Heart Attack Classification Assignment

Objective:

You are tasked with using machine learning classifiers to predict the likelihood of a heart attack. You will be working with a heart dataset to create a classification model using two algorithms: Random Forest and Decision Tree. The key focus will be on proper data preparation, handling missing values, and outlier cleaning before applying the classifiers.

Assignment Instructions:

1. Data Understanding:

- **Explore the dataset**: Familiarize yourself with the structure and contents of the heart dataset. Pay attention to the types of features (e.g., **numerical**, **categorical**), the target variable (e.g., **output**), and any potential issues in the data (**missing values**, **outliers**, **duplicates, Type conversion**).
- **Key columns**: Look out for important features like age, cholesterol levels, blood pressure, and other medical indicators that might predict heart attacks.

2. Data Quality Checks:

- **Check for missing values**: Identify if there are any missing values in the dataset. Use methods like `.isnull()` to inspect the dataset.
- **Handle missing values**: Depending on the nature of the missing data, decide how to handle it. You can either:
    - Replace missing values (e.g., with mean, median, mode).
    - Remove rows or columns with a high percentage of missing data.

3. Outlier Detection and Removal:

- **Visualize outliers**: Use boxplots or scatterplots to identify outliers in numerical features (e.g., age, cholesterol levels).
- **Remove or cap outliers**: Choose appropriate techniques to deal with outliers, such as:
    - Removing extreme values.
    - Capping outliers at a certain threshold (e.g., using the IQR method).

4. Feature Selection:

- **Select relevant features**: Decide which features should be used for training the classification model. Use correlation analysis, feature importance, or domain knowledge to guide your decision.

5. Model Training:

- **Split the data**: Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
- **Train two classifiers**:
    1. **Decision Tree Classifier**:
        - Use the `DecisionTreeClassifier` from `sklearn` to train the model.
    2. **Random Forest Classifier**:
        - Use the `RandomForestClassifier` from `sklearn` to train the model.

6. Model Evaluation:

- **Evaluate performance**: After training both classifiers, compare their performance using metrics such as:
    - **Accuracy**
    - Precision (optional)
    - Recall (optional)
    - F1-Score (optional)
    - Confusion matrix (optional)
    - ROC-AUC curve (optional)

7. Conclusions:

- **Compare models**: Summarize the performance of the Decision Tree and Random Forest classifiers. Which one performs better on the heart dataset? Explain why.
- **Insights**: Provide insights based on the model’s predictions and the features that had the most impact.

Dataset Columns Explanation:

- **Age** : Age of the patient
- **Sex** : Sex of the patient
- **exang**: exercise-induced angina (1 = yes; 0 = no)
- **ca**: number of major vessels (0-3)
- **cp** : Chest Pain type chest pain type:
  Value 1: typical angina
  Value 2: atypical angina
  Value 3: non-anginal pain
  Value 4: asymptomatic
  trtbps : resting blood pressure (in mm Hg)
- **chol** : cholestoral in mg/dl fetched via BMI sensor
- **fbs** : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- **rest_ecg** : resting electrocardiographic results:
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- **thalach** : maximum heart rate achieved
- **target** : 0= less chance of heart attack 1= more chance of heart attack

Trainer: Mohammad AlJadallah
Phone Number: +962786616104