Juan Diego Torres Palomino - 201923251

Sergio Andrés Avendaño Serrano - 201923730

Proyecto 3: Producto de analítica de datos para la predicción del puntaje global del Saber 11 dado ciertos factores socioeconómicos.

Repositorio:

Las pruebas Saber 11 en Colombia son exámenes estandarizados diseñados para evaluar las habilidades de los estudiantes de grado 11 en áreas como matemáticas, ciencias sociales y lectura crítica. El Instituto Colombiano para la Evaluación de la Educación (ICFES) administra estas pruebas, las cuales tienen como objetivo medir el rendimiento académico al concluir la educación media.

Los resultados de las pruebas desempeñan un papel crucial en el proceso de admisión a instituciones de educación superior y también se utilizan para evaluar la calidad del sistema educativo a nivel nacional. Los resultados son proporcionados de manera individual, permitiendo a los estudiantes utilizarlos en sus solicitudes a instituciones de educación superior.

Estas pruebas se llevan a cabo en todo el país, posibilitando la comparación de estudiantes provenientes de diferentes regiones de Colombia. Este enfoque geográfico permite observar y clasificar los resultados de acuerdo con las condiciones socioeconómicas de cada individuo, contribuyendo así a una evaluación más completa y contextualizada del desempeño académico.

Con el fin de que las diferentes secretarias de educación de alcaldías y gobernaciones interesadas puedan evaluar los resultados de los estudiantes en sus departamentos y ver como los diferentes factores socioeconómicos afectan los resultados. Y, del mismo modo, se puedan implementar diferentes estrategias gubernamentales para aumentar la probabilidad de que los estudiantes de sus departamentos obtengan mayores resultados en las pruebas.

Con el fin de poder realizar esto, se realizó el siguiente análisis exploratorio de la base de datos. Se tenían 51 diferentes variables de las cuales se escogieron la siguientes:

'periodo': Periodo de los resultados

'cole\_depto\_ubicacion': Departamento de la sede

'fami\_tieneinternet': ¿Tiene internet?

'cole\_jornada': Jornada de la Sede

'cole\_bilingue': ¿Es colegio bilingüe?

'fami\_estratovivienda': Estrato del examinando

'punt\_global': Puntaje Global

Como esas variables son categóricas, se decidió reemplazar su categoría por un numero para entrenar al modelo. Del mismo modo, con la variable objetivo se decidió tomar intervalos para categorizar dicha variable. Esto fue al realizar cuatro categorías diferentes dependiendo del puntaje obtenido por cada estudiante.

Categoría 1: de 0 a 200

Categoría 2: de 201 a 300

Categoría 3: de 301 a 400

Categoría 4: de 401 a 500

Esto con el fin de discretizar la variable 'punt\_global'. A su vez, se generaron el siguiente objetivo general y los siguientes objetivos específicos.

Se decidió emplear los datos en Antioquia, Atlántico y Cundinamarca en el periodo de 2017 a 2019 para poder tener una gran variedad de situaciones socioeconómicas las cuales nos dan una perspectiva acertada de la realidad del país. Se uso este código para poder depurar los datos con SQL

SELECT "periodo", "cole\_depto\_ubicacion", "fami\_tieneinternet", "cole\_jornada", "cole\_bilingue", "fami\_estratovivienda", "punt\_global"

FROM "entradaproyect3"

WHERE

periodo BETWEEN 20170 AND 20195

AND cole\_depto\_ubicacion IN ('ANTIOQUIA', 'CUNDINAMARCA', 'ATLANTICO')

AND COALESCE("fami\_tieneinternet", '') <> ''

AND COALESCE("cole\_bilingue", '') <> ''

AND COALESCE("fami\_estratovivienda", '') <> '';

**Objetivo General:**

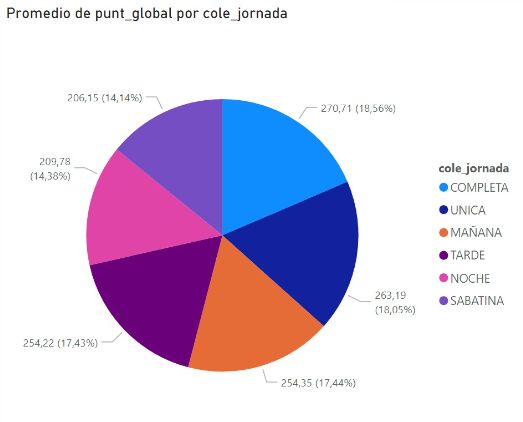
* Realizar una herramienta que pueda predecir un rango de puntaje global del Saber 11 de acuerdo a las condiciones socioeconómicas y así implementar estrategias gubernamentales para mejorarlas

