Informe Parcial II Procesamiento de Datos a Gran Escala

Daniel Sandoval Higuera , Daniela Torres Gómez, Isaac Janica

[1]Ciencia de Datos & Pontifica Universidad Javeriana, Colombia

[2]Ingeniería de Sistemas & Pontifica Universidad Javeriana, Colombia

En este documento se desarrolló un proyecto para poner en práctica, con datos reales, los conceptos aprendidos en la clase de procesamiento de datos a gran escala. Exploramos y brindamos información útil acerca los datos de los resultados de las pruebas de estado Saber 11.

Keywords— Saber 11, pyspark, big-data,, procesamiento distribuido, análisis descentralizado, educación, evaluación académica, machine learning, modelos predictivos.

I. INTRODUCCIÓN

Las pruebas Saber 11, en el contexto educativo de Colombia, representan un pilar fundamental en la evaluación del rendimiento académico de los estudiantes de educación media. Estas pruebas, administradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), abarcan diversas áreas del conocimiento y proporcionan una métrica objetiva para medir el nivel de competencia de los estudiantes en habilidades básicas.

La importancia de este proyecto radica en la posibilidad de identificar tendencias, patrones y factores que influyen en el desempeño de los estudiantes en las pruebas Saber 11. Estos insights no solo pueden beneficiar a las instituciones educativas y al sistema de evaluación, sino también a los encargados de formular políticas educativas y a la sociedad en general, al proporcionar una base sólida para la toma de decisiones informadas en el ámbito educativo.

En este informe, presentaremos nuestra metodología, resultados preliminares y reflexiones sobre el impacto potencial de nuestro trabajo en el mejoramiento continuo de la calidad de la educación en Colombia.

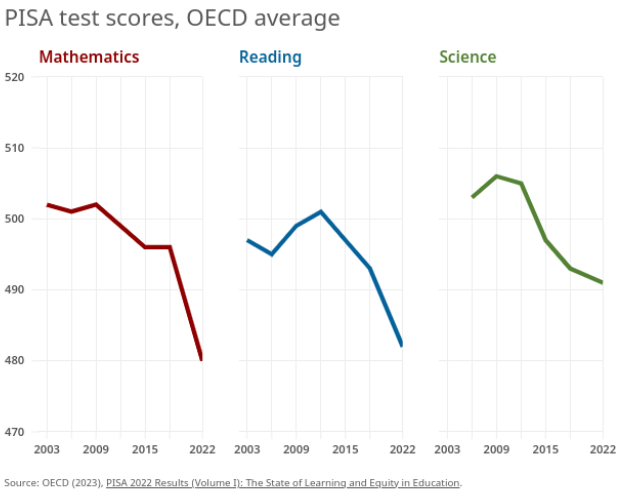
II. ENTENDIMIENTO DEL NEGOCIO Y PLANTEAMIENTO PROBLEMÁTICA

Para empezar, una de las estadísticas más significativas es el hecho de que la prueba es anualmente presentada por casi medio millón de estudiantes del sistema de educación escolar de Colombia, esto se puede evidenciar en al siguiente gráfica:

Muchos padres usan esta preuba como referencia para medir a “los mejores” colegios de Colombia basado en los calendarios que quieran inscribir a sus hijos: Calendario A y Calendario B. El año pasado los mejores 5 colegios calendario A según esta prueba fueron:
1). Colegio Bilingüe Divino Niño (393 puntos, Bucaramanga)
2). Centro Educativo Boston Internacional (386, Barranquilla)
3). Colegio Empresarial de los Andes (382, Neiva)
4). Liceo Campo David (377, Bogotá)
5). Colegio Nuevo Colombo Americano (373, Bogotá)
Por otro lado, los colegios calendario B que mejor se desempeñaron en esta prueba (top 5) fueron:
1). Colegio Bilingüe Diana Oese (389, Cali)
2). Colegio Nuevo Cambridge (378, Floridablanca, Santander)
3). Colegio San Jorge de Inglaterra (369, Bogotá)
4). Aspaen Gimnasio La Fragua (369, Neiva)
5). Colegio San Carlos (365, Bogotá)

En términos de la prueba la página oficial de la entidad que hace estas pruebas nos dice que esta se presenta en: “La jornada de aplicación de la prueba Saber 11° calendario A se realizará en 11.180 sitios dispuestos por el Icfes a partir de las 7 de la mañana; en dos sesiones, cada una de 4 horas y media. Durante la sesión de la mañana las personas responderán 125 preguntas; y en la tarde responderán 118 preguntas.
Las competencias que se evaluarán en esta prueba serán: matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas e Inglés.”

Ahora, si abordamos el tema de la educación escolar con enfoque Internacional nos podemos remitir a la OCDE (Organización para la Cooperación y Desarrollo Económico) la cuál a través de sus pruebas PISA que ha habido una disminución en rendimiento académico de los estudiantes postpandemia en ciertos componentes, esto se puede ver en la siguiente gráfica:



Esta organización mide el nivel educativo de sus países miembros debido a que este es uno de los factores más importantes para el desarrollo es la educación de alta calidad. Cada cierta cantidad de años se hacen las pruebas, estas miden las competencias de los estudiantes en materia de matemáticas, lectura ciencias. Si bien no se puede generalizar que debido a la pandemia hubo una afectación grave en todas las materias evaluadas, si se puede ver que hay una mayor disminución en el componente de matemáticas después de la pandemia. Es importante hacer una revisión de que tanto ha disminuido este componente con respecto a los años anteriores

Ahora, en este informe, nos adentramos en el mundo del procesamiento de datos a gran escala con un enfoque específico en los resultados de las pruebas Saber 11. A través de la aplicación de técnicas de big data y herramientas como pyspark, buscamos analizar exhaustivamente estos datos para extraer información valiosa que pueda contribuir a comprender mejor el panorama educativo colombiano.

El proceso de evaluación del rendimiento académico de los estudiantes mediante las pruebas Saber 11 en Colombia no solo es crucial para las instituciones educativas, sino que también influye en las políticas públicas y en la percepción general sobre la calidad de la educación en el país. Sin embargo, este proceso enfrenta desafíos significativos en términos de análisis y comprensión de los datos generados.

La complejidad de los datos de las pruebas Saber 11, que abarcan múltiples áreas de conocimiento y diversos factores socioeconómicos, requiere un enfoque de procesamiento de datos a gran escala para obtener insights significativos. Por lo tanto, el

objetivo principal de este proyecto es utilizar técnicas de big data y herramientas como pyspark para analizar de manera exhaustiva estos datos y abordar las siguientes problemáticas:

Identificación de patrones y tendencias: ¿Qué patrones se pueden identificar en los resultados de las pruebas Saber 11 a lo largo del tiempo (pre-pandemia y post-pandemia)? ¿Existen tendencias claras en el desempeño de los estudiantes en diferentes áreas del conocimiento?

Influencia de la pandemia en los resultados de las pruebas: ¿Ha afectado la pandemia los resultados de las pruebas, es decir, el desempeño académico de los estudiantes?

Influencia de variables socioeconómicas: ¿Cómo afectan factores como el nivel socioeconómico y el tipo de institución educativa al rendimiento académico de los estudiantes en las pruebas Saber 11? ¿Existen disparidades significativas que requieran atención específica?

Predicción del rendimiento estudiantil: ¿Es posible desarrollar modelos predictivos utilizando datos históricos de las pruebas Saber 11? ¿Qué variables son más relevantes para predecir el desempeño académico de los estudiantes?

Al abordar estas problemáticas, buscamos no solo proporcionar una visión integral del rendimiento académico en Colombia, sino también generar insights accionables que puedan informar políticas educativas y estrategias de intervención destinadas a mejorar la calidad y equidad en la educación del país. En este informe, detallaremos nuestra metodología para abordar estas cuestiones y presentaremos los resultados preliminares obtenidos hasta el momento..

III. COLECCIÓN Y DESCRIPCIÓN DE LOS DATOS

Para este apartado se quiere explicar que este dataset encontrado es la sumatoria de varios datasets, es decir, que no fue necesaria la unión de varios conjuntos de datos debido a que la página de datosabiertos.gov ya proporciona este conjunto de datos final.
EL conjunto de datos se conforma por datos de las pruebas realizadas en los siguientes años: 20181 ,20191, 20194, 20211, 20221 y 20224. Sin embargo, el número de filas por año no son la misma cantidad por año,

```
PERIODO
2019    558751
2022    552841
2018     19798
2021     15528
Name: count, dtype: int64
```

Como se puede observar la mayoría de los datos provienen de los datos 2019 y 2022. Esto no significa que los datos de 2018 y 2021 no tengan cierta incidencia, pero esto si va a verse reflejado en los análisis posteriores.
Ahora, lo que podemos evidenciar es que este conjunto de datos está compuesto de dos grandes conjuntos: Los resultados de las pruebas en 2019 y los resultados de las pruebas en 2022. Esta muestra es bastante significativa y nos va a poder ayudar a comparar los resultados prepandémicos y pospandemicos en los resultados de las pruebas saber 11.
A continuación, se presentan todas las columnas del DataSet con sus respectivos tipos de datos:

Column Name	Descripción de la columna	Data Type
PERIODO	Año en el que se realizó la prueba.	int64
ESTU_TIPODOCUMENTO	Tipo de documento de identificación del estudiante	string
ESTU_CONSECUTIVO	Identificador único para la prueba de cada estudiante.	string
COLE_AREA_UBICACION	Especifica la ubicación de la institución educativa (Rural/Urbano)	string
COLE_BILINGUE	Indica si la institución educativa es bilingüe (Sí o No).	string
COLE_CALENDARIO	Representa el calendario académico seguido por la institución	string
COLE_CARACTER	Describe la naturaleza o carácter de la institución	string
COLE_COD_DANE_ESTABLECIMIENTO	Código DANE asignado al establecimiento educativo.	float64
COLE_COD_DANE_SEDE	Código DANE asignado a la sede de la institución educativa.	float64

