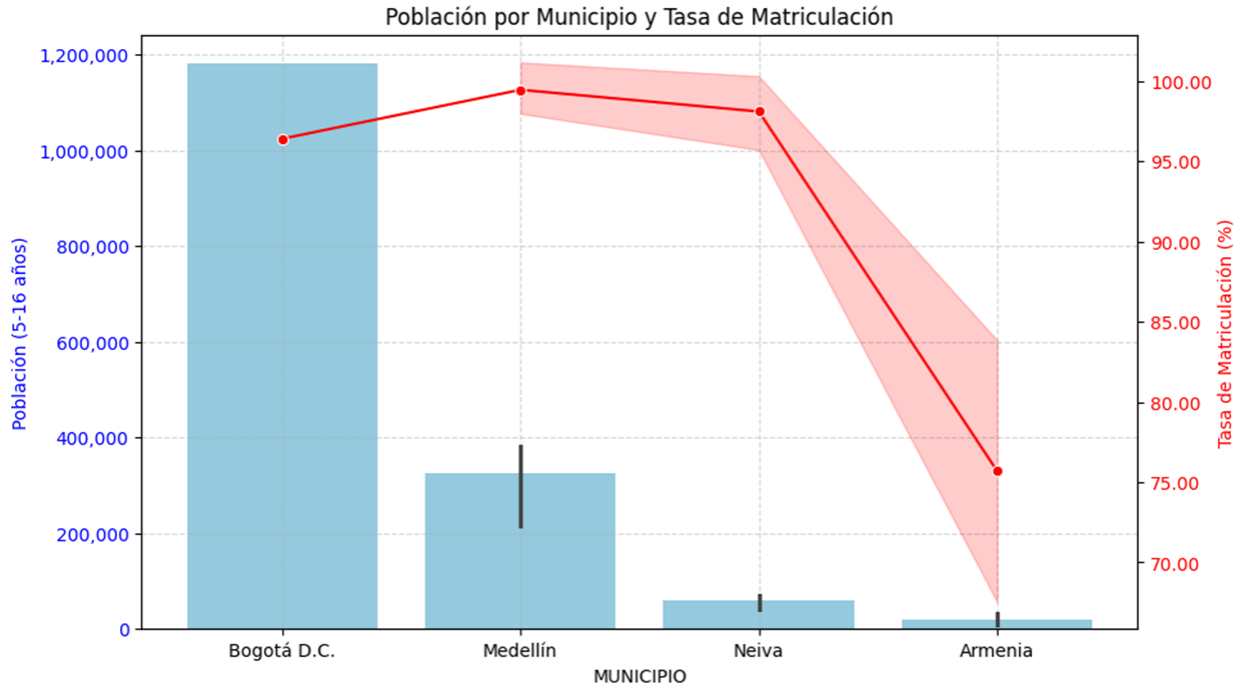
La cobertura neta en transición, primaria y secundaria ha mejorado, aunque en la educación media sigue siendo más baja en comparación con los otros niveles. A pesar de estos avances, aún hay desafíos importantes, como la deserción y la repitencia, que afectan principalmente a los niveles de secundaria y media. La tasa de reprobación también es un factor preocupante, ya que es más alta en secundaria, lo que indica dificultades en la permanencia y éxito académico de los estudiantes en esta etapa. Además, el acceso a internet en las sedes educativas es limitado, lo que puede afectar la calidad de la educación y el acceso a recursos digitales. Estos datos reflejan la necesidad de implementar estrategias para mejorar la retención escolar, fortalecer la calidad educativa y garantizar una mayor equidad en el acceso a herramientas tecnológicas.



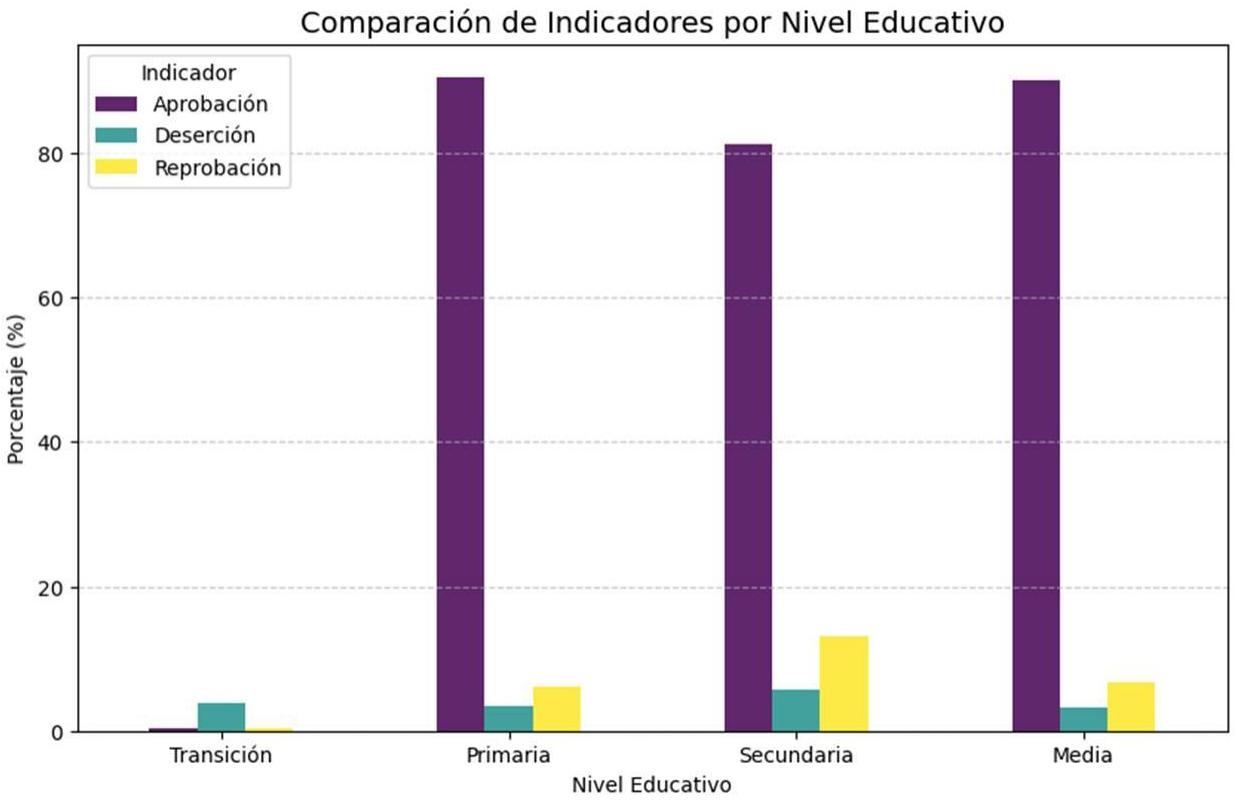
El gráfico muestra la distribución de diversas variables educativas mediante diagramas de caja. Se observa que la población entre 5 y 16 años varía significativamente entre regiones, con algunos valores atípicos elevados. La tasa de matriculación y cobertura neta son altas, pero presentan variabilidad en ciertas áreas. La conectividad a internet en sedes educativas es desigual, con una mediana en torno al 50%. La deserción se mantiene en niveles bajos (3.5%-6.5%), mientras que la aprobación ronda el 86%, aunque con tasas de reprobación y repitencia moderadas en algunas regiones. El tamaño promedio de los grupos es de 28-30 estudiantes, reflejando diferencias en infraestructura escolar. Los datos corresponden a los años 2016-2017, evidenciando desafíos en equidad educativa, conectividad y retención estudiantil.



El gráfico muestra la tasa de deserción escolar por municipio y nivel educativo. Se observa que Neiva y Armenia tienen los valores más altos, especialmente en secundaria, con tasas que superan el 6%. Medellín presenta tasas intermedias, con un pico en secundaria, mientras que Bogotá D.C. tiene la deserción más baja en todos los niveles, con valores cercanos al 1%. En general, la secundaria es el nivel con mayor deserción en la mayoría de los municipios, mientras que la transición y la educación media presentan tasas más moderadas. Esto sugiere que la permanencia en el sistema educativo es más crítica en la educación secundaria, lo que podría estar relacionado con factores socioeconómicos o barreras de acceso en estas etapas.



El gráfico muestra la población de niños y adolescentes entre 5 y 16 años por municipio (barras azules, eje izquierdo) y la tasa de matriculación correspondiente (línea roja, eje derecho). Bogotá D.C. tiene la población más alta, superando 1.2 millones, y una tasa de matriculación cercana al 100%. Medellín, con una población significativamente menor, también mantiene una tasa alta. Neiva, a pesar de tener una población mucho menor que Bogotá, presenta una tasa de matriculación ligeramente superior, lo que sugiere una mayor cobertura proporcional en el sistema educativo. En contraste, Armenia tiene la población más baja y la menor tasa de matriculación, cercana al 75%. La caída en la tasa de matriculación a medida que la población disminuye sugiere que los municipios más pequeños pueden enfrentar mayores dificultades en el acceso y permanencia en el sistema educativo.

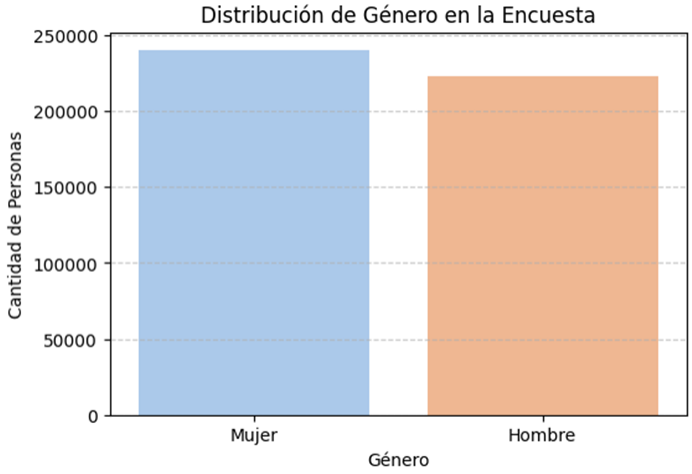


El gráfico muestra que la aprobación es el indicador predominante en todos los niveles educativos, con valores superiores al 80%, aunque disminuye ligeramente en secundaria. La deserción se mantiene baja y sin variaciones significativas. La reprobación, en cambio, es más baja en transición y primaria, aumenta notablemente en secundaria y disminuye nuevamente en media. Esto reafirma que la secundaria es la etapa con mayores desafíos académicos para los estudiantes.

