

FACULTAD DE INGENIERIA



ASIGNATURA: Procesamiento de datos a gran escala

Parcial 2

PROFESOR: John Corredor Franco

AUTORES: Alejandro Salamanca, Andrés Salamanca, Alberto Vigna

Pontificia Universidad Javeriana  
Bogotá  
15 de mayo, 2024

<b>FASE INICIAL .....</b>	<b>4</b>
INTRODUCCIÓN: .....	4
CONJUNTO DE DATOS: .....	5
<i>Información Personal:</i> .....	5
<i>Información de Contacto:</i> .....	5
<i>Información Socioeconómica:</i> .....	6
<i>Información del Colegio:</i> .....	6
<i>Datos de Citación del Examen:</i> .....	7
<i>Resultados:</i> .....	8
PREGUNTAS CLAVE: .....	8
<b>PROCESAMIENTO DE LOS DATOS.....</b>	<b>10</b>
RECOLECCIÓN DE LOS DATOS: .....	10
<i>Manejo de las bases de datos:</i> .....	10
UNIFICACIÓN Y LIMPIEZA DE LOS DATOS .....	12
<i>Carga de los datos desde GitHub:</i> .....	12
LIMPIEZA DE DATOS: .....	13
<i>Proceso de Limpieza de Datos</i> .....	13
1. <i>Cálculo del Porcentaje de Valores Nulos:</i> .....	13
2. <i>Filtrado de Columnas con Menos del 5% de Valores Nulos</i> .....	13
3. <i>Eliminación de Filas con Valores Nulos en Columnas Clave</i> .....	14
4. <i>Contabilización de Valores Nulos Restantes</i> .....	14
<b>EXPLORACIÓN DE LOS DATOS: .....</b>	<b>15</b>
DISTRIBUCIÓN DE LA PUNTUACIÓN GLOBAL .....	15
<i>Promedio de Puntajes:</i> .....	15
<i>Rango de Puntajes:</i> .....	15
<i>Interpretación:</i> .....	16
DISTRIBUCIÓN DE LA PUNTUACIÓN EN CIENCIAS NATURALES .....	16
<i>Mediana de Puntajes:</i> .....	16
<i>Interpretación:</i> .....	16
DISTRIBUCIÓN DE LA PUNTUACIÓN EN INGLÉS .....	16
<i>Mediana de Puntajes:</i> .....	16
<i>Interpretación:</i> .....	16
DISTRIBUCIÓN DE LA PUNTUACIÓN EN CIENCIAS SOCIALES Y CIUDADANAS .....	16
<i>Mediana de Puntajes:</i> .....	16
<i>Interpretación:</i> .....	16
DISTRIBUCIÓN DE LA PUNTUACIÓN EN LECTURA CRÍTICA .....	16
<i>Promedio de Puntajes:</i> .....	16
<i>Interpretación:</i> .....	16
DISTRIBUCIÓN DE LA PUNTUACIÓN EN MATEMÁTICAS .....	17
<i>Promedio de Puntajes:</i> .....	17
<i>Interpretación:</i> .....	17
OBSERVACIONES GENERALES .....	17
<i>Casos Excepcionales:</i> .....	17

<i>Variabilidad:</i> .....	17
<b>PROBLEMA</b> .....	<b>32</b>
<b>IMPLEMENTACIÓN DE TÉCNICAS ML Y RESULTADOS</b> .....	<b>34</b>

## **Tabla de Contenidos**

# Fase Inicial

## Introducción:

El conjunto de datos seleccionado para este análisis comprende los resultados de las pruebas Saber 11 administradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), una entidad autónoma adscrita al Ministerio de Educación Nacional de Colombia. El ICFES desempeña un papel fundamental en la evaluación del sistema educativo colombiano, ofreciendo actividades de evaluación en todos los niveles educativos y colaborando estrechamente con el Ministerio de Educación en la realización de exámenes estatales.

La prueba Saber 11 es reconocida como un referente importante para la admisión a la educación superior en Colombia y se ofrece a estudiantes que cursan su último año de bachillerato. El conjunto de datos abarca los años 2018, 2019, 2020, 2021, 2022 y 2023 (aunque realmente se tienen los registros de las pruebas desde el año 2001, pero esos resultados no serán tomados en cuenta para esta investigación), centrándose específicamente en los exámenes presentados por los colegios de calendario A (colegios que terminan su año escolar en diciembre, identificándose como -2) y los colegios calendario B los cuales culminan su año escolar a mitad de año, imitando el modelo de educación estadounidense y europeo, identificándose como -1.

Este conjunto de datasets que serán estudiados y explicados más adelante en el documento se tomaron con el propósito de evaluar los resultados y en general, el rendimiento de los estudiantes antes, durante y después de la pandemia del COVID-19. Además, gracias a la gran cantidad de información que se recoge en este examen acerca del estudiante, su familia y su situación socioeconómica, nos permite realizar un estudio mucho mas detallado acerca de los resultados, no solo revisando el rendimiento de los estudiantes durante

este espacio temporal (2018-2023) sino también como su situación personal puede afectar positiva o negativamente sus resultados.

Explicando un poco más a detalle el porque se ha decidido trabajar con este espacio temporal, han priorizado los años 2018 y 2019 para representar los resultados obtenidos en condiciones pre-pandémicas, seguidos de los años 2020 y 2021, que reflejan los desafíos y cambios ocasionados por la pandemia del COVID-19. Finalmente, se incluyen los años 2022 y 2023 para observar las posibles tendencias y recuperación en el rendimiento educativo post-pandemia.

El conjunto de datos consta de una extensa cantidad de información, con alrededor de 80 columnas, que proporcionan una riqueza de detalles sobre los estudiantes, sus antecedentes socioeconómicos (y los de sus familiares inmediatos), los colegios en los que estudian, la logística de la citación para el examen y, por supuesto, los resultados obtenidos en la prueba Saber 11.

## Conjunto de Datos:

El conjunto de datos consta de una extensa cantidad de información, con alrededor de 80 columnas, que proporcionan una riqueza de detalles sobre los estudiantes, sus antecedentes socioeconómicos (y los de sus familiares inmediatos), los colegios en los que estudian, la logística de la citación para el examen y, por supuesto, los resultados obtenidos en la prueba Saber 11.

Dentro de este conjunto de datos, se pueden identificar seis secciones principales:

### Información Personal:

Esta sección incluye datos sobre las características demográficas y personales de los estudiantes. Algunos ejemplos de columnas que están en esta sección son:

- **ESTU\_TIPODOCUMENTO:** Tipo de documento de identificación del estudiante.  
Ejemplo: CC, TI, CE
- **ESTU\_NACIONALIDAD:** Nacionalidad del estudiante.  
Ejemplo: Colombiana, Venezolana, Peruana
- **ESTU\_GENERO:** Género del estudiante.  
Ejemplo: M, F
- **ESTU\_FECHANACIMIENTO:** Fecha de nacimiento del estudiante.  
Ejemplo: 2004-06-15, 2003-12-01, 2005-03-22
- **ESTU\_CONSECUTIVO:** Número consecutivo del estudiante.  
Ejemplo: 12345, 67890, 11223

### Información de Contacto:

Esta sección se centra en los datos de contacto de los estudiantes para posibles comunicaciones. Ejemplos de columnas incluyen:

- **STU\_PAIS\_RESIDE:** País de residencia del estudiante.  
Ejemplo: Colombia, Venezuela, Ecuador
- **ESTU\_DEPTO\_RESIDE:** Departamento de residencia del estudiante.  
Ejemplo: Antioquia, Cundinamarca, Valle del Cauca

- **ESTU\_COD\_RESIDE\_DEPTO:** Código del departamento de residencia.  
Ejemplo: 05, 11, 76
- **ESTU\_MCPIO\_RESIDE:** Municipio de residencia del estudiante.  
Ejemplo: Medellín, Bogotá, Cali
- **ESTU\_COD\_RESIDE\_MCPIO:** Código del municipio de residencia.  
Ejemplo: 05001, 11001, 76001

### **Información Socioeconómica:**

Aquí se recopilan datos sobre la situación socioeconómica del estudiante y su familia. Algunas columnas son:

- **AMI\_ESTRATOVIVIENDA:** Estrato socioeconómico de la vivienda del estudiante.  
Ejemplo: 1, 2, 3, 4, 5
- **FAMI\_PERSONASHOGAR:** Número de personas en el hogar.  
Ejemplo: 4, 3, 5, 6, 2
- **FAMI\_CUARTOSHOGAR:** Número de cuartos en la vivienda.  
Ejemplo: 3, 4, 2, 5, 6
- **FAMI\_EDUCACIONPADRE:** Nivel educativo del padre.  
Ejemplo: Primaria, Secundaria, Universitaria, Ninguno, Técnica.
- **FAMI\_EDUCACIONMADRE:** Nivel educativo de la madre.  
Ejemplo: Primaria, Secundaria, Universitaria, Ninguno, Técnica.

### **Información del Colegio:**

Esta sección contiene datos específicos sobre la institución educativa del estudiante. Algunas columnas posibles son:

- **COLE\_CODIGO\_ICFES:** Código ICFES del colegio.  
Ejemplo: 123456, 234567, 345678, 456789, 567890

- **COLE\_NOMBRE\_ESTABLECIMIENTO:** Nombre del establecimiento educativo.  
Ejemplo: Colegio Nacional, Institución Educativa San Juan, Liceo Moderno.
- **COLE\_GENERO:** Género del colegio (si es mixto, solo femenino o solo masculino).  
Ejemplo: Mixto, Femenino, Masculino
- **COLE\_NATURALEZA:** Naturaleza del colegio (público o privado).  
Ejemplo: Público, Privado
- **COLE\_CALEDARIO:** Calendario académico del colegio (A o B).  
Ejemplo: A, B

#### **Datos de Citación del Examen:**

Esta sección detalla la logística y la organización de la citación para la prueba Saber 11.  
Ejemplos de columnas incluyen:

- **ESTU\_COD\_MCPIO\_PRESENTACION:** Código del municipio donde se presentó el examen.  
Ejemplo: 05001, 11001, 76001
- **ESTU\_MCPIO\_PRESENTACION:** Municipio donde se presentó el examen.  
Ejemplo: Medellín, Bogotá, Cali
- **ESTU\_DEPTO\_PRESENTACION:** Departamento donde se presentó el examen.  
Ejemplo: Antioquia, Cundinamarca, Valle del Cauca
- **ESTU\_COD\_DEPTO\_PRESENTACION:** Código del departamento donde se presentó el examen.  
Ejemplo: 05, 11, 76
- **ESTU\_PRESENTACIONESABADO:** Indica si el estudiante presentó el examen un sábado.  
Ejemplo: S, N

## Resultados:

Esta sección incluye los resultados obtenidos por los estudiantes en las diferentes áreas evaluadas por la prueba Saber 11. Algunas columnas son:

- **PUNT\_LECTURA\_CRITICA:** Puntaje en lectura crítica.  
Ejemplo: 65, 70, 80, 75, 60
- **PUNT\_MATEMATICAS:** Puntaje en matemáticas.  
Ejemplo: 80, 85, 90, 75, 70
- **PUNT\_C\_NATURALES:** Puntaje en ciencias naturales.  
Ejemplo: 75, 80, 70, 85, 65
- **PUNT\_SOCIALES\_CIUDADANAS:** Puntaje en sociales y ciudadanas.  
Ejemplo: 85, 90, 75, 80, 70
- **PUNT\_INGLES:** Puntaje en inglés.  
Ejemplo: 70, 75, 80, 85, 60

Esta estructura organizativa facilita la segmentación y el análisis específico de diferentes aspectos relacionados con el rendimiento académico y los factores que pueden influir en él.