**Objetivos Específicos:**

* Desarrollar un modelo Bayesiano
* Implementar un Dashboard funcional
* Emplear el MLflow para comparar modelos

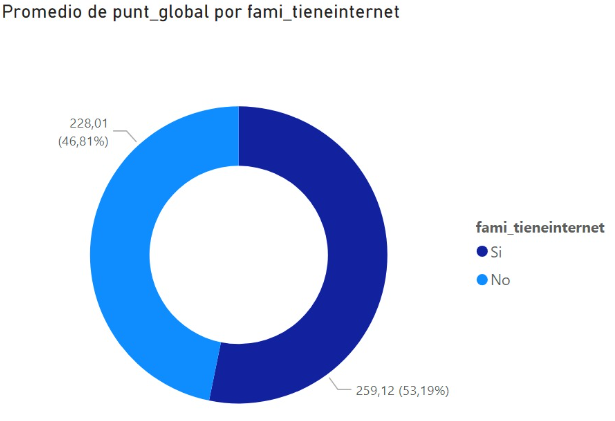
Ya teniendo todo lo anterior, se procede a graficar los datos de las variables seleccionadas para poder tener una pre visualización de los mismos

Grafica 1:



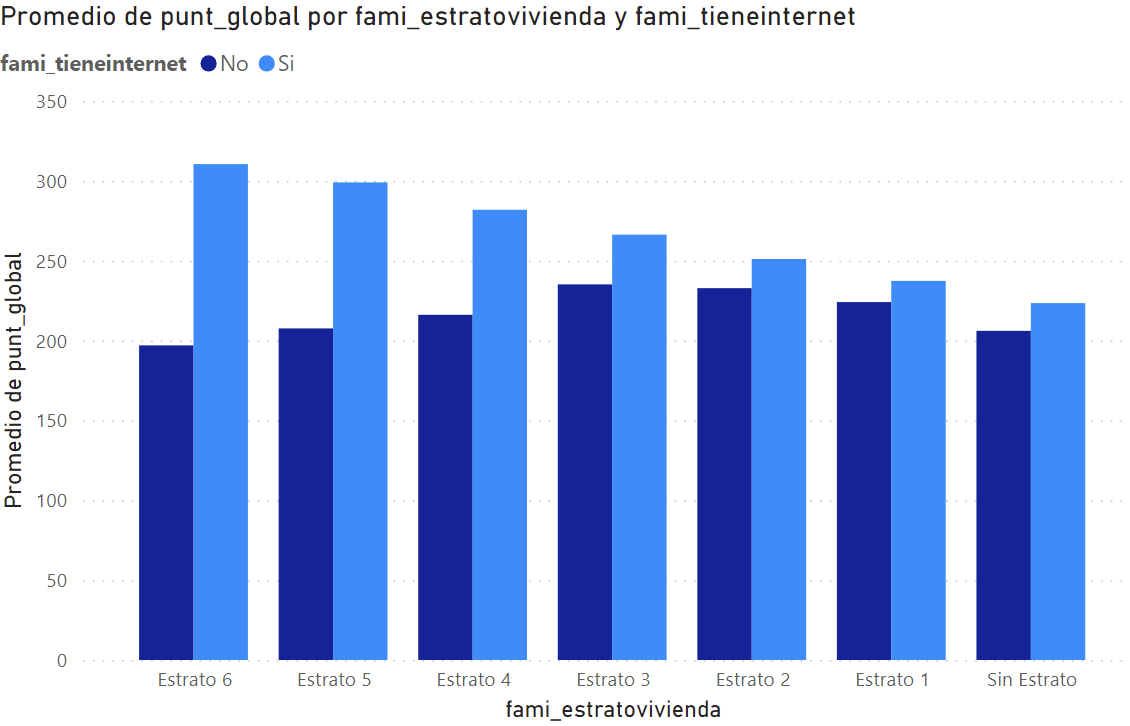
En esta grafica se puede apreciar el promedio de puntaje de los estudiantes de acuerdo a su jornada escolar.

Grafica 2:



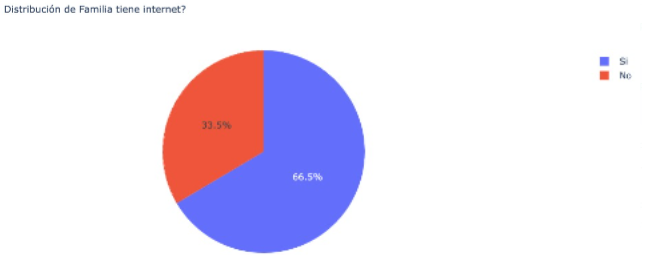
En esta grafica se aprecia el promedio de los puntajes obtenidos mostrando si tienen internet o no.