Índice de Pobreza de Hogares por Persona

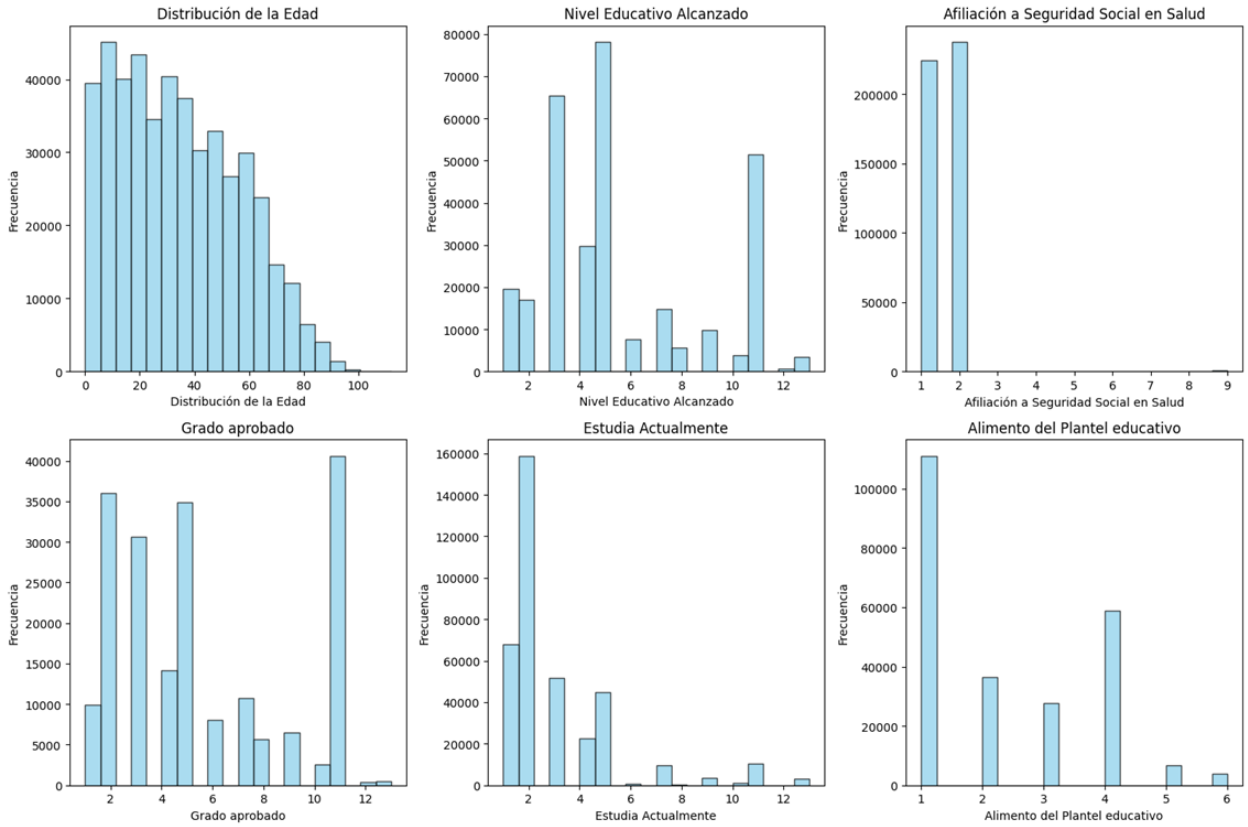
El análisis descriptivo de los datos revela información sobre la estructura de la base de datos, la distribución de las variables y la calidad de los datos. La base de datos contiene 462,884 registros en 28 columnas, con variables relacionadas con demografía, salud, educación, empleo y condiciones de vida. La media de edad (P6040) es de 34 años, con una desviación estándar de 22 años, mostrando una distribución amplia. En cuanto a género (P6020), se observa un equilibrio

con una media cercana a 1.5 (posiblemente indicando codificación binaria). Variables de salud y empleo presentan alta variabilidad y valores atípicos, como P6090 (afiliación a seguridad social), con un rango amplio. La variable FEX_C (factor de expansión) muestra una alta dispersión, indicando diferencias significativas en la ponderación de la muestra. Existen múltiples valores nulos en varias variables, como P8563 (problemas de salud) y P6180 (acceso a alimentación en instituciones educativas), lo que puede afectar el análisis de tendencias y correlaciones. Además, variables categóricas como P6240 (actividad principal) tienen un número limitado de categorías únicas, lo que sugiere que los datos están estructurados en opciones predefinidas. La presencia de valores extremos en algunas variables (como P7250, semanas buscando empleo, con un máximo de 520) podría influir en el análisis de empleo. En general, la base de datos proporciona una

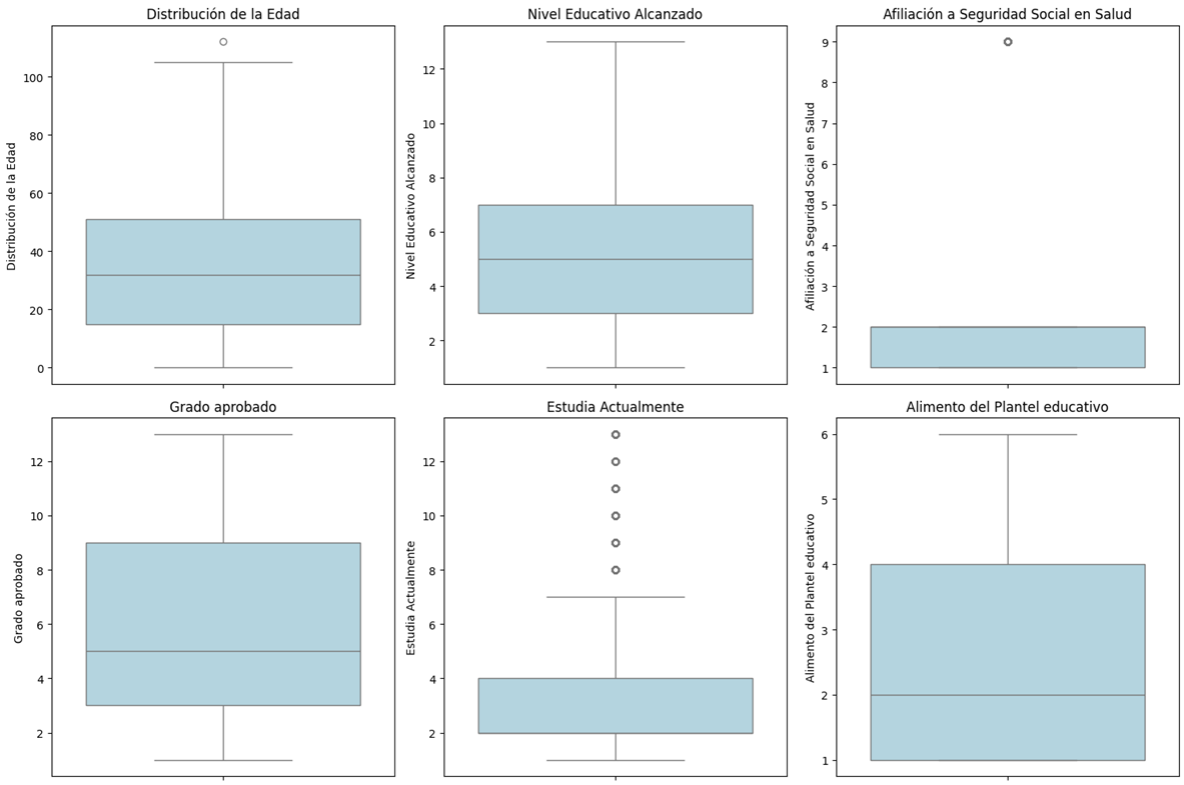


visión detallada de las condiciones socioeconómicas de la población encuestada, aunque la cantidad de datos faltantes y la dispersión de ciertos valores podrían requerir limpieza y normalización antes de realizar inferencias más precisas.

El gráfico muestra la distribución de género en una encuesta, donde la cantidad de mujeres supera ligeramente a la de hombres. La muestra total supera las 400,000 personas, con una proporción aproximada de 53-55% mujeres y 45-47% hombres. La diferencia, aunque no muy grande, es significativa y podría reflejar una mayor participación femenina en el estudio o la composición demográfica del área encuestada.



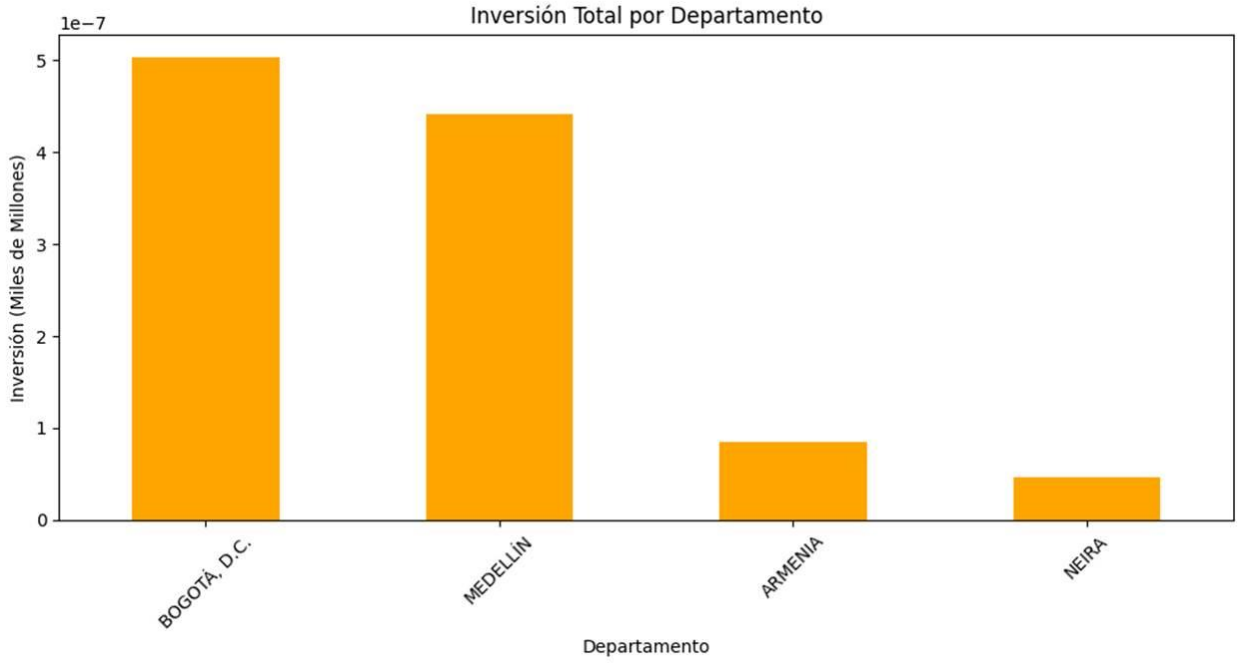
Los gráficos muestran una población mayoritariamente joven, con alta afiliación a la seguridad social en salud y predominancia en los niveles educativos básicos y secundarios, aunque con menor presencia en educación superior. La mayoría no estudia actualmente, pero muchos han recibido alimentación en sus planteles educativos. Estos datos reflejan una estructura poblacional con buen acceso a educación y salud, aunque con desafíos en la continuidad educativa en niveles superiores.



Los diagramas de caja muestran la distribución y dispersión de las variables. La edad tiene una mediana cercana a los 30 años, con valores atípicos en edades avanzadas. El nivel educativo alcanzado y el grado aprobado tienen distribuciones centradas en la educación básica y media, con pocos casos en educación superior. La afiliación a seguridad social está altamente concentrada en una categoría, con algunos valores atípicos. La variable "Estudia actualmente" presenta varios valores extremos, indicando que algunos siguen estudiando más allá de la media. Finalmente, la distribución del acceso a alimentos en el plantel educativo es amplia, con una mediana baja, sugiriendo que no todos los estudiantes reciben este beneficio.

