

# Parcial II: Prueba Saber 11

Daniela T., Isaac J. y Daniel S.



# Contenido

01

**Intro y Entendimiento  
del negocio**

02

**Colección y descripción  
de datos**

03

**Análisis  
Exploratorio**

04

**Limpieza  
de Datos**

05

**Modelado de Datos**

01

# Introducción y Entendimiento del negocio

Una prueba compuesta de varias



# ○ Entendimiento del negocio

El Instituto Colombia para la Evaluación de la Educación fue fundado en 1969

La prueba saber 11 cuenta con 6 módulos:

- Inglés
- Matemáticas
- Lectura Crítica
- Sociales y Ciudadanas
- Ciencias Naturales

La prueba cuenta con un total de 243 preguntas.



Fuente: <https://bogota.gov.co/mi-ciudad/educacion/el-icfes-colombia-la-clasificacion-de-resultados-de-planteles>

# Entendimiento del negocio



## Análisis preliminar de resultados 2023

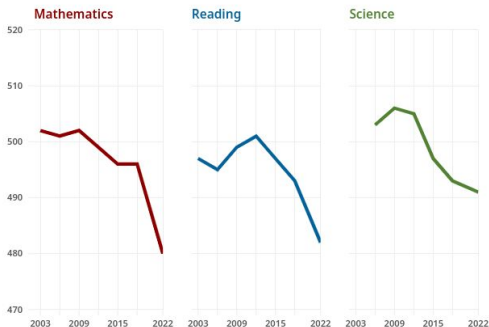
Cantidad de evaluados en el examen Saber 11 A - 2023  
Record en cantidad de estudiantes presentes

La cantidad de estudiantes presentes en 2023 es la más alta observada en cerca de 10 años de aplicación, superando en cerca de 10 mil estudiantes al 2022.



Fuente: Subdirección de análisis y divulgación, ICFES

## PISA test scores, OECD average



Source: OECD (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education.

Casi medio millón de personas evaluadas cada año.

La OCDE reveló una disminución en el desempeño en las pruebas PISA de matemáticas después de la pandemia.

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

***¿Qué patrones se pueden identificar en los resultados de las pruebas Saber 11 a lo largo del tiempo ?***

***¿Existen tendencias claras en el desempeño de los estudiantes en diferentes áreas del conocimiento?***

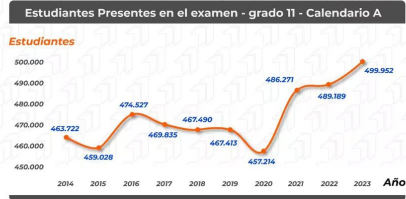
# Entendimiento del negocio



## Análisis preliminar de resultados 2023

Cantidad de evaluados en el examen Saber 11 A - 2023  
Record en cantidad de estudiantes presentes

La cantidad de estudiantes presentes en 2023 es la más alta observada en cerca de 10 años de aplicación, superando en cerca de 10 mil estudiantes al 2022.

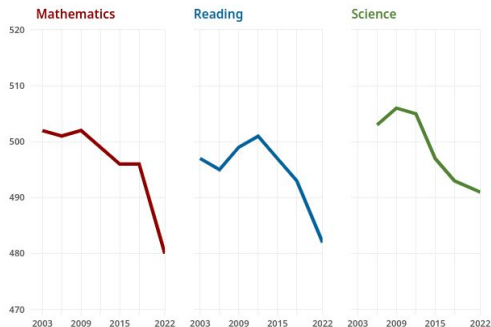


Fuente: Subdirección de análisis y divulgación, ICFES

*¿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?*

*¿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?*

## PISA test scores, OECD average

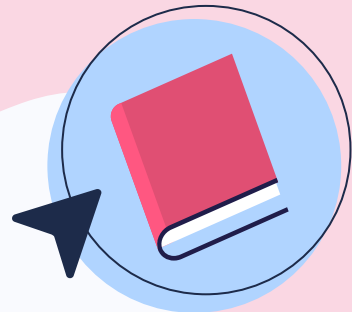


Source: OECD (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education.



02

# Colección de Datos



Muchas columnas



# Tipos de datos

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



| 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_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   |



# Datos Nulos

| Nombre Columna            | Número de Nulos |
|---------------------------|-----------------|
| PERIODO                   | 0               |
| ESTU_TIPODOCUMENTO        | 0               |
| ESTU_CONSECUTIVO          | 0               |
| COLE_AREA_UBICACION       | 1               |
| COLE_BILINGUE             | 198611          |
| COLE_CALEDARIO            | 1               |
| COLE_CHARACTER            | 40747           |
| COLE_COD_DANE_ESTABLECIM  | 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_ESTABLECIMIE  | 1               |
| COLE_NOMBRE_SEDE          | 1               |
| COLE_SEDE_PRINCIPAL       | 1               |
| ESTU_COD_DEPTO_PRESENTACI | 0               |
| ESTU_COD_MCPIO_PRESENTACI | 0               |

