**Universidad de los Andes**

**Facultad de ingeniería**

**Departamento de ingeniería industrial**

**Analítica computacional**

**Proyecto 1**

**01/06/2023**

**Alumno: Pablo José Cortés Sanabria, Isaac Gregory**

**Código: 201111837, 201922386**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Reporte análisis exploratorio de datos**

*1.1 Manipulación de datos*

En conjunto de datos reportados en <https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-11/kgxf-xxbe> cuenta con aproximadamente 7 millones de registros. Con el fin de manipular dicha cantidad de datos, se usó el servicio S3 y ATHENA de AWS, para almacenamiento y consulta de los mismos. En la Ilustración 1 y en la Ilustración 2 se muestran las querys de SQL utilizadas para generar los sets de datos utilizados para visualizaciones y para el entrenamiento del modelo predictivo.

Texto

Descripción generada automáticamente

Ilustración . Query usada para consolidar datos promedio por departamento.

Adicionalmente, para realizar el análisis exploratorio de datos se uso una variante de la query de la Ilustración 2 , limitando el número de datos a 300000 como se ve en la Ilustración 3.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración . Query usada para extraer los datos para el modelo predictivo.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración . Query usada para generar el conjunto de datos usado en el EDA.

* 1. *Análisis exploratorio de datos*

En la se muestran algunas estadísticas de las variables numéricas del conjunto de datos; Puede observarse que la media del puntaje global se ubica en 251.97, con un desempeño promedio bajo, teniendo en cuenta que la máxima puntuación posible obtenida por un estudiante es 500 puntos. Un comportamiento similar se observa en cada una de las áreas (ciencias sociales, ciencias naturales, matemáticas, ingles y lectura critica), siendo el desempeño más bajo en promedio el de inglés.

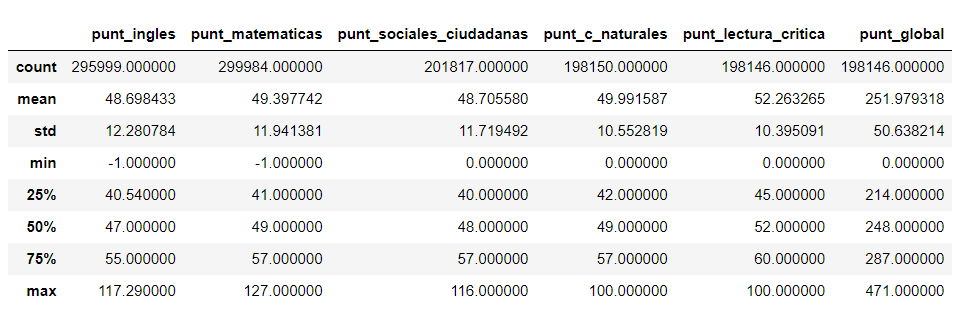


Tabla . Estadísticas resumidas.

Con el fin de realizar las visualizaciones y las predicciones requeridas, el set de datos fue transformado, eliminando datos faltantes, convirtiendo tipos de datos, entre otras operaciones. En la Tabla 2 se observa la estructura final de cada una de las columnas del set de datos.

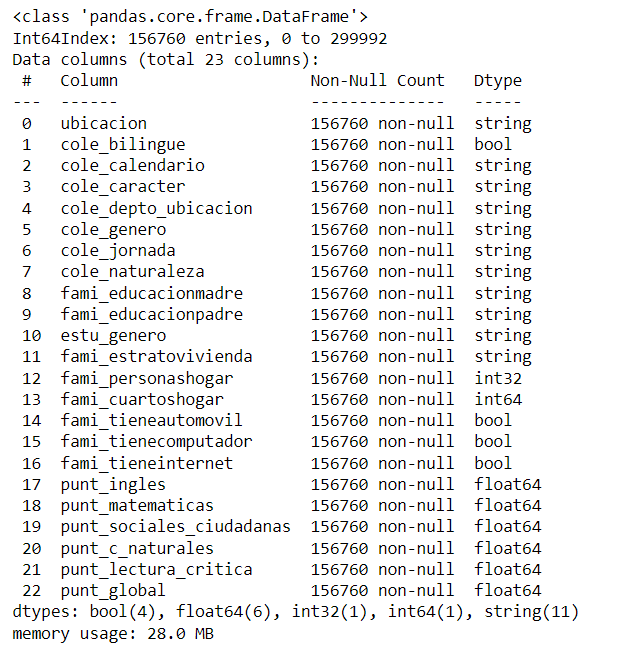


Tabla . Información de las diferentes columnas del conjunto de datos.

