

Proyecto - Procesamiento de datos a gran escala.

Parcial 2 - Proyecto de Aplicación

Procesamiento de Datos a Gran Escala

Santiago Avilés Tibocha

Santiago Botero Pacheco



## Tabla de contenido

<b>1) FASE INICIAL</b>	<b>3</b>
a) Selección de los conjuntos de datos y análisis de su contexto	3
b) Marco teórico	3
c) Preguntas planteadas para el análisis	4
d) Exploración de los datos	4
Dataset Saber PRO	4
Dataset Cobertura Móvil	10
e) Preparación de los datos	15
<b>2) PROBLEMA</b>	<b>17</b>
a) Descripción del problema de analítica a resolver	17
b) Solución a los problemas típicos en datos	17
Problemas de calidad de datos:	17
Normalización de variables	18
Creación de nuevas variables	18
Fusión de los conjuntos de datos	19
c) Analítica descriptiva de los datos	19
<b>3) IMPLEMENTACIÓN DE TÉCNICAS ML Y RESULTADOS</b>	<b>20</b>
a) Entrenamiento y comparación → Resultados finales obtenidos por los modelos	21
b) Respuestas a las preguntas planteadas	24
<b>4) CONCLUSIONES Y FUTURAS LÍNEAS DE TRABAJO</b>	<b>25</b>
<b>5) CONTRIBUCIONES</b>	<b>26</b>
<b>6) REFERENCIAS</b>	<b>26</b>

## 1) FASE INICIAL

### a) Selección de los conjuntos de datos y análisis de su contexto

En un principio, nuestro grupo tuvo dificultades para definir el tema de nuestro proyecto de procesamiento de datos. Exploramos diversos conjuntos de datos disponibles en la plataforma de datos abiertos del gobierno colombiano, abarcando temas como el medio ambiente y el desarrollo tecnológico. Finalmente, nos enfocamos en un conjunto de datos relacionado con la cobertura móvil en diferentes regiones del país, que contaba con más de 400K filas. Este tema nos llamó la atención, ya que nos permitía analizar la brecha tecnológica que existe en ciertas zonas del territorio colombiano.

Para medir esta brecha tecnológica, decidimos utilizar los datos del examen SABER PRO, que evalúa las competencias de los estudiantes universitarios y contaba con 1.22 Millones de filas. Al relacionar los datos de cobertura móvil con los resultados del SABER PRO, nuestro grupo espera poder identificar si existe una correlación entre la disponibilidad de acceso a tecnología móvil y el desempeño académico de los estudiantes en diferentes regiones del país.

Este enfoque nos permitirá explorar la hipótesis de que la falta de acceso a tecnología móvil puede ser un factor que contribuye a la brecha educativa y de desarrollo en ciertas zonas de Colombia. Al analizar estos dos conjuntos de datos, nuestro grupo espera obtener insights valiosos que puedan ser utilizados para informar políticas públicas y estrategias orientadas a reducir la desigualdad tecnológica y mejorar las oportunidades educativas en todo el territorio colombiano.

### b) Marco teórico

**Examen SABER-PRO:** “El examen de Estado de la Calidad de la Educación Superior, Saber Pro, es un instrumento de evaluación estandarizada para la medición externa de la calidad de la educación superior que evalúa las competencias de los estudiantes que están próximos a culminar los distintos programas profesionales universitarios” (Acerca del Examen Saber Pro - Icfes, s. f.)

**5G:** Ofrece velocidades que rivalizan con las redes de fibra óptica más rápidas, con potencial para alcanzar velocidades de hasta 20 Gbps.

**LTE:** Diseñado para ser 10 veces más rápido que el 3G estándar, proporciona comunicación basada en IP de voz y multimedia y streaming a entre 100 Mbit por segundo y 1 Gbit por segundo.

**HSPA+:** Un perfeccionamiento de la tecnología HSPA, que ofrece velocidades de hasta 42 Mbps para las descargas y hasta 11,5 Mbps para las subidas. Suele usarse para redes privadas, relacionadas con actividades militares o satelitales.

**4G:** La última generación de tecnología inalámbrica móvil, que ofrece la mejor velocidad de datos, con una velocidad máxima de descarga de 100 Mbps para alta movilidad y de hasta 1 Gbps para acceso inalámbrico estacionario.

**3G:** La primera tecnología de red móvil que proporciona acceso a Internet de banda ancha de alta velocidad, con velocidades de hasta 14,4 Mbps para descargas y hasta 5,76 Mbps para cargas.

**2G:** Es la primera tecnología de red móvil digital, compatible con SMS e Internet móvil, con velocidades de hasta 236,8 Kbps para descargas y hasta 59,2 Kbps para cargas.

### c) Preguntas planteadas para el análisis

Con todos esto en cuenta, formulamos las siguientes preguntas:

*¿La disponibilidad de acceso a tecnología móvil en una región está relacionada con mejores resultados académicos en el SABERPRO de los estudiantes de esa misma región?*

*¿Las regiones con una mayor variedad de proveedores y tecnologías de conectividad móvil (2G, 3G, 4G, etc.) tienen estudiantes con mejores habilidades y oportunidades para utilizar herramientas y contenidos educativos en línea, lo cual se refleja en sus resultados en el examen SABER PRO?*

### d) Exploración de los datos

Para comenzar la exploración de los datasets de SABER PRO y cobertura móvil, es importante analizar la distribución y características de cada uno de ellos.

## Dataset Saber PRO

En el caso del dataset de SABER PRO, que contiene los resultados de la prueba a nivel municipal, podemos generar histogramas de conteo para revisar cómo se distribuyen los puntajes obtenidos por los estudiantes en las diferentes regiones del país. Esto nos permitirá identificar si existen sesgos o patrones en la distribución, como posibles concentraciones en ciertos rangos de puntaje o la presencia de valores atípicos. Cabe resaltar que la exploración se hizo de tal forma en la cual se separan las columnas cualitativas y las cualitativas, esto para generar visualizaciones acertadas para cada uno de los tipos de variables.

Asimismo, debido al gran volumen de columnas que existen en el dataset, se optó por seleccionar aquellas que no son redundantes y que pueden llegar a generar un buen insight relacionado a las preguntas planteadas en la fase 1.b. En ese sentido las columnas seleccionadas fueron las temporales, aquellas que contienen el periodo donde se desarrolló el examen, las que contienen la calificación bruta por módulo del examen, las que tienen la información demográfica del estudiante y aquellas que tienen la información académica del estudiante. Esto acorde a lo que se muestra en la siguiente imagen:

```
#Seleccionar columnas utiles
cols = ["PERIODO", "ESTU_PAIS_RESIDE", "ESTU_COD_RESIDE_MPIO","ESTU_MPIO_RESIDE","ESTU_NUCLEO_PREGRADO",
"ESTU_PRGM_ACADEMICO", "ESTU_NIVEL_PRGM_ACADEMICO", "ESTU_METODO_PRGM", "ESTU_HORASSEMANATRABAJA",
"ESTU_PRIVADO_LIBERTAD", "ESTU_GENERO", "FAMI_EDUCACIONPADRE", "FAMI_ESTRATOVIVIENDA", "FAMI_TIENEComputador",
"FAMI_TIENEINTERNET", "FAMI_EDUCACIONMADRE", "MOD_RAZONA_CUANTITAT_PUNT", "MOD_COMUNI_ESCRITA_PUNT",
"MOD_LECTURA_CRITICA_PUNT", "MOD_INGLES_PUNT", "MOD_COMPETEN_CIUDADA_PUNT"]
df_resultados = df_resultados_pro.select(cols)
```

Filtro de columnas para posterior exploración

Posterior a esta selección, se procedió a crear las respectivas visualizaciones, histogramas para variables continuas y diagrama de barras para variables discretas:

