Tarea #: 2
Tema: Regresión
Fecha entrega: 11:59 pm Mayo 14 de 2025
Objetivo: Aplicar los conceptos de KNN, regresión y GBM en datos reales.
Entrega: Crear una rama utilizando el mismo repositorio de la tarea 1, crear otra carpeta llamada tarea 2, solucionar el problema y crear un pull request sobre la master donde me debe poner como reviewer (entregas diferentes tienen una reducción de 0.5 puntos). Se puede utilizar chat GPT. No se puede copiar la salida de chatGPT y pegarla.

1 Regresión (40%)

Utilizar kaggle para descargar la base https://www.kaggle.com/competitions/pruebas-del-saber-2025/host/launch-checklist , el caso de uso es que basado en las condiciones del estudiantes vamos a predecir el puntaje que tendrá en las pruebas del saber en lecture critica “PUNT_GLOBAL”.

1. Realizar la exploración de los datos correlación, scatter plots, boxplots e histogramas:
   1.1. ¿Qué variables son importantes para predecir el valor?
   1.2. Existen nulos?, ¿cómo se deben imputar?
   1.3. Crear dummy variables para incluirlas en la correlación
   1.4. Crear una correlación, que variables tienen un efecto positivo en el puntaje y cuales un efecto negativo.
2. Divida los datos en training y testing
   2.1. Aplique las transformaciones más importantes a los datos. (Hint calcular la edad basada en la fecha de nacimiento, agrupar variables categóricas con mucha cardinalidad en grupos).
   2.2. Entrenar un modelos de regresión
   2.3. ¿Cuál es el mejor R squared? Cuál es el MAPE y el MSE.
3. Remueva las variables que nos son relevantes
4. Utilizando los datos de test medir el MAPE y el MSE de test. Qué tan diferentes son las métricas de training. (El menor error del grupo tiene un +1)
5. Describa en palabras que dice el modelo cuales son los principales hallazgos.

Hacer el envio a kaggle y poner el nombre Regression

2 Crear un modelo de KNN (20%)

Utilizar los datos para crear un modelo de KNN que permita predecir el puntaje por estudiante. Utilizar kaggle para descargar la base https://www.kaggle.com/t/efa882e3a6d94bf799278c56ef3c8317 y las mismas transformaciones dle punto anterior

1) Hacer pruebas con 5, 10, 20 y 30 vecinos. Seleccione el numero de vecinos basado en el error de test MSE.

vecinos	MSE train	MAPE train	MSE test	MSE train
5				
10				
20				
30				

2) Describa cual es mejor modelo entre la regresión o el knn.

Hacer el envio a kaggle y poner el nombre Knn con el numero de vecinos que mejor funciono en la predicción.

3 Crear un modelo de GBM (20%)

Entrenar un modelo de GBM y hacer la predicción. Cual es el MSE y MAPE para train y test.

Hacer el envio a kaggle y poner el nombre GBM

4 Crear un modelo de regresión logística (20%)

Utiliza los mismos datos del punto 1, crea una variable Y donde las personas con puntaje mayor a 172 tienen “1” y los demás “0” (‘1’ if PUNT_GLOBAL>172 else ‘0’), la variable Y representa los estudiantes sobresalientes, eliminar la variable PUNT_GLOBAL.
Con el dataset de training:
- Dividir los datos en training 80% y validación 20%.

- Entrenar una regresión logística, cuales son las variables más importantes?.
- Crear una matriz de confusión, cual es la precisión, cuál es el recall, y el accuracy.
- Calcular las mismas métricas para el dataset de validación.