Tarea #: 2
Tema: Aprendizaje No supervisado y Regresión
Fecha entrega: 11:59 pm Septiembre 18 de 2023
Objetivo: Aplicar los conceptos de PCA y regresión 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).

1 PCA (20%)

Cargar el data set de caras que está en la carpeta datos de la tarea 2 (ver notebook https://github.com/jdramirez/UCO_ML_AI/blob/master/src/notebook/PCA.ipynb):

1. Calcular la mean face. Que es la cara con el promedio de los pixeles y visualizarla.
2. Centrar los datos, utilizar PCA. ¿Cuántos componentes se deben utilizar para mantener el 90% de las características?. Crear una tabla para mostrar las primeras 5 caras utilizando, la mean face + los datos reconstruidos utilizando la primera componente, después con 5 componentes, después con las primeras 10 componentes, después con las componentes que explican el 90% de la varianza y por último con el numero de componentes que tiene el 99% de la varianza. ¿Qué se puede concluir de los resultados?

Cara original	MeanFace + 1 comp	MeanFace + 3 comp	MeanFace + 20 comp	MeanFace + 85% comp	MeanFace + 99% comp
1					
2					
3					
4					

2 K-means (20%)

Utilizar las 5 primeras componentes e implementar el algoritmo k-means sin librerías utilizando la distancia de valor absoluto (conocida como la norma 1), crear clase con métodos fit(aprender de los datos) y predict(predecir con los centroides el cluster de un nuevo dato).

1. Crear 7 clusters. Seleccione las 4 caras más cercanas al centroide de cada cluster, describa si son similares y porque estan cerca una de la otra.

Cluster	Centroide	1ra Cara cercana al cluster	2da Cara cercana al cluster	3ra cara más cercana al cluster	4ta cara más cercana al cluster	¿Qué tienen en común?
1						
2						
3						
4						
5						
6						
7						

3 Regresión (60%) .

Utilizar el dataset de la carpeta datos. 'Resultados_Saber_TyT_Gen_ricas_2020-1.csv' (ver origen), el caso de uso es que basado en las condiciones del estudiantes vamos a predecir el puntaje que tendrá en las pruebas del saber. Por supuesto no se pueden utilizar ninguna variable de puntaje en las variables a utilizar o datos que se generen después de presentar el examen. La variable objetivo es `MOD_INGLES_PUNT` que muestra el nivel de inglés.

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.