Diagrama de País donde Reside



Diagrama de Núcleo Profesional

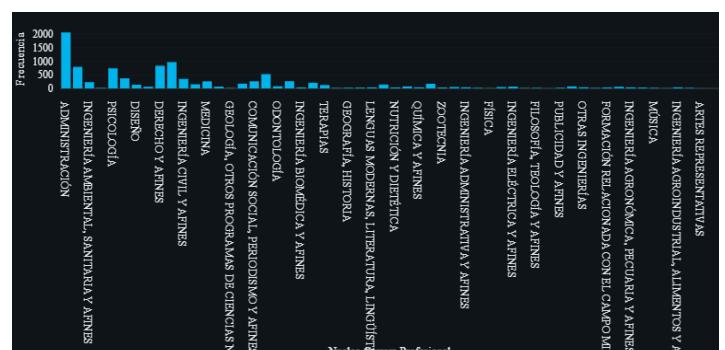


Diagrama de Nivel Académico

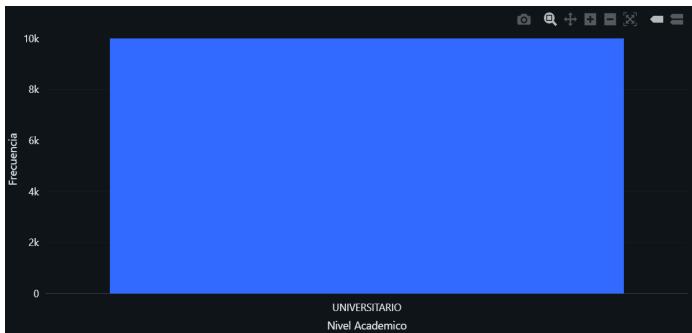


Diagrama de Método de Enseñanza

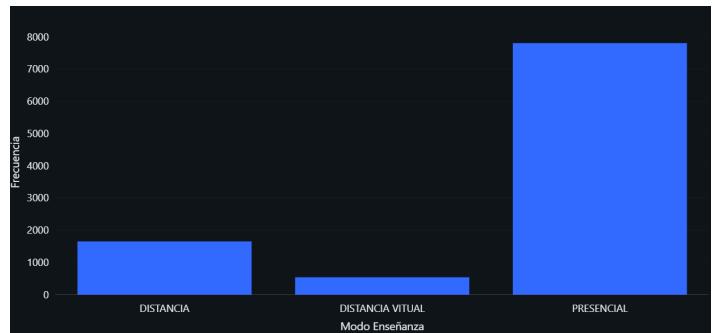


Diagrama de Horas Trabajadas

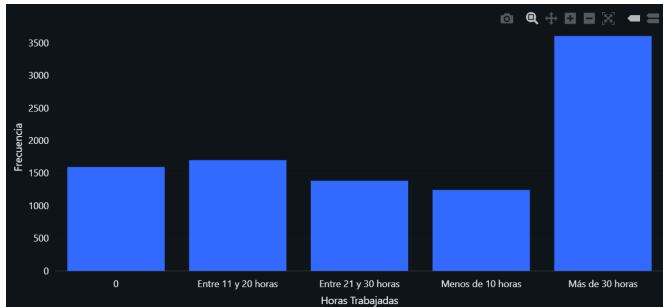


Diagrama Privados de Libertad

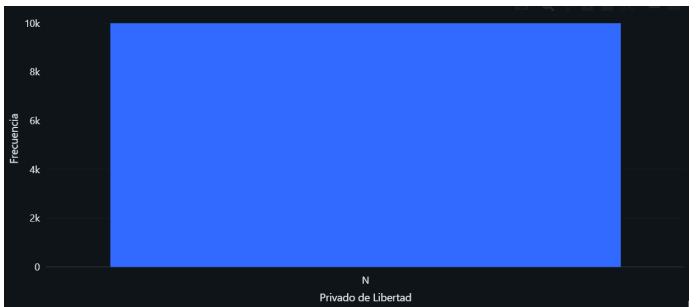


Diagrama de Género

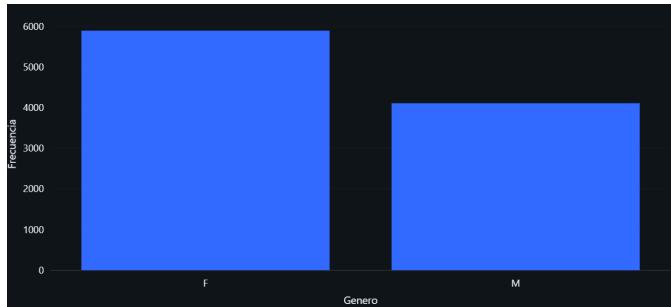


Diagrama de Educación del Padre

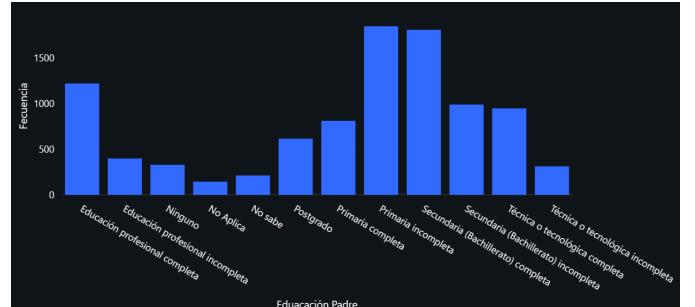


Diagrama de Estrato Vivienda

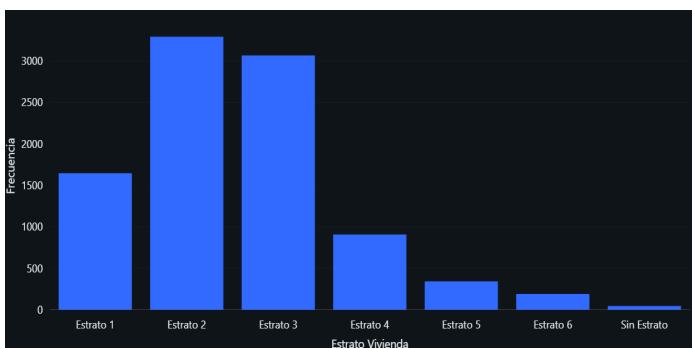


Diagrama Posesión de Computador

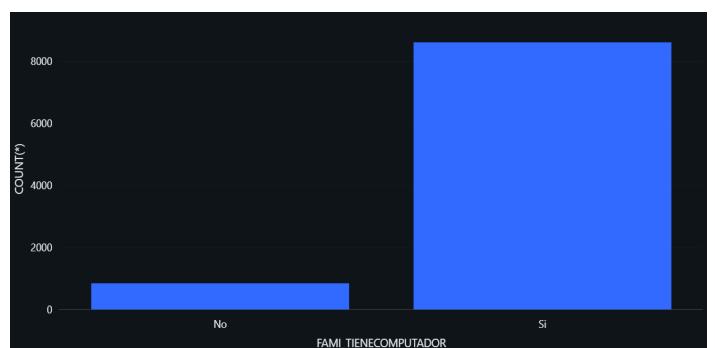


Diagrama Tiene Internet

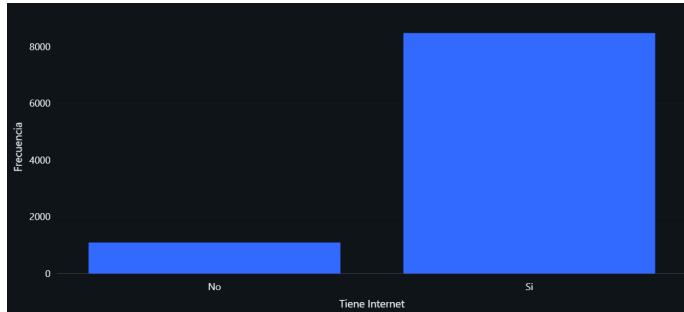


Diagrama Educación Madre

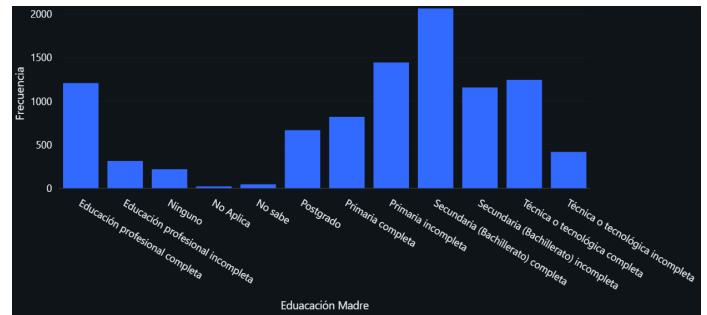
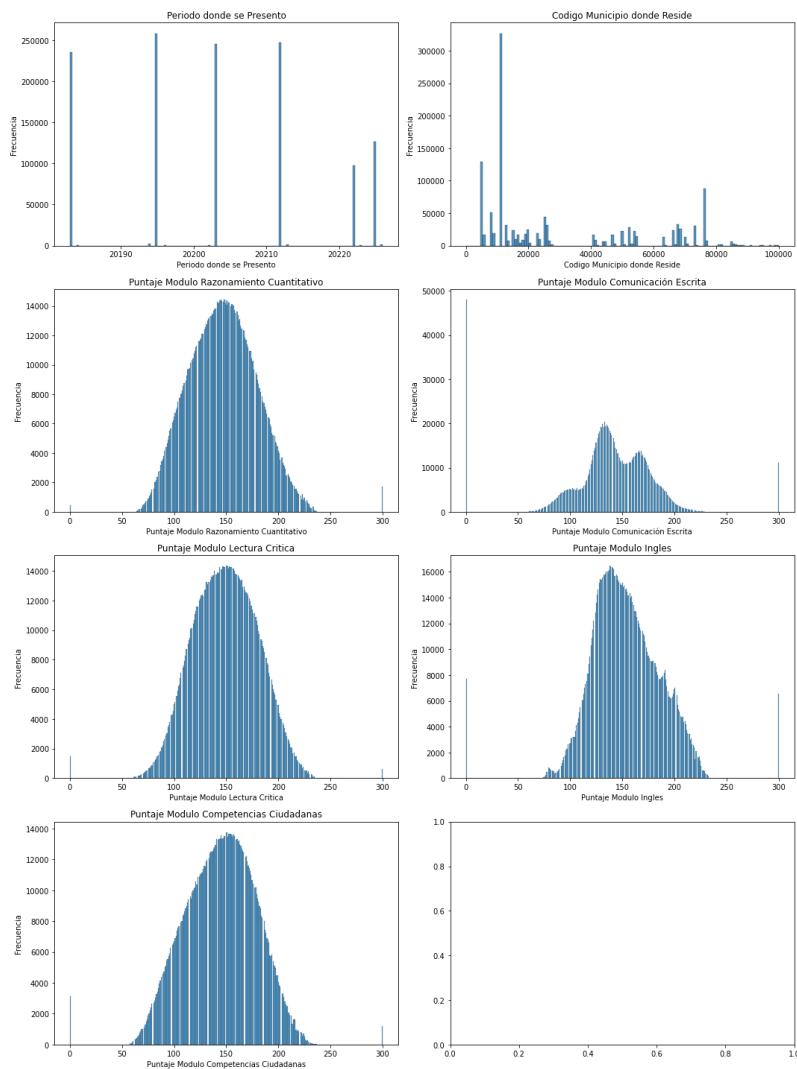


