

Proyecto - Procesamiento de datos a gran escala.

Parcial 2 - Proyecto de Aplicación

Procesamiento de Datos a Gran Escala

Santiago Botero Pacheco

Santiago Avilés Tibocho



Pontificia Universidad
JAVERIANA
Colombia

Tabla de contenido

- 1. Fase Inicial**
- 2. Problema**
- 3. Implementación de Técnicas ML y resultados**
- 4. Referencias**

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) Preguntas

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?

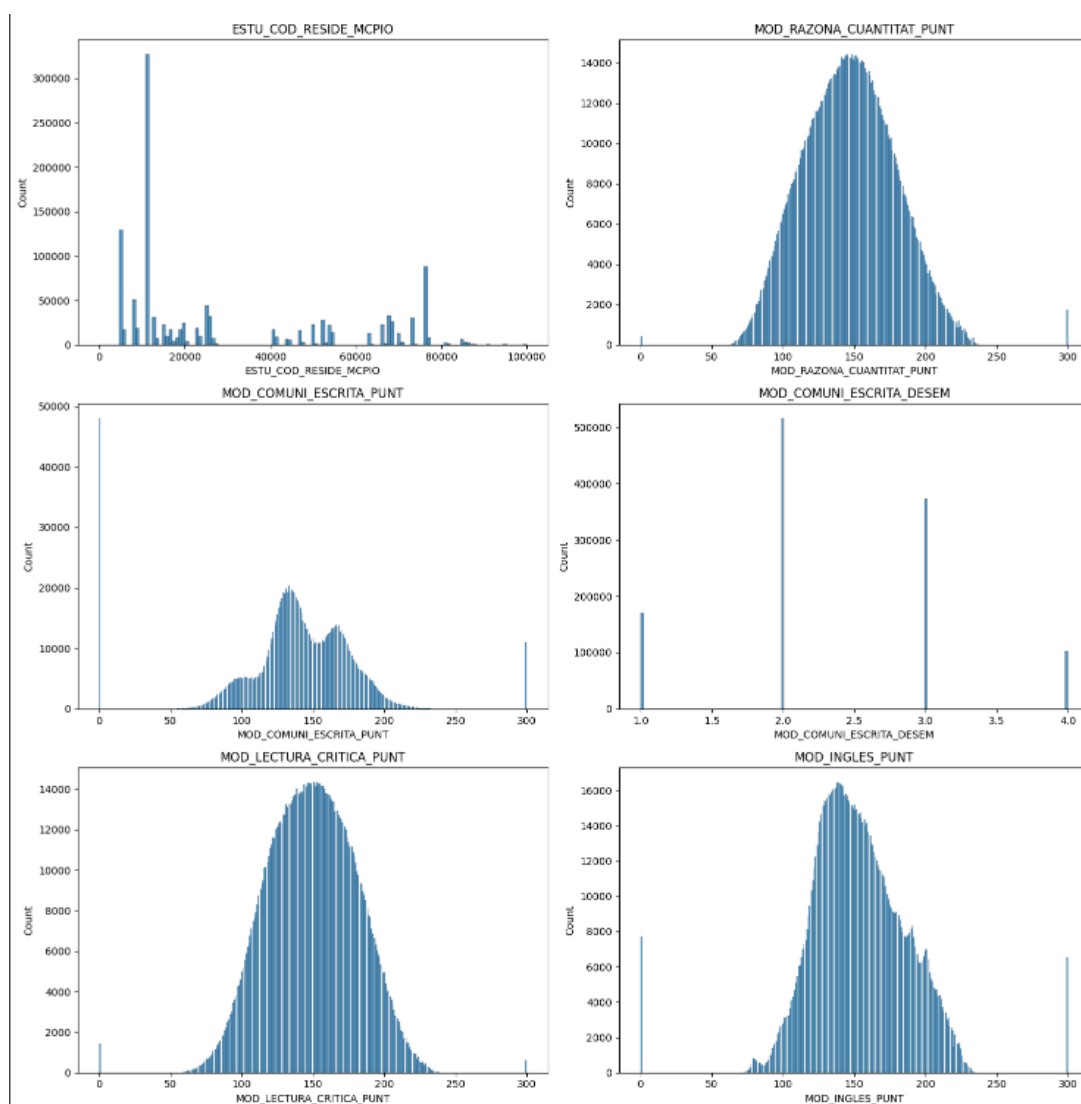
¿Las regiones con una mayor variedad de proveedores y tecnologías de conectividad móvil (2G, 3G, 4G, etc.) tienen estudiantes con mejores habilidades y oportunidades para utilizar

herramientas y contenidos educativos en línea, lo cual se refleja en sus resultados en el examen SABER PRO?

c) 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.