COLE_COD_DEPTO_UBICACION	Código DANE del departamento donde se encuentra la institución educativa.	float64
COLE_COD_MCPIO_UBICACION	Código DANE del municipio donde se encuentra la institución educativa.	float64
COLE_CODIGO_ICFES	Código ICFES asignado a la institución educativa.	float64
COLE_DEPTO_UBICACION	Departamento donde se encuentra la institución educativa.	string
COLE_GENERO	Género o sexo de la institución educativa (Feminino, masculino, mixto)	string
COLE_JORNADA	Representa el turno o horario seguido por la institución	string
COLE_MCPIO_UBICACION	Municipio donde se encuentra la institución educativa.	string
COLE_NATURALEZA	Especifica la naturaleza de la institución educativa	string
COLE_NOMBRE_ESTABLECIMIENTO	Nombre del establecimiento educativo.	string
COLE_NOMBRE_SEDE	Nombre de la sede de la institución educativa.	string
COLE_SEDE_PRINCIPAL	Indica si la sede de la institución es principal (Sí o No).	string
ESTU_COD_DEPTO_PRESENTACION	Código DANE del departamento donde se presentó la prueba.	int64
ESTU_COD_MCPIO_PRESENTACION	Código DANE del municipio donde se presentó la prueba.	int64
ESTU_COD_RESIDE_DEPTO	Código DANE del departamento donde reside el estudiante.	float64
ESTU_COD_RESIDE_MCPIO	Código DANE del municipio donde reside el estudiante.	float64
ESTU_DEPTO_PRESENTACION	Departamento donde se presentó la prueba.	string
ESTU_DEPTO_RESIDE	Departamento donde reside el estudiante.	string
ESTU_ESTADOINVESTIGACION	Indica el estado de investigación del estudiante (por ejemplo	string
ESTU_ESTUDIANTE	Indica si la persona es estudiante	string
ESTU_FECHANACIMIENTO	Fecha de nacimiento del estudiante.	string
ESTU_GENERO	Género o sexo del estudiante.	string
ESTU_MCPIO_PRESENTACION	Municipio donde se presentó la prueba.	string
ESTU_MCPIO_RESIDE	Municipio donde reside el estudiante.	string
ESTU_NACIONALIDAD	Nacionalidad del estudiante.	string
ESTU_PAIS_RESIDE	País donde reside el estudiante.	string
ESTU_PRIVADO_LIBERTAD	Indica si el estudiante está privado de libertad (Sí o No).	string
FAMI_CUARTOSHOGAR	Número de habitaciones en el hogar.	string
FAMI_EDUCACIONMADRE	Nivel educativo de la madre.	string
FAMI_EDUCACIONPADRE	Nivel educativo del padre.	string
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico del hogar.	string
FAMI_PERSONASHOGAR	Número de personas que viven en el hogar.	string
FAMI_TIENEAUTOMOVIL	Indica si el hogar tiene automóvil (Sí o No).	string
FAMI_TIENECOMPUTADOR	Indica si el hogar tiene computadora (Sí o No).	string
FAMI_TIENEINTERNET	Indica si el hogar tiene acceso a internet (Sí o No).	string
FAMI_TIENELAVADORA	Indica si el hogar tiene lavadora (Sí o No).	string

DESEMP_INGLES	Nivel de competencia en inglés.	string
PUNT_INGLES	Puntaje obtenido en la prueba de inglés.	float64
PUNT_MATEMATICAS	Puntaje obtenido en la prueba de matemáticas.	int64
PUNT_SOCIALES_CIUDADANAS	Puntaje obtenido en la prueba de sociales y ciudadanas.	int64
PUNT_C_NATURALES	Puntaje obtenido en la prueba de ciencias naturales.	int64
PUNT_LECTURA_CRITICA	Puntaje obtenido en la prueba de lectura crítica.	int64
PUNT_GLOBAL	Puntaje total obtenido en la prueba.	int64

Nombre Columna	Número de Nulos
PERIODO	0
ESTU_TIPODOCUMENTO	0
ESTU_CONSECUTIVO	0
COLE_AREA_UBICACION	1
COLE_BILINGUE	198611
COLE_CALENDARIO	1
COLE_CARACTER	40747
COLE_COD_DANE_ESTABLECIMIENTO	1
COLE_COD_DANE_SEDE	1
COLE_COD_DEPTO_UBICACION	1
COLE_COD_MCPIO_UBICACION	1
COLE_CODIGO_ICFES	1
COLE_DEPTO_UBICACION	1
COLE_GENERO	1
COLE_JORNADA	1
COLE_MCPIO_UBICACION	1
COLE_NATURALEZA	1
COLE_NOMBRE_ESTABLECIMIENTO	1
COLE_NOMBRE_SEDE	1
COLE_SEDE_PRINCIPAL	1
ESTU_COD_DEPTO_PRESENTACION	0
ESTU_COD_MCPIO_PRESENTACION	0
ESTU_COD_RESIDE_DEPTO	1174
ESTU_COD_RESIDE_MCPIO	1174
ESTU_DEPTO_PRESENTACION	0
ESTU_DEPTO_RESIDE	1174
ESTU_ESTADOINVESTIGACION	0

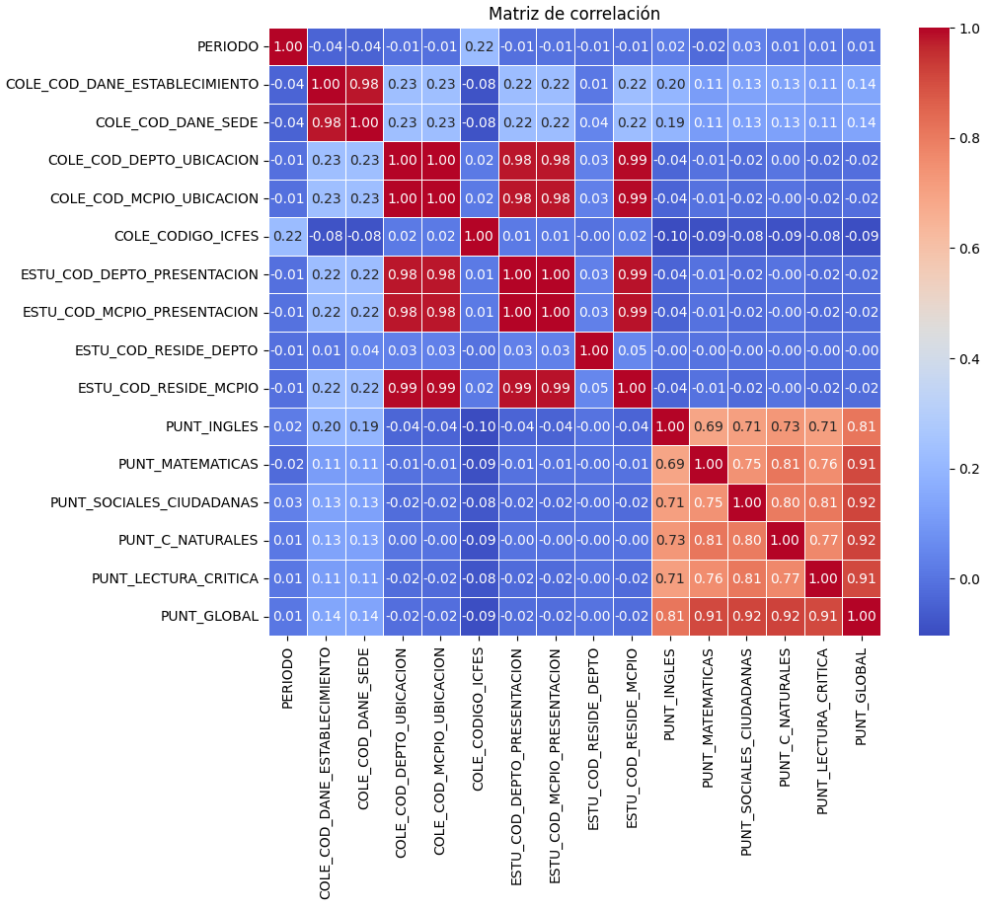
ESTU_ESTUDIANTE	0
ESTU_FECHANACIMIENTO	81
ESTU_GENERO	143
ESTU_MCPIO_PRESENTACION	0
ESTU_MCPIO_RESIDE	1174
ESTU_NACIONALIDAD	0
ESTU_PAIS_RESIDE	0
ESTU_PRIVADO_LIBERTAD	0
FAMI_CUARTOSHOGAR	44642
FAMI_EDUCACIONMADRE	66316
FAMI_EDUCACIONPADRE	66354
FAMI_ESTRATOVIVIENDA	75632
FAMI_PERSONASHOGAR	42645
FAMI_TIENEAUTOMOVIL	46556
FAMI_TIENECOMPUTADOR	44245
FAMI_TIENEINTERNET	67571
FAMI_TIENELAVADORA	44096
DESEMP_INGLES	2128
PUNT_INGLES	2183
PUNT_MATEMATICAS	0
PUNT_SOCIALES_CIUDADANAS	0
PUNT_C_NATURALES	0
PUNT_LECTURA_CRITICA	0
PUNT_GLOBAL	0

IV. ANÁLISIS EXPLORATORIO DE DATOS

Para este Análisis Exploratorio de Datos utilizamos varias gráficas para poder entender estos datos de una mernar más profunda. Este análisis fue hecho también de manera que se pueda ver las pre-pandemia y la post-pandemia así cómo el conjunto de datos en general. A continuación se va a empezar con el conjunto de datos en general.

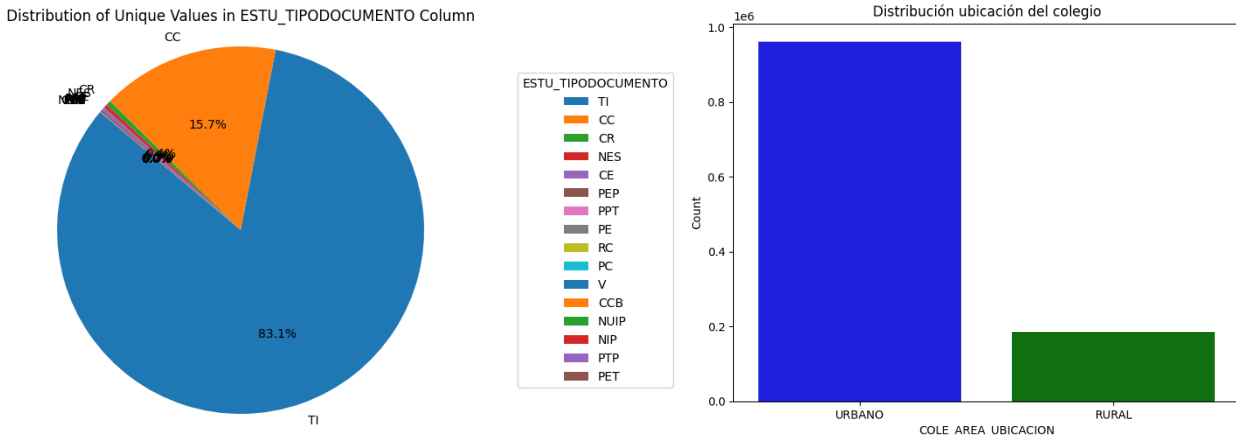
a. Conjunto de datos en general:

A continuación mostramos una matriz de correlación, únicamente de las variables que contienen solo números:



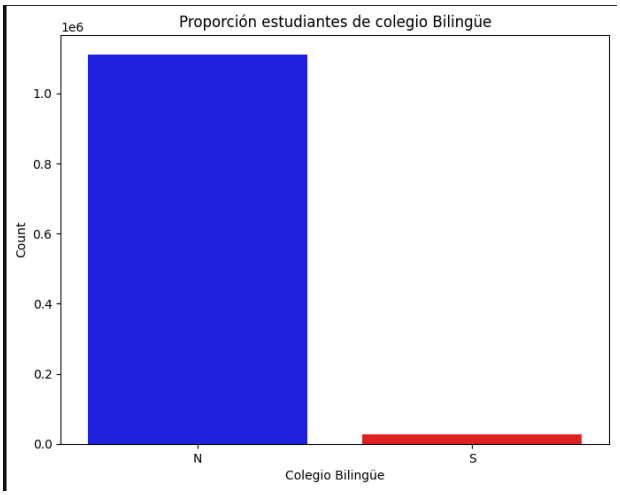
Algo que si bien es obvio, pero importante de resaltar es la fuerte correlación que tienen los resultados de la prueba de matemáticas con las ciencias naturales, así cómo la correlación que hay entre el puntaje de lectura crítica y el puntaje de ciencias sociales y ciudadanas. También resalta el hecho de que hay una baja correlación (a comparación del resto de los puntajes) entre el puntaje de inglés y el puntaje global.