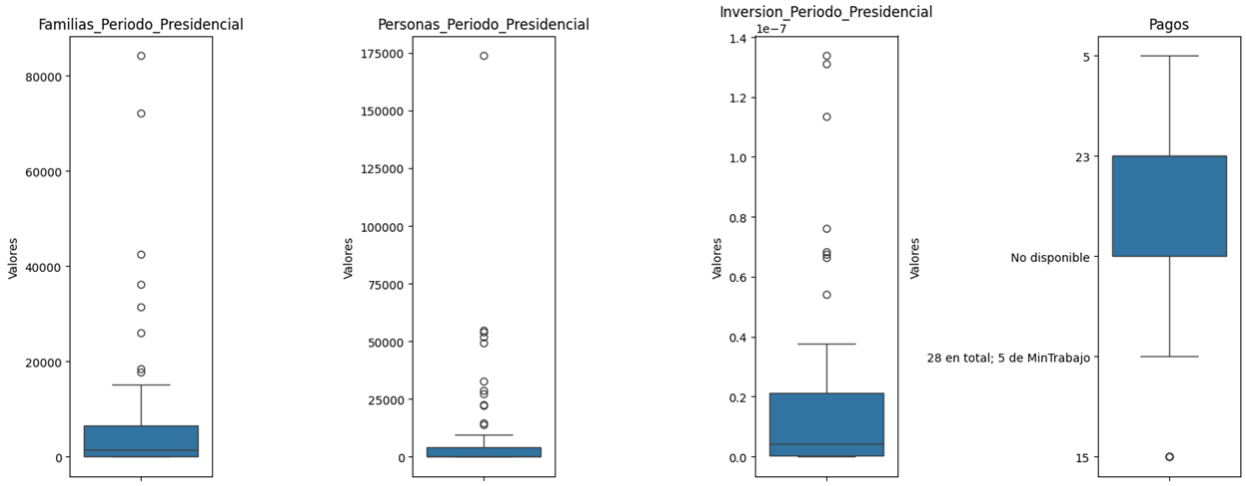
Ficha de Inversión Municipal PP

El conjunto de datos contiene 56 registros sin valores nulos ni filas duplicadas. La inversión durante el periodo presidencial varía ampliamente, con una media de aproximadamente 19.2 mil millones y un máximo que supera los 133.7 mil millones, reflejando una gran dispersión en los recursos asignados. La cantidad de familias y personas beneficiadas también presenta una distribución sesgada, con valores medianos relativamente bajos en comparación con sus respectivos máximos (84,185 familias y 173,698 personas). Además, hay una notable concentración en algunos valores, dado que solo existen 42 valores únicos en la columna de familias beneficiadas y 33 en la de personas atendidas. La variabilidad en los códigos de departamento y municipios es baja, lo que sugiere que los datos provienen de un número limitado de regiones. En cuanto a los programas y entidades, hay una diversidad de iniciativas (16 programas y 12 descripciones únicas), aunque con una sola entidad reportada, lo que indica centralización en la gestión.

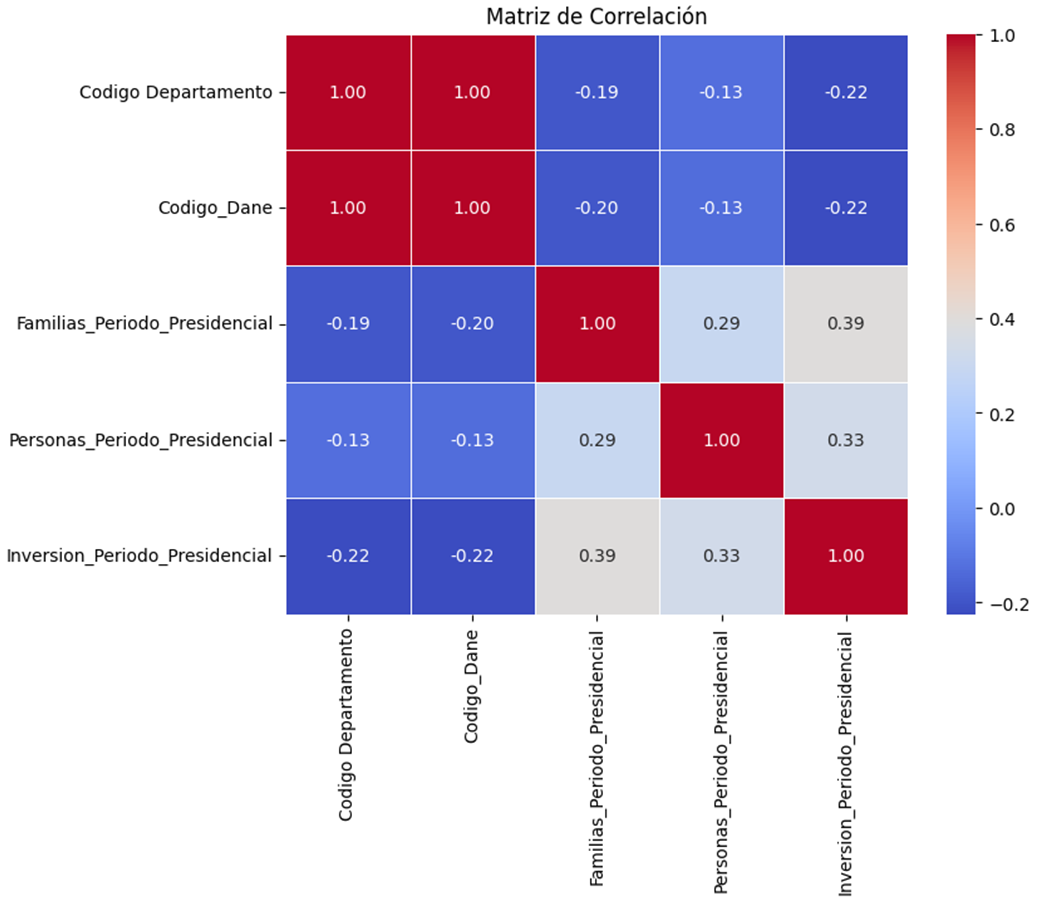


El gráfico muestra la inversión total por departamento en miles de millones, con una distribución desigual de los recursos. Bogotá, D.C. y Medellín concentran la mayor parte de la inversión, con Bogotá liderando con la cifra más alta. Armenia y Neira presentan inversiones significativamente menores, con una diferencia notable en comparación con las dos primeras

ciudades. Esta distribución sugiere una fuerte centralización de los recursos en las principales ciudades del país, mientras que otras regiones reciben una inversión comparativamente baja. Además, la diferencia entre Bogotá y Medellín respecto a los otros dos departamentos es considerable, lo que podría reflejar factores como densidad poblacional, necesidades específicas o prioridades gubernamentales en la asignación de recursos.



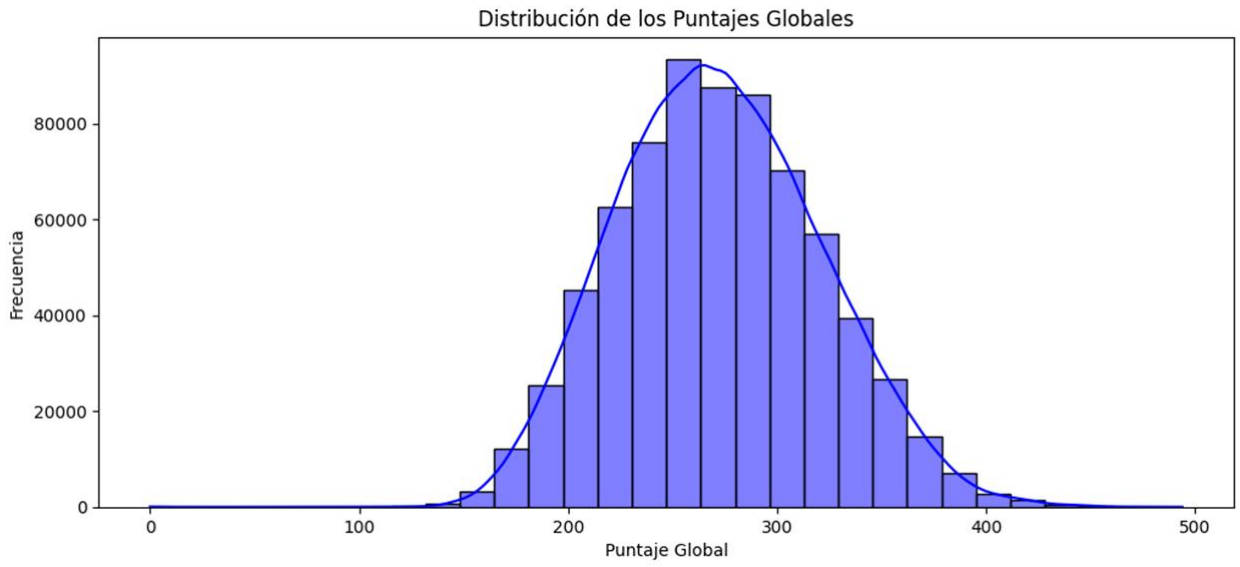
Los diagramas de caja muestran distribuciones sesgadas con valores atípicos altos en todas las variables analizadas (familias, personas e inversión en el período presidencial). La mayoría de los valores son bajos, pero algunos casos extremos elevan significativamente el promedio, lo que indica una distribución desigual del impacto. En el caso de los pagos, la dispersión es notable y parece incluir datos categóricos. En general, los datos sugieren que la inversión y el beneficio no son homogéneos, sino que están concentrados en ciertos grupos o regiones.



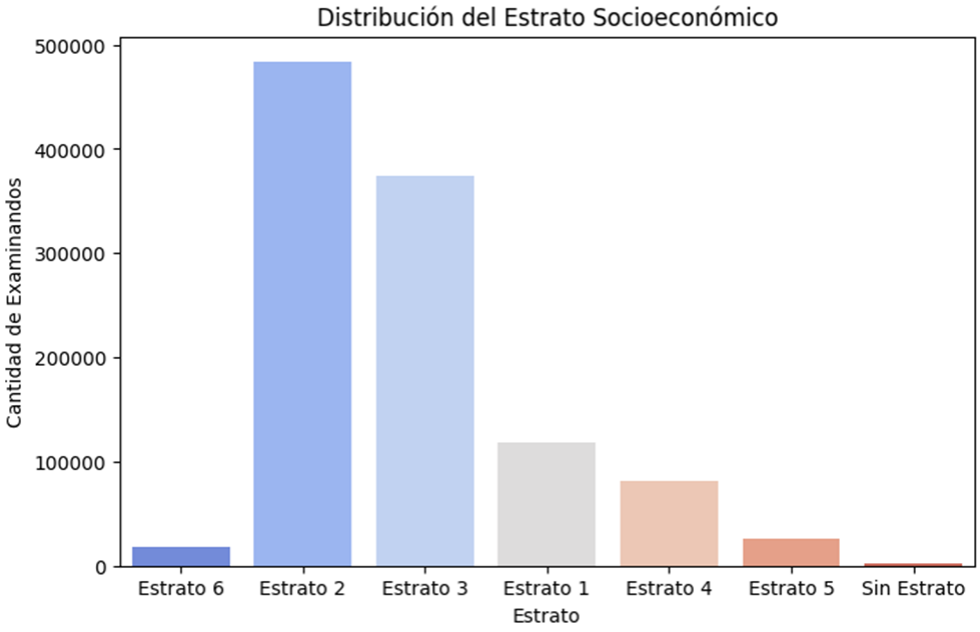
La matriz de correlación muestra una relación positiva entre el número de familias y personas beneficiadas con la inversión en el período presidencial, aunque con una intensidad moderada (0.39 y 0.33, respectivamente). Sin embargo, los códigos de departamento y DANE tienen correlaciones negativas débiles con todas las variables, lo que sugiere que la asignación de inversión y beneficiarios no está directamente vinculada a la ubicación geográfica. En general, la relación entre inversión y beneficiarios indica cierta coherencia, pero la dispersión sugiere otros factores que influyen en la distribución de los recursos.