Diagrama de barras variables discretas representando datos demográficos para los estudiantes que presentaron el SABERPRO



#### Histogramas de variables continuas representando resultados de las distintas temáticas de SABERPRO

Al generar histogramas de los puntajes obtenidos por los estudiantes en estas áreas, podemos observar que la distribución de los datos tiende a seguir una forma de campana, característica de la distribución normal. Esto nos indica que la mayoría de los estudiantes se concentran en torno a un puntaje promedio, con disminuciones graduales hacia los extremos de la distribución donde se hallan los datos atípicos.

Asimismo, podemos ver que algunas variables categóricas solo tienen un valor, lo cual inmediatamente hace que las descartemos, al no proveer información sustancial para el estudio. Podemos observar que hay variedad en cuanto a los datos de cada individuo, ya que las condiciones de vida, condiciones de sus parientes pueden llegar a ser muy distintas. Debido a que estos comportamientos son variables y de la misma forma nos dan información limitada, teniendo en consideración las preguntas planteadas, por ende optamos por el uso de estrategias tanto de agrupación, como de visualización geográfica para obtener información asociada a la conectividad dentro del mismo dataset, al igual que encontrar patrones dentro de las zonas geográficas.

Diagrama de Tener Internet agrupado por Tener dispositivo de Cómputo

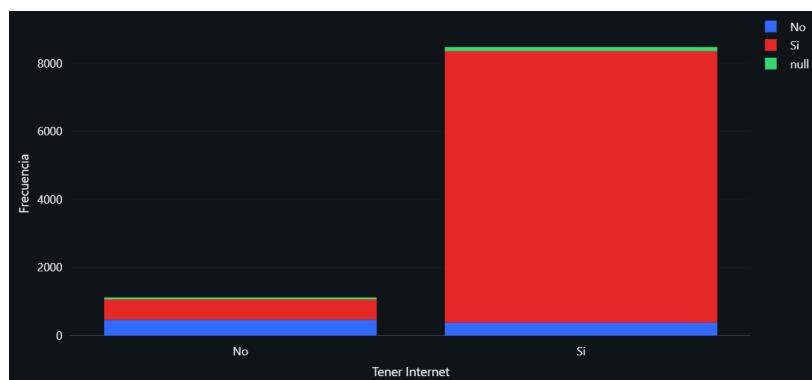


Diagrama de Educación del padre agrupado por Tener Internet

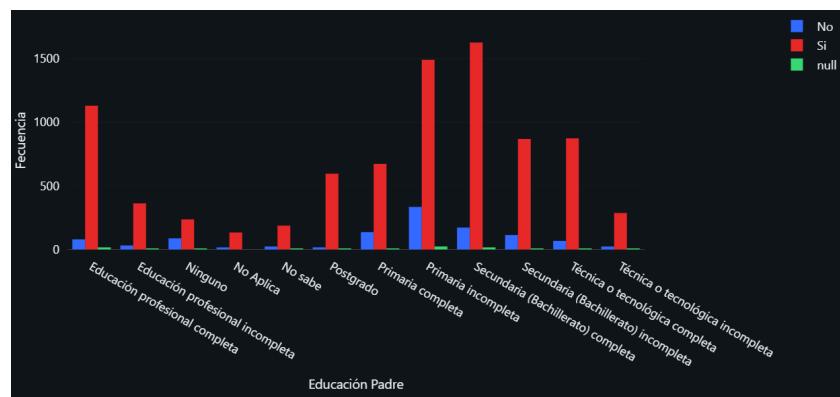
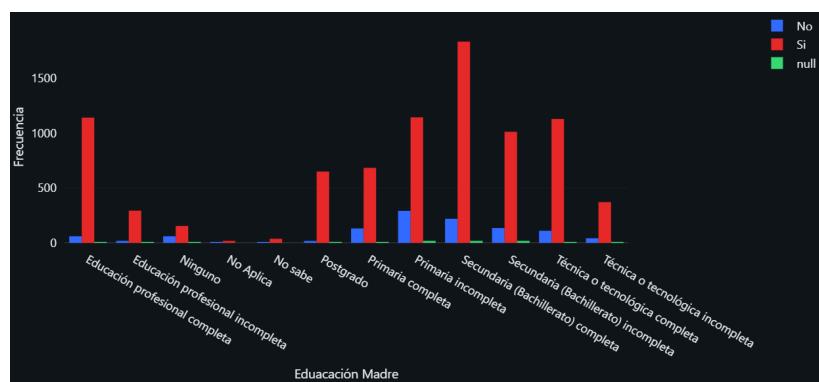
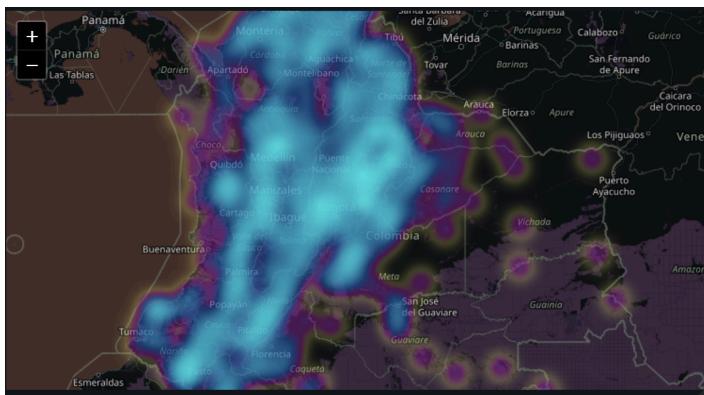


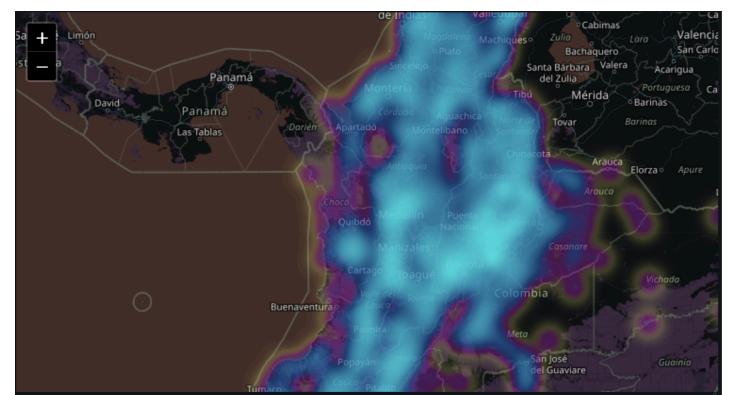
Diagrama de Educación de la madre agrupado por Tener Internet