|                          |       |
|--------------------------|-------|
| ESTU_COD_RESIDE_DEPTO    | 1174  |
| ESTU_COD_RESIDE_MCPIO    | 1174  |
| ESTU_DEPTO_PRESENTACION  | 0     |
| ESTU_DEPTO_RESIDE        | 1174  |
| ESTU_ESTADAINVESTIGACION | 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_LECTURA_CRITICA PUNT_GLOBAL
1                      44          194
4                      71          381
6                      73          372
9                      77          392
11                     80          389
...                   ...          ...
2258328                36          197
2258329                58          290
2258330                47          199
2258331                49          242
2258332                48          233

```

[1095980 rows x 51 columns]

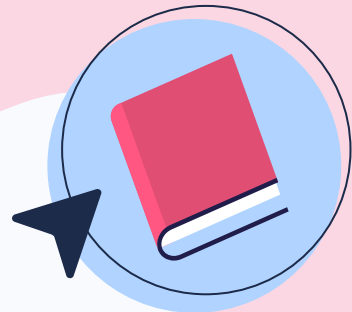
Number of duplicate rows: 1095980



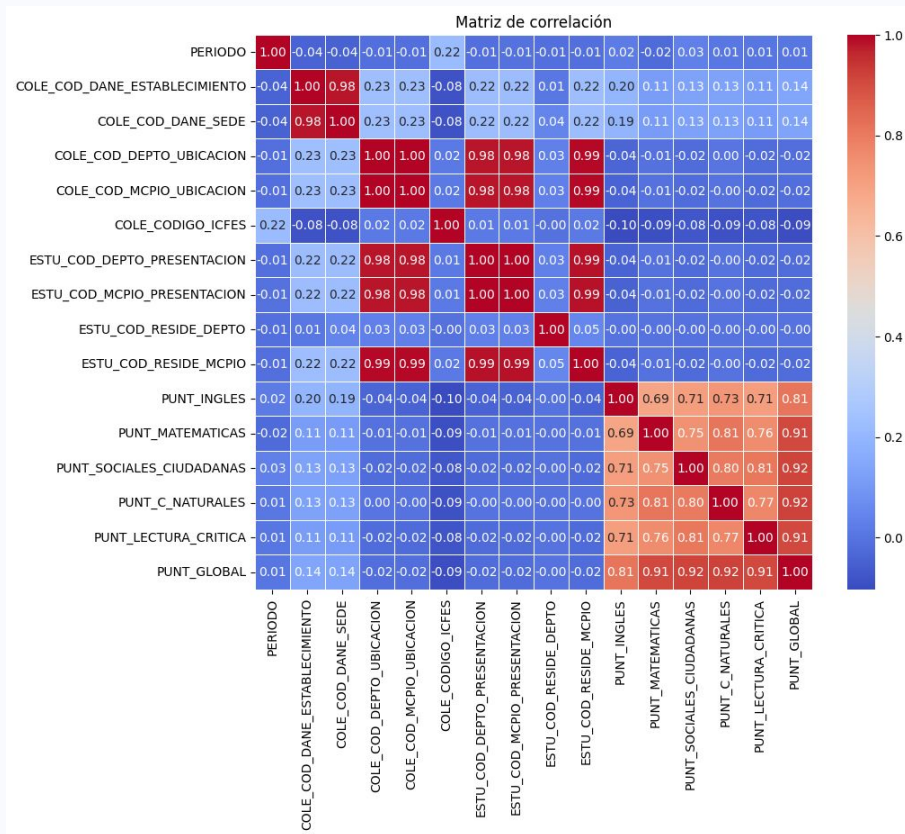
03

# Análisis Exploratorio

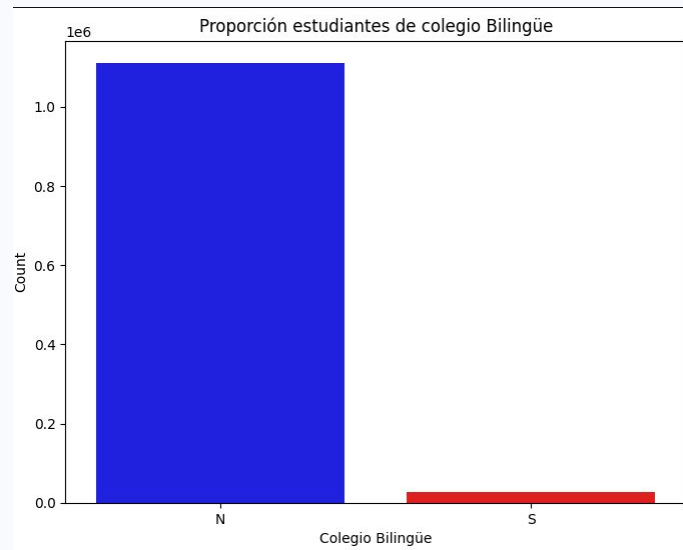
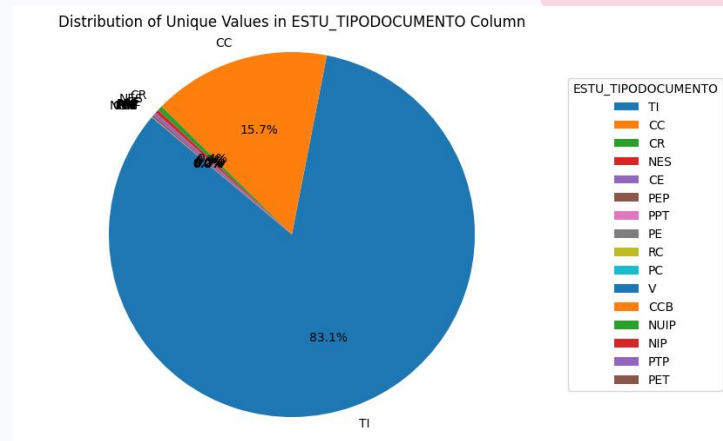
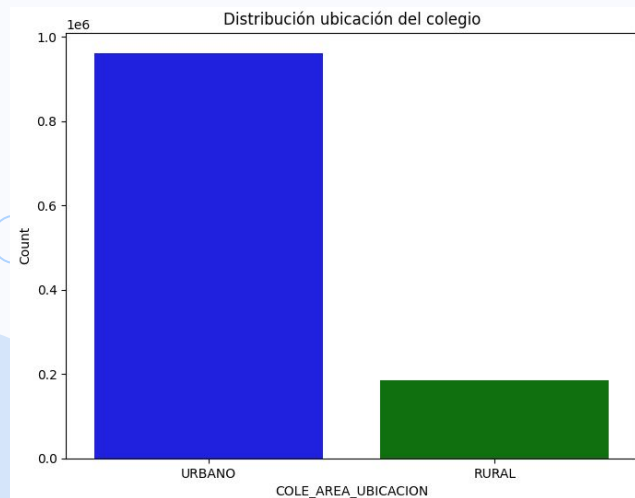
Sorpresas respecto a la pandemia



# Matriz de correlación

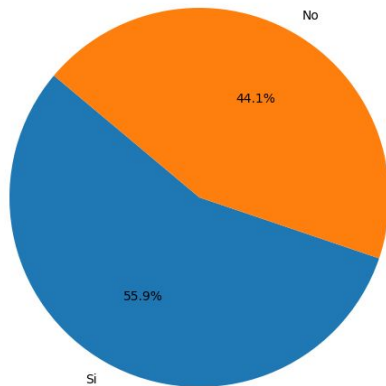


# Estadística Descriptiva

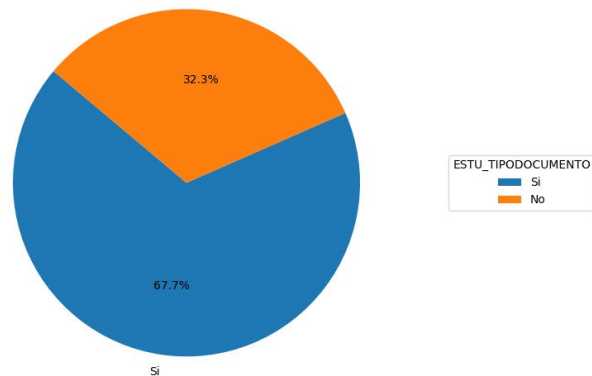


# Estadística Descriptiva

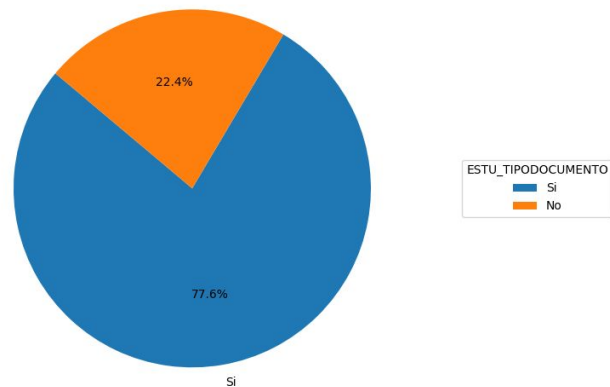
La familia tiene o no computador



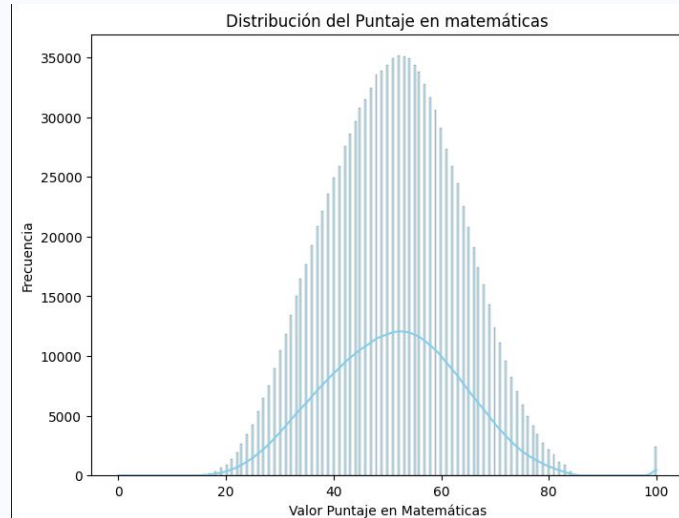
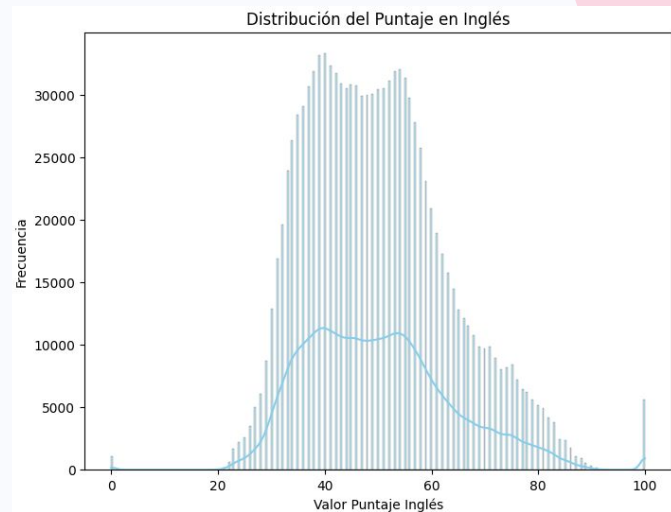
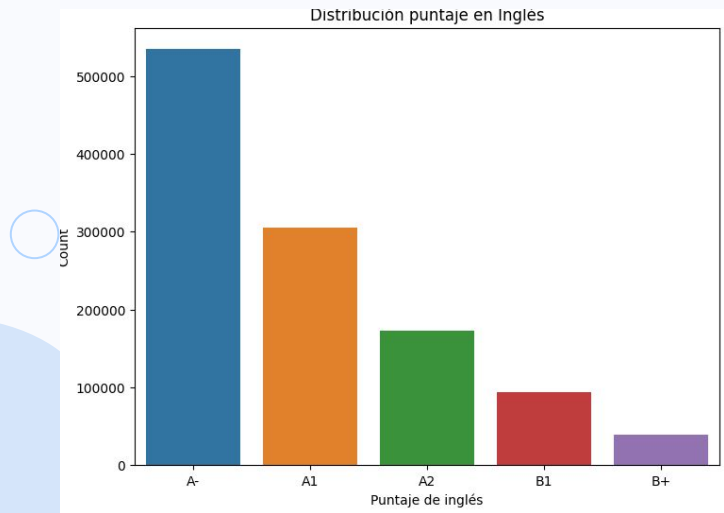
Familia tiene Internet



Familia tiene Lavadora



# Estadística Descriptiva



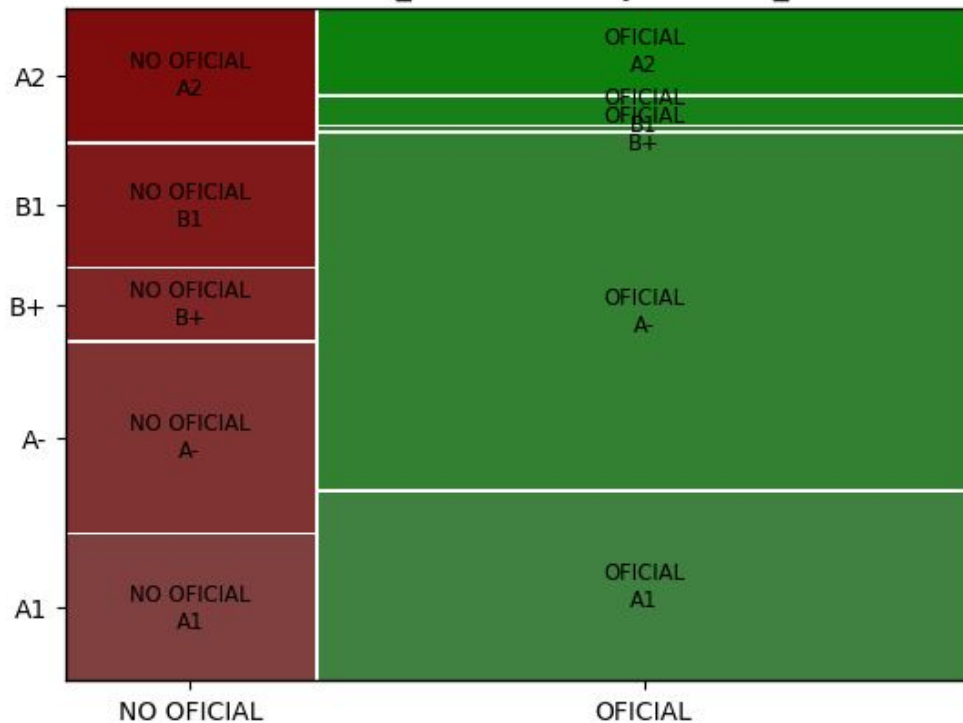


# Coeficiente de Cramer

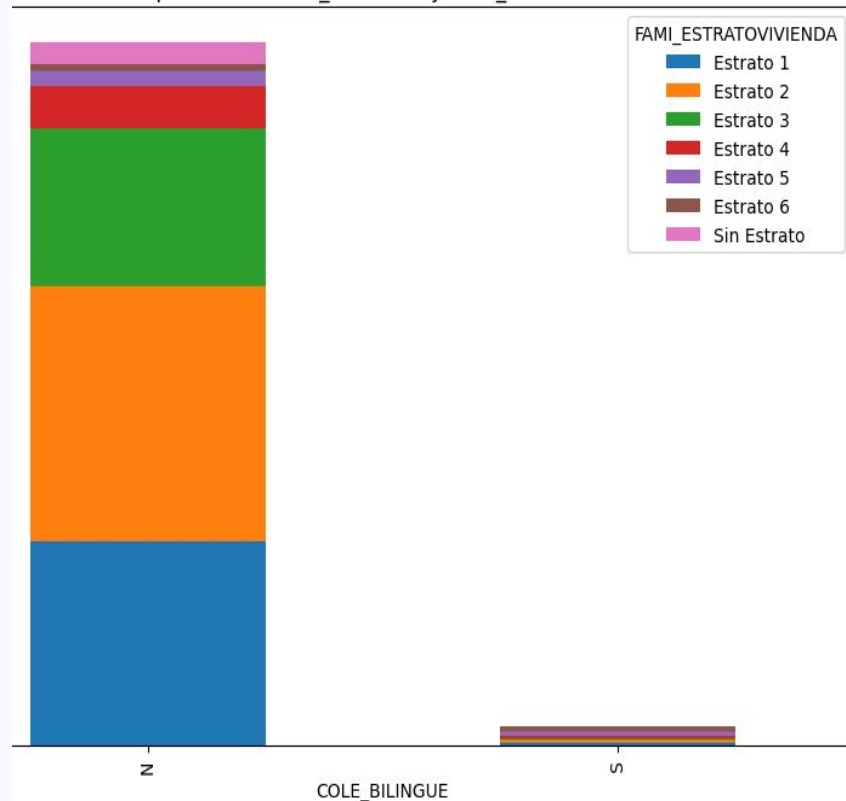
|                         | ESTU_GENERO | ESTU_MCPIO_PRESENTACION | ESTU_MCPIO_RESIDE | ESTU_NACIONALIDAD | ESTU_PAIS_RESIDE | ESTU_PRIVADO_LIBERTAD | FAMI_CUARTOSHOGAR | FAMI_EDUCACIONMADRE | FAMI_EDUCACIONPADRE | FAMI_ESTRATOVIVIENDA | FAMI_PERSONASHOGAR | FAMI_TIENEAUTOMOVIL | FAMI_TIENECOMPUTADOR | FAMI_TIENEINTERNET | FAMI_TIENELAVADORA | DESEMP_INGLES |      |      |      |      |      |      |      |      |      |      |      |      |      |