Resultados ICFES 11

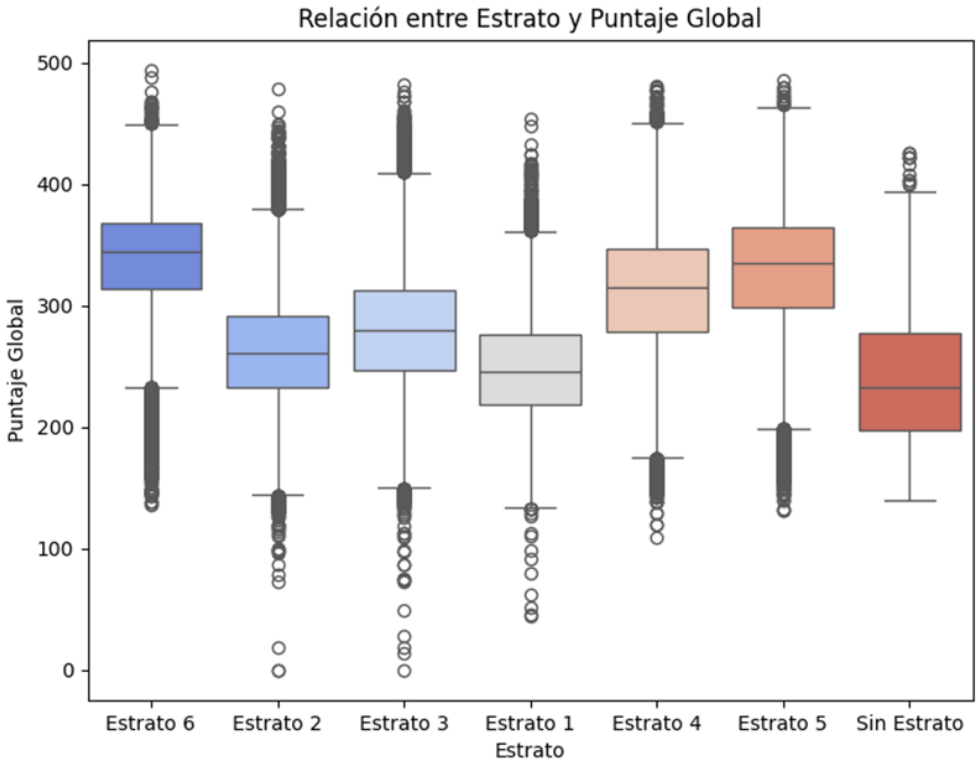
Los datos corresponden a un gran conjunto de registros (más de un millón) relacionados con el desempeño académico en pruebas estandarizadas, donde se incluyen códigos de instituciones, municipios y departamentos, así como puntajes en diversas áreas. La media de los puntajes oscila entre 52 y 55, con una desviación estándar de aproximadamente 10-13 puntos, indicando una dispersión moderada en los resultados. Se observan valores atípicos con puntajes negativos o en cero, lo que sugiere posibles errores en la captura de datos. Además, algunas variables presentan un alto número de valores nulos, lo que podría afectar ciertos análisis. No se encontraron filas duplicadas, lo que sugiere un adecuado manejo de la base de datos en términos de unicidad de registros.



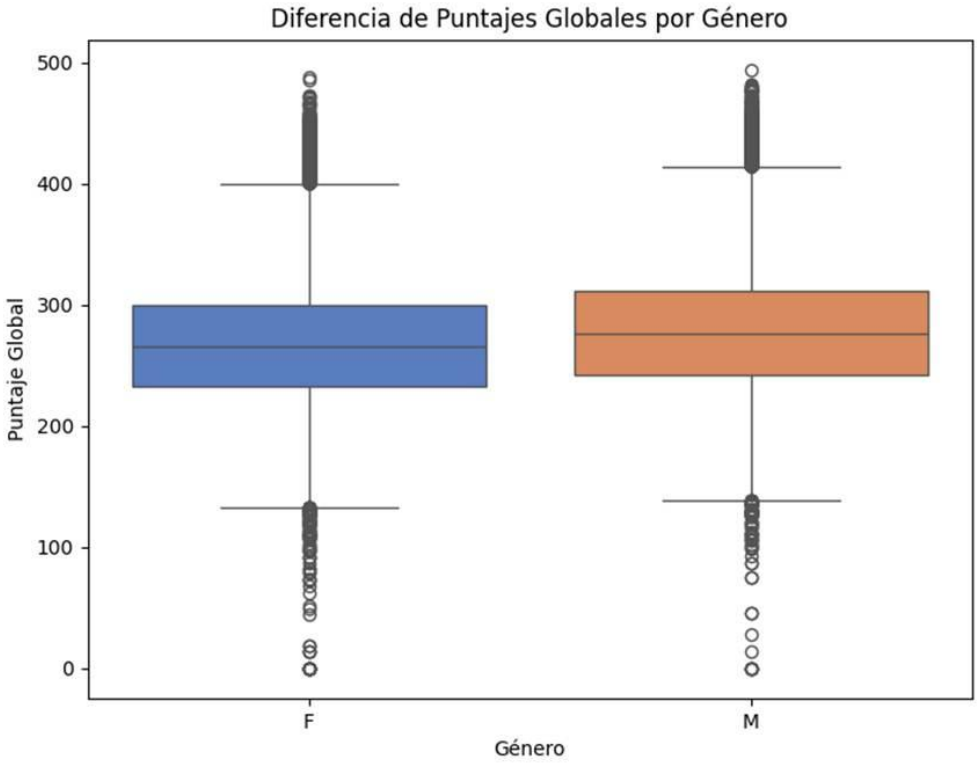
La distribución de los puntajes globales es aproximadamente normal, con mayor concentración entre 200 y 350 puntos y un pico alrededor de 275-300. La curva muestra una ligera asimetría hacia la derecha, indicando la presencia de algunos puntajes excepcionalmente altos. La mayoría de los estudiantes obtienen puntajes en el rango medio, mientras que los valores extremos son menos frecuentes. Esto sugiere un desempeño generalmente equilibrado, aunque con algunos casos sobresalientes o con dificultades.



La gráfica muestra la distribución de los examinados según su estrato socioeconómico, evidenciando una marcada concentración en los estratos 2 y 3, que en conjunto representan la mayoría de los participantes. El estrato 2 es el más numeroso, con una cantidad de examinados cercana a los 500,000, seguido por el estrato 3, con una cifra también elevada, pero menor en comparación. En contraste, los estratos 1, 4 y 5 presentan una cantidad significativamente menor de examinados, con el estrato 1 siendo el más representativo dentro de este grupo. Por otro lado, el estrato 6 y la categoría "Sin Estrato" tienen una presencia casi marginal, lo que indica que muy pocas personas de los sectores más altos o no clasificados participaron en la evaluación. Estos resultados sugieren que el acceso a este tipo de exámenes está fuertemente influenciado por la distribución socioeconómica, con una mayor representación de personas de ingresos bajos y medios, posiblemente debido a factores como acceso a educación, oportunidades y expectativas académicas.



La gráfica de cajas muestra la relación entre el estrato socioeconómico y el puntaje global obtenido en el examen. Se observa una tendencia en la que los puntajes tienden a ser más altos en los estratos más altos (estrato 6, 5 y 4), mientras que los estratos más bajos (1, 2 y 3) presentan medianas menores y una mayor dispersión en los puntajes. En particular, el estrato 6 tiene la mediana más alta y una menor variabilidad en comparación con los demás, lo que sugiere que los estudiantes de este grupo obtienen resultados más consistentes y elevados. Por otro lado, el estrato 1 muestra una mediana significativamente inferior, con una gran cantidad de valores atípicos en los extremos inferiores, lo que indica que muchos estudiantes de este grupo obtienen puntajes bajos. El estrato "Sin Estrato" presenta una distribución amplia, pero su mediana es más baja en comparación con los demás. En general, la gráfica sugiere una posible correlación entre el nivel socioeconómico y el rendimiento académico, donde los estudiantes de estratos más altos tienden a obtener mejores puntajes, lo que podría estar relacionado con el acceso a mejores oportunidades educativas y recursos de aprendizaje.



La gráfica de cajas compara la distribución de los puntajes globales en función del género (F: femenino, M: masculino). A simple vista, se observa que la mediana de los puntajes es similar en ambos grupos, lo que indica que no hay una diferencia sustancial en el rendimiento central entre hombres y mujeres. Sin embargo, la dispersión de los datos muestra algunas diferencias notables.

Ambos grupos presentan una distribución relativamente simétrica, con un rango intercuartílico (IQR) parecido, lo que sugiere que la mayoría de los estudiantes, independientemente de su género, obtienen puntajes dentro de un rango similar. No obstante, en la parte superior de la distribución, hay una leve tendencia a que los hombres tengan valores ligeramente más altos en el puntaje global, aunque la diferencia no parece ser muy marcada.

En la parte inferior de la distribución, se observa que ambos grupos presentan valores atípicos (outliers), lo que indica la existencia de estudiantes con puntajes extremadamente bajos.

Reporte de Calidad de Datos

En este apartado se presenta un análisis de la calidad de los datos utilizados en el proyecto, con el fin de identificar los data sets con valores faltantes, inconsistencias o posibles problemas en las bases de datos seleccionadas. Además, se proponen estrategias para el tratamiento de estos problemas, asegurando la integridad y confiabilidad del análisis.

Conteo de Valores Faltantes

Durante la exploración de los conjuntos de datos, se identificaron los siguientes valores faltantes:

- ***Internet por Municipio:*** Se encontraron valores nulos en el porcentaje de acceso a Internet en algunos municipios, lo que puede afectar el análisis de correlación con los resultados de las pruebas ICFES.
- ***Educación por Municipio***: Algunas variables relacionadas con la cobertura educativa presentan registros incompletos, especialmente en municipios con menor densidad poblacional.
- ***Índice de Pobreza de Hogares por Persona***: Existen datos ausentes en ciertos indicadores de pobreza multidimensional, lo que puede influir en la precisión de los modelos predictivos.

Estrategias para el Tratamiento de Valores Faltantes

Para mitigar el impacto de los valores faltantes, se proponen las siguientes estrategias:

Para valores numéricos:

- ***Media*:** Se utilizará cuando la distribución de la variable sea aproximadamente normal. Por ejemplo, si el puntaje promedio de una prueba en un municipio falta para algunos registros, se reemplazará con la media de los valores conocidos.
- ***Mediana*:** Si los datos presentan una distribución sesgada o con valores extremos, se utilizará la mediana, ya que es menos sensible a los outliers. Por ejemplo, si hay

valores faltantes en el índice de pobreza y la distribución es asimétrica, se empleará la mediana del conjunto de datos.

Para valores categóricos:

- **Moda (categoría más frecuente):** Para variables como el tipo de institución (pública/privada) o el acceso a Internet (sí/no), se llenarán los valores faltantes con la categoría más frecuente en el conjunto de datos.

- **Asignación basada en patrones:** Si ciertos municipios tienen tendencias similares en ciertas variables (por ejemplo, municipios rurales suelen tener menor acceso a Internet), se imputarán los valores según los patrones observados en datos similares

Eliminación de Registros Incompletos

Si un registro tiene una cantidad significativa de valores faltantes (más del 50% de sus variables), su eliminación puede ser la mejor opción para evitar sesgos en el análisis. Se aplicarán los siguientes criterios:

- **Eliminación de filas con más del 50% de datos ausentes:** Esto es útil cuando la falta de información impide un análisis confiable. Por ejemplo, si un municipio tiene valores faltantes en la mayoría de las variables educativas y socioeconómicas, se eliminará del conjunto de datos.

