Taller Árboles de decisión

Instrucciones Generales:

1. **Entregables:**
   - Un cuaderno de Jupyter con el código completo.
2. **Herramientas Necesarias:**
   - Python 3.x instalado.
   - Librerías: `pandas`, `scikit-learn`, `matplotlib`, `seaborn`.
   - Jupyter Notebook o Google Colab.
   - Acceso a internet para descargar datasets.

Fase 1: Selección del Dataset y Planteamiento del Problema

1. **Selecciona un dataset real**:
   - Utiliza el dataset de Kaggle asociado a tu grupo:

     Grupo 1: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

     Grupo 2: https://www.kaggle.com/datasets/ivansher/nasa-nearest-earth-objects-1910-2024

     Grupo 3: https://www.kaggle.com/datasets/priyamchoksi/spotify-dataset-114k-songs

     Grupo 4: https://www.kaggle.com/datasets/yasserh/titanic-dataset/data
2. **Define el problema a resolver**:
   - **Objetivo**: Formular una pregunta clara que puedas responder utilizando un modelo de árbol de decisión.

3. **Cargar los datos**:
    - Descarga y carga el dataset en Python.

Fase 2: Preprocesamiento de Datos

Instrucciones:

1. **Limpieza de datos**:
    - Verifica si existen valores faltantes en el dataset. Si los hay, decide cómo tratarlos (eliminar o imputar).
    - **Ejemplo**: `df.isnull().sum()` te muestra cuántos valores faltan por columna.
2. **Codificación de variables categóricas**:
    - Si el dataset contiene variables categóricas (como 'Género' o 'Clase'), deberás convertirlas en variables numéricas utilizando `pd.get_dummies` o `LabelEncoder`.
3. **División del dataset**:
    - Divide los datos en un conjunto de entrenamiento y otro de prueba (80%/20% o 70%/30%) usando `train_test_split` de `scikit-learn`.

Fase 3: Implementación Básica del Árbol de Decisión

Instrucciones:

1. **Entrenar el modelo**:
    - Usa `DecisionTreeClassifier` o `DecisionTreeRegressor` de `scikit-learn` para entrenar el árbol de decisión con los datos de entrenamiento.
2. **Realiza predicciones** sobre los datos de prueba.
3. **Evalúa el rendimiento del modelo**:
    - Para clasificación, utiliza métricas como precision y matriz de confusión.
    - Para regresión, calcula el error cuadrático medio (MSE) o el error absoluto medio (MAE).

Fase 4: Visualización e Interpretación del Modelo

Instrucciones:

1. **Visualiza el árbol de decisión**:
   - Utiliza `plot_tree` para visualizar la estructura del árbol y analizar cómo el modelo toma decisiones.
2. **Analiza la importancia de las características**:
   - Examina qué características del dataset tienen mayor impacto en las decisiones del modelo.

Fase 6: Comparación con Otros Modelos

Instrucciones:

1. **Compara el rendimiento** del árbol de decisión con otros modelos vistos en clase.
2. **Discute las fortalezas y debilidades** del árbol de decisión frente a otros modelos.