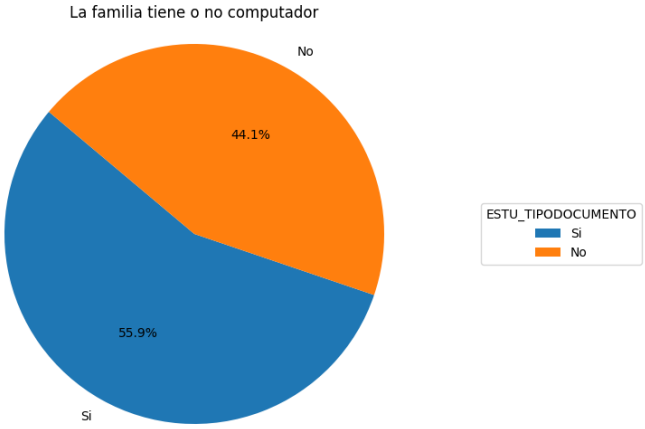
Ahora, empezando con un análisis demográfico de la prueba empezaremos analizando la edad y los lugares geográficos de dónde provienen los estudiantes que presentaron la prueba:



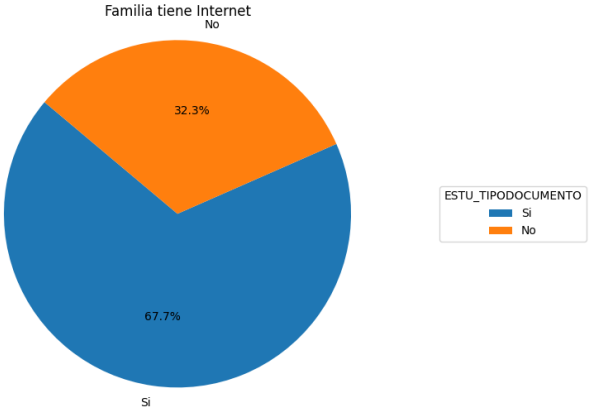
Lo primero que resalta es el hecho de la mayoría de los estudiantes que presentaron la prueba fueron estudiantes menores de edad, esto se concluye porque el 83% tienen tarjeta de identidad. Ahora, otra estadística notable es el hecho de que la mayoría de los estudiantes que presentaron la prueba son de ciudad, a comparación de la población rural que presentó la prueba.

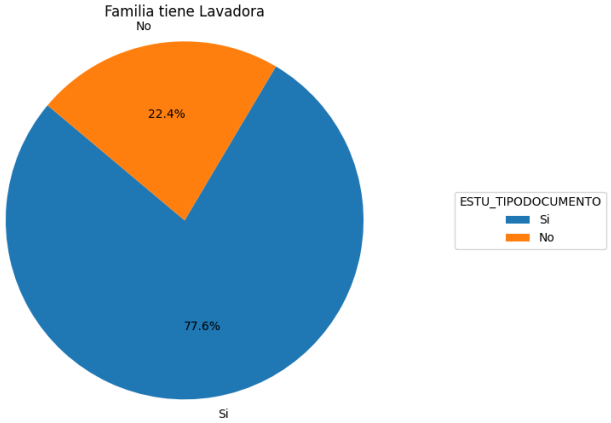
En el siguiente apartado vamos a ver un poco los factores respecto a lo que tiene el estudiante en términos monetario/materiales:



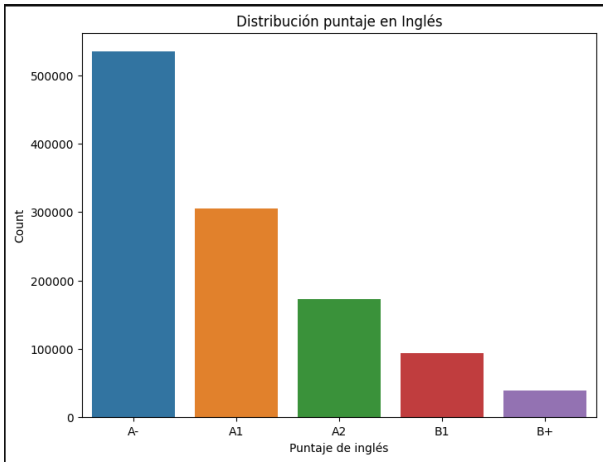


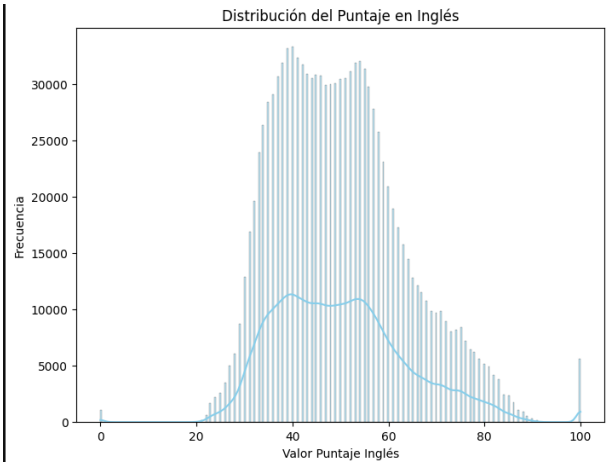
En el primer gráfico podemos ver que la proporción entre estudiantes de colegio Bilingüe a comparación de los que no son de colegio bilingüe es bastante notoria, es decir, que muy poca de la población cuenta con una educación bilingüe. Ahora, también se puede ver que casi la mitad de los estudiantes no tienen un computador en su casa. Esto muestra cierta brecha en desigualdad de oportunidades de acceso para fuentes de consulta y aprendizaje para la prueba. Sin embargo, esto puede ser una conclusión precipitada.



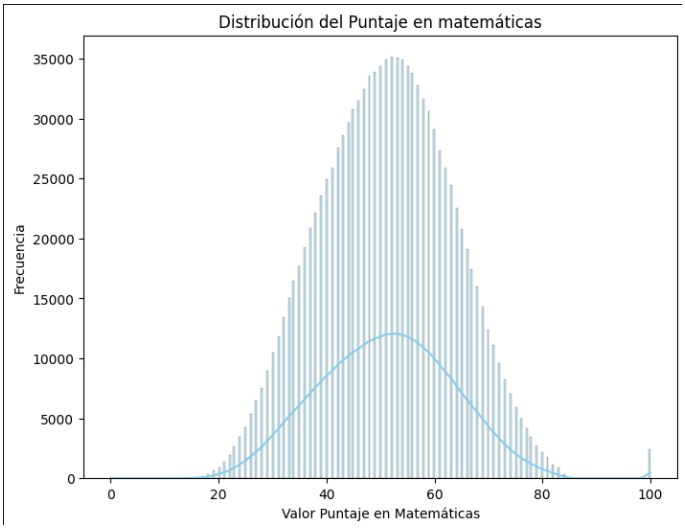


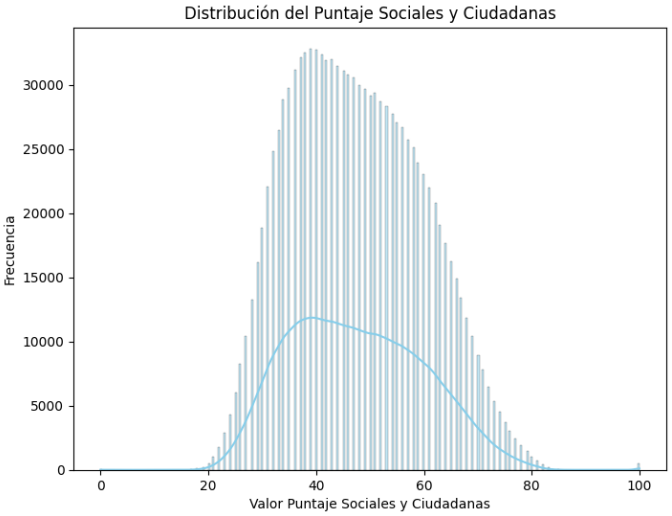
Lo que podemos ver es que si bien la estadística de la proporción de estudiantes que posee un computador en su casa no es la misma que la de los estudiantes que cuentan con internet. Es decir, que los estudiantes pueden contar con el acceso a internet, pero tal vez no con las herramientas para esto, o las herramientas correctas, es decir, que cuentan con un celular más no con un computador el cuál puede proporcionar mejores características para el estudio.



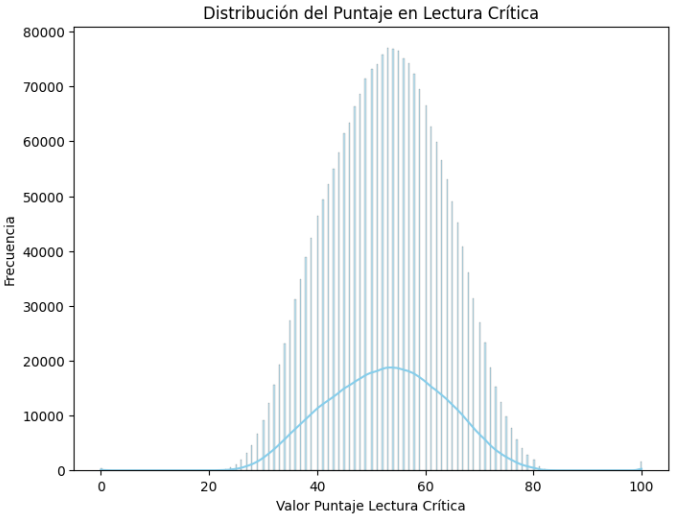


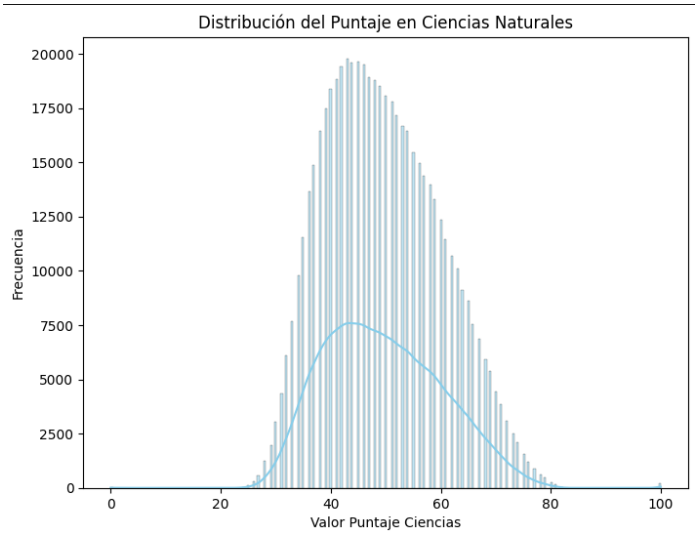
Ahora, las gráficas de arriba ilustran cómo fue el resultado de la prueba de inglés en general y la distribución de este puntaje de manera numérica. Lo que podemos observar es el hecho de que la gráfica de proporción de colegios bilingües muestra porque la mayoría de los puntajes en este idioma son tan bajos. Es decir, la poca cantidad de colegios bilingües se ve traducida en los bajos resultados en la segunda lengua que evalúa esta prueba.