|-------------------------|-------------|-------------------------|-------------------|-------------------|------------------|-----------------------|-------------------|---------------------|---------------------|----------------------|--------------------|---------------------|----------------------|--------------------|--------------------|---------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| ESTU_GENERO             | 0.04        | 0.00                    | 0.00              | 0.02              | 0.03             | 0.03                  | 0.14              | 0.02                | 0.06                | 0.05                 | 0.03               | 0.03                | 0.00                 | 1.00               | 0.05               | 0.06          | 0.01 | 0.01 | 0.02 | 0.06 | 0.08 | 0.06 | 0.07 | 0.02 | 0.05 | 0.04 | 0.04 | 0.04 | 0.05 |
| ESTU_MCPIO_PRESENTACION | 0.05        | 0.42                    | 0.29              | 0.34              | 0.39             | 0.96                  | 0.17              | 0.31                | 0.91                | 0.39                 | 0.99               | 0.96                | 0.03                 | 0.05               | 1.00               | 0.91          | 0.02 | 0.02 | 0.05 | 0.09 | 0.11 | 0.11 | 0.21 | 0.11 | 0.25 | 0.40 | 0.44 | 0.33 | 0.20 |
| ESTU_MCPIO_RESIDE       | 0.05        | 0.50                    | 0.35              | 0.34              | 0.45             | 0.96                  | 0.23              | 0.34                | 0.95                | 0.41                 | 0.98               | 0.99                | 0.04                 | 0.06               | 0.91               | 1.00          | 0.02 | 0.02 | 0.05 | 0.09 | 0.12 | 0.12 | 0.22 | 0.12 | 0.26 | 0.41 | 0.46 | 0.35 | 0.21 |
| ESTU_NACIONALIDAD       | 0.22        | 0.01                    | 0.04              | 0.04              | 0.01             | 0.01                  | 0.01              | 0.02                | 0.02                | 0.04                 | 0.01               | 0.01                | 0.02                 | 0.01               | 0.02               | 0.02          | 1.00 | 1.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.03 | 0.01 | 0.04 | 0.04 | 0.02 | 0.04 | 0.03 |
| ESTU_PAIS_RESIDE        | 0.22        | 0.01                    | 0.04              | 0.04              | 0.01             | 0.01                  | 0.01              | 0.02                | 0.02                | 0.04                 | 0.01               | 0.01                | 0.02                 | 0.01               | 0.02               | 0.02          | 1.00 | 1.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.03 | 0.01 | 0.04 | 0.04 | 0.02 | 0.04 | 0.03 |