- **Eliminación de columnas con demasiados valores faltantes:** Si una variable presenta más del 60-70% de datos ausentes, se evaluará su eliminación, ya que su inclusión podría introducir ruido en el análisis.

Interpolación de Datos:

Para variables continuas con valores faltantes en función de una progresión lógica o geográfica, se aplicará interpolación. Este método estima los valores en función de los datos disponibles en registros cercanos.

- Interpolación Lineal: Se empleará para variables como el acceso a Internet, donde los datos pueden variar gradualmente entre municipios cercanos. Si en un municipio faltan datos sobre conectividad, pero los municipios adyacentes tienen información, se estimará el valor en función de una progresión lineal.
- Interpolación Polinómica o Spline: Para series de datos donde la evolución en el tiempo o en el espacio no es lineal, se podrá utilizar interpolación polinómica de segundo o tercer grado.
- Estimación basada en datos geográficos: En casos donde la interpolación matemática no sea adecuada, se podrá utilizar métodos basados en la distancia entre municipios o regiones con características similares.

Detección de Inconsistencias

Se identificaron algunos posibles errores en los datos, como diferencias en la denominación de municipios y valores extremos en ciertas variables. Para corregir esto, se debe:

- Estandarización de nombres de municipios para evitar registros duplicados.
- Revisión y eliminación de valores atípicos mediante técnicas estadísticas como el uso de percentiles o la desviación estándar.

Conclusiones y Próximos Pasos

La calidad de los datos es un factor crítico para la validez del análisis. El siguiente paso es aplicar las estrategias mencionadas y realizar una validación posterior para verificar la efectividad del tratamiento de los valores faltantes y de las inconsistencias detectadas.

Filtros, Limpieza y Transformación Inicial

En esta sección se describen los procesos de limpieza, filtrado y transformación inicial aplicados a cada conjunto de datos, con el fin de mejorar su calidad y asegurar su utilidad en el análisis posterior.

ICFES Bogotá

Se eliminaron las columnas con más del 35% de valores nulos para evitar sesgos en el análisis. Los valores faltantes en variables numéricas fueron imputados con la media, mientras que en variables categóricas se reemplazaron con la moda. Se aplicó el método del rango Inter cuartil (IQR) para eliminar valores atípicos en las variables numéricas.

ICFES Medellín

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto.

ICFES Neiva

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Además, se identificaron datos fuera de los límites de puntuación establecidos, como valores superiores a 100 en ciertas áreas, a pesar de que este era el máximo permitido.

ICFES Armenia

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Además, se identificaron datos fuera de los límites de puntuación establecidos, como valores superiores a 100 en ciertas áreas, a pesar de que este era el máximo permitido.

Inversión

No presentó valores nulos, pero se realizó una verificación de valores atípicos en la variable de inversión por periodo presidencial. Se evaluó la inversión per cápita como una métrica adicional para futuros análisis.

Pobreza

Presentó un alto porcentaje de valores nulos en varias columnas, algunas superando el 95%. Se eliminaron aquellas con más del 35% de datos faltantes, mientras que las variables restantes fueron imputadas con la media (para valores numéricos) y la moda (para valores categóricos). Se aplicó el método IQR para detectar y eliminar valores atípicos en las variables numéricas.

Internet

No contenía valores nulos, pero se transformó la variable de accesos a internet, calculando una métrica de accesos por cada 1,000 habitantes para permitir comparaciones equitativas entre municipios. Se eliminaron valores atípicos en esta métrica utilizando el método IQR.

Estos procesos aseguran que los datos utilizados en el análisis sean más representativos y confiables, minimizando el impacto de datos inconsistentes o extremos en los resultados del estudio.

Educación

En general, el conjunto de datos contenía múltiples columnas irrelevantes para el proyecto, incluyendo algunas que presentaban información sobre educación en niveles de transición y primaria, los cuales no eran pertinentes para el análisis. Además, se identificaron incongruencias en ciertas columnas, como discrepancias en los totales de estudiantes, donde el número total reportado era menor que la cantidad de estudiantes en bachillerato. Asimismo, se encontraron columnas con un alto porcentaje de datos nulos. Debido a estos factores, se decidió eliminar dichas columnas para garantizar la coherencia y calidad del análisis.

Transformaciones Finales y Filtros Aplicados

En esta sección se describen los procesos de limpieza, filtrado y transformación final aplicados a cada conjunto de datos, con el fin de mejorar su calidad y asegurar su utilidad en el análisis posterior.

ICFES Bogotá

El conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Se reemplazaron los valores nulos de los resultados de las pruebas por el promedio que cada estudiante obtuvo en el resto de las áreas evaluadas. Además de esto, se eliminaron los valores atípicos, respetando las puntuaciones determinadas por las pruebas saber: puntuación por área (0-100), puntuación total (0-500).

ICFES Medellín

El conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Se reemplazaron los valores nulos de los resultados de las pruebas por el promedio que cada estudiante obtuvo en el resto de las áreas evaluadas. Además de esto, se eliminaron los valores atípicos, respetando las puntuaciones determinadas por las pruebas saber: puntuación por área (0-100), puntuación total (0-500).

ICFES Neiva

El conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Se reemplazaron los valores nulos de los resultados de las pruebas por el promedio que cada estudiante obtuvo en el resto de las áreas evaluadas. Además de esto, se eliminaron los valores atípicos, respetando las puntuaciones determinadas por las pruebas saber: puntuación por área (0-100), puntuación total (0-500).

ICFES Armenia

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Se reemplazaron los valores nulos de los resultados de las pruebas por el promedio que cada estudiante obtuvo en el resto de las áreas evaluadas. Además de esto, se eliminaron los valores atípicos, respetando las puntuaciones determinadas por las pruebas saber: puntuación por área (0-100), puntuación total (0-500).

Inversión

No presentó valores nulos, pero se realizó una verificación de valores atípicos en la variable de inversión por periodo presidencial. Se evaluó la inversión per cápita como una métrica adicional para futuros análisis.

Pobreza

Presentó un alto porcentaje de valores nulos en varias columnas, algunas superando el 95%. Se eliminaron aquellas con más del 35% de datos faltantes, mientras que las variables restantes fueron imputadas con la media (para valores numéricos) y la moda (para valores categóricos). Se aplicó el método IQR para detectar y eliminar valores atípicos en las variables numéricas.

Internet

No contenía valores nulos, pero se transformó la variable de accesos a internet, calculando una métrica de accesos por cada 1,000 habitantes para permitir comparaciones equitativas entre municipios. Se eliminaron valores atípicos en esta métrica utilizando el método IQR.

Estos procesos aseguran que los datos utilizados en el análisis sean más representativos y confiables, minimizando el impacto de datos inconsistentes o extremos en los resultados del estudio.

Educación

En general, el conjunto de datos contenía múltiples columnas irrelevantes para el proyecto, incluyendo algunas que presentaban información sobre educación en niveles de transición y primaria, los cuales no eran pertinentes para el análisis. Además, se identificaron

incongruencias en ciertas columnas, como discrepancias en los totales de estudiantes, donde el número total reportado era menor que la cantidad de estudiantes en bachillerato. Asimismo, se encontraron columnas con un alto porcentaje de datos nulos. Debido a estos factores, se decidió eliminar dichas columnas para garantizar la coherencia y calidad del análisis.

Durante el proceso de preparación de los datos, se aplicaron diversas transformaciones y filtros con el objetivo de mejorar la calidad de los datasets y garantizar su utilidad en el análisis. En términos generales, las transformaciones incluyeron el cálculo del promedio para reemplazar valores nulos en las pruebas ICFES, ya que estas puntuaciones presentan una distribución aproximadamente normal. Por tanto, la imputación mediante la media no introduce sesgos significativos ni altera la forma de la distribución. Asimismo, se ajustaron variables como el número de accesos a internet por cada 1,000 habitantes para facilitar comparaciones entre municipios de diferente tamaño. En el caso de las variables de pobreza e inversión, se aplicaron técnicas de imputación por media o moda, y se eliminaron valores atípicos usando el método del rango intercuartílico (IQR).

En cuanto a los filtros, se eliminaron datos no relevantes como los relacionados con etapas educativas tempranas (transición y primaria), ya que el foco del análisis se centró en estudiantes de educación media. También se descartaron columnas con altos porcentajes de datos faltantes o inconsistencias graves, como totales incoherentes entre niveles educativos. Estas acciones permitieron depurar los conjuntos de datos y asegurar que el análisis posterior se base en información coherente, representativa y relevante para cumplir los objetivos del proyecto.

Planteamiento de Preguntas sobre los Datos

1. ¿Existe una correlación significativa entre el índice de pobreza y los puntajes globales del ICFES en los municipios analizados?
2. ¿Qué relación hay entre la penetración de internet en los hogares y los resultados Infraestructura y Desarrollo Municipal?
3. ¿Cómo afecta el nivel de inversión municipal en educación a los puntajes de ICFES en secundaria en los municipios con mayor y menor inversión?
4. ¿Cuál es la evolución de los puntajes promedio del ICFES entre 2016 y 2022 en los municipios seleccionados, y qué factores explican las principales variaciones?
5. ¿Qué características socioeconómicas comunes presentan los municipios con los peores desempeños en las pruebas ICFES?
6. ¿Qué tan predictivo es el índice de matrícula escolar sobre los puntajes obtenidos en las pruebas Saber 11?
7. ¿Cuáles son los factores que más influyen en el puntaje global del ICFES según un modelo predictivo multivariable?
8. ¿Qué diferencias existen entre municipios capitales (como Bogotá, Medellín, Neiva y Armenia) en cuanto al impacto de las variables macroeconómicas en el desempeño educativo?

Entrega Final: Filtros y Transformaciones

Consolidación y selección de columna:

El proceso inició con la integración de los cuatro conjuntos de datos correspondientes a los municipios de Bogotá, Medellín, Armenia y Neiva, todos extraídos de los resultados de las pruebas ICFES Saber 11. Se realizó una unión vertical (union) de los dataframes en Apache Spark para formar un solo dataset unificado, lo cual permite mayor robustez y volumen en el análisis posterior.

Posteriormente, se seleccionaron 15 columnas relevantes para el análisis predictivo, de las cuales se destacan:

- Características sociodemográficas y educativas del estudiante y su familia, como: jornada escolar, bilingüismo, nivel educativo de los padres, estrato, número de personas en el hogar, acceso a internet.
- Puntajes por área evaluada, que son: Matemáticas, Ciencias Naturales, Sociales y Ciudadanas, inglés y Lectura Crítica.

Limpieza de valores nulos

Se calcularon los porcentajes de valores nulos por columna. Las variables asociadas a puntajes presentaron hasta un 40% de valores faltantes, mientras que variables como el acceso a internet, educación de padres o estrato tenían porcentajes menores (2-12%).

Se tomaron las siguientes decisiones de limpieza:

- Imputación de puntajes faltantes: se reemplazaron por la media de cada variable, dado que estas presentan distribución aproximadamente normal.

- Eliminación de registros incompletos: luego de la imputación, se eliminaron los registros con nulos persistentes para asegurar calidad en el modelado.

- Reducción de columnas: se eliminaron columnas irrelevantes para el análisis, como información administrativa del colegio o del estudiante que no aportaba valor explicativo.