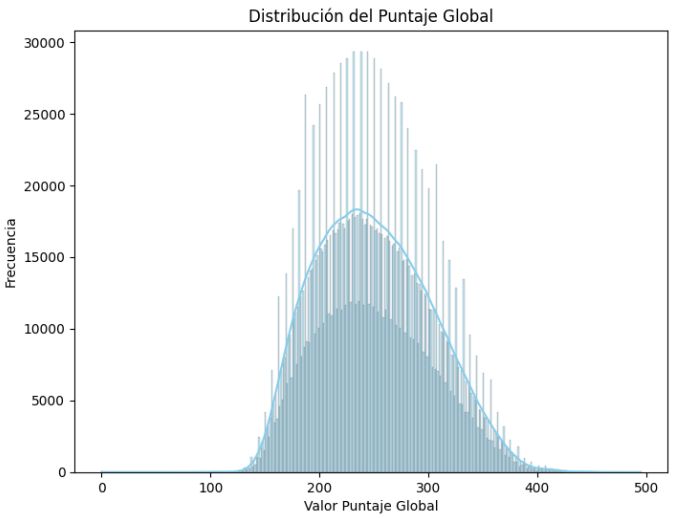


Lo que se puede observar a primera vista es el hecho de que en general la distribución de los puntajes tiende a ser con forma de distribución normal, por lo cual se puede calcular un promedio y una desviación estándar que dé alguna idea de cómo están conformados estos datos numéricos. El promedio del puntaje de la prueba en matemática es de 52.12 con una desviación estándar de 12.41, este resultado puede ser interpretado como bajo porque esto significa que en promedio los estudiantes solo pudieron resolver bien la mitad de la prueba de matemáticas, sin embargo, se ve que esta desviación estándar muestra cierta dispersión de los datos. Por el otro lado, el puntaje de Ciencias sociales y Ciudadanas tienen como media 47.55 con desviación estándar de 12.37. Nuevamente, un resultado bastante bajo.



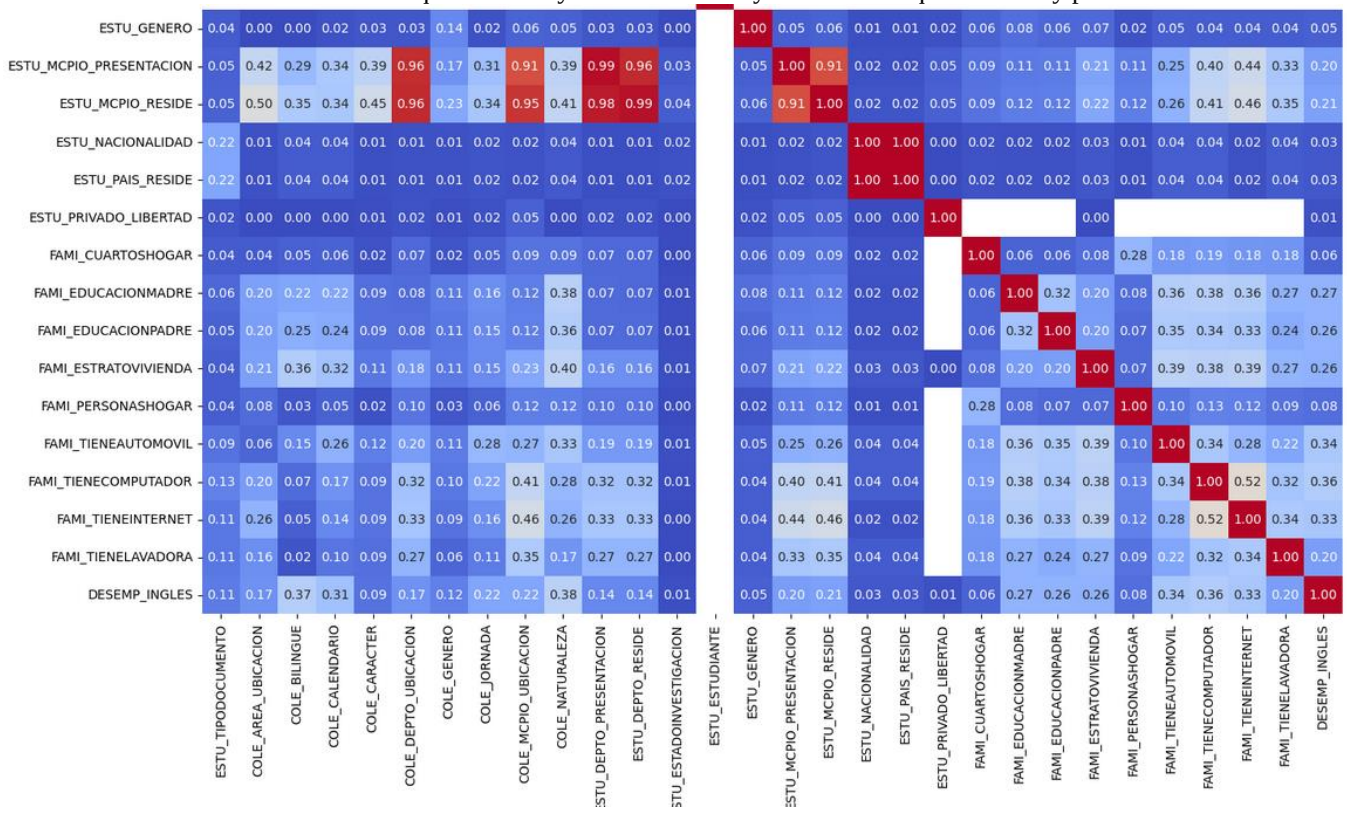


El promedio en lectura crítica es de 52.95 con una desviación estándar de 10.86 y el valor de la media de la prueba de Ciencias es de 49.17 con una desviación estándar de 10.89.

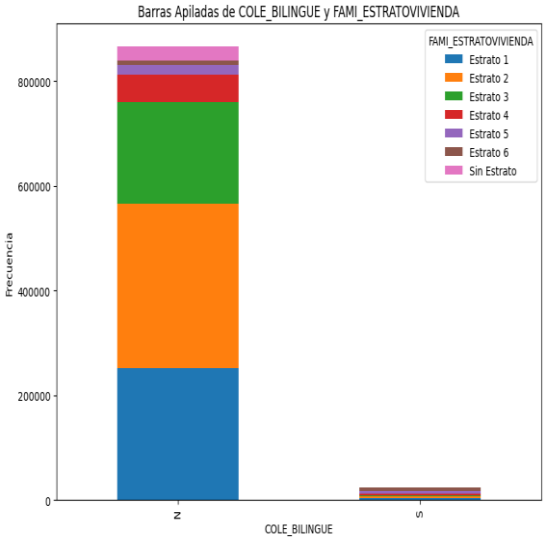


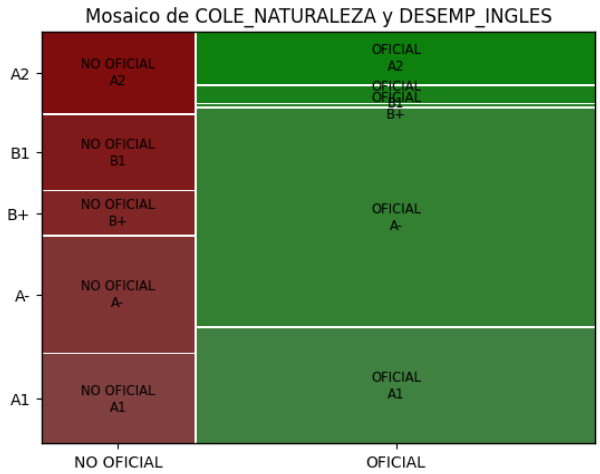
Ahora, el promedio del puntaje global es de 250.96 con una desviación estándar de 53.42, es decir, se tiene una mayor dispersión de los datos, sin embargo, este promedio es preocupante debido a que universidades como la universidad de los Andes, la Pontificia Universidad Javeriana y la Universidad Nacional, en promedio, para ingresar a los programas de pregrado, exigen un puntaje de 300 puntos por lo menos. Es decir, buena parte de la población no podría acceder a esta educación debido a que no cuenta con el puntaje mínimo de ingreso.

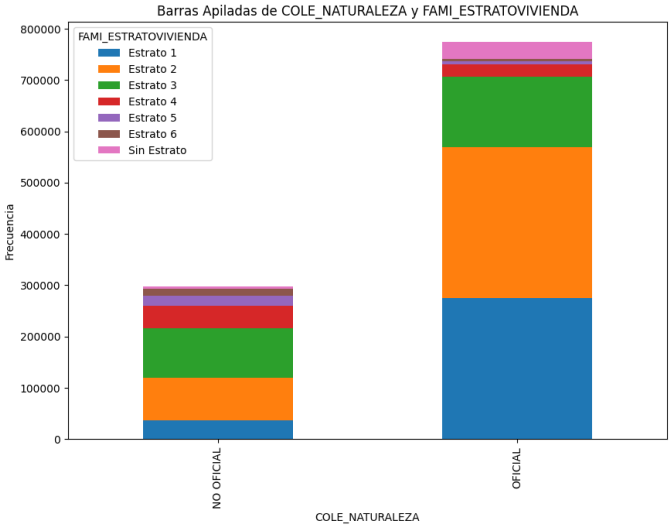
También se realizó un análisis de la relación entre variables categóricas. Se inicio viendo la relación mediante la fórmula de V de Cramér (Cramér's V) la cual es una medida de asociación entre dos variables categóricas basada en el valor del estadístico chi-cuadrado. El valor de 1 vendría a indicar que están muy correlacionados y el valor de 0 que están muy poco relacionados.

