Documentación tarea 2

1. Lectura de datos y limpieza de datos

El puntaje promedio de las mujeres fue de 142.13.

El puntaje promedio de los hombres fue de 145.64.

Aunque la diferencia no es muy grande, los hombres obtuvieron un puntaje global ligeramente superior en promedio. Por eso, el género se considera una variable importante.

Se analizaron los estudiantes según dos criterios:

Si son o no colombianos.

Si estudiaron o no en el extranjero.

Los puntajes promedio fueron:

No colombianos que sí estudiaron en el extranjero: 160.55

No colombianos que no estudiaron en el extranjero: 149.78

Colombianos que sí estudiaron en el extranjero: 158.98

Colombianos que no estudiaron en el extranjero: 149.78

Estudiar en el extranjero está asociado a un puntaje mayor, sin importar la nacionalidad. Por eso, ambas variables son valiosas para el análisis.

La **edad promedio** de los estudiantes fue de **28.96 años**.

La **edad máxima corregida** fue de **80 años**.

A mayor edad, el puntaje tiende a disminuir. La edad es una variable que claramente influye en el puntaje, especialmente en los grupos jóvenes donde el promedio es más alto.

Datos faltantes (nulos)

Se identificaron varias columnas con muchos datos nulos. Por ejemplo:

Columna	**% Nulos**
ESTU_PRESENTACIONCASA	99.99%
ESTU_CURSODOCENTESIES	94.49%
ESTU_CURSOIESEXTERNA	94.50%

Estas columnas no aportan información útil porque casi no tienen datos, por eso deben eliminarse del análisis.

2. Regresión

Los resultados del modelo de regresión indican un desempeño limitado, ya que el coeficiente de determinación (R^2) es de aproximadamente 0.30 tanto en entrenamiento como en prueba, lo que significa que el modelo solo explica el 30% de la variación en el puntaje global.

3. KNN

El mejor desempeño se obtuvo con 5 vecinos, alcanzando un R^2 de 0.4232 y el MSE más bajo de 316.83, lo que indica una mayor capacidad del modelo para explicar la variabilidad del puntaje global y cometer menos errores en las predicciones. A medida que se incrementa el número de vecinos, el desempeño del modelo disminuye notablemente

4. GBM

El modelo GBM logra explicar cerca del 30% de las diferencias en los datos, lo que quiere decir que funciona de forma moderada, pero no es perfecto. El error que comete al predecir es todavía alto, por lo que sus resultados no son muy precisos