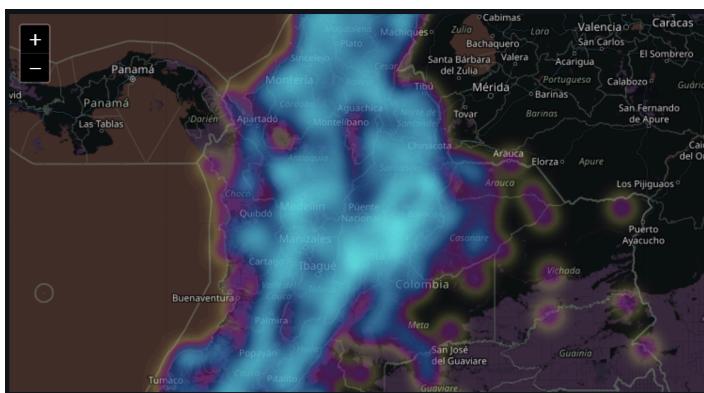
Módulo de Razonamiento Cuantitativo



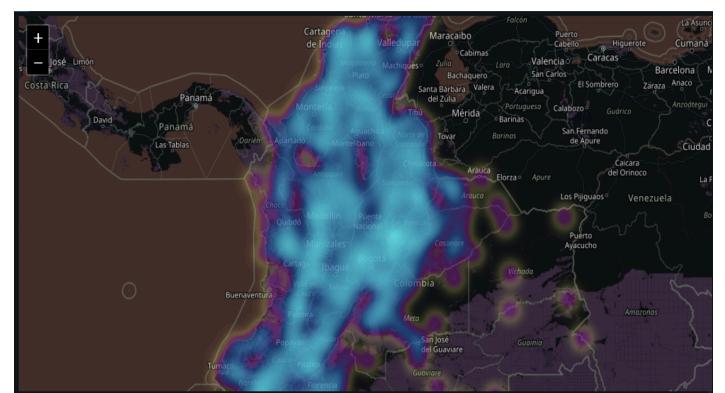
Módulo de Comunicación Escrita



Módulo de Lectura Crítica



Módulo de Inglés



## Módulo de Competencias Ciudadanas

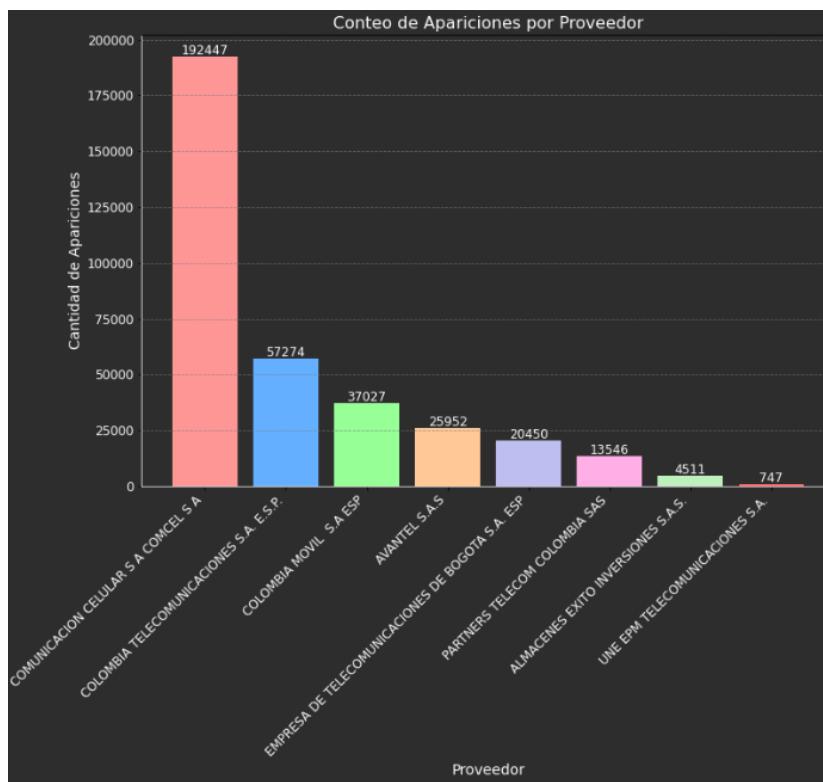


Mapas geográficos con diagrama de calor variando dependiendo del puntaje promedio por zonas en los modelos del SABER PRO

Como se puede evidenciar para las primeras visualizaciones agrupadas, podemos notar que existe una incidencia algo pronunciada en cuanto a la educación de las entidades maternas y el acceso a internet como tal, siendo la educación del padre la que más sale a relucir. Por otro lado podemos notar que si existen personas con acceso a internet, pero sin un sin acceso a un computador, aunque son bajos los casos. Y por último, en la distribución geográfica hecha con la biblioteca *folium*, podemos observar que en todas las gráficas hacia las afueras del país empieza a decrecer el puntaje de los estudiante, siendo estas zonas más modernas y más opacas, además de notar una mayor dispersión en los mapas de los modelos de inglés y competencias ciudadanas.

### Dataset Cobertura Móvil

Por otro lado, el dataset de cobertura móvil también debe ser analizado a profundidad. Mediante histogramas, podremos visualizar la dispersión de los datos de cobertura en las distintas tecnologías (2G, 3G, 4G, etc.) y en las diferentes ciudades. Esto nos ayudará a identificar si existen brechas significativas en la disponibilidad de conectividad móvil entre las regiones, lo cual podría ser un factor relevante a la hora de analizar su posible impacto en el desempeño académico de los estudiantes.



Gráfica acerca de la popularidad de cada proveedor de cobertura móvil en Colombia

El gráfico de barras muestra el conteo de apariciones por proveedor de telecomunicaciones, destacando a COMUNICACIÓN CELULAR S.A. COMCEL S.A. con 192,447 apariciones, seguido por COLOMBIA TELECOMUNICACIONES S.A. E.S.P. con 57,274. Los demás proveedores, como COLOMBIA MÓVIL S.A. E.S.P. y AVANTEL S.A.S., tienen conteos significativamente menores.

Es importante recalcar que estos proveedores son los encargados de decidir e invertir, dependiendo de la infraestructura disponible, qué servicios están disponibles en cada parte del país. Es decir, que, de ser el caso, son los responsables de hacer obsoleto algún tipo de plan para sus clientes, como es el 5G, que actualmente no está disponible en ningún sitio del territorio. Esta decisión afecta la oferta tecnológica y la competitividad del mercado, influenciando directamente la experiencia de los usuarios y la adopción de nuevas tecnologías.

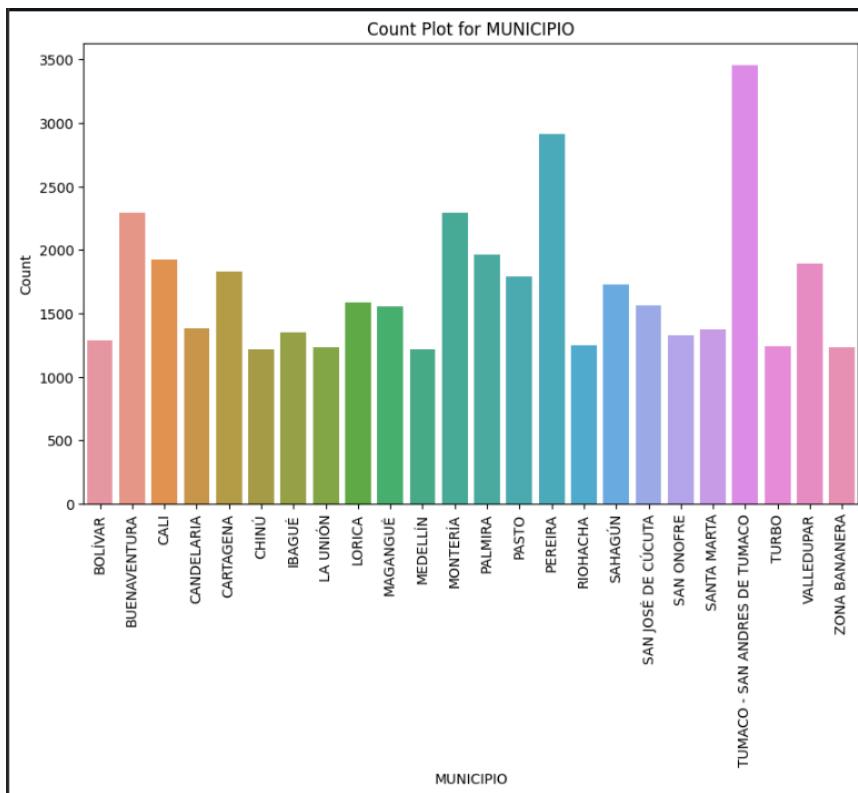
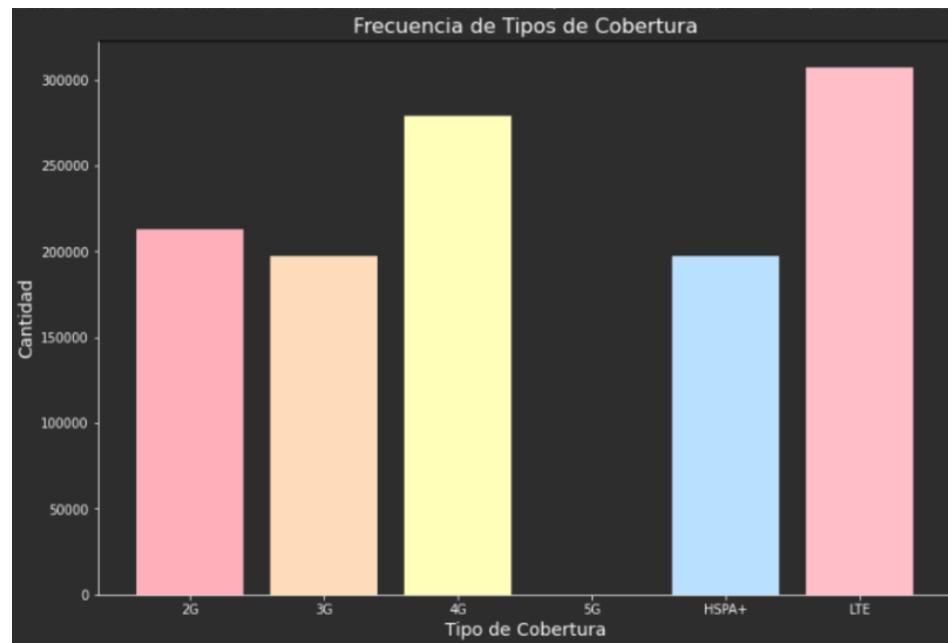


Imagen de los municipios con mayor representación dentro del dataset de cobertura

Al generar los histogramas de las variables de cobertura móvil, como los porcentajes de acceso a tecnologías 2G, 3G y 4G, podemos notar que la distribución de estos datos no es homogénea entre las diferentes regiones del país.