| ESTU_PRIVADO_LIBERTAD   | 0.02        | 0.00                    | 0.00              | 0.00              | 0.01             | 0.02                  | 0.01              | 0.02                | 0.05                | 0.00                 | 0.02               | 0.02                | 0.00                 | 0.02               | 0.05               | 0.05          | 0.00 | 0.00 | 1.00 |      |      |      | 0.00 |      |      |      |      |      | 0.01 |
| FAMI_CUARTOSHOGAR       | 0.04        | 0.04                    | 0.05              | 0.06              | 0.02             | 0.07                  | 0.02              | 0.05                | 0.09                | 0.09                 | 0.07               | 0.07                | 0.00                 | 0.06               | 0.09               | 0.09          | 0.02 | 0.02 |      | 1.00 | 0.06 | 0.06 | 0.08 | 0.28 | 0.18 | 0.19 | 0.18 | 0.18 | 0.06 |
| FAMI_EDUCACIONMADRE     | 0.06        | 0.20                    | 0.22              | 0.22              | 0.09             | 0.08                  | 0.11              | 0.16                | 0.12                | 0.38                 | 0.07               | 0.07                | 0.01                 | 0.08               | 0.11               | 0.12          | 0.02 | 0.02 |      | 0.06 | 1.00 | 0.32 | 0.20 | 0.08 | 0.36 | 0.38 | 0.36 | 0.27 | 0.27 |
| FAMI_EDUCACIONPADRE     | 0.05        | 0.20                    | 0.25              | 0.24              | 0.09             | 0.08                  | 0.11              | 0.15                | 0.12                | 0.36                 | 0.07               | 0.07                | 0.01                 | 0.06               | 0.11               | 0.12          | 0.02 | 0.02 |      | 0.06 | 0.32 | 1.00 | 0.20 | 0.07 | 0.35 | 0.34 | 0.33 | 0.24 | 0.26 |
| FAMI_ESTRATOVIVIENDA    | 0.04        | 0.21                    | 0.36              | 0.32              | 0.11             | 0.18                  | 0.11              | 0.15                | 0.23                | 0.40                 | 0.16               | 0.16                | 0.01                 | 0.07               | 0.21               | 0.22          | 0.03 | 0.03 | 0.00 | 0.08 | 0.20 | 0.20 | 1.00 | 0.07 | 0.39 | 0.38 | 0.39 | 0.27 | 0.26 |
