**Name**: Shubham Dahane  
**Class**: CSE(AI) B div  
**Roll No.**: 282015  
**PRN No.**: 22310312  
**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 using histograms, cleaning and transforming data, and building a classification model to predict heart disease.

**a) Summary Statistics**

Summary statistics provide essential information about the spread and central tendency of the dataset's features:

* **Minimum & Maximum**: Identify the smallest and largest values for each column.
* **Mean**: Represents the average value.
* **Range**: Computed as (Maximum - Minimum) to understand variability.
* **Standard Deviation & Variance**: Measure the dispersion or spread in the dataset.
* **Percentiles**: Calculated at 25th, 50th, and 75th to understand the distribution of data points.

These statistics help detect outliers, understand feature scales, and identify skewness or anomalies in the data.

**b) Feature Distributions (Histograms)**

Histograms were plotted for all numeric features to visualize:

* The shape of feature distributions (e.g., normal, skewed)
* Potential outliers or unusual values
* Concentration and spread of data

This step aids in preliminary understanding of data behavior and guides preprocessing decisions.

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

**Data Cleaning**

* **Missing Values**:
  + 'Ca' column: Missing values replaced with **median**.
  + 'Thal' column: Missing values filled using the **mode**.
* **Dropping Irrelevant Features**: Removed Unnamed: 0 column as it does not contribute to prediction.
* **Duplicates**: Checked and found no duplicate rows.
* **Encoding Categorical Features**:
  + Applied **Label Encoding** to categorical features like ChestPain and Thal.
  + Target variable AHD mapped to binary values: **1 = Yes**, **0 = No**.

**Data Integration**

Only one dataset was used; however, data integration ensures uniformity if multiple datasets are merged.

**Data Transformation**

* Applied **StandardScaler** to normalize selected numeric columns: ['Age', 'RestBP', 'Chol', 'MaxHR', 'Oldpeak'].
* This helps models that rely on distances, though **Random Forest** doesn't strictly require it.

**Data Modeling (Classification)**

* **Model Used**: Random Forest Classifier
  + An ensemble method that builds multiple decision trees and aggregates their outputs for better accuracy and robustness.
* **Train-Test Split**:
  + Split data into **80% training** and **20% testing** sets.
* **Evaluation Metrics**:
  + **Accuracy Score**: Measures overall correctness of predictions.
  + **Classification Report**: Includes **Precision, Recall, and F1-score** for each class.
  + **Confusion Matrix**: Provides a detailed look at correct vs incorrect predictions.

**Feature Importance**

The importance of each feature in the model's decision-making process was plotted, helping interpret which features most influence predictions.

A group of blue and white graphs

AI-generated content may be incorrect.

**Conclusion**

This assignment successfully demonstrated a complete pipeline of data analysis and modeling:

* From summary statistics and visual insights,
* To handling missing values and categorical encoding,
* Followed by classification modeling using Random Forest.

A blue squares with white text

AI-generated content may be incorrect.

The model achieved reliable performance and highlighted key predictive features for heart disease, showcasing the effectiveness of structured ML workflows in real-world datasets.