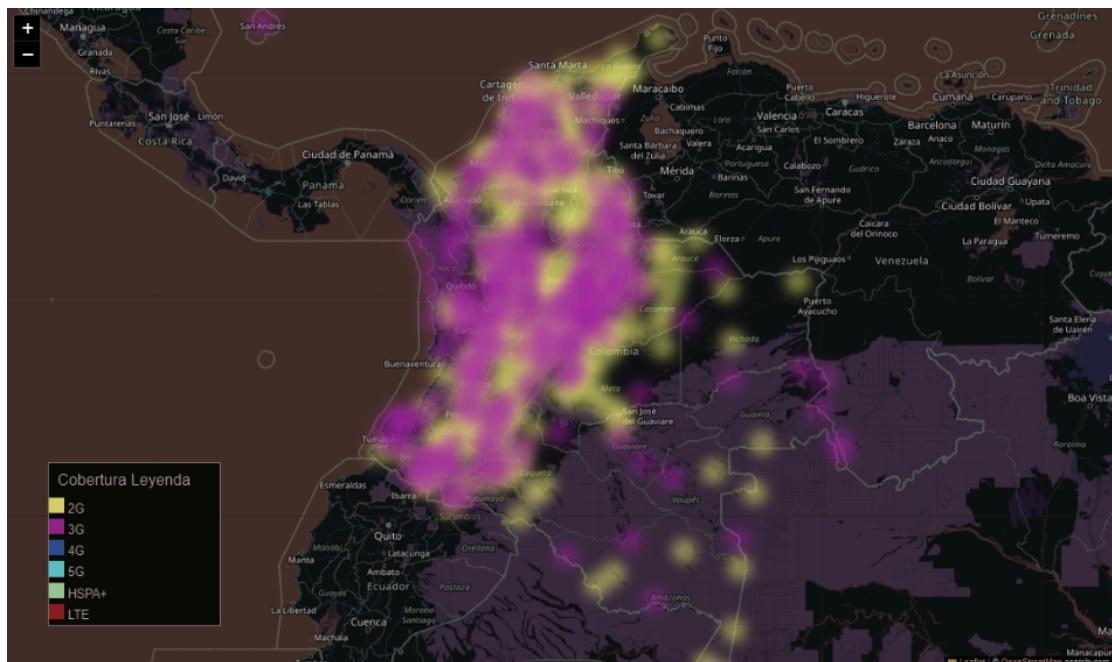
Por un lado, vemos que municipios como Pereira y Tumaco presentan una mayor concentración de datos, lo que sugiere que estos lugares cuentan con una mayor cantidad de registros en el dataset de cobertura móvil. Esto podría indicar que estas ciudades tienen una presencia más significativa en términos de proveedores y redes móviles disponibles, aún si estas no son necesariamente de mejor calidad.



Tipo de cobertura en el país por ofrecimiento de cada proveedor

El siguiente gráfico de barras muestra la frecuencia de diferentes tipos de cobertura de red móvil. En el eje horizontal se presentan los tipos de cobertura: 2G, 3G, 4G, 5G, HSPA+ y LTE. En el eje vertical se indica la cantidad de ocurrencias de cada tipo de cobertura. La barra correspondiente a LTE tiene la mayor cantidad, superando las 300,000 ocurrencias, seguida por 4G con aproximadamente 275,000. Las coberturas 2G y 3G tienen frecuencias menores, alrededor de 200,000 cada una. HSPA+ tiene una frecuencia cercana a 150,000, mientras que 5G no muestra datos en el gráfico.

"Mapa de calor creado con Folium para mostrar las coberturas más populares en cada área del país"



**Uso de Folium para Visualización de Datos:**

Folium es una biblioteca de Python que facilita la creación de mapas interactivos, integrando potentes capacidades de visualización geoespacial con una interfaz simple y fácil de usar. Para nuestra visualización de datos de cobertura de red, utilizamos Folium para crear un heatmap que muestra las áreas donde cada tipo de cobertura es más popular. Esta herramienta nos permitió representar geográficamente la distribución de HSPA+, 2G, 3G, 4G, y LTE, proporcionando una visión clara y detallada de las zonas con mayor y menor cobertura, y permitiéndonos identificar patrones y tendencias en la adopción de tecnologías móviles en diferentes regiones.

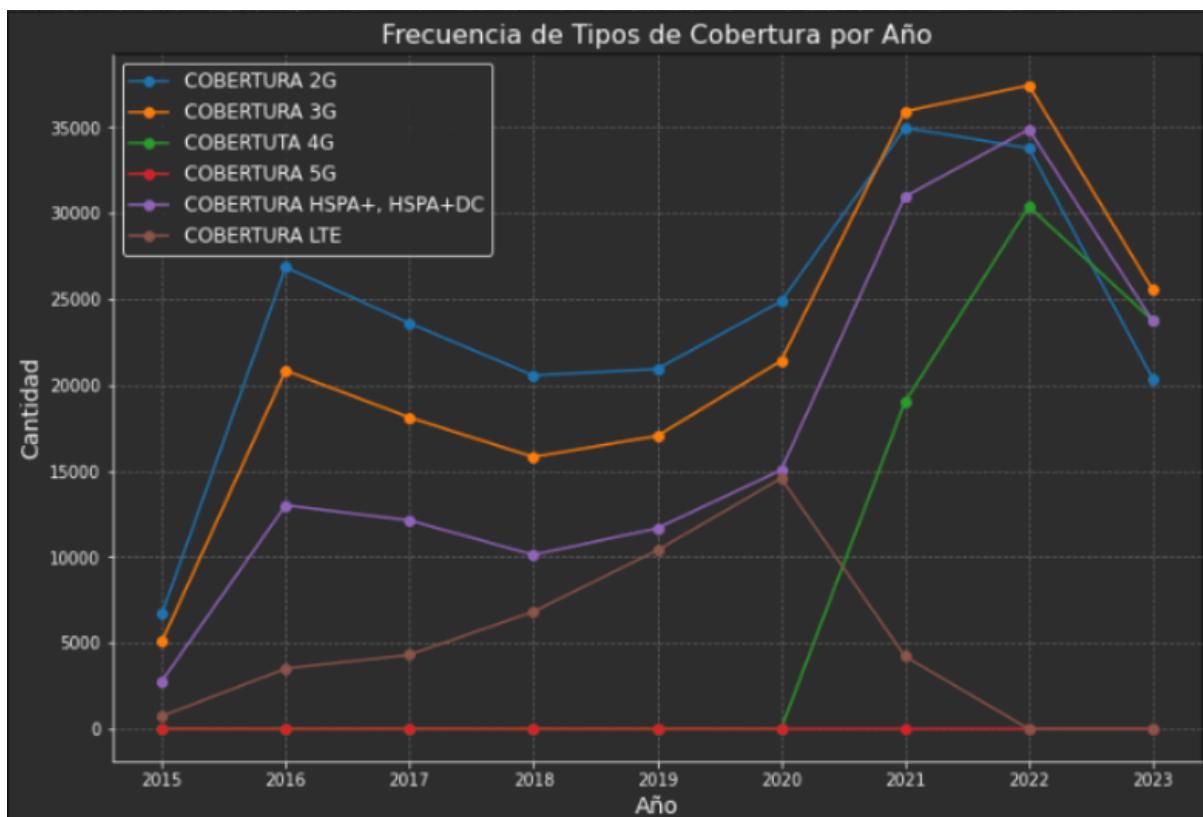
**Importancia en el mundo real:**

El análisis de los datos de cobertura de red revela una significativa brecha tecnológica en el país. A pesar de que los operadores ofrecen mayoritariamente LTE y 4G, la mayoría de la población sigue utilizando 2G y 3G, probablemente debido a su menor costo. Esta disparidad sugiere que las nuevas tecnologías no están al alcance de todos, resaltando la necesidad de políticas que promuevan la inclusión digital. Mientras los proveedores continúan invirtiendo en infraestructuras avanzadas, es crucial considerar las capacidades económicas de la población para cerrar la brecha tecnológica y asegurar que los beneficios de la modernización tecnológica sean accesibles para todos.

**Conexión con los Resultados de Saber Pro:**

Al superponer el mapa de cobertura de red con los resultados geográficos de las pruebas Saber Pro, observamos una posible correlación entre la calidad de la cobertura y el rendimiento académico. Regiones con predominantemente 2G y 3G muestran, en muchos casos, resultados un poco más bajos en Saber Pro, lo que podría indicar que la falta de acceso a tecnologías avanzadas afecta negativamente la educación y el acceso a recursos digitales. Esta conexión subraya la importancia de mejorar la infraestructura tecnológica en áreas con bajos índices de desempeño educativo para potenciar el aprendizaje y reducir desigualdades, sugiriendo una relación directa entre la accesibilidad tecnológica y el desarrollo académico.