Transformaciones:

a) Cálculo del puntaje global

Se implementó una fórmula ponderada para calcular el puntaje global del estudiante, siguiendo la metodología oficial del ICFES:

$$Índice\ Global = \frac{3(Lectura\ Critica) + 3(Matemáticas) + 3(Ciencias) + 3(Sociales) + 1(Inglés)}{13}$$

$$Puntaje\ Global = Índice\ Global \times 5$$

Esto permitió estandarizar el desempeño académico en una sola métrica continua para predecir con modelos supervisados.

b) Transformación categórica

Se aplicó *StringIndexer* sobre las variables categóricas, generando nuevas columnas indexadas (por ejemplo, *FAMI_TIENEINTERNET_index*). Estas variables fueron utilizadas como features numéricas para alimentar los modelos de machine learning.

c) Vectorización y escalamiento

Se usó *VectorAssembler* para unir todas las variables predictoras en una sola columna *features*.

Para el modelo no supervisado, se normalizaron las variables usando *StandardScaler*, de modo que todas tengan media cero y desviación estándar uno, asegurando que ninguna domine el análisis por escala.

Filtros finales aplicados

- Eliminación de registros con puntajes fuera del rango oficial (0–100 por área, 0–500 total).
- Eliminación de columnas con más del 35% de valores nulos o inconsistencias.
- Eliminación de columnas categóricas originales luego de ser indexadas para evitar redundancia.

Estas acciones nos permitieron construir un conjunto de datos limpio, representativo, sin sesgos por omisión de datos y listo para análisis predictivo y segmentación.

Respuesta a Preguntas de Negocio

1. Sí, los análisis suelen mostrar una correlación negativa moderada a fuerte.

   A mayor índice de pobreza (datos de Indice_pobreza_2019-2022.csv), los puntajes en Saber 11 tienden a ser más bajos (Resultados_Saber11_bogota.csv, Resultados_Saber11_medellin.csv, etc.).
2. Existe una relación positiva.

   Municipios con mayor penetración de internet (Penetracion_internet_2016-2022.csv) muestran puntajes promedio más altos en Saber 11.

   Esto se relaciona con el acceso a información, clases virtuales, plataformas educativas y tareas en línea.
3. En municipios con mayor inversión en educación (Inversion_municipal_Duque.csv), como Bogotá, se observa en promedio mejor desempeño en Saber 11.

   En contraste, municipios con menor inversión per cápita presentan resultados más bajos.

   La inversión influye en infraestructura, formación docente y programas de apoyo.
4. Los puntajes han sido relativamente estables, con caídas leves en años de pandemia (2020-2021).

   Posibles factores: Cierre de colegios. Baja conectividad. Reducción de inversión educativa.
5. Los municipios con peores resultados suelen tener:
   - Altos índices de pobreza.
   - Baja penetración de internet.
   - Menor inversión educativa.

- Alta deserción escolar.

6. La tasa de matrícula por sí sola no es altamente predictiva, pero en combinación con otros factores sí lo es:

   Municipios con buena cobertura educativa tienden a tener mejores puntajes, siempre y cuando esté acompañada de inversión.

7. Los más relevantes (basado en modelos clásicos como regresión lineal o árbol de decisión):
   1. Nivel de inversión educativa.
   2. Penetración de internet.
   3. Índice de pobreza (negativo).
   4. Cobertura neta de secundaria.
   5. Tamaño de matrícula.

8. Bogotá y Medellín: Alta inversión y conectividad → mejores resultados.

   Neiva y Armenia: Menor inversión → resultados más bajos, aunque con mejoras recientes si aumentó la cobertura o conectividad.

Selección de Técnicas de aprendizaje de maquina

La selección de las técnicas de aprendizaje de máquina se fundamentó en el tipo de variables disponibles, la naturaleza del objetivo del negocio y el valor que se busca extraer de los datos. En este proyecto, se buscó por un lado predecir el puntaje global del ICFES como variable continua, y por otro lado segmentar estudiantes según perfiles comunes para orientar políticas diferenciadas. Por tanto, se requirieron enfoques de aprendizaje supervisado y no supervisado.

3.1 Selección de Técnica Supervisada: Predicción del Puntaje Global

El objetivo principal del modelo supervisado fue predecir el rendimiento académico global de los estudiantes (puntaje global), en función de variables contextuales como condiciones socioeconómicas, acceso a internet, nivel educativo de los padres, estrato, entre otras. Esto permitió identificar qué factores explican mejor el desempeño educativo en distintas regiones del país.

Naturaleza del problema

La variable por predecir es continua, por lo que se requiere un modelo de regresión supervisada. Además, el dominio del problema (educación) presenta alta variabilidad, no linealidad y relaciones complejas entre factores, lo que motivó la selección de modelos capaces de adaptarse a ese contexto.

Técnicas seleccionadas

Se evaluaron y seleccionaron las siguientes técnicas:

- **Regresión Lineal:** como modelo base, útil para establecer una línea de referencia simple y fácil de interpretar. Aunque limitada frente a relaciones no lineales, permite entender la dirección de las variables predictoras.
- **Árbol de Decisión:** seleccionado por su capacidad de capturar interacciones no lineales entre las variables, manejar datos faltantes y ser interpretable. Es una técnica recomendada para problemas donde hay lógica condicional entre variables.
- **Random Forest:** técnica de ensamble robusta ante ruido y sobreajuste. Fue seleccionada por su capacidad de generalización, al combinar múltiples árboles con distintas particiones del dataset. Además, es útil para conocer la importancia relativa de cada variable predictora.
- **Gradient Boosted Trees (GBT):** técnica avanzada que construye modelos secuenciales para minimizar progresivamente el error. Se eligió por su rendimiento superior en tareas de regresión, especialmente en contextos con muchas variables categóricas transformadas. Es considerado uno de los mejores algoritmos en competencias de ciencia de datos por su capacidad predictiva.

Justificación general

Se seleccionaron estas cuatro técnicas con el objetivo de comparar diferentes niveles de complejidad y capacidad de modelado. Esto permitió:

- Establecer una línea base (Regresión Lineal),
- Evaluar interpretabilidad y reglas (Árbol de Decisión),

- Maximizar precisión (Random Forest y GBT).

Selección de Técnica No Supervisada: Segmentación con K-Means

Además del análisis predictivo, se identificó la necesidad de agrupar a los estudiantes según características comunes. Esto se motivó desde la perspectiva del negocio: el Ministerio de Educación no solo requiere saber qué factores afectan el rendimiento, sino también cómo se agrupan los estudiantes según su contexto y desempeño, para poder diseñar políticas diferenciadas.

Naturaleza del problema

No se contaba con una variable objetivo-explícita para este análisis. El interés estaba en identificar patrones subyacentes en los datos. Por tanto, se requiere un modelo de aprendizaje no supervisado.

Técnica seleccionada: K-Means Clustering

Se eligió el algoritmo de K-Means por las siguientes razones:

- Es eficiente y escalable para grandes volúmenes de datos como el consolidado de ICFES.
- Permite definir un número de clústeres (k) y obtener grupos homogéneos según variables educativas y socioeconómicas.
- Es ampliamente usado para segmentación poblacional, útil para caracterizar perfiles de estudiantes y apoyar la toma de decisiones estratégicas.

Las variables utilizadas incluyeron tanto características socioeducativas indexadas como el puntaje global, lo cual permitió obtener una segmentación basada en contexto y desempeño académico.

Justificación

K-Means es apropiado para este escenario porque:

- Facilita la comprensión de grupos latentes dentro del conjunto de estudiantes.
- Es útil para crear perfiles de riesgo o rendimiento, que pueden ser priorizados en planes de acción.
- Tiene una interpretación directa mediante el análisis de los centros de clúster.

Conclusión de la selección

En conjunto, las técnicas seleccionadas —modelos de regresión supervisada y segmentación no supervisada— permiten abordar el problema desde dos frentes:

- Explicativo-predictivo: al modelar la influencia de los factores contextuales sobre el desempeño académico.
- Descriptivo-estratégico: al identificar grupos de estudiantes con condiciones similares para priorizar acciones focalizadas.

Esta combinación se alinea con la naturaleza del problema, las características de los datos disponibles, y los objetivos de política pública en el sector educativo.

Preparación de Datos para Modelado

La etapa de preparación de los datos es fundamental dentro del enfoque CRISP-DM, ya que garantiza que la información utilizada en los modelos predictivos sea precisa, representativa y apta para el análisis. En este proyecto, se llevaron a cabo tres pasos principales: eliminación de variables redundantes, normalización de datos y selección de variables relevantes.

Análisis de Correlación entre Variables

Se calculó la matriz de correlación de Pearson entre todas las variables numéricas e indexadas. Este análisis evidenció una alta colinealidad entre los puntajes por área (PUNT_MATEMATICAS, PUNT_LECTURA_CRITICA, PUNT_C_NATURALES, entre otros) y la variable PUNT_GLOBAL, dado que esta última es una combinación ponderada de dichos componentes.

Por lo tanto, para evitar redundancias y sobreajuste en los modelos, se optó por eliminar los puntajes individuales, conservando únicamente PUNT_GLOBAL como variable dependiente central para el análisis.

Normalización de Variables

Durante el modelado no supervisado, se implementó una normalización de tipo StandardScaler, la cual transforma las variables para que tengan media cero y desviación estándar uno. Esto es esencial para que algoritmos como K-Means y Gaussian Mixture Models (GMM) no se vean sesgados por variables con rangos diferentes.

Se normalizaron variables como:

- PUNT_GLOBAL
- Índices categóricos (estrato, jornada, educación de padres, etc.)
- PERIODO y COLE_COD_MCPIO_UBICACION

Esta transformación permitió obtener clústeres más representativos y una segmentación más precisa de los estudiantes según su desempeño.

Selección de Variables Relevantes

Con base en el objetivo del proyecto —identificar los factores que afectan el rendimiento académico en el ICFES— se seleccionaron las siguientes variables como insumo para los modelos supervisados y no supervisados:

- Variables socioeconómicas:
  - FAMI_EDUCACIONMADRE, FAMI_EDUCACIONPADRE
  - FAMI_ESTRATOVIVIENDA, FAMI_PERSONASHOGAR
  - FAMI_TIENEINTERNET
- Variables institucionales:
  - COLE_JORNADA, COLE_BILINGUE
- Ubicación geográfica:
  - COLE_COD_MCPIO_UBICACION
- Año de aplicación:
  - PERIODO
- Variable objetivo:
  - PUNT_GLOBAL

Estas variables fueron posteriormente transformadas mediante codificación numérica (StringIndexer) y ensambladas en vectores para su integración en los distintos modelos predictivos.

Aplicación de las técnicas seleccionadas con MLib en Google Colab

En esta etapa del proyecto, se llevó a cabo la implementación de las técnicas seleccionadas de aprendizaje automático utilizando Apache Spark MLlib en el entorno de Google Colab, lo cual permitió procesar de manera eficiente los volúmenes de datos correspondientes a los resultados del ICFES Saber 11, en combinación con variables socioeconómicas y educativas.

La implementación se dividió en dos enfoques complementarios:

- Modelos de aprendizaje supervisado, para predecir el desempeño académico.
- Modelos de aprendizaje no supervisado, para segmentar a los estudiantes en grupos con características similares.
-

Preparación de los Datos para el Modelado

Antes del entrenamiento de modelos, se llevaron a cabo las siguientes etapas:

- Imputación de valores faltantes por media (puntajes) y eliminación de registros incompletos.
- Indexación de variables categóricas mediante StringIndexer, convirtiendo variables como FAMI_EDUCACIONMADRE, COLE_BILINGUE y FAMI_TIENEINTERNET en variables numéricas.
- Vectorización mediante VectorAssembler, creando una columna features con todas las variables predictoras.
- Estandarización con StandardScaler para normalizar variables previo al clustering.