Gráfica3:



En esta gráfica 4 se puede ver el puntaje promedio de las personas de acuerdo a si tienen internet o no y del estrato en el que vive.

Gráfica 4:

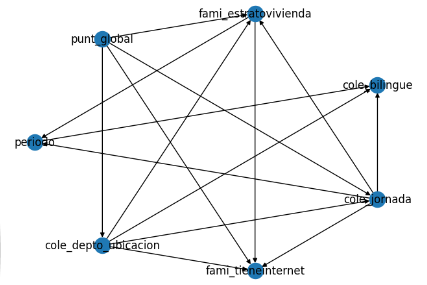


Acá se muestra la distribución de los estudiantes con y sin internet.

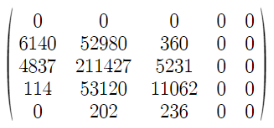
**Modelos:**

Se implemento una estrategia de aprendizaje de puntaje para poder generar un modelo de acuerdo a las variables.

Modelo Bic:

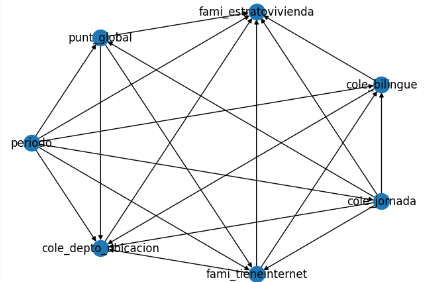


En este modelo se obtuvo una precisión de 0.6396 y un valor de BicScore de -1664070.63. Del mismo modo, se obtuvo la siguiente matriz de confusión

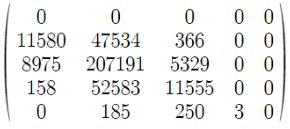


Esta matriz indica que la primera columna de todas las filas contiene exclusivamente ceros, señalando que no hubo instancias predichas como clase 0. En la segunda columna, se observa un número significativo de instancias predichas como clase 1 ([6140, 52980, 360, 0, 0]). La tercera columna presenta instancias predichas como clase 2 ([4837, 211427, 5231, 0, 0]). Sin embargo, las clases 3 y 4 también tienen instancias predichas, aunque todas estas instancias se encuentran en la columna 1, lo que podría indicar un problema en las predicciones. Este patrón plantea interrogantes sobre la precisión y consistencia de las predicciones para las clases 3 y 4, sugiriendo la necesidad de una revisión más detallada.

Modelo K2:

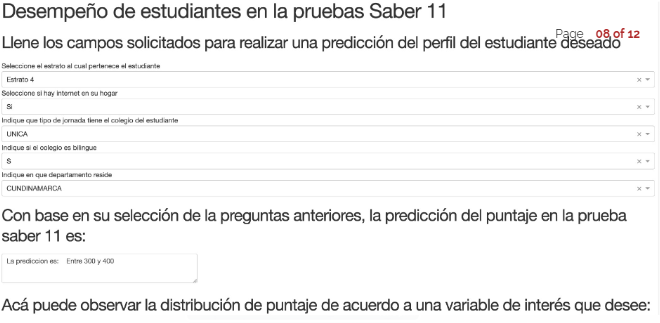


En este modelo se obtuvo una precisión de 0.6407y un valor de K2Score de -1650563.10. Del mismo modo, se obtuvo la siguiente matriz de confusión



Al igual que en la matriz anterior, la primera columna de todas las filas contiene exclusivamente ceros, indicando la ausencia de instancias predichas como clase 0. En comparación con la matriz previa, se observa un incremento en el número de instancias correctamente predichas como clase 1 (47534 frente a 52980). Asimismo, hay un aumento en el número de instancias correctamente predichas como clase 2 (207191 frente a 211427). La tercera columna presenta instancias predichas como clase 2 ([5329]). La cuarta columna muestra algunas instancias predichas como clase 3 ([0, 0, 0, 0, 3]). Este análisis sugiere una mejora en la precisión de las predicciones para las clases 1 y 2, aunque se destaca la necesidad de investigar y entender las instancias predichas como clase 3 en la cuarta columna.

DashBoard





Análisis y conclusiones:

Tomando en cuenta las métricas presentadas durante el informe, se puede ver que el modelo seleccionado fue el K2 ya que tiene una precisión mayor que el otro modelo. Bajo el criterio de BicScore y el K2Score se sabe que, entre más bajo estos valores, la red se ajusta mejor y el BicScore es menor que el K2Score. Pero el modelo seleccionado termina siendo el K2 ya que la precisión tiene más peso. Esto hace que las predicciones sean más exactas y precisas ayudando asi a el objetivo de que los entes gubernamentales tomen mejores decisiones para mejorar los resultados de las pruebas saber.