| FAMI_PERSONASHOGAR      | 0.04        | 0.08                    | 0.03              | 0.05              | 0.02             | 0.10                  | 0.03              | 0.06                | 0.12                | 0.12                 | 0.10               | 0.10                | 0.00                 | 0.02               | 0.11               | 0.12          | 0.01 | 0.01 |      | 0.28 | 0.08 | 0.07 | 0.07 | 1.00 | 0.10 | 0.13 | 0.12 | 0.09 | 0.08 |
| FAMI_TIENEAUTOMOVIL     | 0.09        | 0.06                    | 0.15              | 0.26              | 0.12             | 0.20                  | 0.11              | 0.28                | 0.27                | 0.33                 | 0.19               | 0.19                | 0.01                 | 0.05               | 0.25               | 0.26          | 0.04 | 0.04 |      | 0.18 | 0.36 | 0.35 | 0.39 | 0.10 | 1.00 | 0.34 | 0.28 | 0.22 | 0.34 |
| FAMI_TIENECOMPUTADOR    | 0.13        | 0.20                    | 0.07              | 0.17              | 0.09             | 0.32                  | 0.10              | 0.22                | 0.41                | 0.28                 | 0.32               | 0.32                | 0.01                 | 0.04               | 0.40               | 0.41          | 0.04 | 0.04 |      | 0.19 | 0.38 | 0.34 | 0.38 | 0.13 | 0.34 | 1.00 | 0.52 | 0.32 | 0.36 |
| FAMI_TIENEINTERNET      | 0.11        | 0.26                    | 0.05              | 0.14              | 0.09             | 0.33                  | 0.09              | 0.16                | 0.46                | 0.26                 | 0.33               | 0.33                | 0.00                 | 0.04               | 0.44               | 0.46          | 0.02 | 0.02 |      | 0.18 | 0.36 | 0.33 | 0.39 | 0.12 | 0.28 | 0.52 | 1.00 | 0.34 | 0.33 |
| FAMI_TIENELAVADORA      | 0.11        | 0.16                    | 0.02              | 0.10              | 0.09             | 0.27                  | 0.06              | 0.11                | 0.35                | 0.17                 | 0.27               | 0.27                | 0.00                 | 0.04               | 0.33               | 0.35          | 0.04 | 0.04 |      | 0.18 | 0.27 | 0.24 | 0.27 | 0.09 | 0.22 | 0.32 | 0.34 | 1.00 | 0.20 |