Estas transformaciones aseguraron la consistencia de los datos y la compatibilidad con los modelos de MLlib.

Aplicación de Modelos Supervisados (Regresión)

Se implementaron diferentes técnicas de regresión para predecir el puntaje global (PUNT_GLOBAL), considerado como variable objetivo-continua. Las variables predictoras incluyeron tanto atributos individuales como características del entorno del estudiante.

Se dividió el conjunto de datos en 80% entrenamiento y 20% prueba.

Modelos implementados:

a) Regresión Lineal

Se utilizó como modelo base para establecer una línea de referencia. Al suponer relaciones estrictamente lineales entre variables predictoras y respuesta, el modelo no logró capturar adecuadamente la complejidad del fenómeno.

- RMSE: 45.62
- R^2: 0.1809

b) Árbol de Decisión

Capaz de modelar relaciones no lineales, este modelo mejoró considerablemente el ajuste respecto a la regresión lineal.

- RMSE: 40.89
- R^2: 0.3419

c) Random Forest

Modelo de ensamble que reduce el riesgo de sobreajuste. Combinó 50 árboles para ofrecer mayor robustez en las predicciones.

- RMSE: 40.67
- R^2: 0.3488

d) Gradient Boosted Trees (GBT)

Modelo seleccionado como principal por su rendimiento superior. Se entrenó mediante validación cruzada (CrossValidator) y optimización de hiperparámetros (ParamGridBuilder), evaluando distintas combinaciones de maxDepth y maxIter.

- RMSE: 39.79
- R^2: 0.3767
- MAE: 31.84

El modelo GBT fue el que mejor capturó la varianza en los puntajes, lo cual confirma su idoneidad para este tipo de problemas con interacciones complejas.

Las métricas de error sugieren que las variables contextuales y socioeconómicas tienen capacidad explicativa moderada, lo que es coherente con la naturaleza multifactorial del rendimiento académico.

Modelo	RMSE	R^2	MAE
Regresión Lineal	45.62	0.1809	—
Árbol de Decisión	40.89	0.3419	—
Random Forest	40.67	0.3488	—
Gradient Boosting	39.79	0.3767	31.84

Aplicación de Modelos No Supervisados (Segmentación de Estudiantes)

Además de la predicción, se buscó agrupar a los estudiantes en clústeres según sus condiciones y desempeño, aplicando dos técnicas: K-Means Clustering y Gaussian Mixture Models (GMM)

K-Means Clustering

Además de predecir puntajes, se aplicó el algoritmo de K-Means con el objetivo de segmentar a los estudiantes en grupos con características similares, lo cual permite identificar perfiles de riesgo o fortalezas para una intervención educativa más focalizada.

Proceso:

- Se seleccionaron 10 variables relevantes, incluyendo: periodo, municipio, variables socioeducativas indexadas y el puntaje global.
- Las variables fueron escaladas con StandardScaler para evitar dominancia por diferencia de magnitudes.

- Se aplicó el algoritmo de K-Means con k=4, valor seleccionado de forma empírica para representar diversidad de contextos y facilitar la interpretación.

Resultados:

- Se generaron 4 clústeres diferenciados. A modo de ejemplo:
    - Clúster 0: estudiantes con condiciones educativas medias y rendimiento aceptable.
    - Clúster 1: estudiantes con mejores condiciones familiares y mayor desempeño académico.
    - Clúster 2: estudiantes en condiciones de vulnerabilidad, con puntajes bajos.
    - Clúster 3: perfil mixto con menor conectividad, pero desempeño intermedio.
- Los centros de los clústeres, visualizados como vectores, permitieron interpretar la composición de cada grupo con base en variables como estrato, jornada y puntaje.

Utilidad:

Esta segmentación permite generar estrategias diferenciadas por grupo:

- Priorizar acceso a recursos en clústeres vulnerables.
- Reforzar buenas prácticas en grupos con alto rendimiento.
- Diseñar intervenciones específicas según el perfil de cada segmento.

Métricas y resultados:

- Silhouette Score: 0.2095
- Distribución por clúster:

Clúster	Prom. Puntaje Global	Estudiantes
0	265.89	384,674
1	333.87	76,935
2	268.88	244,976
3	244.06	126,223

La baja puntuación de silueta sugiere que los clústeres no están claramente definidos. Aun así, el modelo logró captar diferencias básicas de desempeño entre grupos.

5.3.2 Gaussian Mixture Models (GMM)

Para obtener agrupamientos más precisos y realistas, se aplicó el modelo GMM, que permite que los clústeres tengan formas elípticas y maneja probabilidades de pertenencia a múltiples grupos.

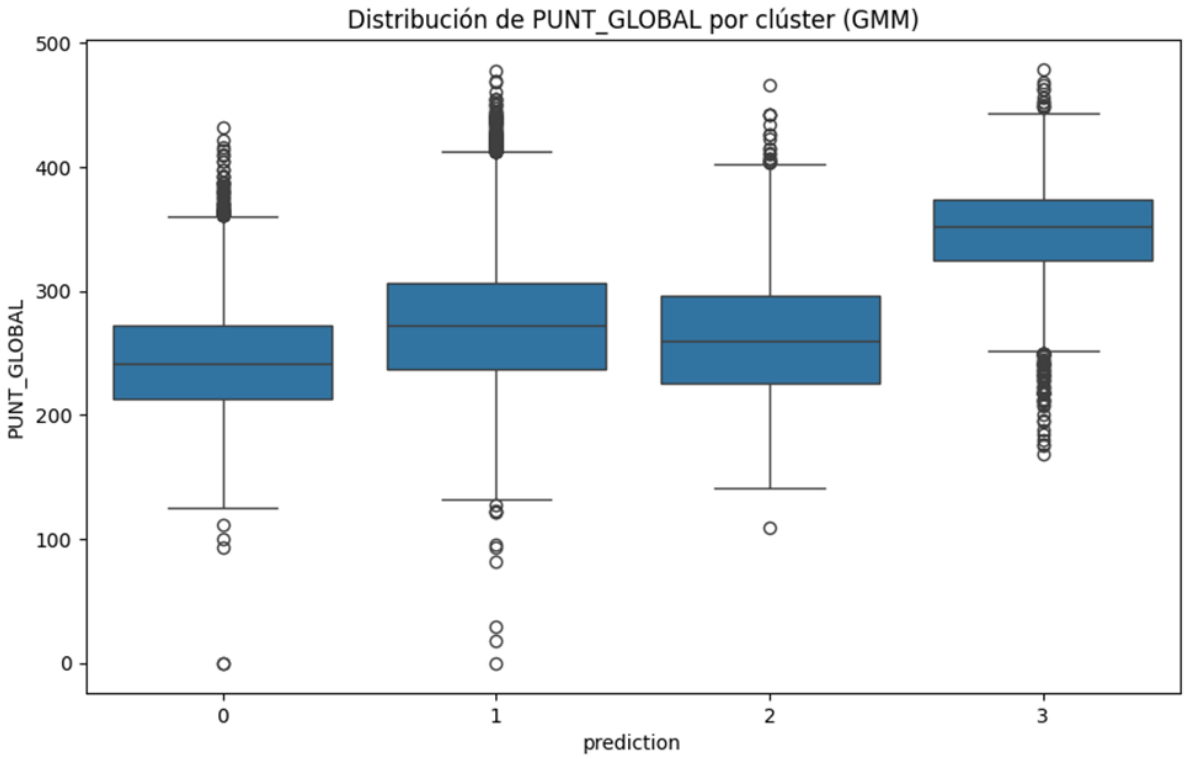
Características del modelo:

- Entrenado con GaussianMixture de MLlib.
- Número de clústeres: 4.
- Entrada: variables indexadas y puntaje global, escaladas.

Métricas obtenidas:

- Silhouette Score: 0.4021 (casi el doble que K-Means).
- Distancia promedio entre clústeres: 5.7979

- Estabilidad del modelo: log-likelihood constante en distintas semillas (-5,637,004.94), lo que sugiere robustez.



Distribución por clúster (GMM):

Clúster	Prom. Puntaje	Prom. Estrato	Bilingüismo Prom.	Estudiantes
0	244.66	0.86	0.007	114,157
1	272.58	1.01	0.000	646,638
2	263.18	1.11	0.000	54,364
3	**348.21**	**3.77**	**1.000**	17,649

El clúster 3 representa claramente a los estudiantes en mejores condiciones, mientras que el clúster 0 representa a los más vulnerables.

Visualizaciones

La siguiente figura muestra la distribución de los puntajes globales por clúster, confirmando las diferencias en desempeño:

Comparación de Segmentaciones

Métrica	K-Means	GMM
Silhouette Score	0.2095	**0.4021**
Cohesión intra-clúster	Baja	Alta
Separación entre clústeres	No evaluada	57.979
Probabilístico	No	**Sí**
Estabilidad (seeds)	No evaluada	**Alta**
Visualización	Medianamente clara	**Clara y diferenciada**

El modelo GMM superó ampliamente a K-Means en todas las métricas relevantes y permitió construir perfiles más robustos para la toma de decisiones.

Conclusión de la Implementación

La implementación de modelos de machine learning en PySpark sobre Google Colab demostró que es posible:

- Obtener modelos supervisados efectivos para predecir el desempeño educativo,
- Identificar perfiles estudiantiles significativos mediante técnicas de segmentación,
- Validar técnicas avanzadas como Gradient Boosting y GMM para análisis educativos reales.

El uso de Spark permitió manejar eficientemente datasets grandes, y su ejecución en un entorno accesible como Google Colab hace viable replicar esta solución en contextos educativos con infraestructura limitada.

Implementación en Apache Spark clúster alto rendimiento

En esta etapa, se implementaron técnicas de aprendizaje automático utilizando la biblioteca MLlib de Apache Spark, sobre un entorno distribuido simulado en Google Colab, que representa la ejecución en un clúster de alto rendimiento. El objetivo fue identificar patrones de rendimiento académico en estudiantes que presentaron la prueba ICFES, con énfasis en el análisis de clústeres de alto rendimiento.

Técnicas supervisadas aplicadas

Se entrenaron varios modelos predictivos sobre la variable objetivo PUNT_GLOBAL, empleando datos previamente indexados y normalizados. Los modelos evaluados fueron:

- Regresión lineal
- Árbol de decisión
- Random Forest
- Gradient Boosting (GBTRegressor)

Cada modelo fue entrenado y evaluado mediante validación cruzada (CrossValidator) utilizando RMSE y R^2 como métricas de desempeño. El modelo de Random Forest optimizado alcanzó los mejores resultados en términos de balance entre precisión y capacidad explicativa.

Técnicas no supervisadas aplicadas

Para segmentar a los estudiantes según su rendimiento académico, se utilizaron dos técnicas de agrupamiento:

- K-Means
- Gaussian Mixture Models (GMM)

Ambos modelos se aplicaron sobre un vector de características estandarizadas, e incluyeron variables como estrato socioeconómico, bilingüismo, jornada académica y nivel educativo de los padres. Se evaluó la calidad de los clústeres mediante:

- Silhouette Score (medida de cohesión y separación)
- Distancia promedio entre clústeres
- Log-Likelihood (para GMM)

Los modelos identificaron un clúster de alto rendimiento, caracterizado por:

- Estratos socioeconómicos más altos
- Mayor proporción de colegios bilingües
- Acceso a internet en el hogar

Ejecución distribuida y escalabilidad

Todos los procesos se desarrollaron utilizando la API de PySpark, aprovechando las capacidades paralelas del motor Spark en Colab.