La relevancia de este conjunto de datos es innegable, ya que representa una vasta cantidad de información que abarca millones de registros. Este volumen de datos ofrece una oportunidad única para realizar un análisis estadístico detallado y exhaustivo, que puede arrojar luz sobre tendencias educativas, desigualdades socioeconómicas y el impacto de eventos externos, como la pandemia de COVID-19, en el rendimiento académico de los estudiantes colombianos.

## Preguntas Clave:

1. ¿Cómo se comparan los resultados promedio de la prueba Saber 11 antes, durante y después de la pandemia? Esta pregunta nos proporcionaría una visión del impacto de la



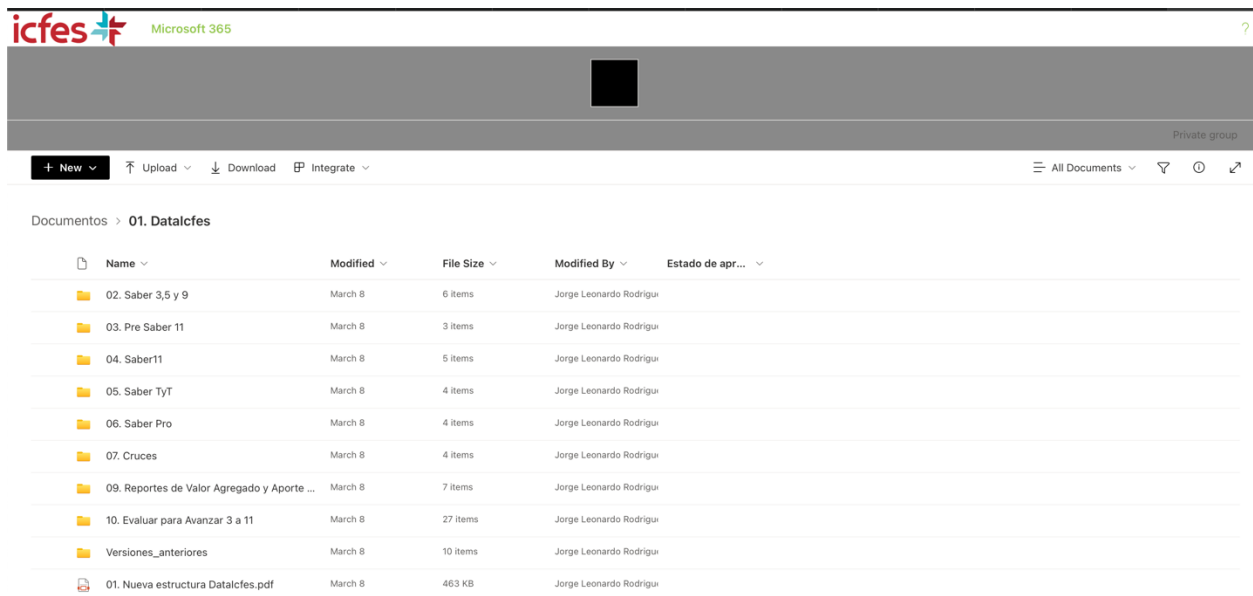
pandemia en el rendimiento académico de los estudiantes. Además, podríamos investigar si los resultados pre-pandémicos vuelven a ser los mismos tras la recuperación mundial de la pandemia.

2. ¿Cuál es el efecto de los factores socioeconómicos, como el nivel de ingresos familiar o el tipo de colegio, en los resultados de la prueba Saber 11? Dado que una parte significativa del conjunto de datos se centra en los factores socioeconómicos, esta pregunta nos permitiría explorar cómo estos factores influyen en el rendimiento general de los estudiantes.
3. ¿Cuál es la composición demográfica de la población estudiada? Al realizar un análisis detallado de los datos a través de diversas gráficas y medidas, como la distribución por género, el tamaño del hogar y el acceso a internet, podemos obtener una mejor comprensión de la población estudiantil. Esta información es crucial para contextualizar los resultados y comprender mejor las posibles disparidades en el rendimiento académico.

# Procesamiento de los Datos

## Recolección de los datos:

Para lograr obtener los datos, tuvimos que entrar a la página del ICFES y solicitar un acceso especial mediante la plataforma, una vez logras llenar todos los cuestionarios te dan una contraseña para que puedas acceder a un SharePoint en el cual se encuentran todos los resultados de todas las pruebas que realiza el estado a sus estudiantes (desde el Saber 3, 5 y 9 hasta el Saber Pro, hecho a universitarios).



Name	Modified	File Size	Modified By	Estado de apr...
02. Saber 3,5 y 9	March 8	6 Items	Jorge Leonardo Rodríguez	
03. Pre Saber 11	March 8	3 Items	Jorge Leonardo Rodríguez	
04. Saber11	March 8	5 Items	Jorge Leonardo Rodríguez	
05. Saber TyT	March 8	4 Items	Jorge Leonardo Rodríguez	
06. Saber Pro	March 8	4 Items	Jorge Leonardo Rodríguez	
07. Cruces	March 8	4 Items	Jorge Leonardo Rodríguez	
09. Reportes de Valor Agregado y Aporte ...	March 8	7 Items	Jorge Leonardo Rodríguez	
10. Evaluar para Avanzar 3 a 11	March 8	27 Items	Jorge Leonardo Rodríguez	
Versiones anteriores	March 8	10 Items	Jorge Leonardo Rodríguez	
01. Nueva estructura DataIcfes.pdf	March 8	463 KB	Jorge Leonardo Rodríguez	

Una vez se hayan descargado los datos de este portal podemos utilizarlo en nuestro código para procesarlos y analizarlos.

## Manejo de las bases de datos:

Los resultados de los exámenes se encontraban divididos en años (2018, 2019, etc), cada uno de esos años tiene 2 periodos, los exámenes a los colegios calendario B (los que terminan en -1) los cuales realmente solo representan a un 15% del total de colegios y que la base de datos de todos estos resultados para estos colegios solo sea de 11MB aproximadamente. El problema consistía en los resultados de los colegios de calendario A, estos al ser mayoría (representando al 85% de los colegios a nivel nacional) tenían tamaños excesivamente grandes, logrando tener hasta 600MB por cada periodo.

Esto no solo ralentizaba enormemente el procesamiento de los datos debido al tamaño de cada uno de los archivos (llegando a tener más de 3GB si sumamos el peso de todos estos) sino que dificultaba su almacenamiento en sistemas de control de versiones como GitHub en los cuales un archivo puede tener como máximo 100MB. El utilizar esta plataforma fue crucial para poder trabajar en este parcial de forma conjunta y poderlo además utilizar en otros dispositivos como fue necesario más adelante.

Para solventar el problema de tamaño se creó un programa aparte el cual tiene el propósito de separar las bases de datos de los resultados de los colegios calendario A para que de esta manera se puedan manejar de forma más cómoda y sobre todo, poderlos cargar al repositorio.

```
1 usage  Alberto Vigna
def split_file(file_path, num_parts=5):
    import os

    # Leer el contenido del archivo
    with open(file_path, 'r', encoding='utf-8') as file:
        lines = file.readlines()

    # La primera línea contiene los nombres de las columnas
    header = lines[0]
    data_lines = lines[1:]

    # Calcular el número de líneas por archivo
    lines_per_part = len(data_lines) // num_parts

    # Obtener el nombre base del archivo original sin la extensión
    base_name = os.path.splitext(os.path.basename(file_path))[0]

    # Crear directorio para archivos divididos si no existe
    output_dir = os.path.join(os.path.dirname(file_path), f'particiones_{base_name}')
    os.makedirs(output_dir, exist_ok=True)

    # Escribir cada parte en un archivo separado
    for i in range(num_parts):
        part_file_path = os.path.join(output_dir, f'{base_name}_parte_{i + 1}.txt')
        start_index = i * lines_per_part
        # Asegurarse de incluir todas las líneas en la última parte
        end_index = (i + 1) * lines_per_part if i != num_parts - 1 else None

        with open(part_file_path, 'w', encoding='utf-8') as part_file:
            part_file.write(header) # Escribir los nombres de las columnas
            part_file.writelines(data_lines[start_index:end_index])

    print(f'Archivo dividido en {num_parts} partes en la carpeta {output_dir}')
```

Una vez ya todos los archivos de los colegios calendario A estuvieron particionados correctamente se procedió a cargarlos en el repositorio de forma organizada para que, cuando sea la hora de utilizarlos sea mucho más fácil localizarlos.

Otra herramienta usada fue el clúster proporcionado por nuestro profesor John Corredor, este clúster compuesto por 4 nodos con unas 64GB de RAM cada uno nos permitía realizar un procesamiento mucho más veloz y eficiente, pero debido a que este se encontraba en mantenimiento y a las restricciones de seguridad que interponía la universidad para poder conectarse de forma remota a este, tuvimos que descartarlo.

## Unificación y limpieza de los datos

Para llevar a cabo un análisis exhaustivo de los resultados del ICFES, es crucial unificar los datos provenientes de diferentes archivos y periodos. Dada la gran cantidad de estudiantes que presentan el examen en los segundos periodos, los datos fueron divididos y almacenados en múltiples archivos. Este proceso permitió manejar los datos de manera eficiente y efectiva, evitando problemas de rendimiento en la carga y manipulación de los mismos. A continuación, se describe detalladamente cómo se cargaron los datos desde GitHub hasta el notebook y cómo se unieron posteriormente.

### Carga de los datos desde GitHub:

Como se explicó anteriormente, los archivos pesados tuvieron que ser particionados en partes (5 partes para ser exactos), estos se cargaron de forma individual a un repositorio en GitHub el cual posee toda la información de este parcial.

A continuación, se procedió a cargar cada uno de las rutas (5 por cada año) a una lista para poder extraerlos, luego Utilizando la biblioteca *pandas*, cada archivo se leyó directamente desde su URL. Este enfoque permitió acceder a los datos de manera remota sin necesidad de descargarlos previamente al sistema local. Una vez leídos todos los archivos, se unieron en un solo DataFrame utilizando la función *pd.concat*. Este paso fue crucial para consolidar todos los datos en una única estructura, facilitando así el análisis posterior.

```
# URLs de los archivos divididos
paths2018 = ["https://raw.githubusercontent.com/Betico1928/Parcial2---Procesamiento-de-Datos-a-Gran-Escala/main/Datasets/Datos/Pre-Pandemia/2018/2018-2/particiones_SB11_20182/SB11_20182_parte_1.txt",
            "https://raw.githubusercontent.com/Betico1928/Parcial2---Procesamiento-de-Datos-a-Gran-Escala/main/Datasets/Datos/Pre-Pandemia/2018/2018-2/particiones_SB11_20182/SB11_20182_parte_2.txt",
            "https://raw.githubusercontent.com/Betico1928/Parcial2---Procesamiento-de-Datos-a-Gran-Escala/main/Datasets/Datos/Pre-Pandemia/2018/2018-2/particiones_SB11_20182/SB11_20182_parte_3.txt",
            "https://raw.githubusercontent.com/Betico1928/Parcial2---Procesamiento-de-Datos-a-Gran-Escala/main/Datasets/Datos/Pre-Pandemia/2018/2018-2/particiones_SB11_20182/SB11_20182_parte_4.txt",
            "https://raw.githubusercontent.com/Betico1928/Parcial2---Procesamiento-de-Datos-a-Gran-Escala/main/Datasets/Datos/Pre-Pandemia/2018/2018-2/particiones_SB11_20182/SB11_20182_parte_5.txt"]

# Creación del DataFrame combinado
dataframes = [] # Lista para almacenar cada DataFrame temporal

for path in paths2018:
    # Leer cada archivo desde la URL y agregarlo a la lista de DataFrames
    df_temp = pd.read_csv(path, delimiter=';', header=0, encoding='utf-8', engine='python')
    dataframes.append(df_temp)

# Concatenar todos los DataFrames en uno solo
DF_Resultados_2018_2 = pd.concat(dataframes)

# Mostrar las primeras 10 filas del DataFrame final
DF_Resultados_2018_2
```

Este proceso se repitió con cada uno de los años, logrando así que tengamos un dataframe unificado para cada uno de los años. Cada uno de estos contaba con unos 600.000 registros aproximadamente, generando en total unos 3'600.000 de registros en total.