| DESEMP_INGLES           | 0.11        | 0.17                    | 0.37              | 0.31              | 0.09             | 0.17                  | 0.12              | 0.22                | 0.22                | 0.38                 | 0.14               | 0.14                | 0.01                 | 0.05               | 0.20               | 0.21          | 0.03 | 0.03 | 0.01 | 0.06 | 0.27 | 0.26 | 0.26 | 0.08 | 0.34 | 0.36 | 0.33 | 0.20 | 1.00 |

Coeficiente de Cramer

Mosaico de COLE\_NATURALEZA y DESEMP\_INGLES

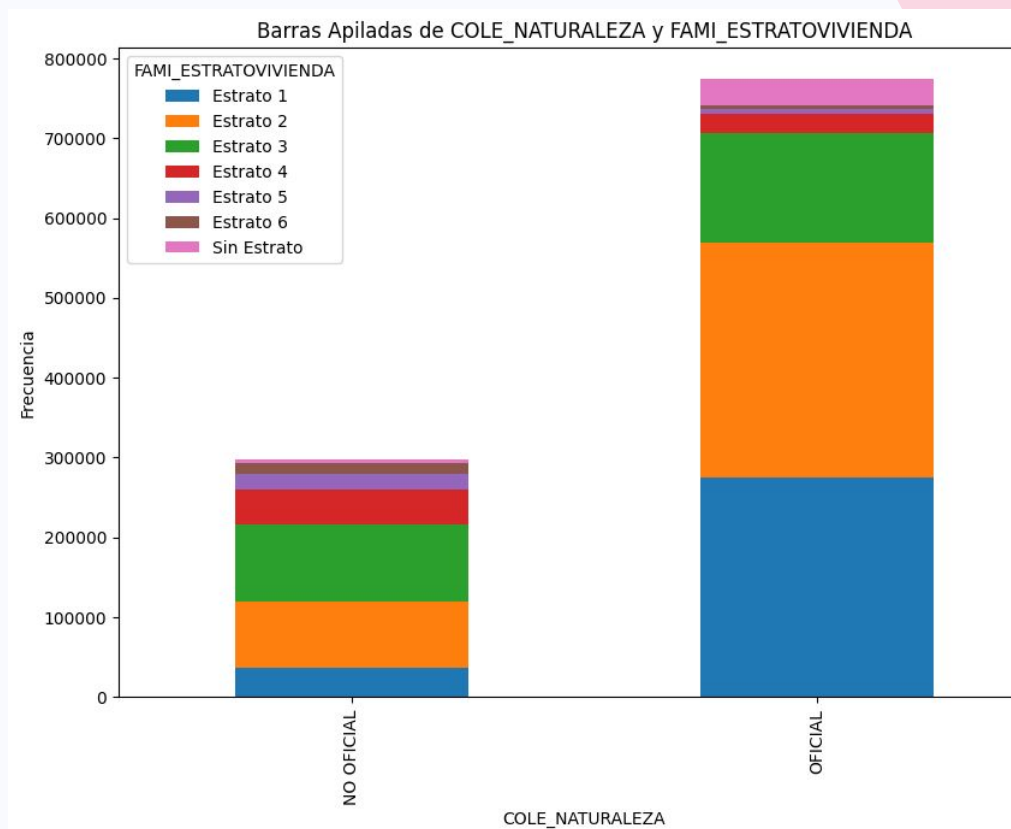
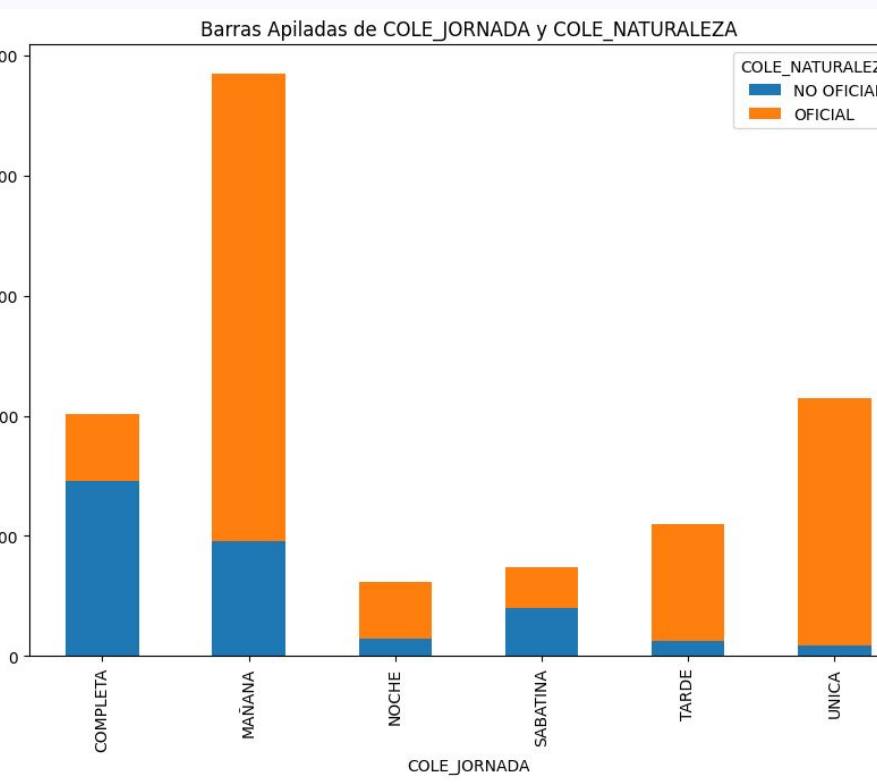


Barras Apiladas de COLE\_BILINGUE y FAMI\_ESTRATOVIVIENDA

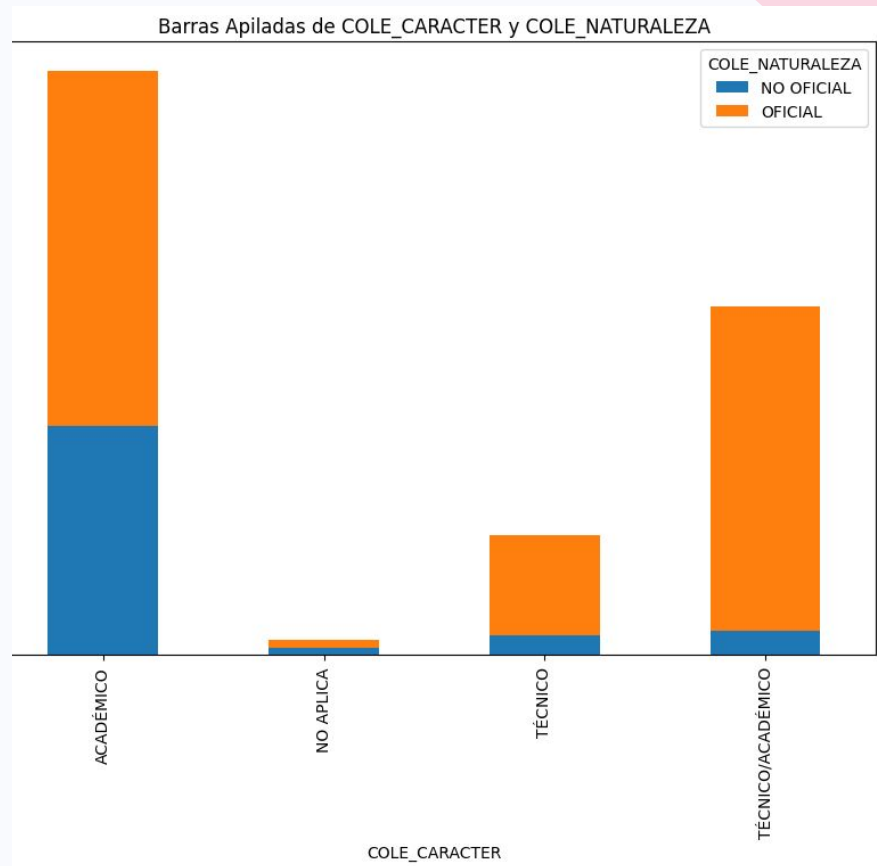
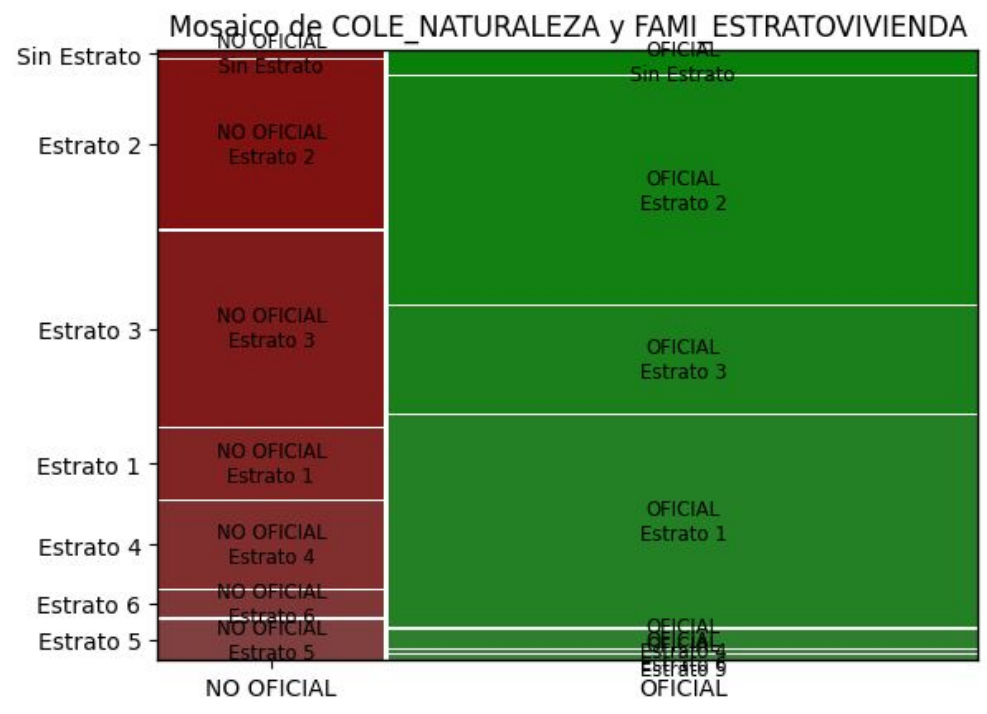


Estadística Descriptiva



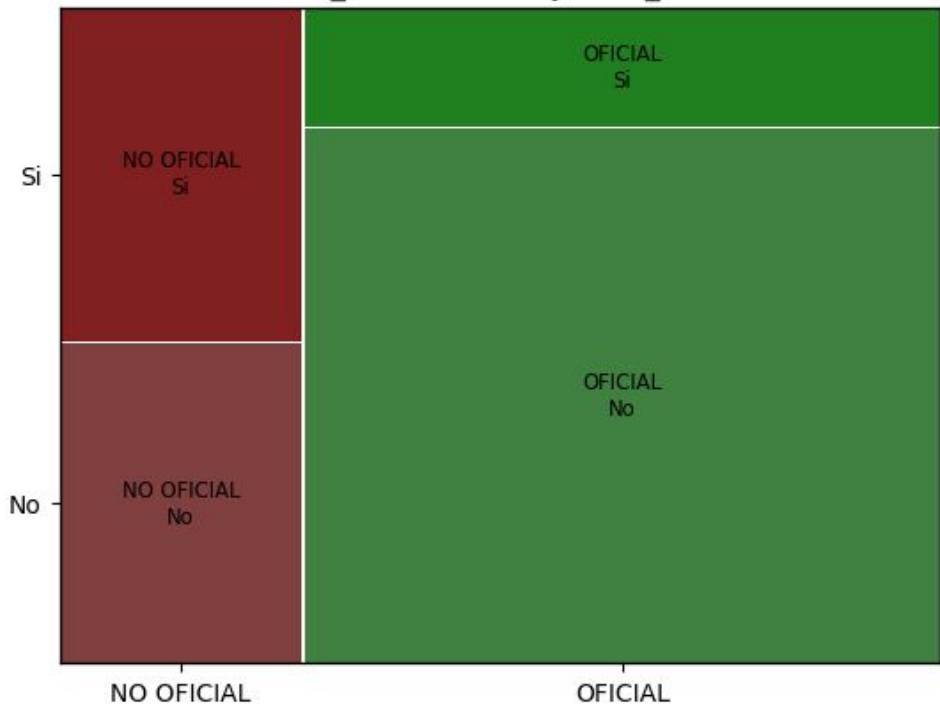


# Estadística Descriptiva

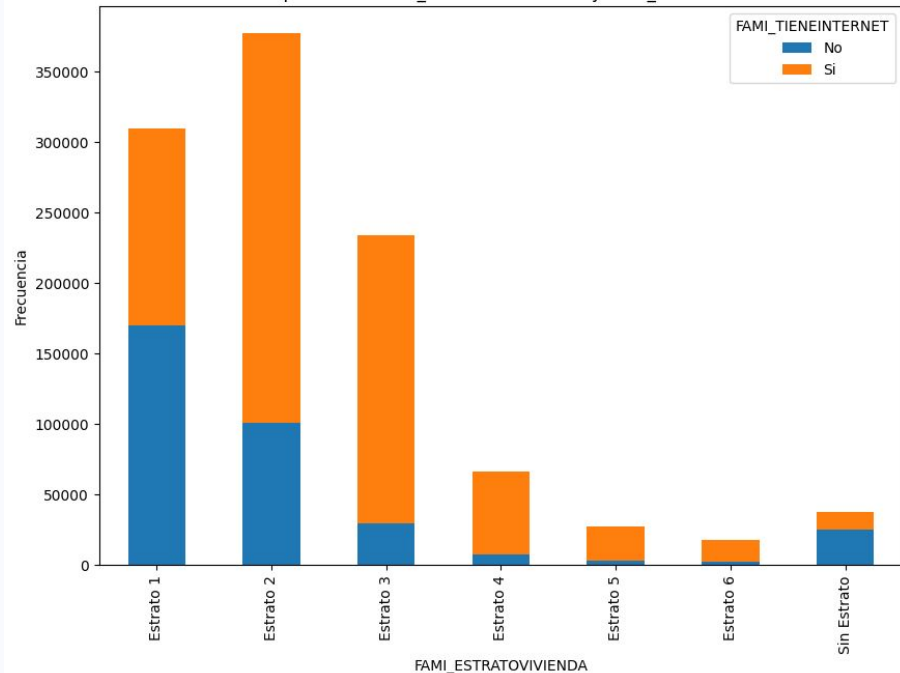


# Estadística Descriptiva

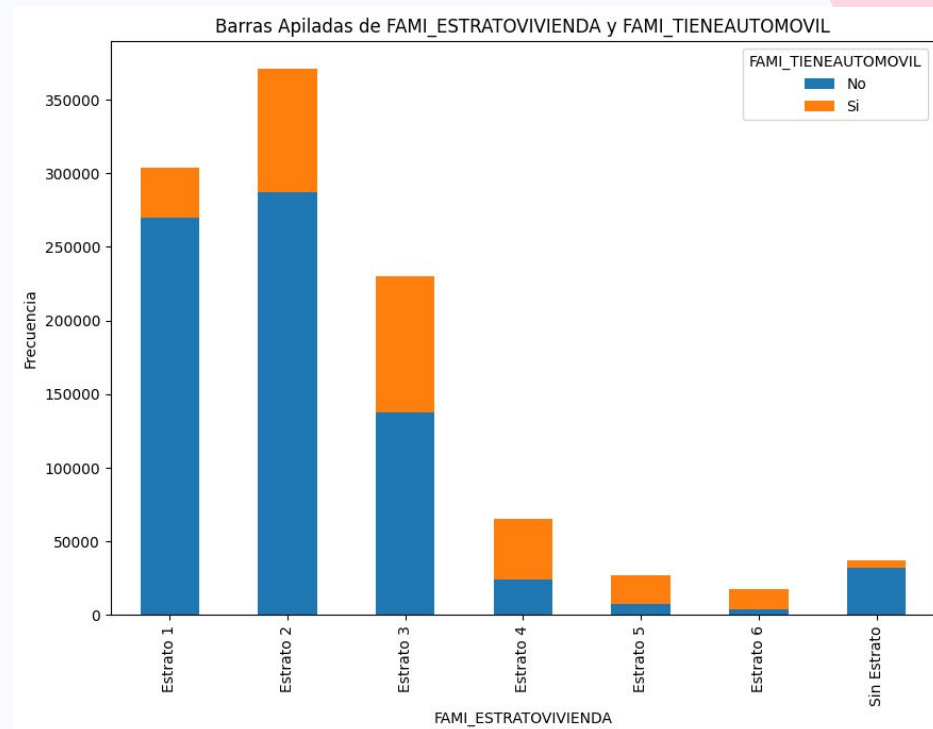
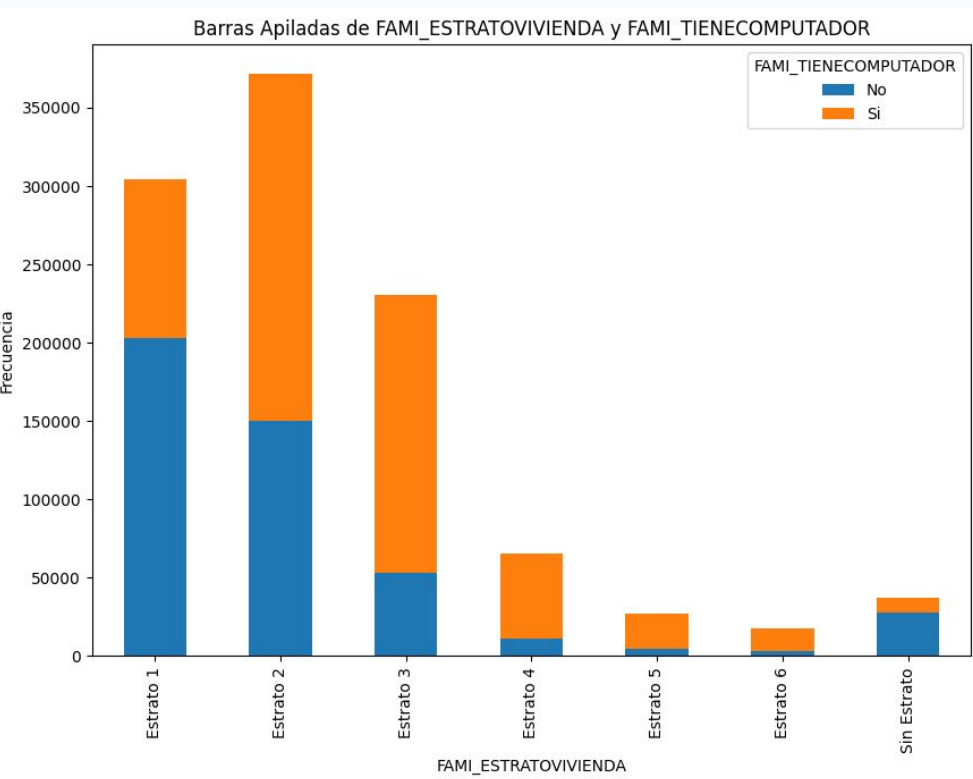
Mosaico de COLE\_NATURALEZA y FAMI\_TIENEAUTOMOVIL



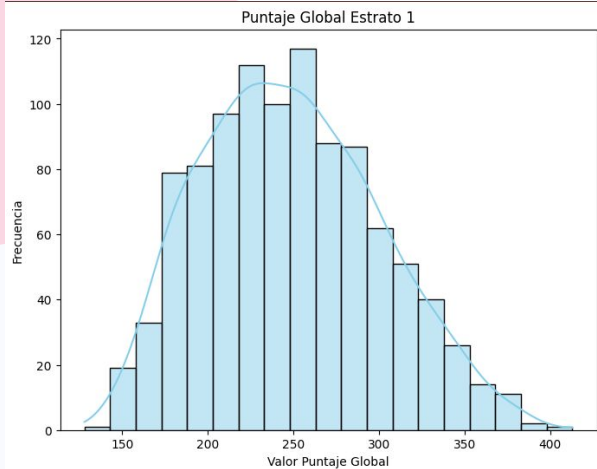
Barras Apiladas de FAMI\_ESTRATOVIVIENDA y FAMI\_TIENEINTERNET



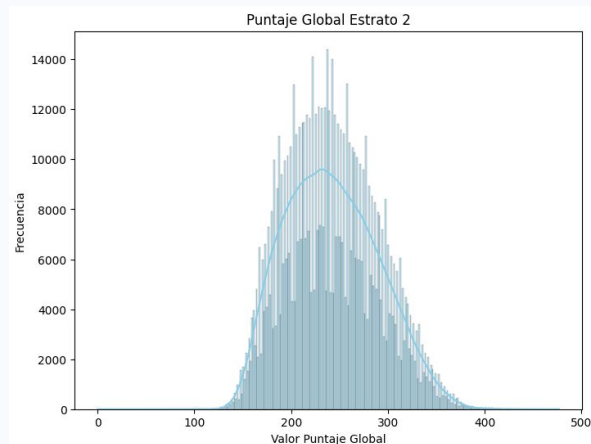
# Estadística Descriptiva



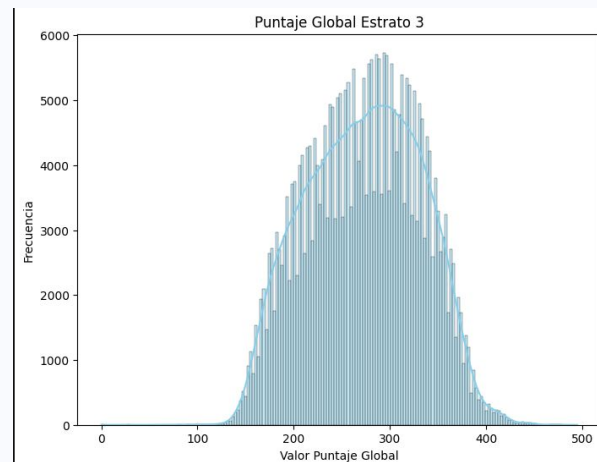
# Estadística Descriptiva



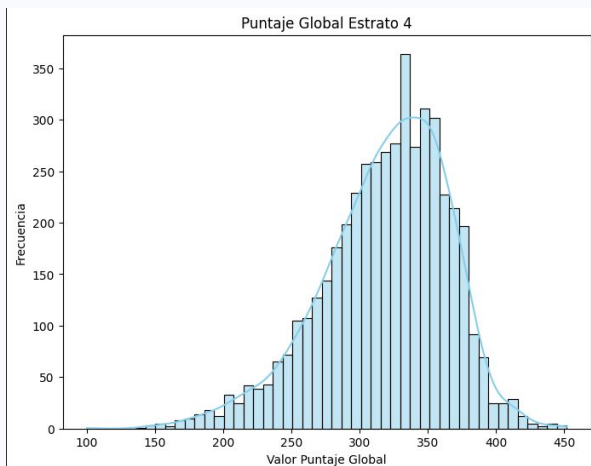
[Stage 167:> (0 + 12) / 12]  
Media: 248.387541625857



[Stage 215:> (0 + 12) / 12]  
Media: 242.23879865598646



[Stage 175:> (0 + 12) / 12]  
Media: 272.9642371412486

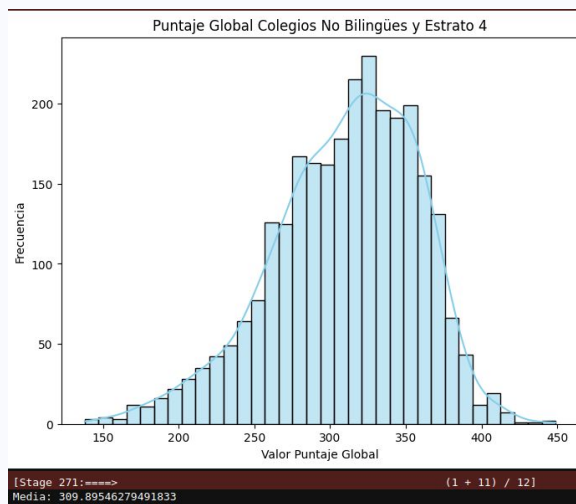
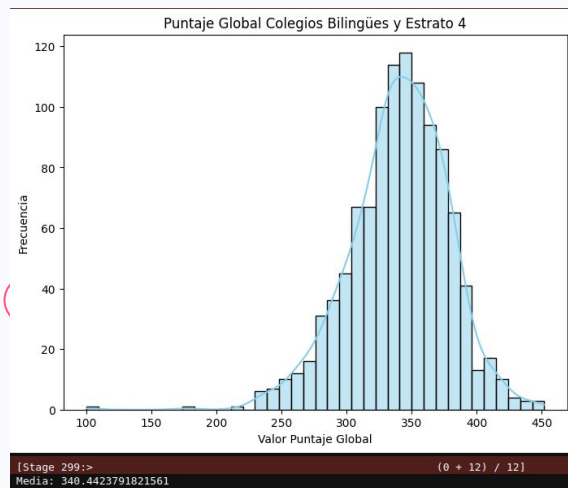
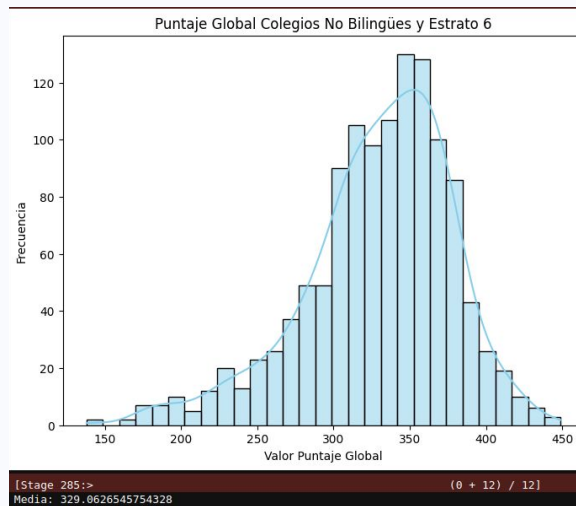
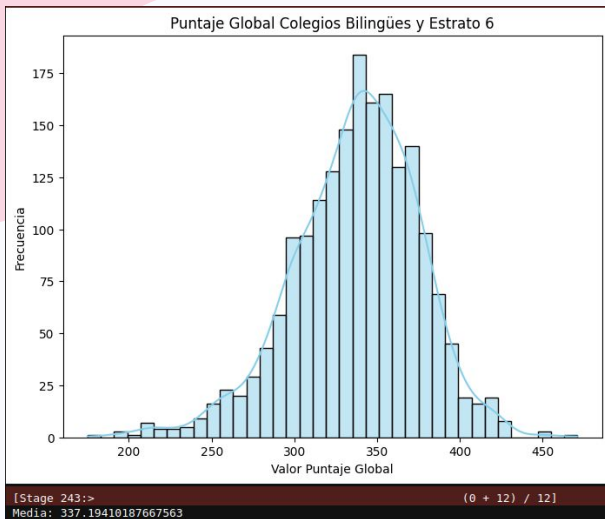


[Stage 179:> (0 + 12) / 12]  
Media: 318.832223334043004

# Estrato

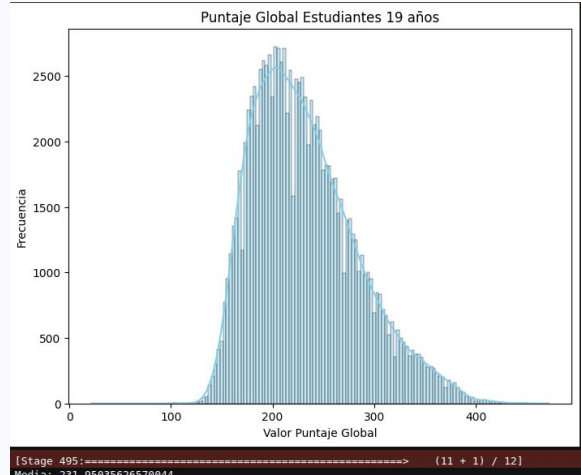
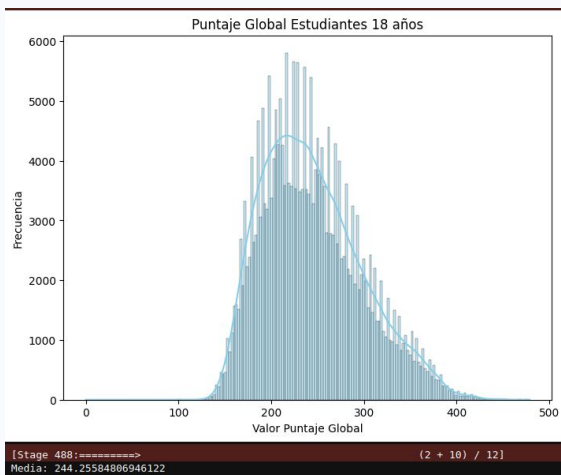
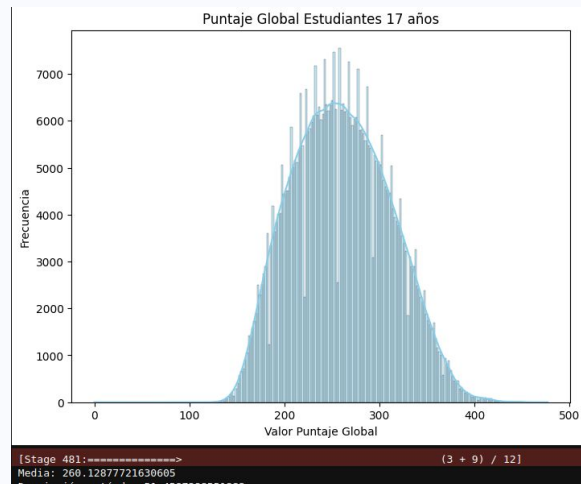
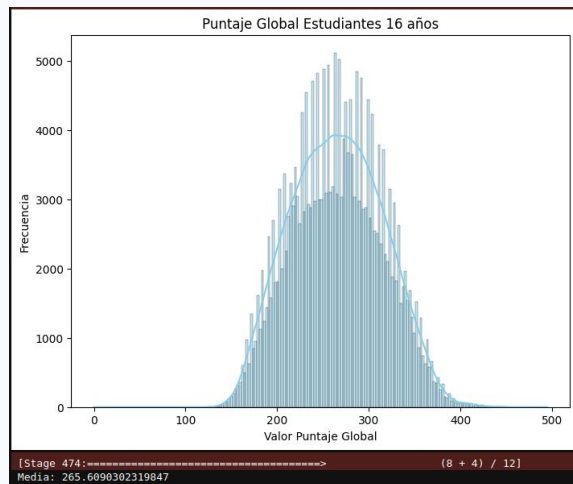






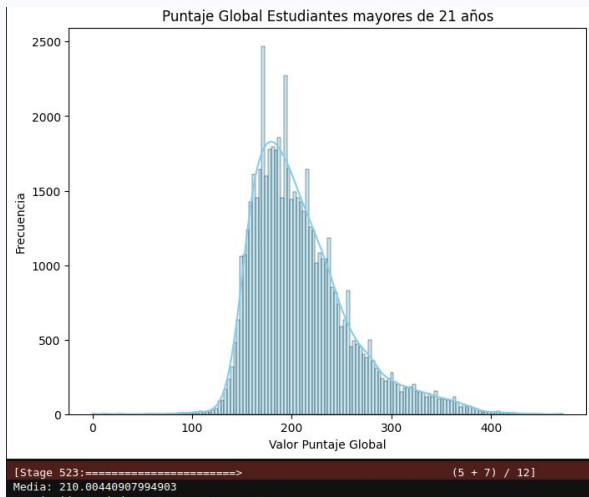
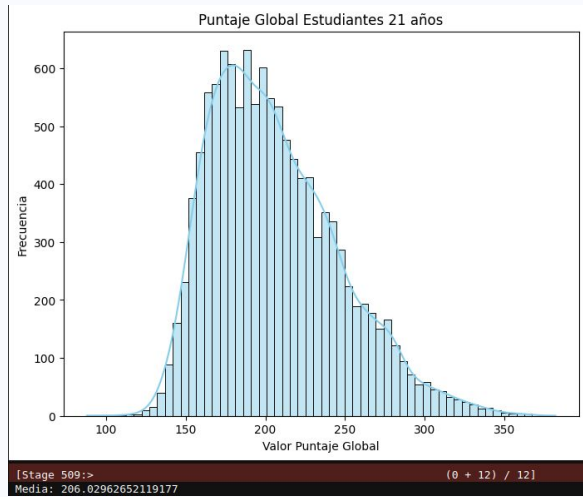
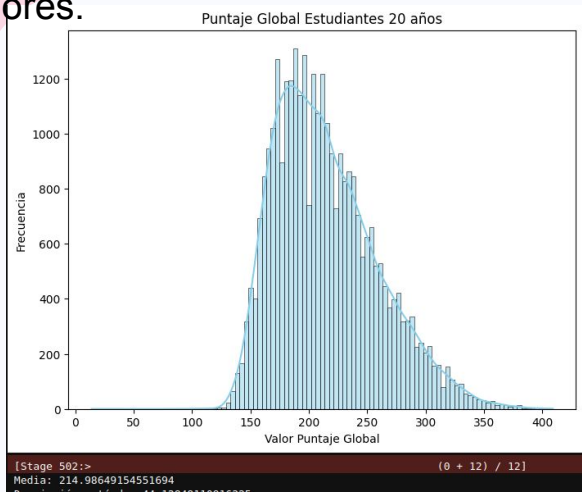
# Colegio Bilingüe y Estrato

# Edad





ordenadores.

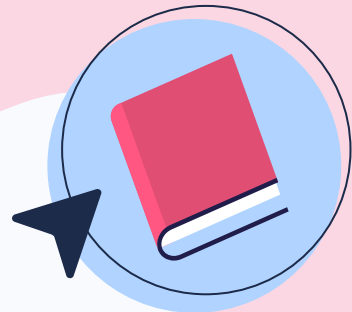


Edad

04

# Limpieza de Datos

Sorpresas respecto a la pandemia



# Limpieza de Datos

cmd 50

