Assignment 3
Name: Srinidhi Soundarrajan
Roll No: 281064
Batch: A3

Statement

In this assignment, we aim to:
a) Visualize the data using Python by plotting the graphs for Assignment No. 1 and 2. Consider a suitable dataset. Use Scatter Plot, Bar Plot, Box Plot, Pie Chart, and Line Chart.

Objective

1. Understand how to compute summary statistics for a dataset using Python.
2. Learn how to visualize data distributions using histograms.
3. Develop skills in data preprocessing, transformation, and integration to prepare data for machine learning models.
4. Build a classification model using the cleaned and transformed dataset.
5. Implement various data visualization techniques to represent the dataset effectively.

Resources Used

- **Software:** Visual Studio Code
- **Libraries:** Pandas, Matplotlib, Seaborn, Scikit-Learn, NumPy

Dataset Used

The dataset used in this assignment is a **Heart Disease Prediction Dataset**, which contains patient-related health attributes like age, cholesterol levels, chest pain type, and more. This dataset is used for classification tasks related to heart disease diagnosis.

Data Analysis and Preprocessing

1. Data Collection and Exploration

- **Load Data:** The dataset is imported into a Pandas DataFrame using pd.read_csv().

- **Check for Missing Values:** We use isnull().sum() to identify missing data.
- **Summarize Data:** We compute summary statistics using describe(), which includes mean, standard deviation, minimum, and maximum values.

2. Data Cleaning and Transformation

- **Handle Missing Values:** We use imputation techniques (e.g., mean or median replacement).
- **Convert Categorical Features:** Label encoding is applied to categorical columns such as Chest Pain Type.
- **Normalize Data:** Features like cholesterol levels and age are normalized to improve model performance.

3. Summary Statistics Computation

- Compute **minimum, maximum, mean, range, standard deviation, variance, and percentiles** for numerical features using describe() and other Pandas functions.

4. Data Visualization

- **Bar Plot:** Displays the distribution of chest pain types.
- **Histogram:** Visualizes the distribution of the age feature.
- **Scatter Plot:** Shows the relationship between cholesterol levels and age.
- **Box Plot:** Identifies outliers in cholesterol levels.
- **Pie Chart:** Represents the proportion of patients with and without heart disease.
- **Line Chart:** Plots trends in a selected numerical feature.

5. Model Building (Classification)

- **Train-Test Split:** The dataset is split into training and testing sets using train_test_split().
- **Model Selection:** A **Decision Tree Classifier** is implemented using DecisionTreeClassifier().
- **Model Evaluation:** Accuracy and confusion matrix are used to assess model performance.

Advantages of Pandas and Machine Learning in Data Analysis

1. **Pandas** simplifies data handling and provides powerful functions for data cleaning and transformation.
2. **Visualization tools** help in understanding feature distributions and data patterns.
3. **Machine learning models** allow for automated classification and prediction.

Disadvantages of Pandas and Data Processing

1. **Memory-intensive** operations may slow down processing for large datasets.
2. **Preprocessing complexity** increases with unstructured data.

Conclusion

This assignment provided insights into using Pandas for structured data analysis, including reading, preprocessing, and summarizing datasets. We visualized the dataset using histograms, bar plots, scatter plots, and other charts. Additionally, we built a classification model to predict heart disease based on the given features. These techniques are essential for real-world data analysis and machine learning applications.