## Limpieza de Datos:

En el análisis de los resultados del ICFES, la limpieza y el tratamiento de datos son fundamentales para garantizar la calidad y la precisión del análisis. En este caso, se llevaron a cabo varias operaciones clave para identificar y manejar valores nulos, y seleccionar las columnas relevantes para el análisis.

### Identificación de Problemas en los Datos

1. **Valores Nulos:** Los valores nulos pueden surgir por diversas razones, como errores en la captura de datos o la falta de información en ciertas variables. La identificación de estos valores es esencial para decidir cómo manejarlos adecuadamente.
2. **Datos Duplicados:** Los datos duplicados pueden inflar artificialmente el tamaño del conjunto de datos y sesgar los resultados del análisis. Es fundamental identificar y eliminar estos duplicados para asegurar la integridad de los datos.
3. **Valores Atípicos:** Los valores atípicos, o outliers, son observaciones que se encuentran significativamente alejadas de la mayoría de los datos y pueden distorsionar los resultados del análisis estadístico y los modelos predictivos.

### Proceso de Limpieza de Datos

A continuación, se detalla el proceso de limpieza de datos que se realizó:

#### 1. Cálculo del Porcentaje de Valores Nulos:

Primero, se calculó el porcentaje de valores nulos para cada columna del DataFrame final. Esto permitió identificar qué columnas tenían una proporción significativa de datos faltantes y necesitaban ser tratadas.

```
total_rows = final_df.count()

# Calcula el porcentaje de valores nulos para cada columna
Porcentaje_nulos = final_df.select([((count(when(isnan(c) | col(c).isNull(), c)) / total_rows) * 100).alias(c) for c in final_df.columns]).show()
```

#### 2. Filtrado de Columnas con Menos del 5% de Valores Nulos

Basado en el porcentaje de valores nulos calculado anteriormente, se seleccionaron solo las columnas que tenían menos del 5% de datos faltantes. Esto asegura que las columnas con demasiados valores nulos no interfieran en el análisis.

```
1 filtered_columns = [
2     c for c in final_df.columns
3     if ((count(when(isnan(c) | col(c).isNull(), c)) / total_rows) * 100) < 5
4 ]
```

### 3. Eliminación de Filas con Valores Nulos en Columnas Clave

Se identificaron varias columnas clave que son cruciales para el análisis. Las filas que tenían valores nulos en estas columnas específicas fueron eliminadas para mantener la integridad del análisis.

```
borrar = ["ESTU_DEPTO_RESIDE", "ESTU_COD_RESIDE_DEPTO", "ESTU_MCPID_RESIDE", "ESTU_COD_RESIDE_MCPID", "FAMI_PERSONASHOGAR", "FAMI_CUARTOSHOGAR",  
"FAMI_TRABAJOLABORPADRE", "FAMI_TRABAJOLABORMADRE", "FAMI_TRABAJOLABORMADRE", "FAMI_TIENECOMPUTADOR", "FAMI_TIENELAVADORA", "FAMI_TIENEHORNOMICROOGAS",  
"FAMI_TIENEAUTOMOVIL", "FAMI_TIENEMOTOCICLETA", "FAMI_TIENECONSOLAVIDEJUEGOS", "FAMI_SITUACIONECONOMICA", "ESTU_HORASSEMANATRABAJA",  
"ESTU_HORASSEMANATRABAJA", "ESTU_TIPOREMUNERACION", "PUNT_INGLES", "PERCENTIL_INGLES", "DESEMP_INGLES", "DESEMP_INGLES", "PERCENTIL_GLOBAL"]  
final_df = final_df.dropna(subset=borrar)
```

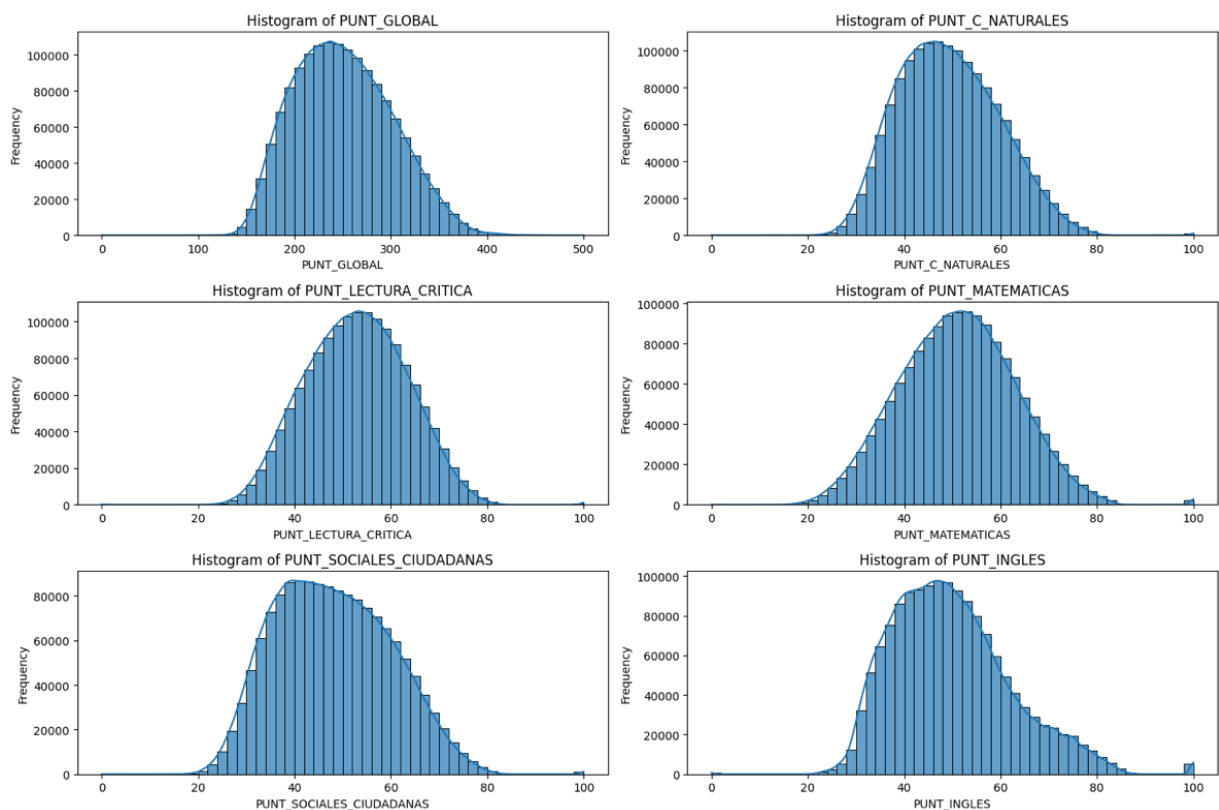
### 4. Contabilización de Valores Nulos Restantes

Después de eliminar las filas con valores nulos en las columnas clave, se contabilizaron los valores nulos restantes en cada columna para asegurar que el DataFrame esté listo para el análisis.

```
valores_nulos = final_df.select([((count(when(isnan(c) | col(c).isNull(), c))) ).alias(c) for c in final_df.columns)).show()
```

## Exploración de los datos:

Comencemos analizando la distribución de los puntajes del Saber 11 en nuestro conjunto de datos. Graficaremos la distribución de las columnas de puntuación global, puntuación en ciencias sociales, puntuación en ciencias naturales, puntuación en lectura crítica, puntuación en matemáticas y puntuación en inglés.



### Distribución de la Puntuación Global

La primera gráfica muestra la distribución de la puntuación global de los estudiantes. Observamos que:

**Promedio de Puntajes:** La puntuación global promedio de los estudiantes se encuentra entre 230 y 260 puntos.

**Rango de Puntajes:** Los puntajes más bajos oscilan alrededor de 100 puntos, mientras que los más altos están entre 380 y 400 puntos.

**Interpretación:** Esto sugiere que, en promedio, los estudiantes de calendario A logran alcanzar aproximadamente la mitad de los puntos totales posibles en el examen.

### **Distribución de la Puntuación en Ciencias Naturales**

En la gráfica de ciencias naturales, notamos que:

**Mediana de Puntajes:** La mayoría de los estudiantes obtienen una puntuación alrededor de 50 puntos.

**Interpretación:** Esto indica que la mayoría de los estudiantes responden correctamente aproximadamente la mitad de las preguntas en esta área.

### **Distribución de la Puntuación en Inglés**

En la gráfica de inglés, observamos una tendencia similar a la de ciencias naturales:

**Mediana de Puntajes:** La mediana de los puntajes se encuentra alrededor de 50 puntos.

**Interpretación:** Los estudiantes logran responder correctamente aproximadamente la mitad de las preguntas en la sección de inglés.

### **Distribución de la Puntuación en Ciencias Sociales y Ciudadanas**

La distribución de los puntajes en ciencias sociales y ciudadanas muestra que:

**Mediana de Puntajes:** La mayoría de los estudiantes tienen una puntuación cercana a 50 puntos.

**Interpretación:** Al igual que en ciencias naturales e inglés, los estudiantes tienden a responder correctamente la mitad de las preguntas en esta sección.

### **Distribución de la Puntuación en Lectura Crítica**

La gráfica de lectura crítica revela un rendimiento ligeramente mejor:

**Promedio de Puntajes:** La puntuación promedio se sitúa alrededor de 60 puntos.

**Interpretación:** Esto sugiere que los estudiantes están mejor preparados en lectura crítica en comparación con otras áreas.



## **Distribución de la Puntuación en Matemáticas**

La distribución de los puntajes en matemáticas también muestra un rendimiento mejorado:

**Promedio de Puntajes:** La puntuación promedio se encuentra alrededor de 60 puntos.

**Interpretación:** Los estudiantes tienden a estar más preparados en matemáticas, logrando puntajes más altos en comparación con otras áreas.

