**Name: Aditee Mohanty**

**Class: CSE(AI) B div**

**Roll no.: 282023**

**PRN No.: 22310631**

**Subject: Machine Learning**

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

**Objective**

This assignment involves performing basic data analysis operations such as computing summary statistics, visualizing data, data cleaning, transformation, integration, and building a classification model using Python.

**a) Summary Statistics**

**Summary statistics** are essential descriptive measures that provide insights into the distribution and spread of data in each feature. The key statistics computed include:

* **Minimum and Maximum**: Show the smallest and largest values in each column.
* **Mean**: The average of the values.
* **Range**: Difference between the maximum and minimum values.
* **Standard Deviation**: Indicates the amount of variation or dispersion in a dataset.
* **Variance**: The square of the standard deviation, representing data spread.
* **Percentiles**: Help in understanding the distribution by identifying cut-off points below which a percentage of observations fall (e.g., 25th, 50th, 75th).

These statistics help in identifying outliers, skewness, and potential issues in the data before applying any modeling techniques.

**b) Feature Distributions (Histograms)**

**Histograms** are graphical representations that show the frequency distribution of numeric data. Each bar in a histogram represents the number of data points that fall within a particular range of values (known as a bin).

Histograms help in:

* Understanding the distribution (normal, skewed, etc.)
* Identifying outliers or anomalies
* Visualizing data concentration and spread

In this assignment, all numeric features are plotted using histograms to understand their distributions.

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

**Data Cleaning**

This is a critical step that ensures data quality and reliability for analysis. The following techniques were used:

* **Handling Missing Values**: Missing numeric values were filled using the mean of the respective columns.
* **Encoding Categorical Variables**: Categorical (non-numeric) columns were converted to numeric form using **Label Encoding** so that they can be used in machine learning models.

**Data Integration**

Although not explicitly shown here (as only one dataset was used), data integration refers to combining data from multiple sources and ensuring consistency across datasets.

**Data Transformation**

Transformation steps include:

* Converting categorical variables into numerical format
* Normalizing or scaling the data (if required, though not used here because tree-based models like Random Forest don’t require feature scaling)

**Data Modeling (Classification)**

For classification, a **Random Forest Classifier** was used:

* **Random Forest** is an ensemble learning method that operates by constructing multiple decision trees and outputs the class that is the mode of the classes from individual trees.
* **Train-Test Split**: The dataset was split into 80% training and 20% testing to evaluate model performance.
* **Model Evaluation**:
  + **Accuracy Score** measures how many predictions were correct.
  + **Classification Report** includes metrics like **precision**, **recall**, and **F1-score**, which provide detailed evaluation of the model's performance across different classes.

**Conclusion**

This assignment demonstrated how to perform comprehensive data analysis, including statistical summarization, visualization, preprocessing, and classification modeling. The Random Forest model provided a reliable classification framework, making it a strong candidate for structured tabular data like the one used in this heart disease dataset.