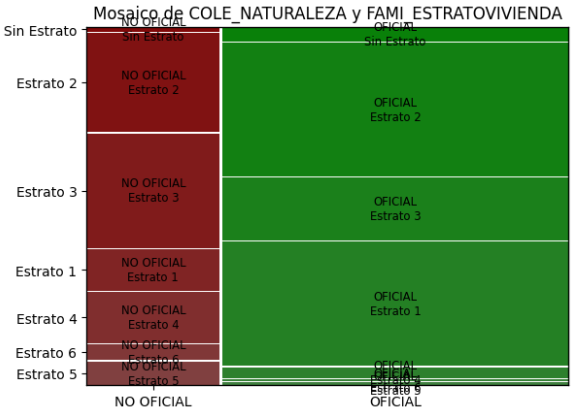


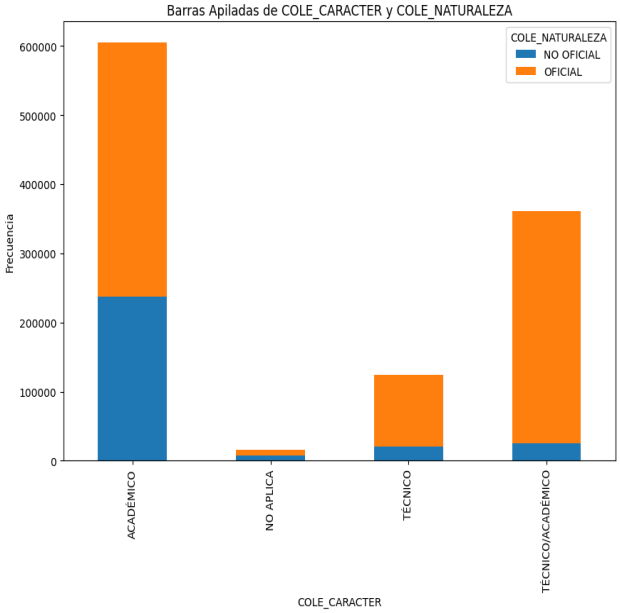
Luego de identificar las mayores correlaciones se graficaron pares de columnas que creemos contienen una relación interesante:

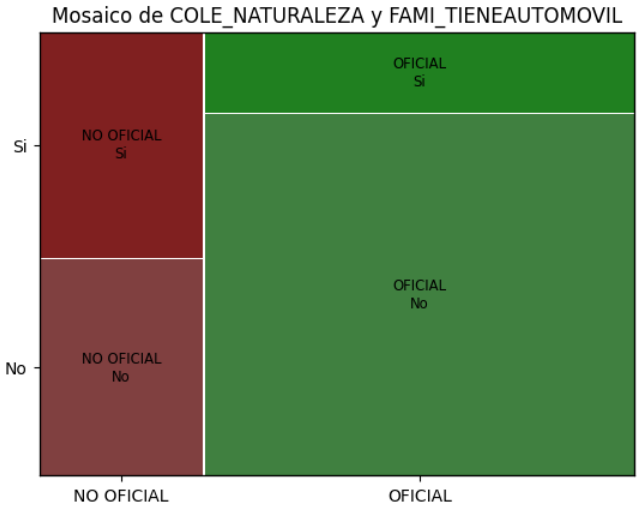


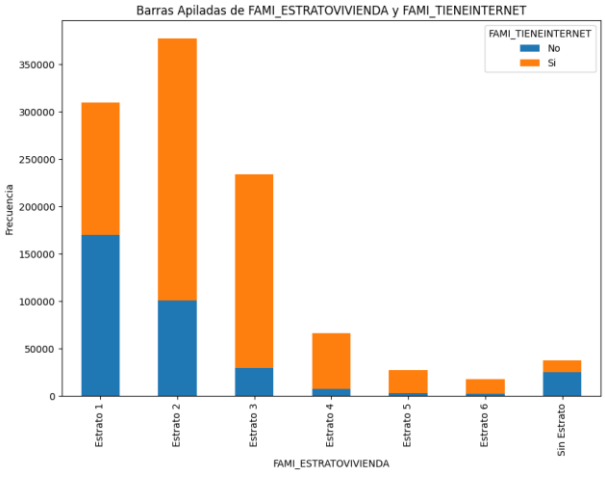


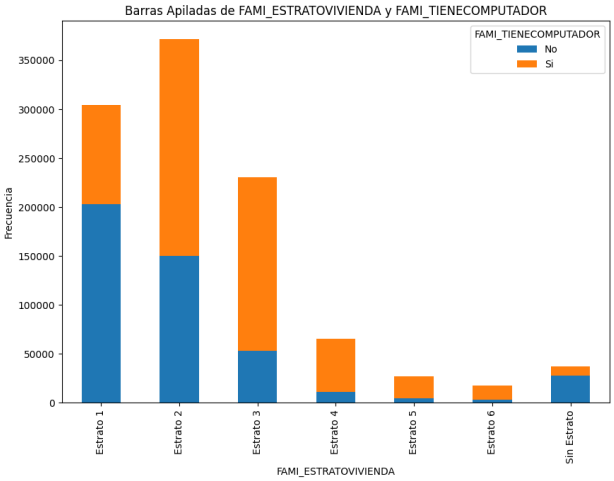












Estás gráficas no solo nos ayudan a entender relaciones entre los estudiantes que realizaron la prueba Icfes, sino que también son una brindan información relevante sobre la educación en Colombia.

V. LIMPIEZA DE DATOS

Para la limpieza de los datos solo se tuvo enfoque en las columnas que iban a ser utilizadas para el modelado, es decir, para la columna de si el colegio es o no bilingüe, el estrato de la vivienda y la edad (una variable que fue creada).

La columnas eliminadas fueron las siguientes:

```python
from pyspark.sql.functions import col


columns_to_drop = [
    'ESTU_CONSECUTIVO',
    'COLE_CARACTER',
    'COLE_COD_DANE_ESTABLECIMIENTO',
    'COLE_COD_DANE_SEDE',
    'COLE_CODIGO_ICFES',
    'ESTU_COD_RESIDE_DEPTO',
    'ESTU_DEPTO_RESIDE',
    'ESTU_ESTADOINVESTIGACION',
    'ESTU_PRIVADO_LIBERTAD',
    'ESTU_PAIS_RESIDE',
    'ESTU_NACIONALIDAD'
]

df = df.drop(*columns_to_drop)

```

Para la creación de la variable se utilizó el año del periodo en el que se presentó la prueba y se resto junto con el año de nacimiento del estudiante:

```python
from pyspark.sql.functions import substring, col, mean

# Extract the year from the 'PERIODO' column
df = df.withColumn('PERIODO', df['PERIODO'].substr(1, 4).cast('int'))

# Calculate age by subtracting birth year from the year in 'PERIODO' column
df = df.withColumn('EDAD', df['PERIODO'] - df['AÑO_NACIMIENTO'])

# Calculate the average age
average_age = df.select(mean('EDAD')).collect()[0][0]

# Print the average age
print("Edad promedio:", average_age)

# Filter the DataFrame for rows where 'ESTU_TIPODOCUMENTO' is equal to 'TI'
filtered_df = df.filter(col('ESTU_TIPODOCUMENTO') == 'TI')

# Calculate the average age from the filtered DataFrame
average_age_ti = filtered_df.select(mean('EDAD')).collect()[0][0]

# Print the average age of people with 'ESTU_TIPODOCUMENTO' equal to 'TI'
print("Edad promedio cuando ESTU_TIPODOCUMENTO = TI:", average_age_ti)

```

Posteriormente, para la limpieza de la edad lo que se hizo fue rellenar los valores nulos con el promedio. Sin embargo, este promedio fue obtenido a través de la variable de tipo de documento, debido a que se obtuvo el promedio de edad de las personas que tenían Tarjeta de Identidad y el promedio de edad de las personas que tenía cédula. Ya con esto se logró el objetivo de reducir los nulos:

```
▸ (2) Spark Jobs

Null count in EDAD column: 0

1

Command took 51.89 seconds -- by daniela.torresg@javeriana.edu.c
```

Para la limpieza de si el colegio es bilingüe o no lo que se hizo fue usar las jornadas que tenían los colegios para poder rellenar el valor nulo con N debido a que es muy raro ver colegios bilingües con jonadas en la noche, única, en la tarde o sabatina, así que todos los valores nulos en Colegio Bilingüe que tuvieran alguna jornada de estas, eran rellenados con el valor de no bilingüe:

```
▸ cole_counts: pyspark.sql.dataframe.DataFrame = [COLE_BI

+-------------+------------+------+
|COLE_BILINGUE|COLE_JORNADA| count|
+-------------+------------+------+
|            N|    COMPLETA|270565|
|            N|      MAÑANA|716979|
|            N|       NOCHE| 84221|
|            N|    SABATINA| 80056|
|            N|       TARDE|168383|
|            N|       UNICA|330516|
|            S|    COMPLETA| 24717|
|            S|      MAÑANA| 10166|
|            S|       NOCHE|   419|
|            S|    SABATINA|   632|
|            S|       TARDE|   510|
|            S|       UNICA|  3414|
+-------------+------------+------+

Command took 45.94 seconds -- by daniela.torresg@javeriana.edu.c
```

Otra limpieza que se hizo fue tomar a todo colegio que fuera de carácter Oficial en dónde el valor de colegio bilingüe fuera nulo, era declarado no bilingüe debido a que la proporción de colegios públicos bilingües es muy baja:

```
1  from pyspark.sql.functions import when
2
3  # Define the condition
4  condition = (df['COLE_NATURALEZA'] == 'OFICIAL') & (df['COLE_BILINGUE'].isNull())
5
6  # Update the 'COLE_BILINGUE' column based on the condition
7  df = df.withColumn('COLE_BILINGUE', when(condition, 'N').otherwise(df['COLE_BILINGUE']))
8

▸ df: pyspark.sql.dataframe.DataFrame = [PERIODO: integer, ESTU_TIPODOCUMENTO: string ... 38 more fields]

Command took 0.19 seconds -- by daniela.torresg@javeriana.edu.co at 5/14/2024, 10:50:11 PM on My Cluster
```

Al finalizar el ejercicio de limpieza dentro de esta columna se logró reducir el número de nulos a 1800 cuando inicialmente se contaban con 198 mil valores faltantes:

```
column_name = "COLE_BILINGUE"

null_count = df.filter(col(column_name).isNull()).count()

print(f"Nulos en  {column_name} son: {null_count}")
```

```
[Stage 507:>                                                (0 + 12) / 12]
Nulos en  COLE_BILINGUE son: 1870
```

Otra limpieza importante que se hizo fue la de la columna del estrato, para esta se utilizó una técnica similar para hacerlo, pero se necesito tomar ciertos parámetros específicos para cada tipo de nulo:

```
from pyspark.sql.functions import expr

# Creamos un diccionario para mapear el estrato con un valor numérico
mapping = {'Estrato 1': 1, 'Estrato 2': 2, 'Estrato 3': 3, 'Estrato 4': 4, 'Estrato 5': 5, 'Estrato 6': 6}

# Creo el mapa a usuar
when_exprs = [when(df['FAMI_ESTRATOVIVIENDA'] == k, v) for k, v in mapping.items()]

expr_expr = expr("CASE " + " ".join([f"WHEN FAMI_ESTRATOVIVIENDA = '{k}' THEN {v}" for k, v in mapping.items()]) + " END")

# Aplicamos la limpieza
df = df.withColumn("FAMI_ESTRATOVIVIENDA", expr_expr)

# Print the updated DataFrame
df.select('FAMI_ESTRATOVIVIENDA').show()
```