## **Observaciones Generales**

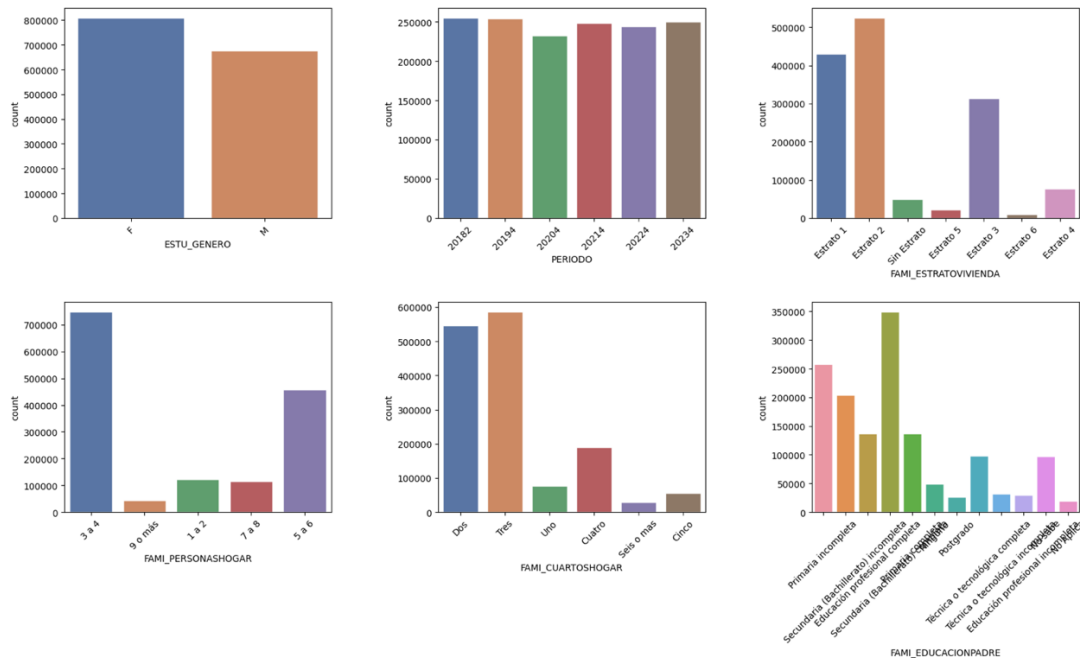
**Casos Excepcionales:** Se destacan algunos casos excepcionales donde unos pocos estudiantes obtienen puntajes perfectos de 100 en algunas de las secciones mencionadas.

**Variabilidad:** La variabilidad de los puntajes en cada área sugiere diferencias en el nivel de preparación de los estudiantes en las distintas áreas del conocimiento.

El análisis de la distribución de los puntajes del Saber 11 nos permite identificar patrones en el rendimiento de los estudiantes. Mientras que el promedio de la puntuación global indica que los estudiantes logran aproximadamente la mitad de los puntos posibles, áreas como lectura crítica y matemáticas muestran un mejor rendimiento relativo. Estas observaciones son útiles para orientar esfuerzos educativos y mejorar la preparación de los estudiantes en las áreas que presentan mayores desafíos.

## Análisis de Variables Categóricas

Después de un análisis exhaustivo de las variables numéricas, procedemos a explorar las variables categóricas utilizando diagramas de barras. Esta visualización nos permite entender mejor las características demográficas y socioeconómicas de los evaluados en el conjunto de datos del examen Saber 11.



### Distribución por Género

La primera gráfica muestra la distribución de los evaluados por género. Observamos una tendencia interesante:

- **Predominancia de Mujeres:** Hay una mayor cantidad de mujeres entre los evaluados. Este hallazgo puede suscitar preguntas sobre posibles disparidades de género en el acceso a la educación o diferencias en el desempeño académico entre hombres y mujeres.

### Distribución por Periodo de Presentación del Examen

La siguiente gráfica analiza la distribución de los estudiantes según el periodo en que presentaron el examen:

- **Constancia en la Mayoría de los Periodos:** En la mayoría de los periodos, se mantiene un número constante de evaluados.

- **Disminución en el Año 2020:** Observamos una disminución notable en 2020, posiblemente debido a los efectos disruptivos de la pandemia de COVID-19 en la educación. Este fenómeno sugiere la necesidad de investigar cómo eventos externos pueden influir en la participación en los exámenes estandarizados.

## Distribución por Estrato Socioeconómico

La gráfica de distribución de los diferentes estratos socioeconómicos muestra:

- **Predominancia de Estratos 1 y 2:** La mayoría de los evaluados pertenecen a los estratos 2 y 1, seguidos por el estrato 3.
- **Diversidad Socioeconómica:** Esta distribución indica la diversidad de los participantes en el examen Saber 11 y resalta la importancia de considerar el contexto socioeconómico al interpretar los resultados del examen.

## Distribución del Número de Personas en el Hogar

Al explorar el número de personas que viven en el hogar del evaluado, encontramos que:

- **Mayoría con 3 a 4 Personas:** La mayoría de los evaluados reportan vivir con 3 a 4 personas, reflejando dinámicas familiares comunes en la población estudiantil. Esto puede influir en el acceso a recursos educativos y el apoyo familiar.

## Distribución del Número de Habitaciones en el Hogar

La gráfica sobre el número de habitaciones en el hogar del evaluado revela:

- **Mayoría con 2 a 3 Habitaciones:** La mayoría de los evaluados tienen entre 2 y 3 habitaciones en su hogar, proporcionando una visión sobre las condiciones de vivienda de los estudiantes y su posible impacto en el entorno de estudio.

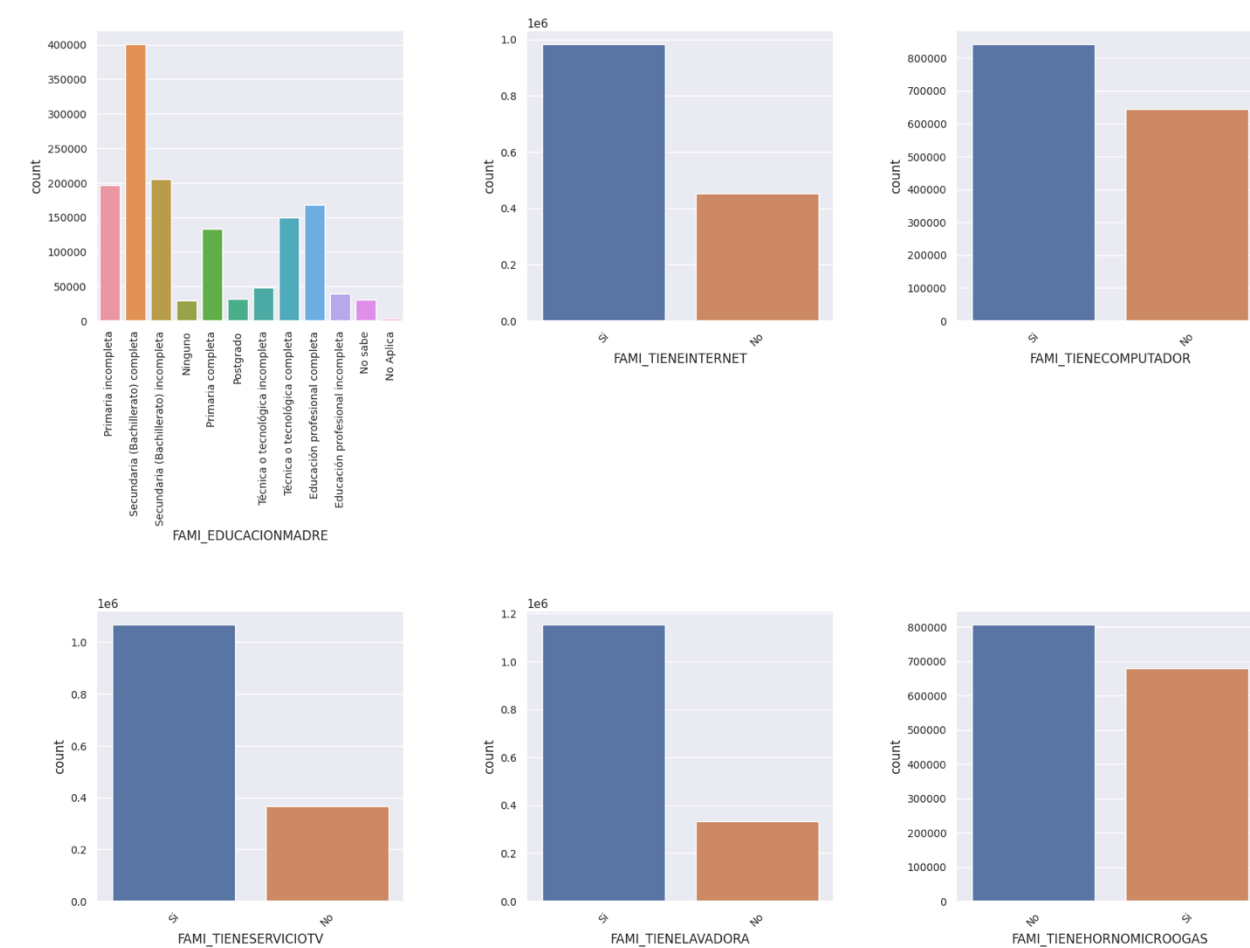
## Nivel Educativo del Padre

Finalmente, la gráfica que muestra el nivel educativo del padre de los evaluados indica:

- **Mayoría con Educación de Bachillerato o Primaria:** La mayoría de los padres tienen educación de bachillerato o primaria, lo que sugiere una diversidad en los antecedentes educativos de los padres. Esto plantea interrogantes sobre cómo el nivel educativo de los padres puede influir en el rendimiento académico de los estudiantes.

Desde la predominancia de mujeres hasta la diversidad en el nivel educativo de los padres, cada variable ofrece insights valiosos para comprender mejor el contexto en el que los estudiantes rinden el examen. Estos hallazgos son cruciales para interpretar correctamente los resultados y orientar políticas educativas que aborden las necesidades y desafíos específicos de los estudiantes que debemos seguir evaluando, con eso podemos ver el siguiente conjunto de gráficas.

Continuando con la exploración de las variables categóricas en el conjunto de datos del examen Saber 11, se presentan a continuación varias gráficas adicionales que nos ayudan a comprender mejor las características demográficas y socioeconómicas de los evaluados.



## Educación de la Madre

La primera gráfica muestra la distribución del nivel educativo de las madres de los evaluados:

- **Diversidad en la Educación:** Observamos una diversidad en los niveles educativos, con una predominancia de madres que han completado la educación secundaria (bachillerato) y primaria. Esto sugiere que las madres de los evaluados tienen diversos antecedentes educativos, lo cual podría influir en el rendimiento académico de los estudiantes.

## Acceso a Internet

La gráfica de acceso a Internet en los hogares muestra:

- **Mayoría con Acceso a Internet:** La mayoría de los estudiantes reportan tener acceso a Internet en sus hogares. Este acceso es crucial para las actividades educativas y el aprendizaje en línea, especialmente en tiempos de educación remota.

## Posesión de Computadoras

En cuanto a la posesión de computadoras:

- **Acceso Moderado a Computadoras:** Un número considerable de familias posee computadoras, aunque hay una proporción significativa que no tiene acceso a ellas. Esto podría afectar la capacidad de los estudiantes para realizar tareas y acceder a recursos educativos en línea.

## Acceso a Servicios de TV

La siguiente gráfica analiza la posesión de servicios de televisión:

- **Mayoría con Servicios de TV:** La mayoría de las familias reportan tener acceso a servicios de televisión. Este dato refleja la penetración de los medios de comunicación en los hogares y su posible influencia en el tiempo de ocio de los estudiantes.