Para complementar el diagrama geográfico anterior, queremos mostrar un diagrama que muestra la popularidad de cada tecnología en cada año para así mostrar la tendencia de cambio de cada una a lo largo de los años.



“Diagrama de frecuencia de cobertura por año”

El gráfico de líneas muestra la frecuencia de diferentes tipos de cobertura de red móvil a lo largo de los años, desde 2015 hasta 2023. Como podemos ver, es lo más común en el público general las redes 2G, 3G y HSPA+. Sin embargo, al ser HSPA+ una red prominentemente privada, no cuenta dentro del estudio de folium geográfico. De esta manera, podemos comprobar también cómo, aunque 4G está ganando rápidamente tracción en el mercado, no ha sido suficiente para alcanzar a sus tecnologías anteriores, especialmente en zonas alejadas del centro del país.

#### e) Preparación de los datos

Después de la exploración inicial de los datasets de SABER PRO y cobertura móvil, procedimos a revisar la calidad de los datos para garantizar que fueran adecuados para los análisis posteriores.

En el caso del dataset de SABER PRO, encontramos que había algunos valores nulos, principalmente en las columnas relacionadas con los puntajes de los estudiantes a nivel municipal.

Al analizar la proporción de estos datos faltantes, determinamos que no eran significativos en comparación con el total de registros. Específicamente, los valores nulos en las columnas más relevantes, como 'MOD\_RAZONA\_CUANTITAT\_PUNT', 'MOD\_COMUNI\_ESCRITA\_PUNT', 'MOD\_LECTURA\_CRITICA\_PUNT' y 'MOD\_INGLES\_PUNT', representaban menos del 1% del dataset completo. Dado que esta cantidad no era relevante, decidimos eliminar estos registros con datos faltantes, asegurando así una alta integridad en el conjunto de datos.

```
Proporción de valores nulos en la columna 'ESTU_PAIS_RESIDE': 0.0
Proporción de valores nulos en la columna 'ESTU_COD_RESIDE_MCPPIO': 0.0027752491808989243
Proporción de valores nulos en la columna 'ESTU_NUCLEO_PREGRADO': 0.0
Proporción de valores nulos en la columna 'ESTU_PRGM_ACADEMICO': 0.0
Proporción de valores nulos en la columna 'ESTU_METODO_PRGM': 0.0
Proporción de valores nulos en la columna 'ESTU_HORASSEMANATRABAJA': 0.0455221379113169
Proporción de valores nulos en la columna 'ESTU_GENERO': 9.776632697660509e-05
Proporción de valores nulos en la columna 'FAMI_EDUCACIONPADRE': 0.05846262040046402
Proporción de valores nulos en la columna 'FAMI_ESTRATOVIVIENDA': 0.045222266404203786
Proporción de valores nulos en la columna 'FAMI_TIENEComputador': 0.0532812663647686
Proporción de valores nulos en la columna 'FAMI_TIENEINTERNET': 0.03898000966161349
Proporción de valores nulos en la columna 'FAMI_EDUCACIONMADRE': 0.03878529857343319
Proporción de valores nulos en la columna 'MOD_RAZONA_CUANTITAT_PUNT': 0.0
Proporción de valores nulos en la columna 'MOD_COMUNI_ESCRITA_PUNT': 0.0061329884947929165
Proporción de valores nulos en la columna 'MOD_LECTURA_CRITICA_PUNT': 0.0
Proporción de valores nulos en la columna 'MOD_INGLES_PUNT': 0.00010105259006825568
Proporción de valores nulos en la columna 'MOD_COMPETEN_CIUDADA_PUNT': 0.0
Proporción de valores nulos en la columna 'AÑO': 0.0
```

Proporción de datos nulos en el dataset del SABERPRO

Por otro lado, en el dataset de cobertura móvil también identificamos algunos valores nulos, principalmente en las columnas que indican los niveles de cobertura para las diferentes tecnologías (2G, 3G, 4G, etc.) a nivel de centros poblados. Sin embargo, al final realmente no fué necesario utilizar ninguna de las columnas que tenían los valores nulos, garantizando que el dataset final tuviera una calidad adecuada para los análisis posteriores.

```
(spark_cobertura.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in spark_cobertura.columns]).show())
▶ (2) trabajos de Spark
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|AÑO|PROVEEDOR|COD MUNICIPIO|MUNICIPIO|COBERTURA 2G|COBERTURA 3G|COBERTURA HSPA+, HSPA+DC|COBERTURA 4G|COBERTURA LTE|COBERTURA 5G|
+---+---+---+---+---+---+---+---+---+---+
| 0|    0|     0|     0|     0|     0|          0|     0|     0|     0|
```

Cantidad de nulos en el dataset de Cobertura Móvil

Mediante este proceso de revisión y limpieza de los datos, logramos asegurar que ambos conjuntos de datos, tanto el de SABER PRO como el de cobertura móvil, contarán con una integridad y calidad suficiente para llevar a cabo un análisis robusto y confiable de la posible relación entre la disponibilidad de conectividad móvil y el desempeño académico de los estudiantes en Colombia.

## 2) PROBLEMA

### a) Descripción del problema de analítica a resolver

El problema que buscamos resolver usando analítica, es intentar hallar un modelo de ML que sea capaz de dar, con cierto grado de precisión, predicciones de los resultados que obtendría un estudiante cualquiera dependiendo de si la región en la que vive tiene mejor acceso a conectividades de mayor calidad. Para esto, después de todo un proceso de preparación y procesamiento de los datos, aplicaremos técnicas de Machine Learning como Regresión con *Random Forest* o *Gradient Boosting* y observaremos qué tan correcto sería poder hacer predicciones.

Uno de los principales retos de este análisis es que, como pudimos ver en la correlación de los datos, esta no era muy grande. Esto significa que la relación entre ambos datasets, el de cobertura móvil y el de resultados del SABER PRO, puede no ser lo suficientemente significativa para poder realizar predicciones confiables. Esto se debe a que el desempeño académico de los estudiantes puede depender de muchas otras variables, más allá de la disponibilidad y acceso a conectividad de internet móvil en su región y por ende un mayor acceso a información.

### b) Solución a los problemas típicos en datos

#### Problemas de calidad de datos:

En el proceso de preparación de los datos, identificamos algunos problemas de calidad que debemos abordar. En el caso del dataset de SABER PRO, encontramos valores nulos principalmente en las columnas relacionadas con los puntajes de los estudiantes a nivel municipal. Sin embargo, como mencionamos anteriormente, estos valores nulos representaban menos del 1% del total de registros, por lo que decidimos eliminarlos sin que ello afectara significativamente la integridad del conjunto de datos.

De manera similar, en el dataset de cobertura móvil también detectamos algunos valores nulos, pero estos se concentraban en columnas que no eran críticas para nuestro análisis. Por lo tanto, optamos por eliminar estos registros con datos faltantes, asegurando que el dataset final tuviera una calidad adecuada.

### Normalización de variables

Específicamente, utilizamos el método de Standard Scaler para estandarizar las variables en ambos conjuntos de datos. Este proceso consistió en transformar los valores de cada variable, restando la media y dividiendo por la desviación estándar. Al aplicar Standard Scaler, logramos que todas las variables tuvieran una media de 0 y una desviación estándar de 1. Este paso es necesario para poder trabajar con los datos en igualdad de condiciones, evitando que algunas variables dominaran el análisis debido a diferencias de escala.

La normalización de los datos mediante Standard Scaler es un paso fundamental en la preparación de los datasets, ya que permitió que pudiéramos proceder con el análisis de machine learning de manera más efectiva y sin sesgos relacionados con las unidades o rangos de las variables.

### Creación de nuevas variables

Sin embargo, sí tuvimos que crear nuevas variables derivadas a partir de los datos existentes. En el caso del dataset de SABER PRO, aplicamos técnicas de one-hot encoding a algunas columnas categóricas, como 'ESTU\_GENERO' y 'ESTU\_PRGM\_ACADEMICO'. Esto nos permitió transformar estas variables en un formato numérico, facilitando así su integración y análisis conjunto con los datos de cobertura móvil.

De igual manera, se creó una variable nueva, la cual tiene como propósito ayudar al desarrollo de los modelos de IA. La variable en cuestión es GLOBAL, la cual como su nombre indica es el puntaje global que el estudiante obtuvo en el examen SABER-PRO. Cabe resaltar que esta variable se calculó con un promedio simple entre todos los módulos, ya que la forma en la que califican el examen no es de conocimiento público. Por consiguiente el puntaje más alto que se puede obtener bajo este parámetro es 300 puntos, teniendo toda la prueba perfecta.

Esta variable nueva se creó con el fin de ser nuestra variable objetivo, ya que a pesar de no contar con la precisión contextual necesaria, igual nos da un apoyo para estandarizar el desempeño dentro del examen de una manera general.

```
df_combinado_2 = df_combinado_1.withColumn("GLOBAL",
                                             (df_combinado_1["MOD_RAZONA_CUANTITAT_PUNT"] +
                                              df_combinado_1["MOD_COMUNI_ESCRITA_PUNT"] +
                                              df_combinado_1["MOD_LECTURA_CRITICA_PUNT"] +
                                              df_combinado_1["MOD_INGLES_PUNT"] +
                                              df_combinado_1["MOD_COMPETEN_CIUDADA_PUNT"]) / 5)

display(df_combinado_2)
```

Creación de la variable GLOBAL

### Fusión de los conjuntos de datos

Para llevar a cabo la fusión de los datasets de SABER PRO y cobertura móvil, primero subimos ambos conjuntos de datos a un repositorio de GitHub. Esto nos permitió acceder a los datos de manera centralizada y facilitar su manipulación.

