Assignment 3

Name: Sakshi Vyavahare
Roll No: 281023
Batch: A1

Statement

In this assignment, we aim to:
a) Visualize the data using Python by plotting graphs for Assignment 1 and 2.
b) Consider a suitable dataset and use various visualization techniques:

- Scatter Plot
- Bar Plot
- Box Plot
- Pie Chart
- Line Chart

Objectives

1. Compute summary statistics for a dataset using Python.
2. Visualize data distributions using histograms.
3. Develop data preprocessing, transformation, and integration skills for machine learning.
4. Build a classification model using a cleaned and transformed dataset.
5. Implement diverse data visualization techniques to effectively represent data.

Resources Used

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

---

Dataset Used

Heart Disease Prediction Dataset
This dataset includes patient health attributes such as age, cholesterol, chest pain type, etc., and is used for classification tasks to predict heart disease diagnosis.

Data Analysis and Preprocessing

1. Data Collection and Exploration

- Loaded dataset using pd.read_csv()
- Checked missing values using isnull().sum()
- Summary statistics obtained using describe()

2. Data Cleaning and Transformation

- Missing values handled using mean/median imputation
- Categorical columns (e.g., chest pain type) encoded using LabelEncoder()
- Features normalized to improve model performance

3. Summary Statistics Computation

Used functions like min(), max(), mean(), std(), var(), and quantile() to compute:

- Minimum, Maximum
- Mean, Range
- Standard Deviation, Variance
- Percentiles (25th, 50th, 75th)

Data Visualization

1. Bar Plot: Distribution of chest pain types
2. Histogram: Age distribution
3. Scatter Plot: Relationship between age and cholesterol
4. Box Plot: Outlier detection for cholesterol levels
5. Pie Chart: Heart disease vs. no heart disease proportion
6. Line Chart: Trend of cholesterol levels with age

Visualizations created using Matplotlib and Seaborn.

Model Building (Classification)

- Dataset split using train_test_split()
- Model: Decision Tree Classifier from Scikit-learn
- Evaluated using:
    - Accuracy Score
    - Confusion Matrix
    - Classification Report

Advantages of Pandas and Machine Learning

1. Simplified data manipulation and cleaning
2. Easy-to-use visualization tools for pattern recognition
3. ML models offer automated prediction and classification

Disadvantages

1. High memory usage for large datasets
2. Preprocessing is more complex for unstructured data

Conclusion

This assignment strengthened our understanding of structured data analysis using Pandas. We explored real-world datasets, performed preprocessing, and visualized them using various graphs. Additionally, we built a classification model to predict heart disease. These skills are foundational for real-world data science and machine learning projects.