## Posesión de Lavadoras

La gráfica sobre la posesión de lavadoras muestra:

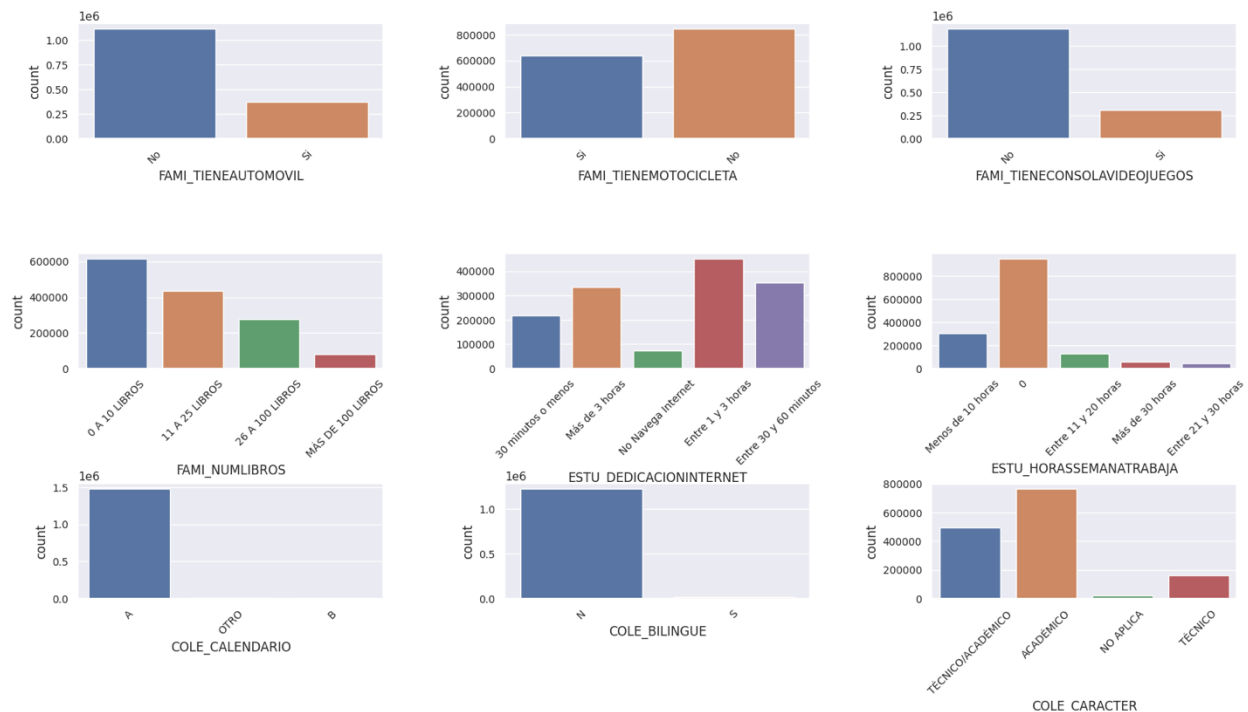
- **Mayoría con Lavadoras:** Un gran número de familias posee lavadoras, lo cual indica un cierto nivel de acceso a electrodomésticos que facilitan las tareas domésticas.

## Posesión de Hornos Microondas

Finalmente, la gráfica sobre la posesión de hornos microondas revela:

- **Acceso Moderado a Hornos Microondas:** Similar a la posesión de computadoras, una proporción considerable de familias tiene hornos microondas, aunque muchas también carecen de este electrodoméstico.

Continuando con los diagramas de barra, este otro nos permite medir un poco mejor la situación económica y la calidad del entretenimiento de los familiares de los estudiantes:



## Posesión de Automóviles y Motocicletas

Al analizar la posesión de vehículos:

- **Pocas Familias con Automóviles:** La mayoría de las familias no cuentan con un automóvil.
- **Mayor Posesión de Motocicletas:** Hay una proporción mayor de familias que poseen motocicletas en comparación con aquellas que tienen automóviles. Esto sugiere que las motocicletas son una forma de transporte más común en esta población, posiblemente debido a su menor costo y mayor accesibilidad.

## Posesión de Consolas de Videojuegos

En cuanto a las consolas de videojuegos:

- **Baja Posesión de Consolas:** La mayoría de las familias no cuentan con una consola de videojuegos. Esto podría deberse a que las consolas no se consideran una necesidad y pueden estar fuera del alcance económico de muchas familias.

## Cantidad de Libros en el Hogar

La gráfica de la cantidad de libros en el hogar muestra:

- **Pocos Libros en el Hogar:** La mayoría de los estudiantes reporta tener entre 0 a 10 libros. A medida que aumenta el número de libros, se observa una disminución en el número de estudiantes con acceso a ellos. Esto indica que muchas familias tienen recursos limitados para adquirir libros, lo cual podría afectar el hábito de lectura y el acceso a materiales de aprendizaje en el hogar.

## Tiempo Dedicado a Internet

En cuanto al tiempo que los estudiantes dedican a navegar por Internet:

- **Diversidad en el Uso de Internet:** La mayoría de los estudiantes dedican entre 1 a 3 horas diarias a Internet, seguidos por aquellos que navegan más de 3 horas diarias. Esta distribución refleja la diversidad en el uso de Internet entre los estudiantes, influenciado por factores como la necesidad de estudiar, el acceso a recursos y las actividades recreativas.

## Situación Laboral de los Estudiantes

La gráfica sobre la situación laboral de los estudiantes muestra:

- **Mayoría No Trabaja:** La mayoría de los estudiantes reporta que no trabaja, lo cual es positivo ya que pueden dedicar más tiempo a sus estudios. Este dato es relevante para entender el contexto en el que los estudiantes se preparan para el examen Saber 11 y cómo la carga laboral puede influir en su rendimiento académico.
- 

## Tipo de Colegio

Finalmente, analizamos las características de los colegios:

- **Predominancia de Colegios de Calendario A:** La mayoría de los colegios son de calendario A, no son bilingües y se clasifican como colegios meramente académicos o técnico-académicos. Este dato proporciona una visión general del tipo de instituciones

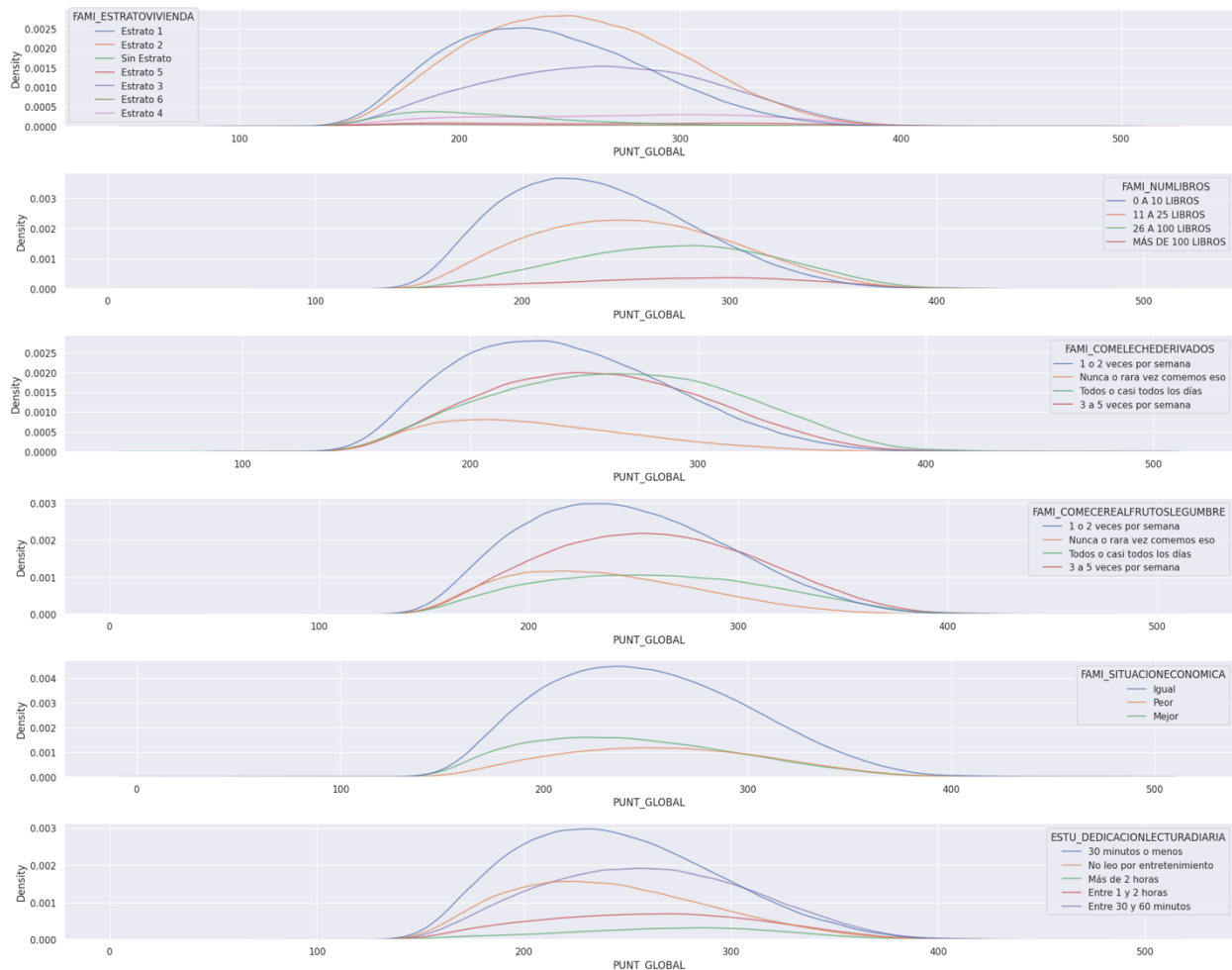
educativas a las que asisten los evaluados y puede ayudar a contextualizar los resultados del examen Saber 11.

El análisis de estas variables categóricas adicionales proporciona una comprensión más profunda de las características demográficas y socioeconómicas de los evaluados en el examen Saber 11. Estos insights son esenciales para interpretar correctamente los resultados del examen y para diseñar intervenciones educativas que aborden las necesidades específicas de los estudiantes.



# Análisis de Posibles Causas que Afectan los Puntajes Globales usando Diagramas de Densidad

En esta sección, se graficaron las posibles causas que pueden afectar una puntuación global satisfactoria en el examen Saber 11. Utilizando gráficas de densidad, se analizaron diversas características y su relación con los puntajes globales. Las gráficas de densidad son ventajosas porque permiten visualizar la distribución de una variable continua y comparar fácilmente diferentes grupos en un mismo gráfico, facilitando la identificación de tendencias y patrones. A continuación, se describen las observaciones de cada una de estas características:



## **Estrato Socioeconómico de la Vivienda (FAMI\_ESTRATOVIVIENDA)**

La gráfica de densidad muestra la distribución de los puntajes globales según el estrato socioeconómico de la vivienda:

- **Observaciones:** Los estudiantes de estratos más altos tienden a obtener mejores puntajes hasta el estrato 3 o 4.
- **Interpretación:** Este resultado sugiere que el contexto socioeconómico puede tener un impacto significativo en el rendimiento académico, probablemente debido a un mejor acceso a recursos educativos y un entorno más favorable para el estudio.