Una vez en GitHub, creamos un nuevo entorno de trabajo en Databricks, una plataforma de análisis de datos en la nube. Desde este entorno, cargamos los datasets de SABER PRO y cobertura móvil ya procesados, y procedimos a realizar el proceso de merging entre ellos.

El proceso en cuestión se planteó de forma en la cual pudiéramos utilizar una columna o varias columnas en común entre los dos datasets. Para este caso en particular, tanto el dataset que contiene los resultados del SABER PRO, y el dataset que contiene la información de cobertura tienen en común dos columnas, la de código de municipio y año en el cual se tomó el dato.

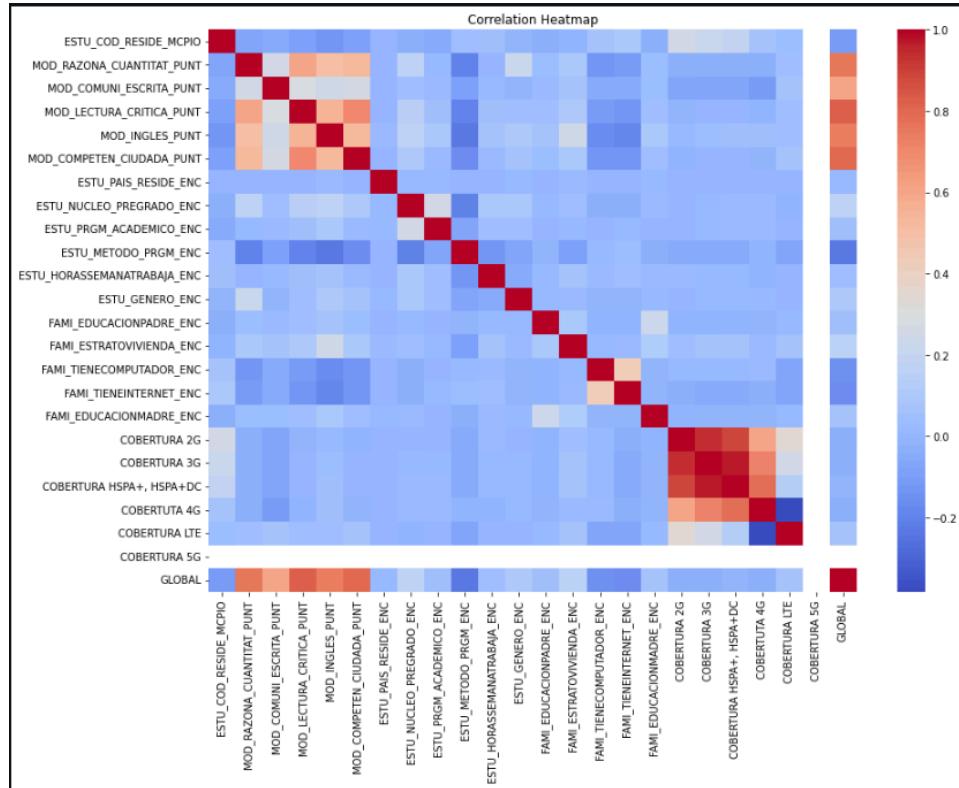
Aunque los dos datasets tuvieran estas dos columnas en común, la granularidad de las mismas no era igual, o en otras palabras, el contexto de las columnas no era el mismo. Esto pasa debido a que en un dataset el registro es por persona, mientras que en otro el registro es por empresa que da servicio de conectividad en el municipio. Por lo que, para estandarizar los datos se usó técnicas de agrupación, para que así el dataset de conectividad tuviera un solo registro por año y por municipio, obviamente sumando los demás registros para tener un historial de cuantos proveedores de conexión hay por municipio en una cierta ventana de tiempo.

De esta manera, logramos integrar ambos conjuntos de datos en una sola estructura, preparando el terreno para realizar los análisis posteriores y explorar la posible relación entre la disponibilidad de conectividad móvil y el rendimiento académico de los estudiantes en Colombia.

### c) Analítica descriptiva de los datos

Una de las herramientas que utilizamos para analizar las relaciones entre las variables fue la creación de un heatmap de correlación. Un heatmap es una representación visual de una

matriz de correlación, donde se muestra la fuerza y dirección de la relación entre las diferentes variables. En nuestro caso, el heatmap que generamos, incluyó todas las variables relevantes de ambos datasets, tanto las relacionadas con los puntajes del SABER-PRO como las de cobertura móvil.



Heatmap relacionando ambos datasets

Al analizar el heatmap, observamos que no encontramos ninguna correlación superior a 0.3, lo cual indica que no hay relaciones fuertes entre las variables analizadas. Esto supondría que nuestra hipótesis puede ser preliminarmente rechazada para nuestra sorpresa y que el acceso a conectividad de internet móvil y mejores resultados académicos no están directamente relacionados. Posteriormente, haremos un análisis más a profundidad con modelos de machine learning que nos puedan ayudar a tener un mejor entendimiento de la situación.

### 3) IMPLEMENTACIÓN DE TÉCNICAS ML Y RESULTADOS

**a) Entrenamiento y comparación → Resultados finales obtenidos por los modelos**

Para la implementación de los modelos de ML, teniendo en consideración nuestra problemática decidimos abordarla desde la regresión, optamos por hacer un pequeño paso de preprocesamiento, ya que ningún modelo recibe datos brutos y más con la biblioteca de pyspark que decidimos usar.

Para este paso se utilizaron técnicas de normalización y ensamblaje de vectores estandarizados. Esto con el fin de generar un modelo lo más limpio posible, sin parcialidades ni ruido excesivo. De esta forma, se implementaron rutinas que pudieran tomar el dataset bruto, normalizar sus datos y posteriormente transformarlo en un vector de características.

Una vez pasada esta etapa, y teniendo en consideración que todos los modelos entrenados a continuación utilizan todas las columnas que tengan alguna evidencia de conectividad, ya sea conexión LTE o acceso a computador personal, se procedió a subdividir los datos, unos para entrenar y otros para validar. Estos fueron los resultados obtenidos:

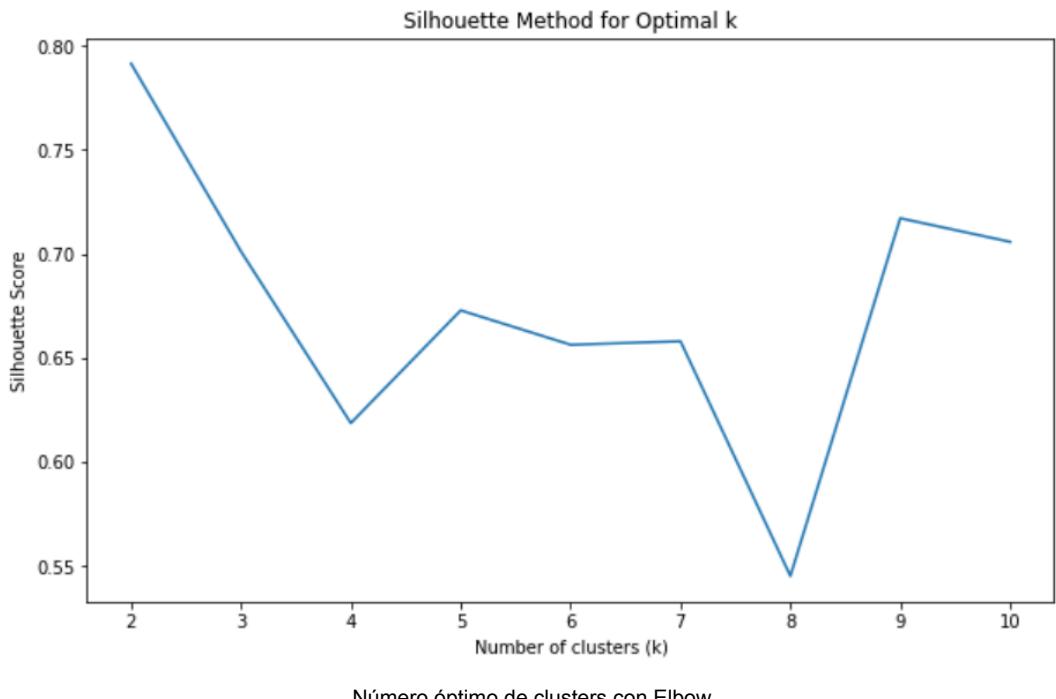
Técnica	Bibliotecas	Hiper-parámetros	Resultados
Regresión Lineal Multidimensional	pyspark.ml.regression	maxIter: 100 regParam: 0.0 elasticNetParam: 0.0 tol: 1e-6 solver: "auto" standardization: True fitIntercept: True weightCol: None	RMSE: 24.401951399879696  R2: 0.047324707925299414
Regresión con Random Forest	pyspark.ml.regression	numTrees: 20 maxDepth: 5 impurity: "variance" featureSubsetStrategy: "auto" minInstancesPerNode: 1 minInfoGain: 0.0 maxBins: 32 seed: None	RMSE: 24.1503940570205  R2: 0.06695345578474199
Regresión con Gradient	pyspark.ml.regression	maxIter: 20 maxDepth: 3 stepSize: 0.1	RMSE: 23.90258687586869