```
1 from pyspark.sql.functions import col
2
3
4 columns_to_drop = [
5     'ESTU_CONSECUTIVO',
6     'COLE_CARACTER',
7     'COLE_COD_DANE_ESTABLECIMIENTO',
8     'COLE_COD_DANE_SEDE',
9     'COLE_CODIGO_ICFES',
10    'ESTU_COD_RESIDE_DEPTO',
11    'ESTU_DEPTO_RESIDE',
12    'ESTU_ESTADONVESTIGACION',
13    'ESTU_PRIVADO_LIBERTAD',
14    'ESTU_PAIS_RESIDE',
15    'ESTU_NACIONALIDAD'
16 ]
17
18 df = df.drop(*columns_to_drop)
19
```

## ► (2) Spark Jobs

Null count in EDAD column: 0



1

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

Python

```
1 from pyspark.sql.functions import substring, col, mean
2
3 # Extract the year from the 'PERIODO' column
4 df = df.withColumn('PERIODO', df['PERIODO'].substr(1, 4).cast('int'))
5
6 # Calculate age by subtracting birth year from the year in 'PERIODO' column
7 df = df.withColumn('EDAD', df['PERIODO'] - df['AÑO_NACIMIENTO'])
8
9 # Calculate the average age
10 average_age = df.select(mean('EDAD')).collect()[0][0]
11
12 # Print the average age
13 print("Edad promedio:", average_age)
14
15 # Filter the DataFrame for rows where 'ESTU_TIPODOCUMENTO' is equal to 'TI'
16 filtered_df = df.filter(col('ESTU_TIPODOCUMENTO') == 'TI')
17
18 # Calculate the average age from the filtered DataFrame
19 average_age_ti = filtered_df.select(mean('EDAD')).collect()[0][0]
20
21 # Print the average age of people with 'ESTU_TIPODOCUMENTO' equal to 'TI'
22 print("Edad promedio cuando ESTU_TIPODOCUMENTO = TI:", average_age_ti)
23
```

# Limpieza de Datos