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.



Histogramas 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.

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.

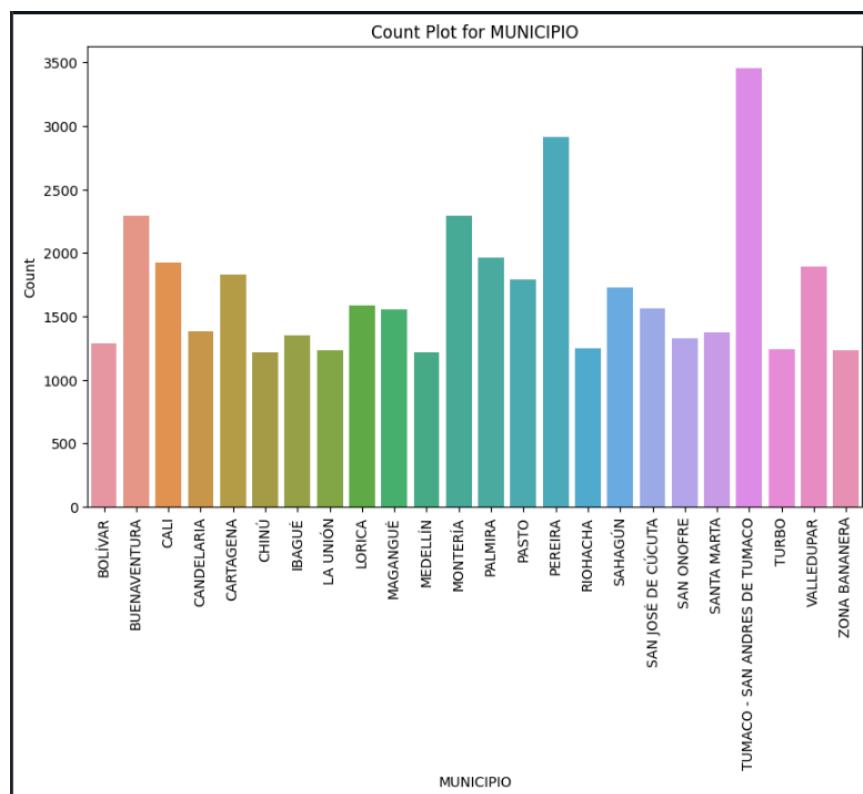


Imagen de los municipios con mayor representación dentro del dataset de cobertura

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.

d) 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_HORASSEMANTRABAJA': 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': 0.05328182663647686
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_CIUDDADA_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 debimos 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. Esto fue fundamental 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 fue un paso crucial 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_CIUADADA_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.

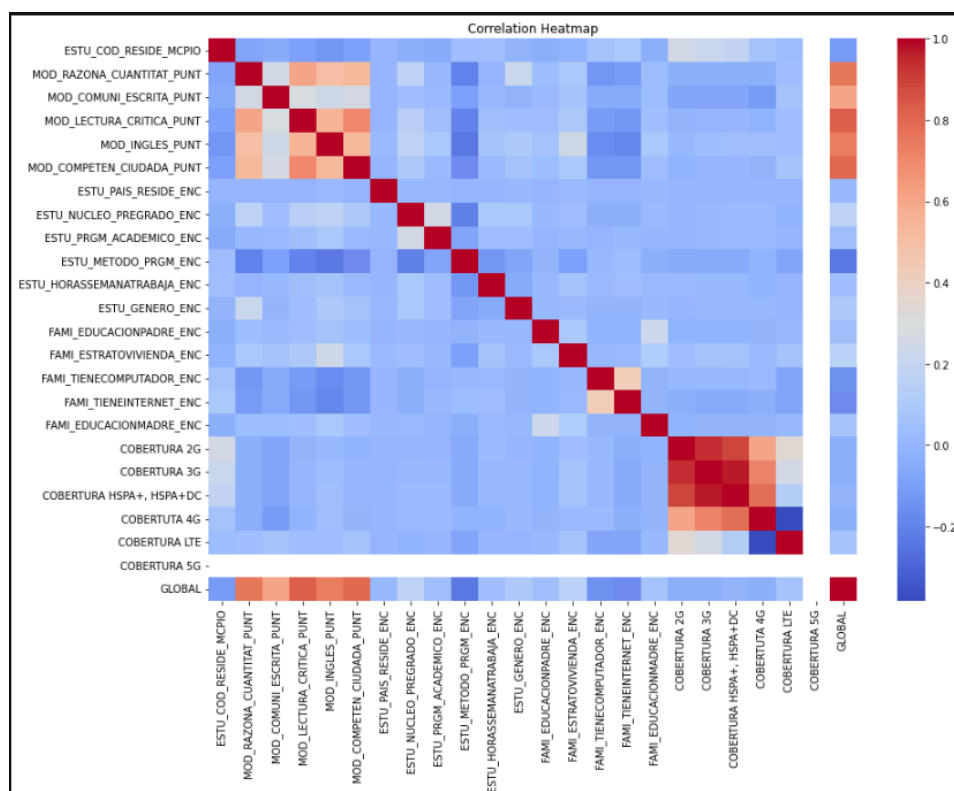
El proceso en cuestión se planteó de forma en la cual pudiéramos utilizar una columna o varias columnas en común entre los dos datasets. Para este caso en particular, tanto el dataset que contiene los resultados del SABER PRO, y el dataset que contiene la información de cobertura tienen en común dos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