Boost		impurity: "variance" featureSubsetStrategy: "auto" minInstancesPerNode: 1 minInfoGain: 0.0 maxBins: 32 seed: None	R2: 0.08393186023917332
-------	--	--	----------------------------

Como podemos observar las métricas de los 3 modelos son bastante malas, por no decir que en sí estos modelos no tienen la capacidad de representar absolutamente ningún dato del dataset. Esto se pudo deber a diferentes factores, pudo ser el proceso de normalización, o asimismo el uso de ciertos hiper-parámetros que afectarán la precisión del modelo. Sin embargo, la más fuerte de todos nuestras hipótesis recae en la correlación, ya que como pudimos ver en la matriz de correlación, representada como un heatmap, las evidencias de conectividad tenían muy poca correlación con nuestra variable objetivo, explicando así porque los modelos de regresión son ineficaces al predecir correctamente el valor continuo, presentando errores muy grandes como se muestra en el RMSE.

Teniendo en consideración estos resultados pudimos saltar a la conclusión prematura de que es posible que los datos no se puedan predecir de una forma precisa, teniendo en consideración factores individuales que afecten a los individuos que presenten el examen. Sin embargo, decidimos seguir intentando, pero cambiando el enfoque del problema, de una predicción a una clasificación. En este sentido, y con el fin de ver si existía alguna forma de clasificar a los individuos, tuvimos que hacer más preprocesamiento de los datos, ya que como tal el modelo no predice valores continuos, sino un solo valor que representa el cluster en el que está.

Para este paso de preprocesamiento, nuevamente tuvimos que utilizar las mismas técnicas de ensamblaje de vectores de características y normalización de datos, sin embargo, para este modelo en específico tuvimos que cambiar la variable objetivo, ya que GLOBAL solo tiene datos continuos. Sabiendo esto, procedimos a agrupar los valores continuos en intervalos utilizando la biblioteca Bucketizer que provee Pyspark para hacer esta división de forma automática. Utilizamos 8 divisiones teniendo en consideración el método Elbow mostrado a continuación, donde se muestra que el número de divisiones óptimo son 8.



Teniendo todos los datos ya procesados y sabiendo específicamente cuántos clusters vamos a hacer, solo procede entrenar el modelo. Cabe resaltar que este modelo es de carácter no supervisado, por lo que siguiendo con esa filosofía, utilizamos los hiper-parámetros básicos o predeterminados del modelo. Sabiendo esto, estos fueron los resultados:

Técnica	Bibliotecas	Resultados
K-means	pyspark.ml sklearn	Silhouette: 0.54  Calinski-Harabasz: 537.00  Davies-Bouldin: 0.58

Los resultados expuestos para este modelo, a diferencia de los de regresión, son bastante prometedores, ya que las métricas, aunque no perfectas, son buenas, teniendo en consideración que estas denotan que muy pocos grupos se sobreponen y que estos están bien definidos y son cohesivos.

En una futura implementación es posible que mover los hiper-parámetros, o usar otro número de centroides pueda aportar a la mejora en estas métricas.

### b) Respuestas a las preguntas planteadas

*¿La disponibilidad de acceso a tecnología móvil en una región está relacionada con mejores resultados académicos en el SABERPRO de los estudiantes de esa misma región?*

Acorde a la información que obtuvimos con la técnica de Clustering, existe una relación entre la disponibilidad de acceso a tecnología móvil en una región y la capacidad de agrupar de manera considerablemente acertada los resultados del SABERPRO si se conoce qué tipo y calidad de tecnología se cuenta.

Esta pregunta es importante porque busca establecer una posible correlación entre el acceso a tecnología móvil y el rendimiento académico de los estudiantes. Si se demuestra que existe una relación positiva, podría sugerir que la disponibilidad de herramientas tecnológicas móviles contribuye a mejorar el aprendizaje y el desempeño en pruebas académicas como el SABERPRO.

*¿Las regiones con una mayor calidad de tecnologías de conectividad móvil (2G, 3G, 4G, etc.) tienen estudiantes con mejores habilidades y oportunidades para utilizar herramientas y contenidos educativos en línea, lo cual se refleja en sus resultados en el examen SABER PRO?*

Dados los resultados que obtuvimos a lo largo del proyecto, no tenemos información estadística suficiente para poder responder la pregunta. Según nuestras propias observaciones, concluimos que esto se puede deber a dos situaciones. La primera es que, debido a que tomamos los datos a nivel municipal, no se logra identificar realmente de manera significativa las brechas tecnológicas existentes. Ese podría ser un experimento a futuro, siguiendo algún conjunto de datos de información académica que entienda a niveles mucho más sectorizados los resultados y condiciones de los estudiantes.

La segunda condición que observamos tenía que ver con el hecho de que los resultados del SABERPRO de los usuarios tienen una cantidad mucho mayor de variables que influyen en sus resultados y, aunque la conectividad es una de ellas y se encuentra relacionada, no tiene un peso suficientemente grande como para poder afirmar que, dadas estas condiciones y herramientas, se obtienen mejores resultados globales.

Para cualquiera de las dos preguntas, explorar esta relación es relevante porque:

- Identificaría factores externos que influyen en el rendimiento académico, más allá de los factores individuales o institucionales.
- Brindaría información valiosa para el diseño de políticas educativas y la asignación de recursos tecnológicos en diferentes regiones.
- Podría conducir a intervenciones específicas para mejorar el acceso a la tecnología móvil y, potencialmente, mejorar los resultados académicos.
- Contribuiría a comprender mejor el impacto de la tecnología en el proceso de aprendizaje y en el desarrollo de habilidades evaluadas por el SABERPRO.

Esta pregunta puede ser útil en el proceso de análisis de datos educativos, ya que permitiría identificar patrones y relaciones que podrían ser aprovechados para mejorar el sistema educativo y promover la equidad en el acceso a recursos tecnológicos que faciliten el aprendizaje.

#### **4) CONCLUSIONES Y FUTURAS LÍNEAS DE TRABAJO**

Inicialmente, los modelos de regresión como Random Forest y Gradient Boosting no arrojaron resultados óptimos para predecir los puntajes del SABER PRO a partir de los datos de cobertura móvil, sugiriendo la existencia de factores adicionales que influyen en el rendimiento académico. No obstante, las técnicas de agrupamiento (clustering) revelaron patrones coherentes en los datos, permitiendo agrupar a los estudiantes en función de su conectividad móvil y desempeño en el examen SABER PRO. Los resultados de clustering fueron prometedores, con un Silhouette de 0.54, un Calinski-Harabasz de 537.00 y un Davies-Bouldin de 0.58.

Estos hallazgos indican que, si bien no hay una correlación lineal directa, existe un vínculo entre la disponibilidad de conectividad móvil y los resultados académicos de los estudiantes. Esta relación, aunque sutil, se manifiesta a través de patrones y agrupaciones identificados mediante técnicas de análisis avanzadas. A pesar de no haber encontrado modelos predictivos robustos, los resultados del análisis de clustering sugieren que la brecha tecnológica y de conectividad puede ser un factor que contribuye a las disparidades en los resultados académicos a nivel regional en Colombia y por ende puede ser agrupada y clasificada en términos de su acceso a estas tecnologías.

Una línea de investigación futura podría centrarse en centros poblados específicos, donde la brecha tecnológica sea más evidente, para obtener una comprensión más profunda de cómo la falta de conectividad afecta el desempeño académico de los estudiantes.

## 5) CONTRIBUCIONES

++++++++++ **Contribuciones** +++++++

**Santiago Botero Pacheco:** Contribución con el dataset de Conectividad, colaboración con los modelos de ML y elaboración del documento y presentación.

**Santiago Avilés Tibocha:** Contribución con el dataset de SABER PRO, construcción de modelos ML y elaboración del documento y presentación.

++++++++++

## 6) REFERENCIAS

- ❖ Cobertura móvil por tecnología, departamento y municipio por proveedor | Datos Abiertos Colombia. (2024, February 23).  
[https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Cobertura-m-vil-por-tecnolog-a-departamento-y-muni/9mey-c8s8/data\\_preview](https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Cobertura-m-vil-por-tecnolog-a-departamento-y-muni/9mey-c8s8/data_preview)
  
- ❖ Resultados únicos Saber Pro | Datos Abiertos Colombia. (2023, August 23).  
[https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-Pro/u37r-hjmu/about\\_data](https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-Pro/u37r-hjmu/about_data)
  
- ❖ Acerca del examen Saber Pro - Icfes. (s. f.). Icfes.  
<https://www.icfes.gov.co/acerca-del-examen-saber-pro>