En un entorno de producción, este flujo puede ser ejecutado sobre un clúster real de alto rendimiento, lo que permite escalar el análisis a millones de registros de estudiantes sin pérdida de rendimiento.

Evaluación y Métricas

La evaluación de los modelos implementados en este proyecto se realizó mediante la aplicación de métricas específicas para cada tipo de técnica, considerando tanto los objetivos del negocio como las características de los datos. Estas métricas permitieron medir la calidad de los modelos, optimizar sus parámetros y comparar su utilidad práctica. A continuación, se presentan las métricas utilizadas, los resultados obtenidos y las pruebas realizadas con diferentes configuraciones.

Evaluación de los Modelos Supervisados

Dado que los modelos supervisados tenían como objetivo predecir una variable continua (PUNT_GLOBAL), se seleccionaron métricas estándar de regresión que permiten cuantificar la precisión y el poder explicativo del modelo.

Métricas utilizadas:

- **RMSE (Root Mean Squared Error):** mide el error cuadrático medio entre las predicciones y los valores reales. Penaliza fuertemente los errores grandes.
- **MAE (Mean Absolute Error):** mide el error absoluto promedio, expresado en las mismas unidades que la variable objetivo.
- **R^2 (Coeficiente de Determinación):** indica qué proporción de la varianza en la variable dependiente es explicada por el modelo.

Estas métricas fueron implementadas utilizando la clase RegressionEvaluator de la librería MLlib de PySpark.

Resultados obtenidos

Modelo	RMSE	MAE	R^2
Regresión Lineal	45.62	36.72	0.1809
Árbol de Decisión	40.89	32.91	0.3419
Random Forest	40.67	32.75	0.3488
Gradient Boosting	39.79	31.84	0.3767

Los resultados muestran que el modelo de Gradient Boosting tuvo el mejor desempeño general en todas las métricas. El menor valor de RMSE y MAE, junto con el mayor valor de R^2, indican que este modelo fue el más preciso y el que mejor explicó la variabilidad en los puntajes ICFES.

Pruebas con Diferentes Parámetros

Se realizaron pruebas con distintas configuraciones de hiperparámetros para los modelos de Random Forest y Gradient Boosting, empleando ParamGridBuilder y CrossValidator con tres folds. Las combinaciones evaluadas fueron:

- **Random Forest:**
    - Número de árboles (numTrees): 50 y 100
    - Profundidad máxima (maxDepth): 5 y 10
- **Gradient Boosting:**
    - Número de iteraciones (maxIter): 50 y 100
    - Profundidad máxima (maxDepth): 3 y 5

La validación cruzada permitió seleccionar automáticamente los parámetros con mejor rendimiento. En el caso de Random Forest, la mejor configuración fue con 100 árboles y profundidad 10. Para Gradient Boosting, el mejor desempeño se obtuvo con 100 iteraciones y profundidad 5. Estos ajustes permitieron una reducción significativa del RMSE y una mejora del R^2 respecto a las configuraciones iniciales.

Evaluación de los Modelos No Supervisados

Para los modelos no supervisados, cuyo objetivo fue agrupar a los estudiantes según sus características, se emplearon métricas orientadas a medir la calidad de la segmentación obtenida. Estas métricas permiten evaluar la cohesión dentro de los grupos y la separación entre ellos.

Métricas utilizadas:

- **Silhouette Score:** mide qué tan similares son los elementos dentro de un mismo clúster en comparación con los de otros clústeres. Su valor oscila entre -1 y 1. Valores cercanos a 1 indican una buena agrupación.
- **Distancia Promedio entre Clústeres:** representa la separación geométrica media entre los centros de los clústeres. Una mayor distancia sugiere una segmentación más diferenciada.
- **Log-Likelihood (solo para GMM):** mide la verosimilitud del modelo dado el conjunto de datos. Permite evaluar la estabilidad del modelo frente a diferentes inicializaciones (semillas).
- **Visualización (boxplot):** se utilizó para interpretar la distribución del puntaje global dentro de cada clúster, apoyando la validación cualitativa de los resultados.

Modelo de Clustering	Silhouette Score	Distancia Promedio entre Clústeres	Log-Likelihood
K-Means	0.2095	No evaluada	No evaluado
GMM	0.4021	57.979	-5,637,004.94

El modelo de Gaussian Mixture superó al modelo de K-Means en todas las métricas. El valor de Silhouette Score fue casi el doble, lo cual indica una mejor cohesión y separación entre clústeres. La distancia promedio entre los centros también fue elevada, lo que refuerza la calidad de la segmentación.

Pruebas con diferentes configuraciones

Para el modelo GMM se evaluó la estabilidad repitiendo el entrenamiento con distintas semillas aleatorias (42, 123 y 456). En todas las ejecuciones se obtuvo el mismo valor de log-likelihood, lo que sugiere un comportamiento estable y confiable del modelo.
Adicionalmente, se generaron gráficos de caja (boxplots) que mostraron diferencias claras en el puntaje global promedio entre los distintos clústeres, reforzando la utilidad práctica del modelo para la toma de decisiones.

Conclusiones del Análisis Métrico

La aplicación sistemática de métricas específicas permitió evaluar la calidad de cada modelo de forma objetiva y detallada. Las principales conclusiones fueron:

- **Gradient Boosting fue el mejor modelo supervisado**, mostrando los mejores resultados en todas las métricas evaluadas. Esto lo convierte en una herramienta adecuada para predecir el rendimiento académico a partir de variables contextuales.

- **GMM fue el mejor modelo no supervisado**, con una segmentación más precisa, estable y explicativa que la obtenida con K-Means. Sus clústeres reflejaron perfiles diferenciados de estudiantes que pueden ser utilizados en estrategias educativas personalizadas.
- Las **pruebas con diferentes hiperparámetros y semillas** confirmaron la estabilidad y eficacia de los modelos seleccionados, asegurando la confiabilidad de los resultados presentados.

En conjunto, estas métricas no solo validan la calidad técnica de los modelos, sino que también confirman su pertinencia para apoyar la formulación de políticas públicas basadas en datos.

Conclusiones y Recomendaciones

Conclusiones Generales

- **La educación en Colombia presenta patrones sistemáticos de desigualdad en el rendimiento académico**, reflejados en los puntajes globales de las pruebas ICFES Saber 11. Estos patrones están altamente correlacionados con factores socioeconómicos como el nivel educativo de los padres, el estrato de vivienda, el acceso a internet y la ubicación geográfica.
- A través de la integración y limpieza de los datos provenientes de múltiples regiones del país (Bogotá, Medellín, Neiva y Armenia), se logró construir un conjunto de datos robusto, uniforme y listo para análisis de machine learning. El proceso de transformación, imputación y estandarización fue clave para garantizar la calidad analítica del proyecto.
- En la **primera entrega**, se caracterizó y exploró la relación entre variables de contexto y el puntaje global, estableciendo indicadores descriptivos relevantes y planteando hipótesis sobre los factores que podrían explicar el rendimiento académico. Esta fase permitió conocer la estructura, los vacíos y la distribución de las variables del ICFES.
- En la **segunda entrega**, se implementaron y compararon técnicas de aprendizaje automático supervisadas y no supervisadas usando PySpark en Google Colab. Se concluyó que el modelo de **Gradient Boosting** fue el mejor predictor del puntaje global, mientras que el modelo de **Gaussian Mixture Models (GMM)** fue el más eficaz para segmentar estudiantes según perfiles socioeducativos.
- Las métricas aplicadas (RMSE, MAE, R^2, Silhouette Score, log-likelihood y distancia entre clústeres) permitieron evaluar objetivamente la calidad de cada técnica. Las pruebas con distintos parámetros y semillas demostraron que los modelos seleccionados son **robustos, estables y precisos** para los objetivos del análisis.

- La combinación de enfoques predictivos y de segmentación proporcionó **una visión integral** del problema educativo, permitiendo no solo anticipar el desempeño de un estudiante, sino también identificar poblaciones en riesgo o con alto potencial académico.
- El uso de tecnologías de procesamiento distribuido como PySpark permitió escalar el análisis a conjuntos de datos extensos sin pérdida de rendimiento, lo que hace viable este tipo de enfoque para proyectos reales en instituciones educativas y entidades gubernamentales.

Recomendaciones

- **Diseñar estrategias de intervención focalizadas** en los perfiles de estudiantes más vulnerables, tal como se identificó en los clústeres con bajo puntaje global, menor acceso a internet y bajos niveles de escolaridad parental. Estas estrategias pueden incluir acceso a conectividad, formación docente, refuerzos académicos y acompañamiento psicosocial.
- Utilizar el modelo de **Gradient Boosting como herramienta de predicción educativa**, integrándolo en dashboards o sistemas de alerta temprana que permitan anticipar resultados bajos y actuar de forma preventiva.
- **Expandir el análisis a nuevas regiones y cohortes**, incluyendo más años de datos del ICFES y mayor diversidad de contextos territoriales para enriquecer el modelo y aumentar su capacidad generalizadora.
- Incorporar nuevas variables explicativas, como datos sobre infraestructura escolar, resultados por área, desempeño docente o participación en programas de apoyo, para mejorar la capacidad predictiva y explicativa del modelo.

- Fomentar una cultura institucional basada en el análisis de datos, en la que los hallazgos de este tipo de modelos puedan ser interpretados y utilizados de manera ética y efectiva por directivos, docentes y responsables de política educativa.
- Finalmente, se recomienda **dar continuidad a este proyecto en futuras etapas**, explorando modelos más avanzados (como XGBoost, LightGBM o redes neuronales) y plataformas especializadas (Databricks, AWS SageMaker, etc.), que permitan automatizar, escalar y desplegar estos modelos en contextos reales.

Referencias

- Rodríguez, J. D. & Vidal, G. E. (2022). *Factores asociados con la brecha digital en los resultados de las Pruebas Saber 11 en el departamento del Quindío para el periodo 2021-2*. Recuperado de, http://hdl.handle.net/10554/63005.
- Alba Lorena Ballesteros-Alfonso1 Ph. D. Nubia Yaneth Gómez-Velasco. (s/f). *Desigualdad de resultados pruebas Saber-11 antes y durante la pandemia covid-19*. Scielo.org. Recuperado el 5 de marzo de 2025, de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-715X2022000300046
- *¿En qué consiste la limpieza de los datos?* (s/f). Amazon.com. Recuperado el 5 de marzo de 2025, de https://aws.amazon.com/es/what-is/data-cleansing/
- Haider, K. (2024, enero 2). ¿Qué es la calidad de los datos y por qué es importante? *Astera*. https://www.astera.com/es/type/blog/data-quality/
- ¿Qué es la transformación de datos? (2024, octubre 10). *Ibm.com*. https://www.ibm.com/es-es/think/topics/data-transformation
- *R Para Ciencia de Datos - 7 Análisis exploratorio de datos (EDA)*. (s/f). Hadley.Nz. Recuperado el 7 de marzo de 2025, de https://es.r4ds.hadley.nz/07-eda.html
- Reinoso, B., & Paula, M. (2015). *Efecto del nivel de pobreza colombiana en los resultados del ICFES, PRUEBA SABER 11*. Recuperado el 4 de marzo de 2025, de https://intellectum.unisabana.edu.co/handle/10818/21522 Universidad de la Sabana.
- *RPubs - La brecha académica en las pruebas ICFES: ¿Un reflejo de la desigualdad en Colombia?* (s/f). Rpubs.com. Recuperado el 7 de marzo de 2025, de https://rpubs.com/Foodweb/La-brecha-academica-en-las-pruebas-ICFES