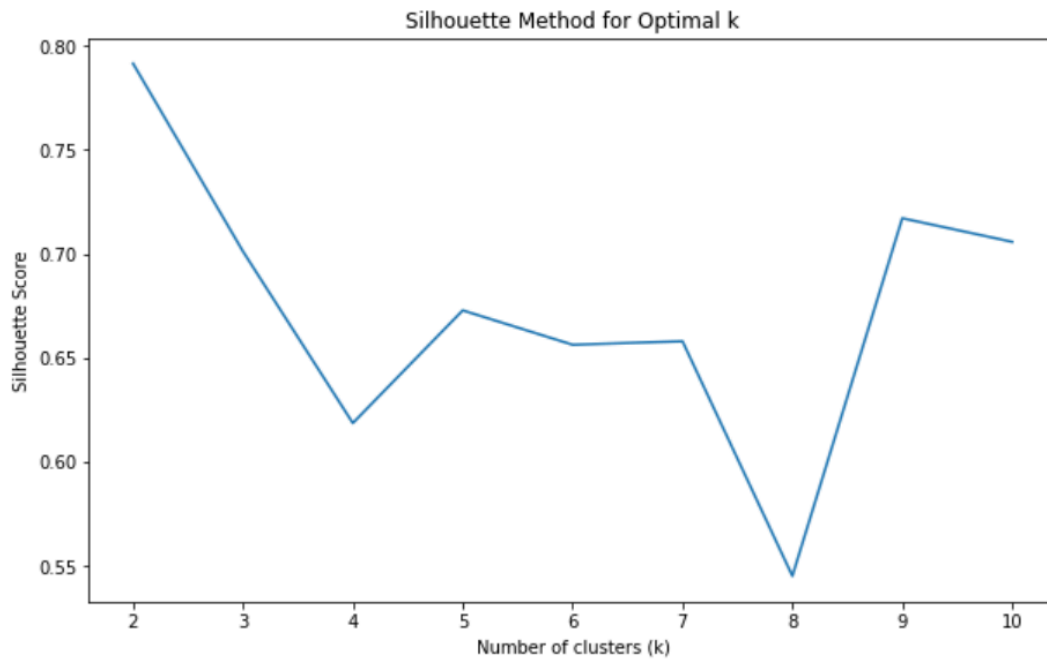
Técnica	Bibliotecas	Hiper-parámetros	Resultados
Regresión Lineal Multidimensional	pyspark.ml.regression	maxIter: 100 regParam: 0.0 elasticNetParam: 0.0 tol: 1e-6 solver: "auto" standardization: True fitIntercept: True weightCol: None	RMSE: 24.401951399879696 R2: 0.047324707925299414
Regresión con Random 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 Boost	pyspark.ml.regression	maxIter: 20 maxDepth: 3 stepSize: 0.1 impurity: "variance" featureSubsetStrategy: "auto" minInstancesPerNode: 1	RMSE: 23.90258687586869 R2: 0.08393186023917332

		minInfoGain: 0.0 maxBins: 32 seed: None	
--	--	---	--

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 todas 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.



Número óptimo de clusters con Elbow

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 superponen 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.

c) 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.

+++++ **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.*

+++++

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-de-partamento-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