Lo primero que se hizo fue convertir esta variable a un tipo numérico debido a que se va a utilizar el promedio de estos valores para la posterior limpieza.

```
from pyspark.sql.functions import col

# Filtro dónde 'COLE_NATURALEZA' es 'NO OFICIAL'
filtered_df = df.filter(col('COLE_NATURALEZA') == 'NO OFICIAL')

# Contar las ocurrencias en dónde 'FAMI_ESTRATOVIVIENDA' in the filtered DataFrame
estrato_counts = filtered_df.groupBy('FAMI_ESTRATOVIVIENDA').count()

# Print el conteo
estrato_counts.show()

# Calculo la media del dataframe filtrado 'FAMI_ESTRATOVIVIENDA' in the filtered DataFrame
mean_estrato = filtered_df.selectExpr('avg(FAMI_ESTRATOVIVIENDA)').collect()[0][0]
print(mean_estrato)
```

```
+--------------------+------+
|FAMI_ESTRATOVIVIENDA| count|
+--------------------+------+
|                NULL| 25083|
|                   1| 36505|
|                   6| 15870|
|                   3|100416|
|                   5| 22679|
|                   4| 45652|
|                   2| 86185|
+--------------------+------+

[Stage 330:=================================================>     (11 + 1) / 12]
2.9330474086174414
```

En este apartado podemos ver que el promedio de lo estudiantes de colegios NO OFICIALES, es decir, colegios privados, es de estrato 3, en este sentido, lo que se hace posteriormente es que cada estudiante que tenga el estrato nulo y que sea de colegio privado, se le asigna el valor de estrato 3 por lo que la media en este tipo de estudiantes es este valor.

```
from pyspark.sql.functions import col

# Filtrar las filas del DataFrame donde 'COLE_NATURALEZA' es 'OFICIAL'
filtered_df = df.filter(col('COLE_NATURALEZA') == 'OFICIAL')

# Contar las ocurrencias de cada valor en 'FAMI_ESTRATOVIVIENDA' en el DataFrame filtrado
estrato_counts = filtered_df.groupBy('FAMI_ESTRATOVIVIENDA').count()

# Imprimir los conteos
estrato_counts.show()

# Calcular la media de 'FAMI_ESTRATOVIVIENDA' en el DataFrame filtrado
mean_estrato = filtered_df.selectExpr('avg(FAMI_ESTRATOVIVIENDA)').collect()[0][0]
print(mean_estrato)
```

```
+--------------------+------+
|FAMI_ESTRATOVIVIENDA| count|
+--------------------+------+
|                NULL| 88541|
|                   1|274696|
|                   6|  3973|
|                   3|137457|
|                   5|  7031|
|                   4| 23712|
|                   2|294552|
+--------------------+------+

[Stage 336:=================================================>     (11 + 1) / 12]
1.9287449370870262
```

En este caso vemos que el promedio del estrato de los estudiantes que pertenecen a un Institución Educativa Oficial (colegio público), son de estrato 2. Por lo cual, cualquier estudiante que pertenezca a una de estas instituciones y tenga el estrato en valor nulo, se le imputa el valor de estrato 2.

Dentro del cuaderno se puede ver que inicialmente se contabana 75 mil nulos dentro de esta columna, después de la limpieza se redujeron a 1200.

VI. MODELADO

Se planteó realizar una clasificación binaria a partir de 6 variables principales.

- Municipio
- Estrato de Vivienda
- Área del colegio
- Colegio Bilingüe
- Genero

- Edad

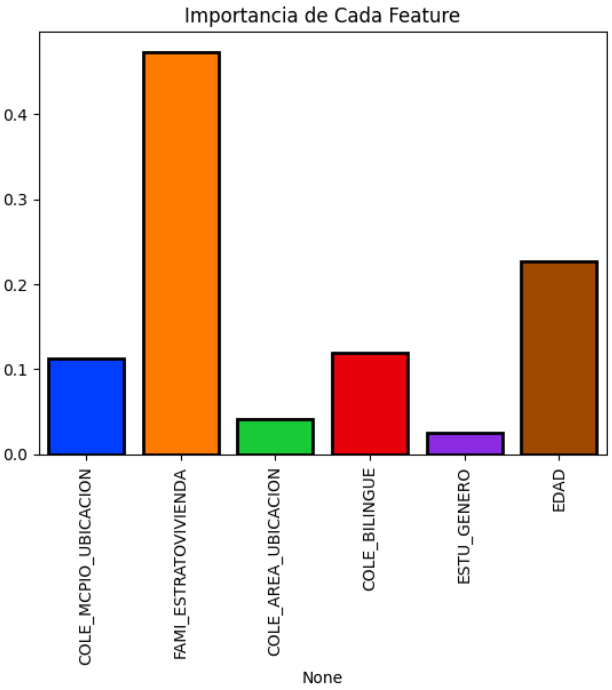
El objetivo del modelamiento es predecir si dadas unas condiciones un estudiante obtiene un puntaje mayor o menor a 300 en el resultado final de las pruebas. Esto es muy útil debido a que muchas universidades tienen en cuenta este puntaje como treshold para aceptar o no a un estudiante.

Realizamos 3 modelos para resolver este problema:

- **Decision Tree**
- **Random Forest**
- **Resgresión Lógistica**

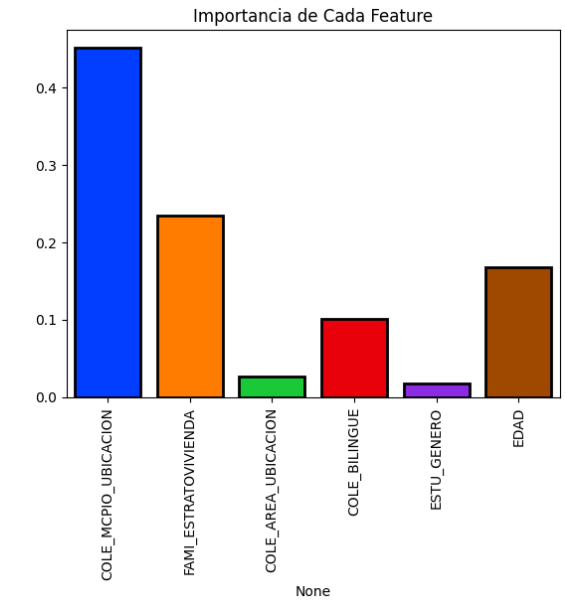
Decision Tree:

```
El accuracy en train es:  0.8207968315070613
El accuracy en test es:  0.8222151501412479
```



Random Forest:

```
accuracy en train 0.8315942515481715
accuracy en test 0.824691347260489
```

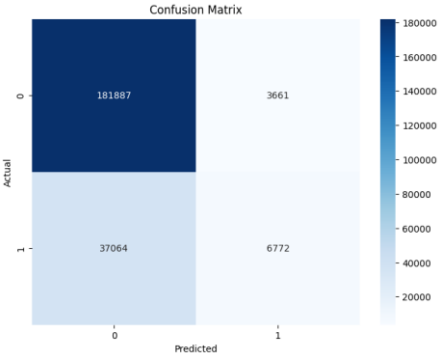


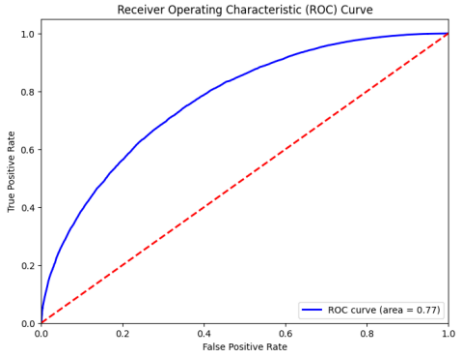
Resgresión Lógistica:

```
Best parameters: {'classifier__C': 0.01}
Accuracy: 0.8225
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.98      0.90    185548
           1       0.65      0.15      0.25     43836

    accuracy                           0.82    229384
   macro avg       0.74      0.57      0.57    229384
weighted avg       0.80      0.82      0.78    229384
```





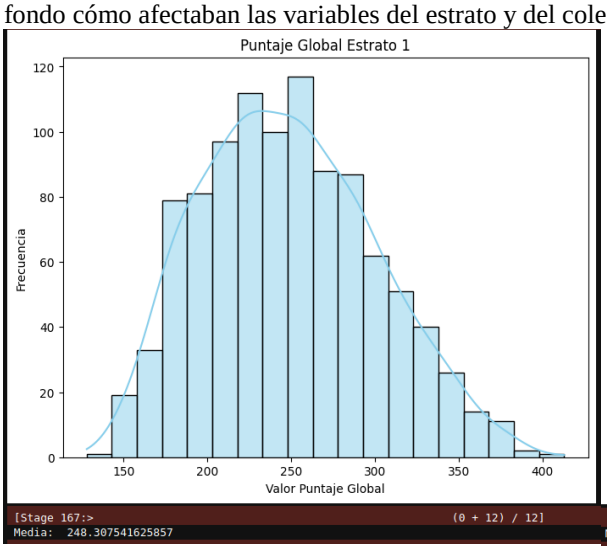
Conclusiones de modelado:

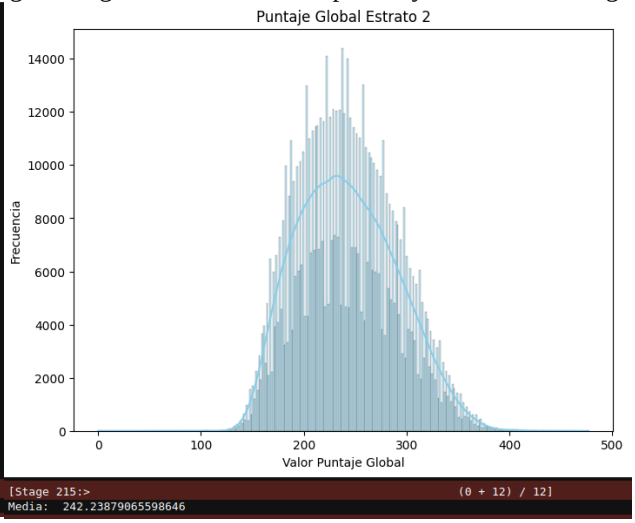
	Decision Tree	Random Forest	Regresión Logística
Accuracy	**0.8222**	**0.8246**	**0.8225**

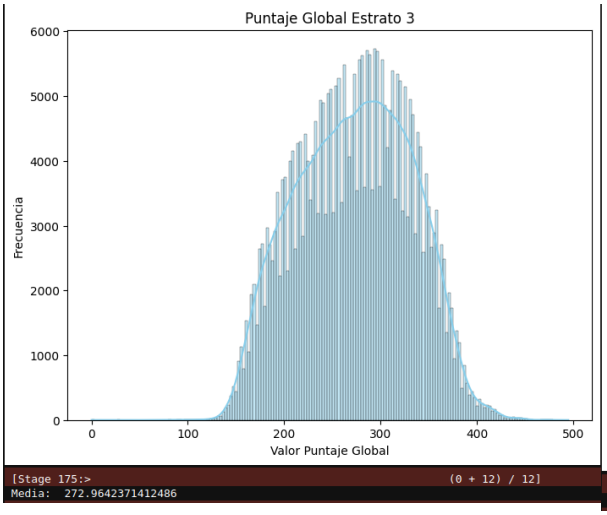
Concluimos que los modelos tienen un accuracy similar, sin embargo, el ganador fue el Random Forest. Dependiendo del cliente en específico podemos revisar metricas más detalladas como recall, precisión y el área bajo la curva.