## **Número de Libros en el Hogar (FAMI\_NUMLIBROS)**

La relación entre la cantidad de libros en el hogar y los puntajes globales se grafica de la siguiente manera:

- **Observaciones:** Los estudiantes que reportan tener más libros en casa tienden a obtener puntajes más altos.
- **Interpretación:** Este hallazgo indica que el acceso a material de lectura puede ser un factor clave en el desarrollo de habilidades críticas y en el rendimiento académico general.

## **Consumo de Leche y Derivados (FAMI\_COMELECHEDERIVADOS)**

Se grafica la distribución de los puntajes globales en función del consumo de leche y derivados:

- **Observaciones:** Las distribuciones de los puntajes son bastante similares independientemente de la cantidad de consumo de estos productos.
- **Interpretación:** Esta categoría no parece tener un impacto significativo en los puntajes globales.

## **Consumo de Cereales, Frutos y Legumbres (FAMI\_COMECEREALFRUTOSLEGUMBRE)**

La gráfica correspondiente al consumo de cereales, frutos y legumbres muestra lo siguiente:

- **Observaciones:** Las distribuciones de los puntajes globales son parecidas entre los diferentes niveles de consumo.

- **Interpretación:** Esta categoría tampoco afecta en gran medida los puntajes.

### **Situación Económica Percibida (FAMI\_SITUACIONECONOMICA)**

Al analizar la percepción de la situación económica de las familias, se observa lo siguiente:

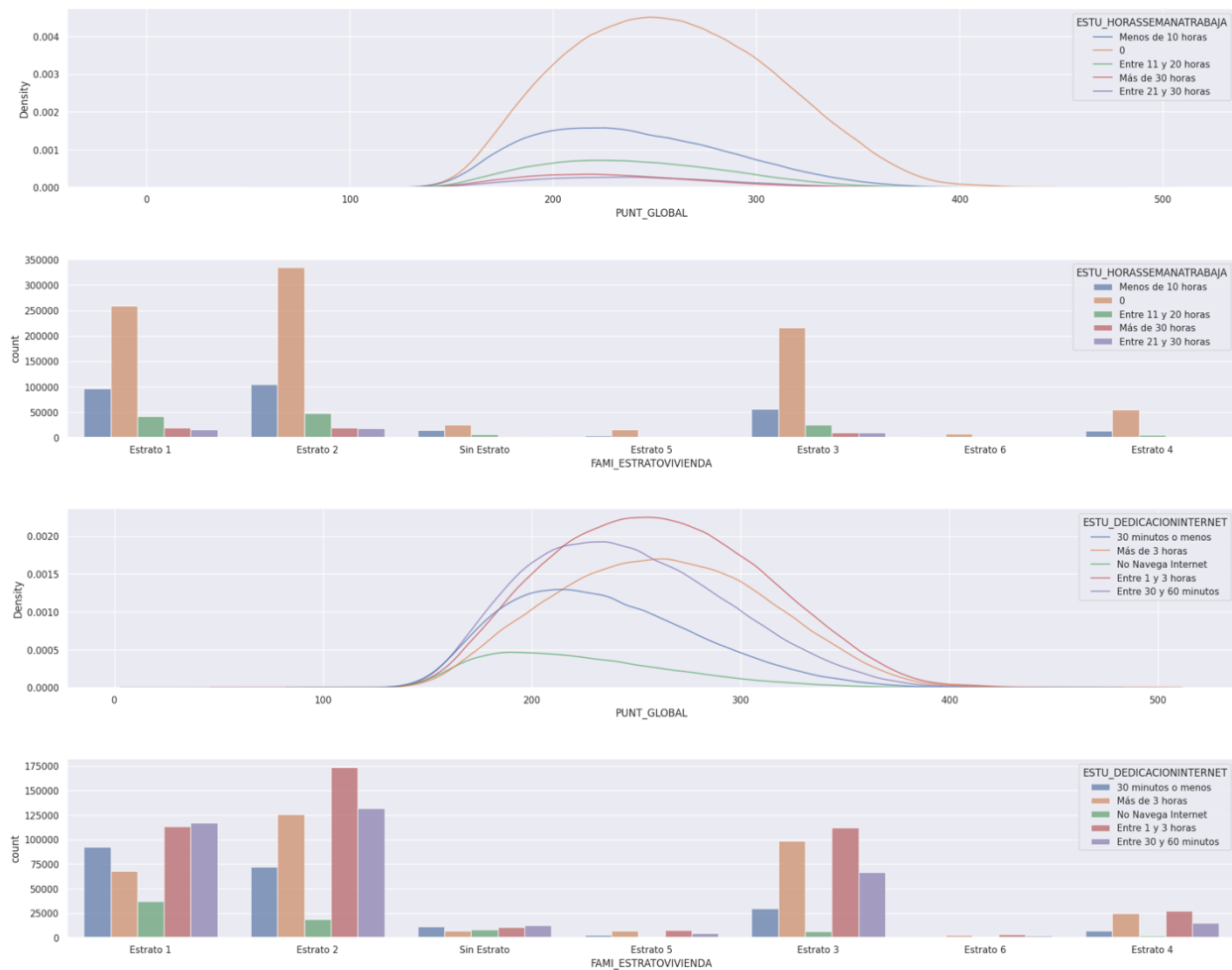
- **Observaciones:** Las distribuciones de los puntajes globales son similares independientemente de cómo los estudiantes perciben su situación económica.
- **Interpretación:** La percepción de la situación económica por sí sola no tiene un impacto significativo en los puntajes globales.

### **Dedicación a la Lectura Diaria (ESTU\_DEDICACIONLECTURADIARIA)**

La relación entre la dedicación a la lectura diaria y los puntajes globales se grafica de la siguiente manera:

- **Observaciones:** Los estudiantes que dedican más tiempo a la lectura diaria tienden a obtener mejores puntajes.
- **Interpretación:** Este hallazgo subraya la importancia de la lectura regular en el desarrollo de habilidades críticas y en el éxito académico.

En esta sección, se analiza la relación entre la dedicación al uso de Internet y las horas de trabajo semanal con los puntajes globales en el examen Saber 11. Utilizando gráficas de densidad y de barras, se investigan cómo estas variables pueden influir en el rendimiento académico de los estudiantes.



## Dedicación al Uso de Internet (ESTU\_DEDICACIONINTERNET)

La gráfica de densidad muestra la relación entre la dedicación al uso de Internet y los puntajes globales:

- **Distribución Variada:** La dedicación al uso de Internet muestra una distribución variada de puntajes.

- **Observaciones:** Aunque algunos estudiantes que pasan más tiempo en Internet obtienen buenos puntajes, también hay casos donde un uso excesivo se relaciona con puntajes más bajos.
- **Interpretación:** Esto sugiere que el tipo de actividades realizadas en Internet (educativas vs. recreativas) puede influir en el rendimiento académico. Un uso equilibrado y enfocado en actividades educativas podría ser beneficioso, mientras que un uso excesivo para entretenimiento podría ser perjudicial.

### **Horas de Trabajo Semanal (ESTU\_HORASSEMANATRABAJA)**

La gráfica de densidad muestra la relación entre las horas de trabajo semanal y los puntajes globales:

- **Distribución de Puntajes:** Los estudiantes que trabajan más horas a la semana tienden a tener puntajes más bajos.
- **Observaciones:** La gráfica indica que las responsabilidades laborales pueden afectar negativamente el tiempo y la energía disponibles para el estudio.
- **Interpretación:** El trabajo excesivo reduce el tiempo de estudio y descanso, impactando negativamente en el rendimiento académico. Los estudiantes con menos horas de trabajo semanal pueden dedicar más tiempo y energía a sus estudios, obteniendo mejores resultados.

### **Análisis de Estratos Socioeconómicos y Uso de Internet**

La gráfica de barras analiza cómo la dedicación al uso de Internet varía según los estratos socioeconómicos:

- **Distribución por Estrato:** La mayoría de los estudiantes que pasan más tiempo en Internet pertenecen a los estratos 1 y 2.
- **Observaciones:** La dedicación al uso de Internet varía según el estrato socioeconómico, reflejando el acceso a la tecnología y los recursos disponibles en los hogares.
- **Interpretación:** Los estudiantes de estratos socioeconómicos más bajos pueden tener acceso limitado a recursos educativos en línea, afectando su rendimiento académico.

## Análisis de Estratos Socioeconómicos y Horas de Trabajo

La gráfica de barras muestra la relación entre las horas de trabajo semanal y los estratos socioeconómicos:

- **Distribución por Estrato:** La mayoría de los estudiantes que trabajan más horas pertenecen a los estratos 1 y 2.
- **Observaciones:** Los estudiantes de estratos más bajos tienden a trabajar más horas, lo que podría ser una necesidad económica.
- **Interpretación:** La necesidad de trabajar para apoyar a sus familias puede limitar el tiempo que los estudiantes de estratos socioeconómicos más bajos tienen para estudiar, impactando negativamente en sus puntajes globales.

El análisis de la dedicación al uso de Internet y las horas de trabajo semanal revela insights importantes sobre cómo estas variables afectan los puntajes globales del examen Saber 11. Mientras que el uso equilibrado de Internet puede ser beneficioso, el uso excesivo para entretenimiento puede ser perjudicial. Además, las responsabilidades laborales excesivas pueden reducir significativamente el tiempo y la energía disponibles para el estudio, impactando negativamente en el rendimiento académico. Estos hallazgos subrayan la necesidad de políticas y programas que apoyen a los estudiantes en la gestión equilibrada de su tiempo de estudio y trabajo, así como el acceso a recursos educativos en línea.

Para entender mejor cómo la pandemia pudo haber afectado los puntajes globales del examen Saber 11, se graficaron los puntajes globales separados por los diferentes periodos de presentación. El objetivo era observar cómo la media de los puntajes cambió a lo largo de los años, especialmente durante la pandemia, y comparar esos cambios con los periodos pre y post-pandemia.

Sorprendentemente, se observó que la media de los puntajes globales aumentó durante la época de la pandemia. A continuación se muestra una tabla con los promedios de los puntajes globales para cada periodo:

2018	250.731
2019	246.146
2020	248.262
2021	245.999
2022	250.087
<b>2023</b>	<b>252.505</b>

**Periodo 2018-2 y 2019-4:**

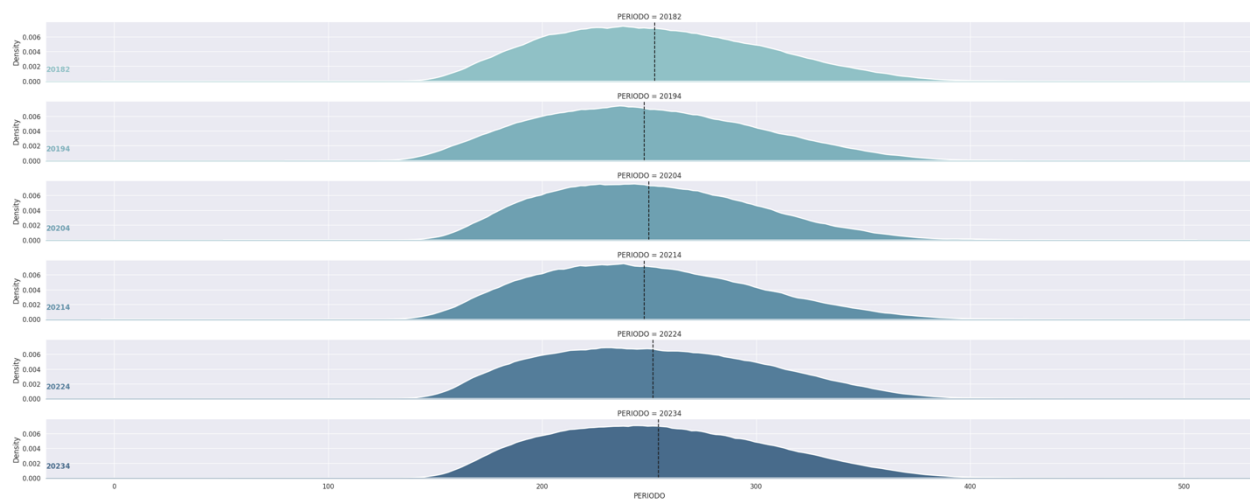
Antes de la pandemia, en el segundo semestre de 2018 y el cuarto semestre de 2019, los puntajes promedio fueron 250.73 y 246.15, respectivamente. Estos valores reflejan una base estable de puntajes que puede ser utilizada como referencia para evaluar el impacto de la pandemia.

