Ejercicio: Modelado Predictivo con Regresión Lineal y Regresión Logística

En este ejercicio, aplicarás dos técnicas fundamentales de regresión para resolver problemas de predicción tanto en variables continuas como categóricas. A continuación, se detallan los pasos que debes seguir para resolver el ejercicio:

1. Exploración de Datos (EDA)

- Realiza una exploración inicial del conjunto de datos. Esto incluye la inspección de la estructura del dataset (número de filas, columnas, tipos de datos), la detección de valores nulos o atípicos (outliers), y el análisis descriptivo de las variables (medidas estadísticas básicas como media, mediana, desviación estándar, etc.).
- Visualiza las relaciones entre las variables mediante gráficos (distribuciones, diagramas de dispersión, matrices de correlación, etc.) para identificar posibles patrones o tendencias.
- Comenta cualquier observación relevante sobre la distribución de las variables, la presencia de colinealidad o cualquier otra característica que pueda influir en los modelos.

2. División del Dataset

- Divide el dataset en dos subconjuntos: training y test. Utiliza una proporción adecuada, como 80%-20% o 70%-30%, dependiendo del tamaño del dataset.
- Asegúrate de que la división sea aleatoria y que mantenga la representatividad de la variable objetivo en ambos subconjuntos.
- Justifica la elección de la proporción utilizada y evalúa si es necesario aplicar técnicas adicionales, como la estratificación, en función de la distribución de la variable objetivo.

3. Modelado con Regresión Lineal

- Construye un modelo de regresión lineal simple utilizando el subconjunto de training para predecir una variable continua. Asegúrate de incluir la normalización de las variables si es necesario.
- Evalúa el desempeño del modelo en el subconjunto de test utilizando métricas adecuadas (MSE, RMSE, R^2, etc.).
- Discute los resultados obtenidos y cualquier limitación del modelo de regresión lineal en este contexto.

4. Modelado con Regresión Logística

- Aplica la técnica de regresión logística sobre el conjunto de datos de training para predecir una variable categórica (binaria). Recuerda que este método es adecuado para problemas de clasificación.
- Evalúa el desempeño del modelo en el conjunto de test utilizando métricas adecuadas (precisión, recall, F1-score, curva ROC, etc.).

- Analiza los coeficientes obtenidos para entender la influencia de las variables predictoras en la probabilidad del resultado.

5. Comparación de Resultados

- Compara los resultados obtenidos de cada modelo (Regresión Lineal para predicción continua y Regresión Logística para clasificación) con un modelo trivial (baseline). Específicamente:

- **Para Regresión Lineal:** Compara el rendimiento del modelo predictivo con un modelo que simplemente predice siempre la media de la variable objetivo. Evalúa la precisión del modelo en relación a este baseline utilizando métricas como el R^2 o el MSE.
- **Para Regresión Logística:** Compara el rendimiento del modelo con un modelo que asigna una probabilidad igual a $1/n$ a cada clase, donde n es el número de clases. Evalúa este baseline usando métricas como la exactitud, el AUC-ROC, o el log-loss.

- Analiza las ventajas y desventajas de cada modelo en función de su rendimiento en comparación con el baseline y su aplicabilidad al tipo de problema planteado. Considera también aspectos como la interpretabilidad, la complejidad y la facilidad de implementación.