**Name:** Aayush Avinash Bokde  
**Class:** CSE(AI) B div  
**Roll No.:** 282016  
**PRN No.:** 22310316  
**Subject:** Machine Learning

**Assignment 2 – Data Analysis and Model Building Report**

**Objective**

The objective of this assignment is to explore and analyze the Heart Disease dataset using Python. The tasks include computing summary statistics, visualizing feature distributions, cleaning and transforming the data, and building a classification model to predict heart disease.

**a) Summary Statistics**

Summary statistics offer a foundational understanding of the dataset’s characteristics:

* **Minimum & Maximum:** Identifies the range of values for each numerical feature.
* **Mean:** Average value of each feature.
* **Range:** Computed as (Maximum - Minimum) to represent variability.
* **Standard Deviation & Variance:** Indicators of dispersion across data values.
* **Percentiles (25th, 50th, 75th):** Show how data is distributed across features.

These metrics are essential for identifying outliers, skewed distributions, and general patterns in data spread.

**b) Feature Distributions (Histograms)**

Histograms were generated for all numerical features to visualize:

* The shape of each feature's distribution (e.g., symmetric, skewed).
* The presence of outliers or irregularities.
* The overall concentration and spread of values.

These plots help guide decisions on feature scaling, encoding, and treatment of extreme values.

**c) Data Cleaning, Integration, Transformation, and Model Building**

**Data Cleaning**

* **Missing Values Handling:**
  + Column 'Ca': Missing entries filled with the median of the column.
  + Column 'Thal': Missing values replaced with the mode.
* **Irrelevant Column Removal:**
  + Dropped 'Unnamed: 0' as it doesn’t aid model learning.
* **Duplicate Rows:** Checked and confirmed there are no duplicate entries.
* **Encoding Categorical Variables:**
  + Applied Label Encoding to 'ChestPain' and 'Thal'.
  + Target variable 'AHD' was encoded as: **1 = Yes**, **0 = No**.

**Data Integration**

Only a single dataset was used, so no data merging was required. However, data was checked for uniformity and consistency.

**Data Transformation**

* Selected numeric features ('Age', 'RestBP', 'Chol', 'MaxHR', 'Oldpeak') were standardized using StandardScaler to normalize values.
* This step supports algorithms sensitive to scale differences, although Random Forest does not require it.

**Data Modeling (Classification)**

* **Algorithm Used:** Random Forest Classifier
  + An ensemble method that builds multiple decision trees and combines their results for improved accuracy.
* **Data Split:**
  + 80% of the data was used for training and 20% for testing.
* **Evaluation Metrics:**
  + **Accuracy Score:** Evaluates overall model performance.
  + **Classification Report:** Includes metrics like Precision, Recall, and F1-score for both classes.
  + **Confusion Matrix:** Visualizes true vs predicted labels to assess model error types.

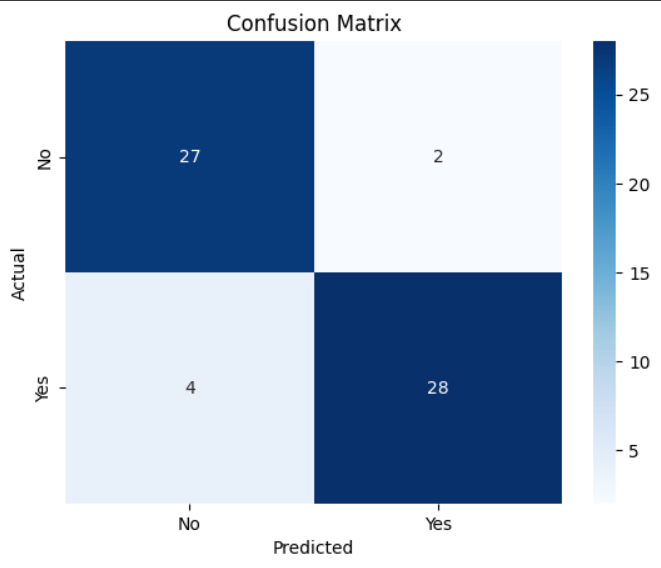
**Feature Importance**

Feature importances were extracted and visualized using a bar chart to interpret which features most influenced the model's predictions. This helps in feature selection and understanding model reasoning.

**Conclusion**

This assignment completed a full machine learning pipeline, including:

* **Statistical exploration** and **feature visualization**,
* **Handling missing values**, **feature encoding**,
* **Normalization** and **classification modeling**,
* Achieving reliable results using **Random Forest Classifier**.



The model effectively predicted heart disease and provided insights into the most influential features, demonstrating the value of structured machine learning workflows on health-related datasets.