#### **Periodo 2020-4 y 2021-4:**

Durante la pandemia, en el cuarto semestre de 2020 y el cuarto semestre de 2021, los puntajes promedios fueron 248.26 y 246.00, respectivamente. Es interesante notar que, aunque se esperaba una caída significativa en los puntajes debido a las interrupciones educativas causadas por la pandemia, los promedios no disminuyeron drásticamente. De hecho, en 2020-4, hubo un ligero aumento en comparación con 2019-4.

#### **Periodo 2022-4 y 2023-4:**

Después de la pandemia, en el cuarto semestre de 2022 y el cuarto semestre de 2023, los puntajes promedio fueron 250.09 y 252.51, respectivamente. Estos promedios muestran una recuperación y una tendencia al alza en los puntajes globales. El aumento en 2023-4 hasta 252.51 sugiere una posible adaptación y mejora en las estrategias educativas post-pandemia.



# Problema

## Descripción del Problema de Analítica a Resolver

El problema de analítica que se quiere resolver es el de predecir la puntuación global del examen Saber 11 de un estudiante basándose en factores socioeconómicos presentes en el dataset. Esto se enmarca dentro de un problema de regresión, donde el objetivo es predecir un valor numérico continuo (la puntuación global) utilizando diversas características del estudiante y su entorno.

## Técnicas de Analítica Propuestas

Para abordar este problema, se han seleccionado dos técnicas de machine learning: Random Forest Regressor y Gradient Boosting Regressor. Ambas técnicas pertenecen a la familia de modelos de árboles de decisión y son adecuadas para manejar tanto relaciones no lineales como interacciones complejas entre las variables.

### Random Forest Regressor

El Random Forest Regressor es un método de ensemble learning que construye múltiples árboles de decisión durante el entrenamiento y promedia sus resultados para obtener la predicción final. Las ventajas de este método incluyen:

**Robustez a overfitting:** Al promediar los resultados de varios árboles, el modelo reduce el riesgo de sobreajuste.

**Manejo de la variabilidad:** Random Forest maneja bien la variabilidad en los datos y es menos sensible a los cambios en los datos de entrenamiento.

**Importancia de características:** El modelo puede proporcionar una medida de la importancia de cada característica en la predicción.

### Gradient Boosting Regressor

El Gradient Boosting Regressor también es un método de ensemble learning, pero en lugar de construir árboles independientes, construye árboles secuencialmente, donde cada árbol intenta corregir los errores del anterior. Las ventajas de este método incluyen:

**Alta precisión:** Gradient Boosting puede proporcionar predicciones muy precisas ajustando el modelo de manera más precisa a los datos de entrenamiento.

**Flexibilidad:** El modelo puede ajustarse mediante varios hiperparámetros, permitiendo un control fino sobre el proceso de aprendizaje.



**Importancia de características:** Al igual que el Random Forest, este modelo también puede proporcionar una medida de la importancia de cada característica.

## **Problemas de Calidad de Datos:**

**Datos faltantes:** Afortunadamente, en el dataset había muy pocos valores faltantes. Esto se debe a que el dataset estuvo bien mantenido y los datos de presentación se recopilan mediante una encuesta con valores predefinidos que deben ser completados. Como resultado, no se encontraron errores significativos ni datos inconsistentes. Para los pocos valores faltantes, se decidió eliminarlos dado que su cantidad era insignificante en comparación con el tamaño total del dataset.

**Datos fuera de rango o con errores:** No se encontraron datos fuera de rango o con errores debido a la naturaleza predefinida de las respuestas en la encuesta.

**Datos inconsistentes entre distintas fuentes:** No se identificaron inconsistencias gracias a la uniformidad en la recopilación de datos.

## **Normalización de Variables:**

Dado que todas las variables seleccionadas fueron variables categóricas, no fue necesario reescalar los valores. Sin embargo, se realizó una transformación adicional para convertir las variables categóricas en un formato numérico que pudiera ser utilizado por los modelos de inteligencia artificial. Para esto, se aplicó un Ordinal Encoder, que asigna a cada categoría un valor numérico único por ejemplo, si tenemos una variable categórica "Tamaño" con las categorías "Pequeño", "Mediano" y "Grande", el Ordinal Encoder podría asignar los valores 0, 1 y 2 respectivamente. Esta técnica es útil cuando hay un orden implícito en las categorías, es decir, cuando existe una relación de orden entre ellas, permitiendo así que los datos sean utilizados en los algoritmos de aprendizaje automático.

## **Creación de Nuevas Variables (Variables Derivadas):**

Dado que la gran mayoría de las variables eran categóricas y no se encontraron relaciones lineales entre ellas, no se identificaron oportunidades para crear nuevas variables derivadas. La creación de variables adicionales podría haber sido útil en el caso de variables numéricas para capturar relaciones más complejas o interacciones entre características, pero en este contexto, no fue necesario realizar este paso.

## Implementación de Técnicas ML y resultados

Primero, se entrenó el modelo de Random Forest utilizando las siguientes columnas como características:

```
feature_cols = ['FAMI_ESTRATOVIVIENDA', 'FAMI_PERSONASHOGAR', 'FAMI_CUARTOSHOGAR',  
                'FAMI_EDUCACIONPADRE', 'FAMI_EDUCACIONMADRE', 'FAMI_TRABAJOLABORPADRE',  
                'FAMI_TRABAJOLABORMADRE', 'FAMI_TIENEINTERNET', 'FAMI_TIENESERVICIOTV',  
                'FAMI_TIENECOMPUTADOR', 'FAMI_TIENELAVADORA', 'FAMI_TIENEAUTOMOVIL',  
                'FAMI_TIENEMOTOCICLETA', 'FAMI_TIENECONSOLAVIDEOJUEGOS', 'FAMI_NUMLIBROS',  
                'FAMI_COMELECHEDERIVADOS', 'FAMI_COMECARNEPESCADOHUEVO',  
                'FAMI_COMECEREALFRUTOSLEGUMBRE', 'FAMI_SITUACIONECONOMICA',  
                'ESTU_DEDICACIONLECTURADIA', 'ESTU_DEDICACIONINTERNET',  
                'ESTU_HORASSEMANATRAABA']
```

El primer paso para entrenar el modelo fue convertir estas columnas en números, como se discutió anteriormente. Estas columnas representan los factores socioeconómicos que forman parte del conjunto de características.

```
# Convert string columns to numerical representations  
indexers = [StringIndexer(inputCol=col, outputCol=col+"_index", handleInvalid="keep") for col in string_cols]  
pipeline = Pipeline(stages=indexers)  
indexed_data = pipeline.fit(data).transform(data)
```

La métrica utilizada para comprobar el rendimiento del modelo fue RMSE (Root Mean Squared Error). El RMSE es una medida de la diferencia entre los valores predichos por el modelo y los valores reales. Se calcula tomando la raíz cuadrada de la media de los cuadrados de todos los errores. Matemáticamente, el RMSE se define como:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_{ri})^2}$$

donde  $y_i$  son los valores reales y  $y_{ri}$  son los valores predichos por el modelo. El RMSE es útil porque proporciona una medida de la precisión del modelo, siendo más penalizador para grandes errores. Un RMSE más bajo indica un mejor rendimiento del modelo.

Una vez entrenado el modelo de Random Forest, se obtuvo un RMSE de 43.58157631482367. Esto significa que, en promedio, las predicciones del modelo están a aproximadamente 43.58 puntos de los valores reales de la puntuación global en la prueba Saber 11. Este valor proporciona una idea de cuán preciso es el modelo al predecir los resultados basándose en los factores socioeconómicos. Un RMSE de 43.58 indica que el modelo tiene un margen de error

moderado en sus predicciones, lo cual puede ser aceptable dependiendo del contexto y de los objetivos específicos del análisis.

Luego, se entrenó un Gradient Boosting Random Forest Regressor (GBRF). Aunque también es un modelo de árboles de decisión, Gradient Boosting difiere del Random Forest en la forma en que construye los árboles:

Random Forest: Construye múltiples árboles de decisión de manera independiente y agrega sus predicciones (promedio o voto mayoritario).

Gradient Boosting: Construye árboles secuencialmente, donde cada nuevo árbol intenta corregir los errores de los árboles anteriores. Este método se enfoca en los errores más grandes, ajustando el modelo de manera más precisa a los datos de entrenamiento.

El modelo de Gradient Boosting se entrenó con los mismos datos socioeconómicos que el Random Forest y se evaluó utilizando la misma métrica de RMSE. El GBRF obtuvo un RMSE de 42.685460384742704. Este valor, ligeramente menor que el del Random Forest, indica que el modelo de Gradient Boosting tiene una mejor precisión, ya que sus predicciones están, en promedio, a aproximadamente 42.69 puntos de los valores reales. Esto sugiere que el GBRF es un poco más eficiente en capturar las relaciones entre las variables socioeconómicas y la puntuación global del Saber 11.