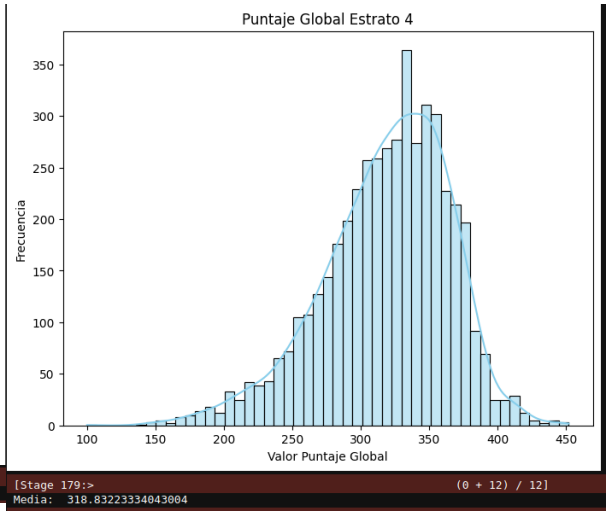
Aunque el Random Forest superó a los otros modelos en precisión, su complejidad podría dificultar su implementación en entornos donde se requiere una comprensión más clara del proceso de toma de decisiones.

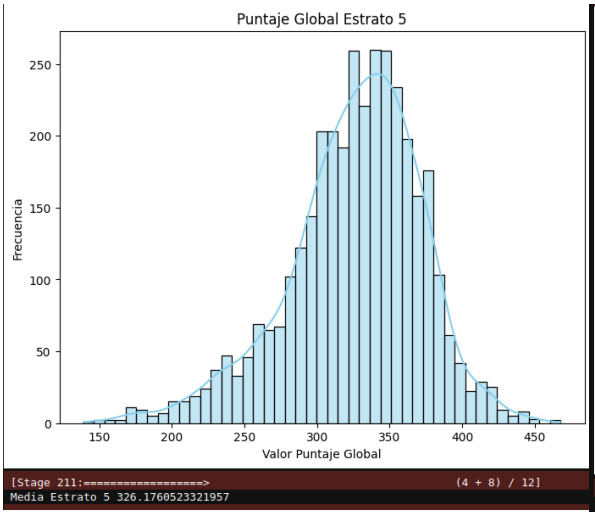
Ahora, después de ver los importances de tanto el modelo de Decision Tree y Randome Forest se quiso explorar un poco más a fondo cómo afectaban las variables del estrato y del colegio bilingüe el resultado de la prueba y se encontró lo siguiente:

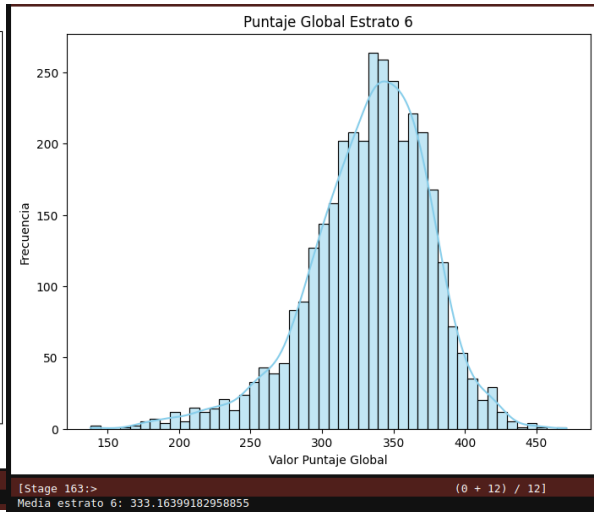






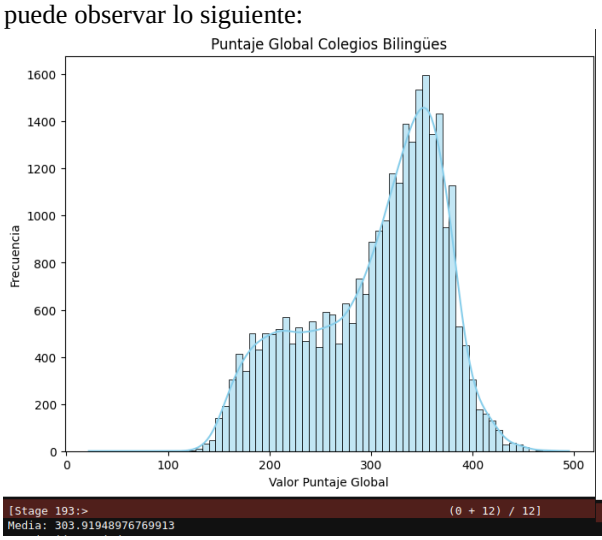


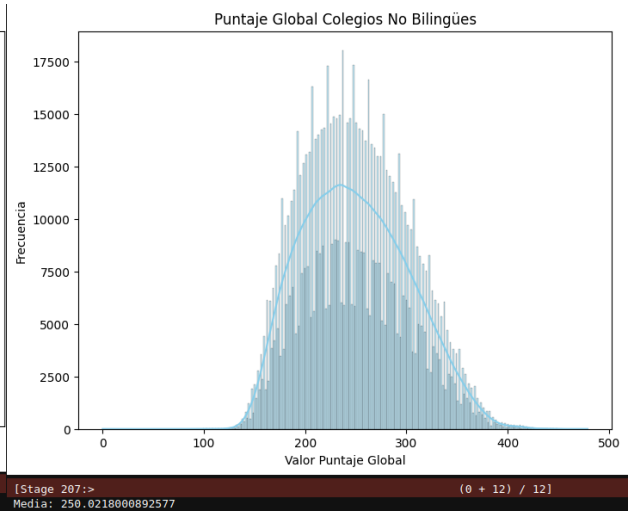




Cómo se puede observar, es evidente que conforme se incrementa el estrato, la media del puntaje aumenta, no solo esto, sino que la distribución completa muestra que es una media que es representativa debido a que hay un mayor cúmulo o concentración de estudiantes de los estratos altos en puntajes de entre 300 y 400. Ahora, la brecha en términos del promedio es abismal: 85 puntos, esto es la diferencia del puntaje entre una persona de estrato 1 a comparación de una personas de estrato 6. Esto signfica que las probabilidades de que una persona de estrato 1 pueda entrar a una universidad privada son muchísimo más bajas que los de uan personas de estrato 6, no solo por el costo de la matrícula, sino porque no cuenta con el puntaje mínimo que le exigiría la institución: 300.

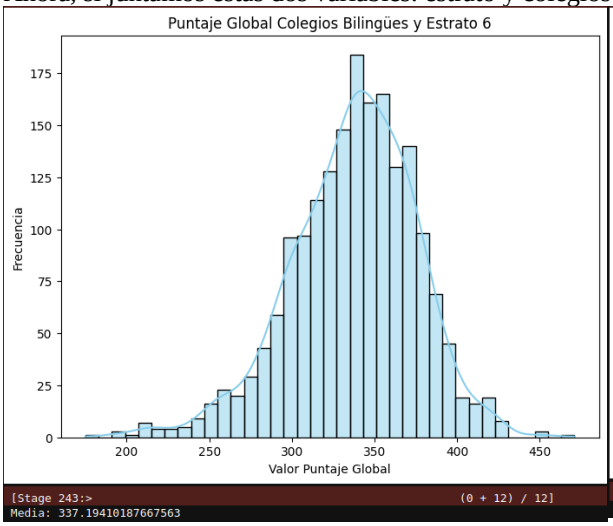
Ahora, si revisamos la segunda variable que se tomó en cuenta en los importances, si el colegio del estudiante era bilingüe o no se puede observar lo siguiente:

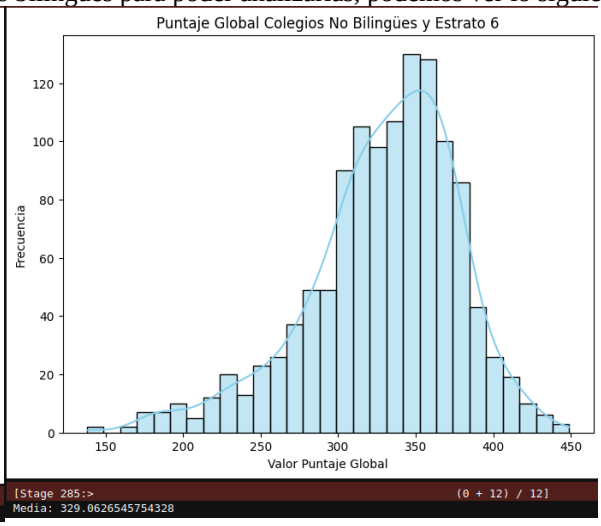


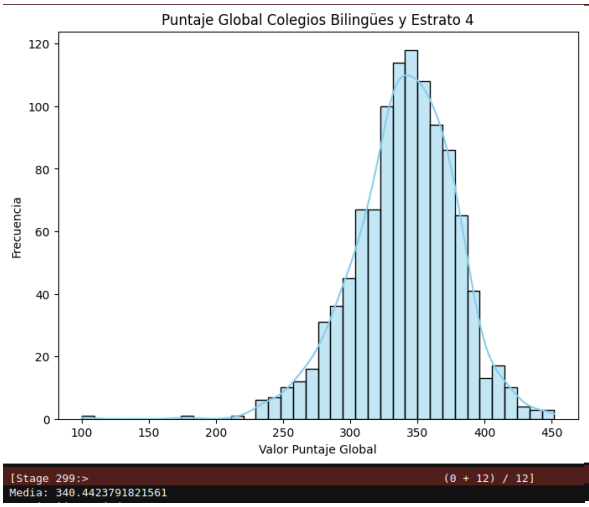


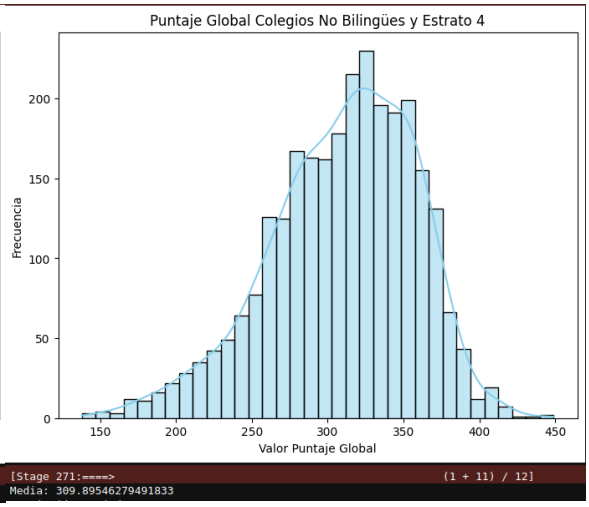
En promedio, los colegios bilingües tienen un mejor desempeño que los no bilingües, esta diferencia es de 50 puntos. Esto significa que un estudiante va a tener una mayor probabilidad de tener un puntaje suficiente para una universidad si está en un colegio bilingüe.