```
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

Cmd 80

```
1 from pyspark.sql.functions import col
2
3 # Get the count of null values in the 'COLE_BILINGUE' column
4 null_count = df.filter(col('COLE_BILINGUE').isNull()).count()
5
6 # Print the count
7 print("Null count in COLE_BILINGUE column:", null_count)
8
```

► (2) Spark Jobs

Null count in COLE\_BILINGUE column: 0

► 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.co

Cmd 76

# Limpieza de Datos

```
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 usar
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()
```

```
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                  | 1274696|
| 6                  | 3973|
| 3                  | 137457|
| 5                  | 7031|
| 4                  | 23712|
| 2                  | 294552|
+-----+-----+
```

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

```
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)
```

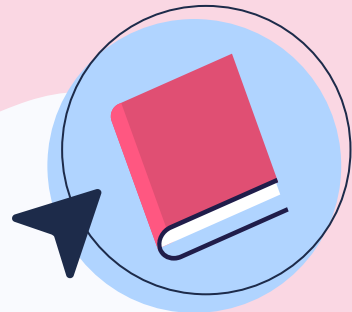
```
+-----+-----+
|FAMI_ESTRATOVIVIENDA| count|
+-----+-----+
| NULL              | 25083|
| 1                  | 36505|
| 6                  | 15870|
| 3                  | 100416|
| 5                  | 22679|
| 4                  | 45652|
| 2                  | 86185|
+-----+-----+
```

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

05

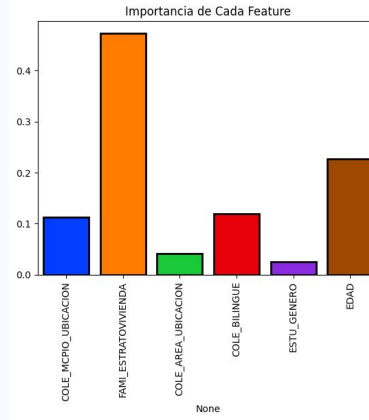
# Modelado

Decision Tree, Random Forest, Regresión



# Decision Tree

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



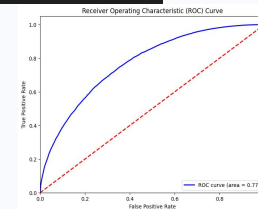
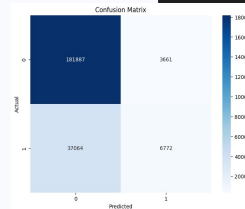
Clasificación binaria a partir de 6 variables:

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

# Regresión Logística

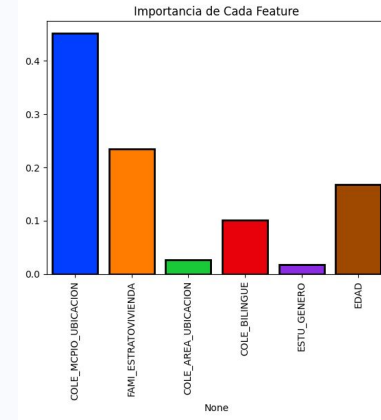
Best parameters: {'classifier\_C': 0.01}  
Accuracy: 0.8225  
Classification Report:

|              | 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  |



# Random Forest

accuracy en train 0.8315942515481715  
accuracy en test 0.824691347260489



# Conclusiones del modelado



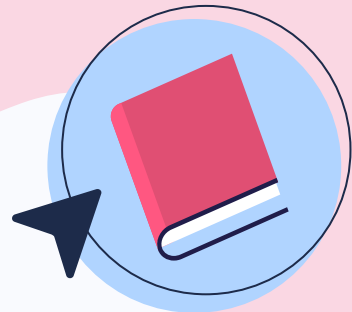
|          | Decision Tree | Random Forest | Regresión Logística |
|----------|---------------|---------------|---------------------|
| Accuracy | 0.8222        | 0.8246        | 0.8225              |



06

# Conclusiones

No todo es lo que parece

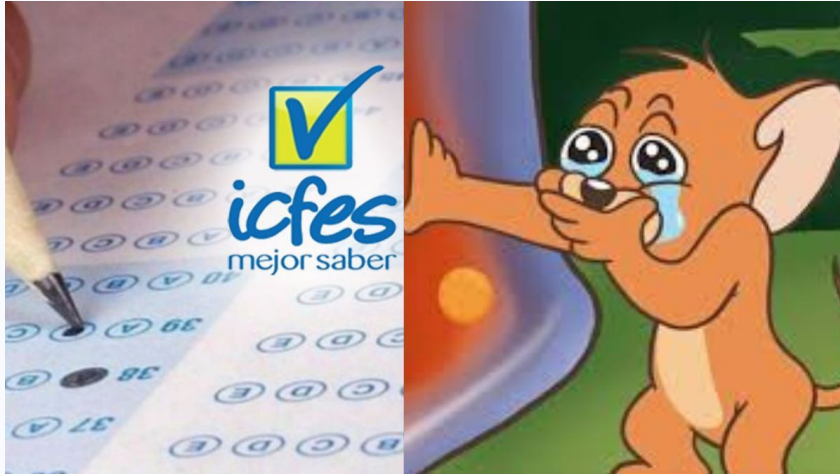




# Conclusiones

- La pandemia no afectó los resultados de la pruebas, no se evidencia un cambio sustancial.
- Los factores que pueden influenciar el puntaje de manera sustancial son el estrato, la edad de la persona que lo presenta y si el colegio en el que estudió es bilingüe o no.
- El bajo rendimiento la prueba de idioma puede tener como causa la baja cantidad de colegios bilingües en el país.
- Hay una brecha de desigualdad respecto al acceso a internet y a un computador, lo cual puede representar cierta ventaja de algunas estudiantes respecto a otros al momento de estudiar.
- Hay una fuerte correlación y el puntaje global de la prueba.
- Hay casos en donde el pertenecer colegio bilingüe puede influir en el puntaje del estudiante de manera significativa y puede incluso tener un efecto igual o mayor que el estrato en el puntaje global de la prueba.

# Gracias!



**CREDITS:** This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