Modelo	Descripción	RMSE
Random Forest Regressor	Construye múltiples árboles de decisión de manera independiente y agrega sus predicciones (promedio o voto mayoritario). Este enfoque reduce el riesgo de sobreajuste y mejora la precisión general del modelo al promediar los resultados de varios árboles.	43.58157631482367
Gradient Boosting Random Forest Regressor	Construye árboles de decisión secuencialmente, donde cada nuevo árbol intenta corregir los errores de los árboles anteriores. Este método enfoca en los errores más grandes y ajusta el modelo de manera más precisa a los datos de entrenamiento.	42.685460384742704

Aunque ambos modelos presentan un buen rendimiento, el Gradient Boosting Random Forest Regressor muestra una mejor precisión con un RMSE más bajo. Esto sugiere que es más eficaz en capturar las relaciones complejas entre las variables socioeconómicas y la puntuación global

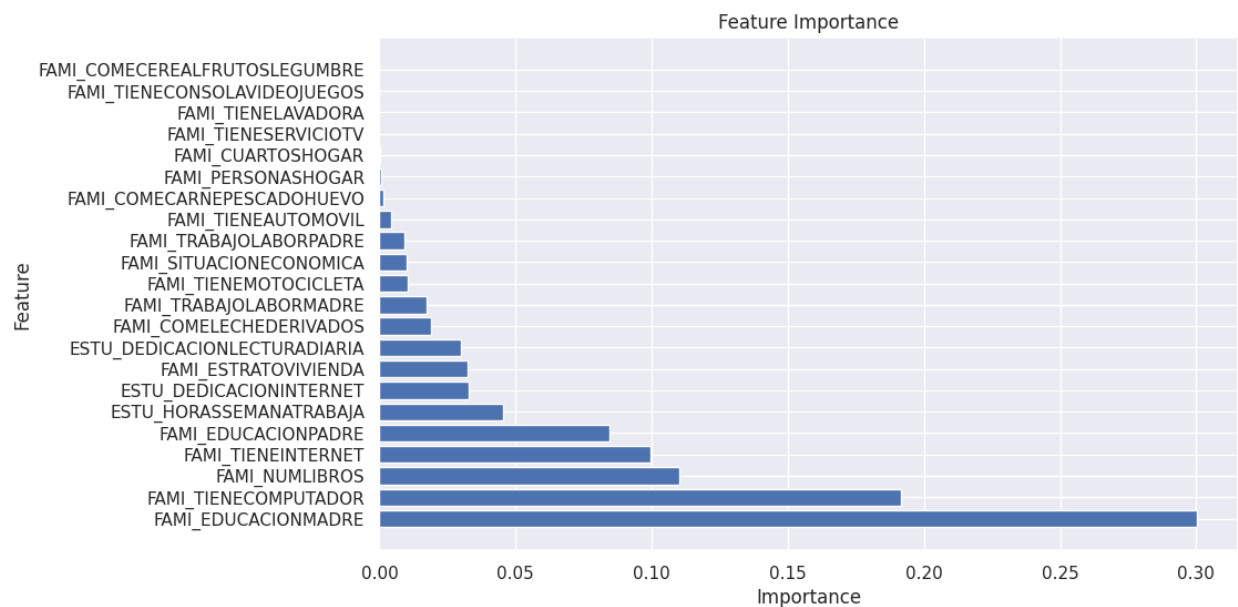
del Saber 11. Sin embargo, la diferencia en RMSE no es muy grande, por lo que ambos modelos pueden considerarse efectivos para este tipo de predicción.

El modelo de Random Forest Regressor tiene la capacidad de identificar las características más importantes que influyen en la predicción. Para el modelo entrenado, las siguientes variables fueron identificadas como las más importantes:

**Educación de la Madre (FAMI\_EDUCACIONMADRE)**

**Posesión de Computador (FAMI\_TIENECOMPUTADOR)**

**Número de Libros en el Hogar (FAMI\_NUMLIBROS)**



## Conclusión

La identificación de las características más importantes nos proporciona una visión valiosa sobre los factores que más influyen en el rendimiento de los estudiantes en la prueba Saber 11:

**Educación de la Madre:** Esta variable resultó ser la más importante. Esto sugiere que la educación de la madre tiene un impacto significativo en el rendimiento académico de los estudiantes. Es posible que las madres con mayor nivel educativo proporcionen un entorno más propicio para el estudio, ayudando con las tareas y fomentando hábitos de estudio efectivos.

**Posesión de Computador:** La segunda variable más importante indica que el acceso a un computador es crucial para el rendimiento académico. En la era digital, el acceso a un computador facilita la investigación, el aprendizaje en línea y la realización de tareas escolares, especialmente durante la pandemia cuando la educación presencial se vio interrumpida.

**Número de Libros en el Hogar:** El número de libros en el hogar también es una variable significativa. La disponibilidad de libros refleja un entorno que valora la lectura y el aprendizaje, lo cual puede estimular el desarrollo cognitivo y el rendimiento académico de los estudiantes.

### **Implicaciones**

Estos hallazgos tienen importantes implicaciones para las políticas educativas y las intervenciones dirigidas a mejorar el rendimiento académico:

- **Apoyo a la Educación Parental:** Invertir en la educación de los padres, especialmente de las madres, puede tener efectos indirectos positivos en el rendimiento de los estudiantes.
- **Acceso a Tecnología:** Garantizar que todos los estudiantes tengan acceso a computadoras puede ser crucial para nivelar el campo de juego educativo, especialmente en tiempos de educación remota.
- **Fomento de la Lectura:** Promover la lectura en el hogar y aumentar el acceso a libros puede mejorar significativamente el rendimiento académico.

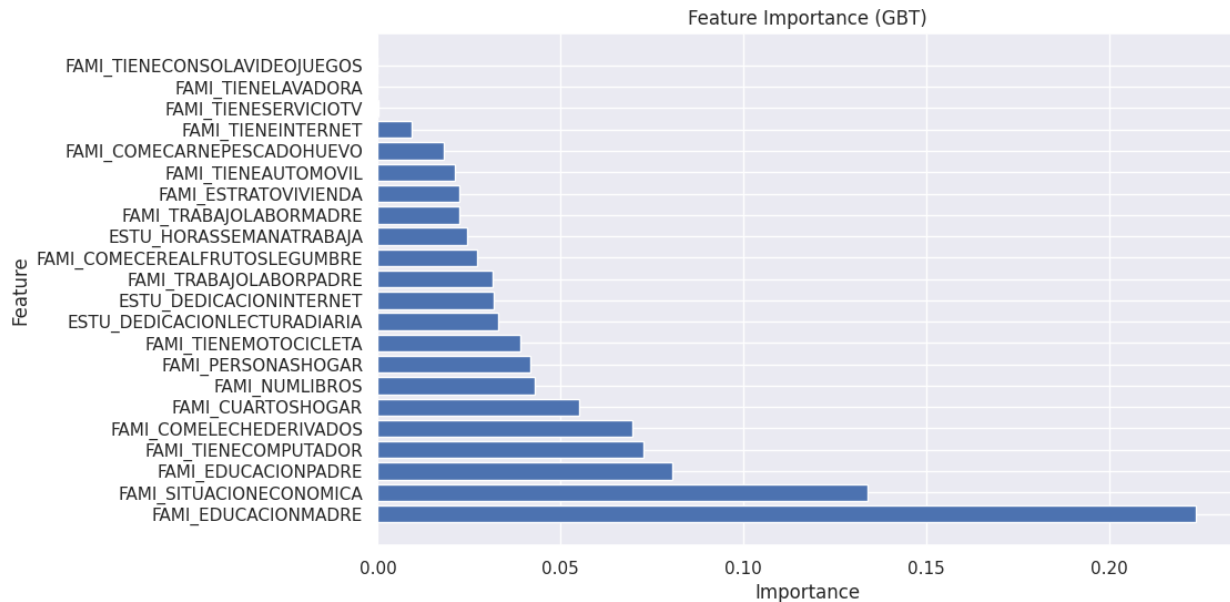
### **Importancia de las Características en el Modelo de Gradient Boosting Regressor**

El modelo de Gradient Boosting Regressor también tiene la capacidad de identificar las características más importantes que influyen en la predicción. Para el modelo entrenado, las siguientes variables fueron identificadas como las más importantes:

**Educación de la Madre (FAMI\_EDUCACIONMADRE)**

**Situación Socioeconómica Percibida (FAMI\_SITUACIONECONOMICA)**

**Educación del Padre (FAMI\_EDUCACIONPADRE)**



## Conclusión

La identificación de las características más importantes por el modelo de Gradient Boosting Regressor proporciona una visión complementaria sobre los factores que influyen en el rendimiento de los estudiantes en la prueba Saber 11:

**Educación de la Madre:** Similar al modelo de Random Forest, la educación de la madre sigue siendo la característica más importante. Esto refuerza la idea de que el nivel educativo de la madre tiene un impacto significativo en el rendimiento académico de los estudiantes, probablemente debido a su papel crucial en el apoyo educativo y la creación de un entorno de aprendizaje positivo en el hogar.

**Situación Socioeconómica Percibida:** La percepción de la situación socioeconómica del hogar es la segunda característica más importante. Esto sugiere que los estudiantes que perciben una situación económica más estable pueden tener menos estrés y más recursos para dedicar al estudio, lo cual mejora su rendimiento académico.

**Educación del Padre:** La educación del padre también es una característica importante, indicando que ambos padres juegan un papel significativo en la educación del estudiante. Un mayor nivel educativo de los padres generalmente implica más apoyo académico y un ambiente que valora la educación.

## Implicaciones

Estos hallazgos tienen importantes implicaciones para las políticas educativas y las intervenciones dirigidas a mejorar el rendimiento académico:

- **Educación de los Padres:** Invertir en la educación de los padres puede tener un efecto multiplicador en el rendimiento académico de los hijos. Programas de educación para adultos y apoyo educativo pueden ser beneficiosos.
- **Mejorar la Situación Socioeconómica:** Políticas y programas que mejoren la situación socioeconómica de las familias, como subsidios, becas y apoyo financiero, pueden reducir las barreras económicas y permitir a los estudiantes centrarse en sus estudios.
- **Ambiente Educativo en el Hogar:** Promover un ambiente de aprendizaje en el hogar, con acceso a recursos educativos y apoyo parental, puede mejorar significativamente el rendimiento académico de los estudiantes.

## **Respuestas a las Preguntas Planteadas**

### **1. ¿Cómo se comparan los resultados promedio de la prueba Saber 11 antes, durante y después de la pandemia?**

#### **Análisis de Resultados Promedio:**

Para responder a esta pregunta, analizamos los datos de los puntajes globales de la prueba Saber 11 para los periodos antes, durante y después de la pandemia. Se observó lo siguiente:

- **Antes de la pandemia (2018-2019):** Los puntajes promedio eran relativamente estables, con una media alrededor de 250 puntos.
- **Durante la pandemia (2020-2021):** Hubo una ligera variación en los puntajes promedio. Sorprendentemente, en 2020, los puntajes promedio aumentaron, alcanzando 248.26 puntos en 2020 y 245.99 puntos en 2021. Esto puede estar relacionado con cambios en las condiciones de evaluación o en la preparación de los estudiantes debido al confinamiento.
- **Después de la pandemia (2022-2023):** Los puntajes promedio volvieron a niveles similares a los pre-pandemia, con una media de 250.08 puntos en 2022 y 252.51 puntos en 2023.

#### **Conclusión:**

La pandemia no causó una disminución significativa en los puntajes promedio de la prueba Saber 11. De hecho, hubo un aumento leve durante el periodo de pandemia, posiblemente debido a las condiciones de estudio en casa y el cambio en las dinámicas de evaluación.

### **2. ¿Cuál es el efecto de los factores socioeconómicos, como el nivel de ingresos familiar o el tipo de colegio, en los resultados de la prueba Saber 11?**

#### **Análisis de Factores Socioeconómicos:**

Usamos modelos de regresión para identificar las características más influyentes en el puntaje global de la prueba Saber 11. Los modelos de Random Forest y Gradient Boosting destacaron varias variables importantes:

- **Educación de los Padres (Madre y Padre):** Tanto la educación de la madre como la del padre son factores significativos. Estudiantes con padres de mayor nivel educativo tienden a obtener mejores resultados.
- **Posesión de Computador:** Los estudiantes que tienen acceso a un computador en casa también tienden a obtener mejores puntajes.
- **Situación Socioeconómica Percibida:** Los estudiantes que perciben una situación socioeconómica estable tienen mejores rendimientos académicos.

### **Conclusión:**

Los factores socioeconómicos juegan un papel crucial en el rendimiento académico. La educación de los padres, el acceso a tecnología y la percepción de estabilidad económica son determinantes clave en los puntajes de la prueba Saber 11.