Adicionalmente, como era de esperar, existe una fuerte correlación entre los distintos puntajes de las 5 áreas y el puntaje global; esto puede apreciarse en la Ilustración 4; aprovechando lo anterior, a continuación se muestra el análisis exploratorio únicamente para la variable puntaje global, entendiendo que los resultados pueden extrapolarse fácilmente a los puntajes de las áreas individuales (ciencias sociales, ciencias naturales, matemáticas, inglés y lectura critica).

* + 1. *Dependencia del puntaje global respecto al bilingüismo*

En la Ilustración 5 se observa un diagrama de cajas para el puntaje global en función de si el colegio es bilingüe o no; puede observarse que los colegios bilingües tienden en general a tener un mejor desempeño en el puntaje global.

*1.2.2 Relación entre el puntaje global y el departamento*

En la se observa el diagrama de cajas para el puntaje global comparado con cada uno de los 32 departamentos del país, más el distrito capital. Puede observarse que los departamentos con un puntaje promedio más alto son Bogotá, Santander, Boyacá y Cundinamarca, mientras que los departamentos con peor desempeño en promedio son la Guajira, Amazonas, Vaupés y Choco.

Patrón de fondo

Descripción generada automáticamente

Ilustración . Correlación existente entre las variables tipo puntaje.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Relación entre el puntaje global y el bilingüismo del colegio.

Gráfico

Descripción generada automáticamente

Ilustración . Diagrama de cajas para el puntaje global en función del departamento.

* + 1. *Relación entre el calendario del colegio y el puntaje global*

En la se muestra el diagrama de cajas para el puntaje global, en función del calendario académico del colegio. Puede observarse claramente que los colegios con calendario B tienen un rendimiento superior en las pruebas saber.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Relación entre el puntaje global y el calendario académico.

* + 1. *Relación entre el género del colegio y el puntaje global*

En la Ilustración 8 se observa la relación entre el puntaje global y el genero del colegio; puede observarse que los colegios masculinos o femeninos tienen un rendimiento significativamente superior a los colegios mixtos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Relación entre el puntaje global y el genero del colegio.

* + 1. Influencia de la educación de los padres en el puntaje global

En la Ilustración 9 se muestra la influencia de la educación de los padres en los resultados globales del estudiante; Se observa que entre mayor sea el nivel académico alcanzado por los progenitores, mayor será el puntaje global obtenido por los alumnos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Relación entre el resultado global y el nivel educativo del padre del alumno.

* + 1. *Dependencia del puntaje global del acceso a internet.*

En la se muestra la relación entre el puntaje global y el acceso a internet del núcleo familiar del estudiante. Es claro que los estudiantes con acceso a internet tienden a tener mejores resultados en las pruebas saber.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Relación entre el acceso a internet y el puntaje global.

* + 1. *Dependencia del puntaje global del estrato socioeconómico*

En la se muestra el diagrama de violín para la variable puntaje global, en función del estrato socioeconómico del núcleo familiar. Es claro que entre mayor poder adquisitivo (y por ende mayor estrato socioeconómico), los estudiantes tienden a tener mejores resultados en las pruebas saber. Es notable sin embargo que en el estrato 6, la distribución presenta un área de densidad de probabilidad importante por debajo de los 250 puntos.

Gráfico, Gráfico de embudo

Descripción generada automáticamente

Ilustración . Diagrama de violín del puntaje global en función del estrato socioeconómico.

1. Modelo de predicción

En el presente trabajo, se propone un modelo de predicción basado en random forest para predecir el resultado del puntaje global de un estudiante en función de las siguientes variables.