Ahora, si juntamos estas dos variables: estrato y colegios bilingües para poder analizarlas, podemos ver lo siguiente:



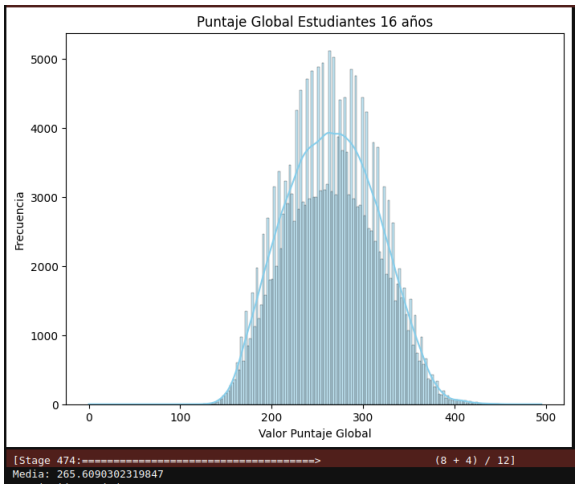


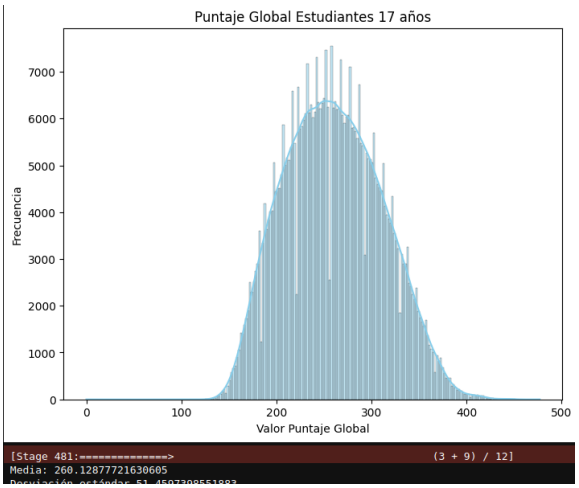


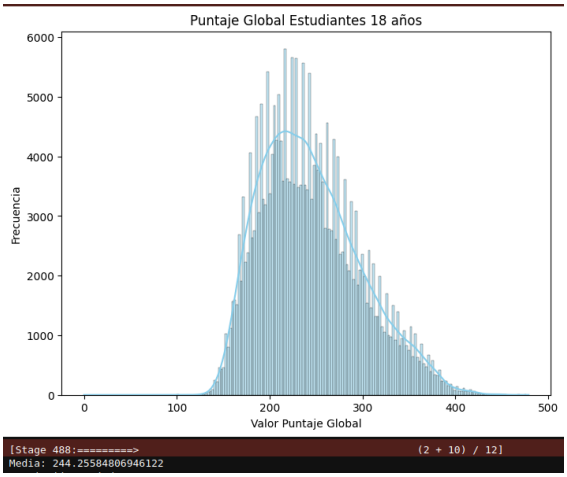


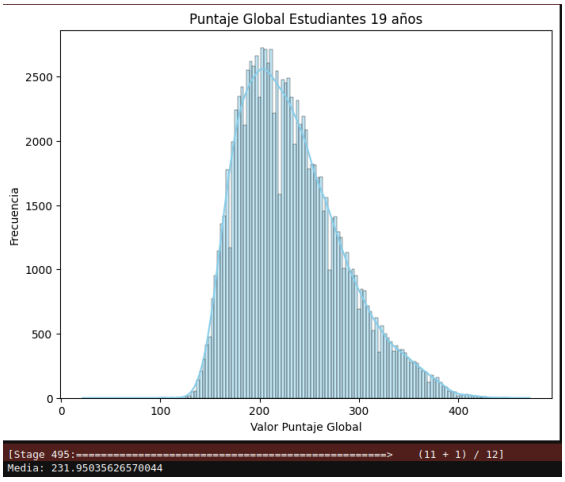
En estos cuatro casos: colegio bilingüe estrato 4, colegio con bilingüe estrato 4, colegio bilingüe estrato 6 y colegio con bilingüe estrato 6 vemos que tanto afecta el hecho de que un estudiante tenga cierto estrato y cómo el estar en un colegio bilingüe incide en su puntaje. En el caso del estrato 6 vemos que la diferencia de puntaje entre un estudiante de este estrato en un colegio bilingüe y un colegio no bilingüe es de tan solo 10 puntos. Mientras que, en el estrato 4 si un estudiante es o no de un colegio bilingüe influye en 30 puntos sobre el resultado de su prueba, no solo eso, sino que un estudiante de estrato 4 en un colegio bilingüe, en promedio, supera a un estudiante de estrato 6 en un mismo tipo de colegio. En otras palabras, es más costo-eficiente para una persona de estrato 4 estar en un colegio bilingüe, que una persona de estrato 6.

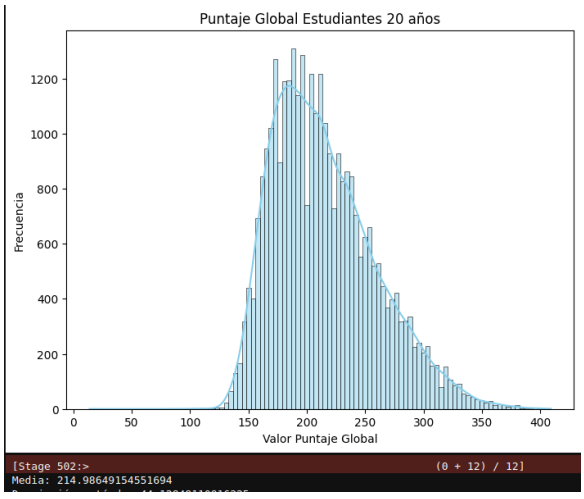
El último análisis que se hizo fue respecto a cómo la edad tiene una correlación con el puntaje global de la prueba. En este sentido se hizo esto con las edades más comunes con las que se presentan este examen y también con las que pueden llegar a ser atípicas. Es decir, edades como 16, 17, 18, 19, 20 y 21 años, y luego con aquellas personas que tenían más de 21 años.

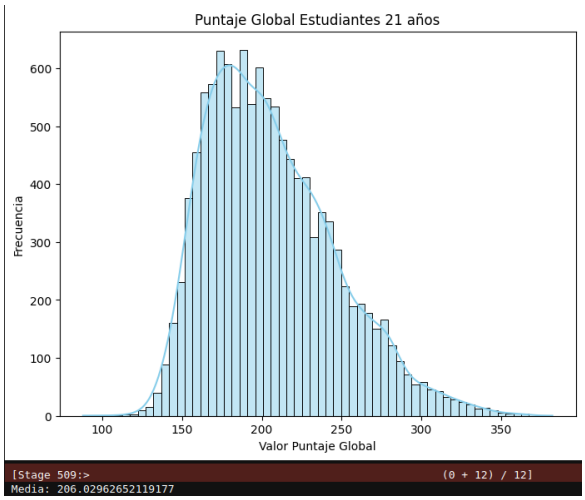












Lo que estas graficas nos muestran es el hecho de que hay una correlación o colinealidad negativa, al menos en el intervalo entre los 16 y los 21 años, en dónde a mayor edad de las personas obtienen menores puntajes. Esto puede ser debido a que muchas de estas personas sean lo que en los colegios se conoce como: personas que perdieron en el año. Es decir, personas que por su bajo rendimiento académico les tocó ver dos veces un mismo año escolar, es decir, que estamos hablando de personas con un bajo rendimiento escolar que se ve reflejado en estas pruebas, sin embargo, esta hipótesis no puede ser comprobada debido a que no hay ninguna pregunta dentro de la encuentra que hace la prueba, la cual nos dé un indicio de que esta persona haya peridido un año escolar.

Sin embargo, lo que sí es claro es que, en este intervalo, a mayor edad, menor es el puntaje de la prueba, sin embargo, cuándo se hace la revisión de las personas mayores de 21 años se obtiene un promedio mayor en esta prueba que las personas de 21 años. Estas pueden ser personas que no acabaron el bachillerato de manera “normal” o habitual, y que muy probablemente comenzaron a trabajar desde una edad temprana y en algún punto de su vida decidieron terminar la educación secundaria.

VII. CONCLUSIONES

Las herramientas de procesamiento distribuido son fundamentales para manejar grandes volúmenes de datos. Al intentar ejecutar varios algoritmos mencionados en este documento, encontramos limitaciones significativas en la capacidad de nuestros ordenadores.

Pyspark facilita la computación distribuida sin necesidad de preocuparse por su complejidad subyacente, convirtiéndose en una herramienta crucial para aprovechar al máximo grandes volúmenes de datos.

Durante el análisis exploratorio de datos, descubrimos patrones interesantes que resaltan la importancia no solo de modelos analíticos complejos, sino también de gráficos estadísticos simples que pueden proporcionar valiosa información.

Las variables de el estrato, si un estudiante pertenece a un colegio bilingüe o no, y la edad, no factores que influyen en el resultado de la prueba. En este sentido, se puede ver una colinealidad positiva entre el estrato y el puntaje global, es decir que, a mayor estrato, se puede obtener un mayor puntaje global. También, hay una relación entre el puntaje global y el hecho de que el estudiante pertenezca a un colegio, es tan fuerte esta influencia en el puntaje que hay casos de donde el hecho de pertenecer a un colegio bilingüe puede potenciar el puntaje de un estudiante más que el estrato. Por último, la edad tiene una colinealidad negativa con el puntaje en el intervalo de los 16 a 21 años. En este sentido, el hecho de presentar este examen a una edad temprana puede tener un efecto positivo en los resultados del estudiante. Sin embargo, esto no signfica que haya estudiantes de edades más avanzadas que no puedan obtener puntajes sobresalientes.

Se llevó a cabo una limpieza de datos enfocada en las variables relevantes para el modelado, seguida por la implementación de un modelo de clasificación binaria para predecir si un estudiante obtendría un puntaje superior o inferior a 300 en las pruebas. Se probaron tres modelos distintos: Árbol de decisión, Bosque Aleatorio y Regresión Logística, concluyendo que el Bosque Aleatorio demostró el mejor rendimiento en términos de precisión. La experimentación con nuevos algoritmos fue enriquecedora, evidenciando que, en este caso, los resultados de los modelos son comparables, lo que sugiere que se debe enfocar en la obtención de más características de los datos.

REFERENCIAS:

- OECD. (2022). PISA 2022 Results. [Informe]. Recuperado de https://www.oecd.org/publication/pisa-2022-results/
- Ministerio de Educación de Colombia. (s. f.). Estudiantes colombianos obtienen buenos resultados en PISA 2022. [Comunicado de prensa]. Recuperado de https://mineducacion.gov.co/1759/w3-printer-357732.html#:~:text=Durante%20la%20sesi%C3%B3n%20de%20la,sociales%20y%20ciudadanas%20e%20Ingl%C3%A9s.
- El País. (2023). Icfes en Colombia 2023: Estos son los puntajes mínimos para ingresar a estas universidades en Colombia. [Artículo periodístico]. Recuperado de https://www.elpais.com.co/educacion/icfes-en-colombia-2023-estos-son-los-puntajes-minimos-para-ingresar-a-estas-universidades-en-colombia-2803.html#:~:text=Es%20esencial%20seleccionar%20las%20universidades,Icfes%20es%20de%20250%20puntos.