Assignment 3

Name: Yashraj Nimbalkar

Roll No: 281039
Batch: A2

Statement

The focus of this assignment is to:

a) Generate visual representations using Python for datasets from Assignment 1 and 2.
b) Work with a suitable dataset and apply multiple data visualization techniques such as:

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

Objectives

1. Generate statistical summaries for datasets using Python libraries.
2. Represent data distributions through histogram visualizations.
3. Enhance skills in data preprocessing, integration, and transformation for machine learning readiness.
4. Construct a classification model after preparing the dataset.
5. Apply various graphing methods to illustrate patterns and insights clearly.

Tools and Libraries Used

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

Dataset Description

Heart Disease Prediction Dataset
The dataset includes medical attributes like age, cholesterol levels, and chest pain type, and is intended for predicting the likelihood of heart disease using classification techniques.

Data Analysis and Preparation :

1. Data Import and Overview
   - Loaded dataset using pd.read_csv()
   - Checked for null values using isnull().sum()
   - Used describe() to generate statistical summaries

2. Data Cleaning & Transformation
   • Handled missing data with techniques like mean or median filling
   • Encoded categorical features (e.g., chest pain type) using LabelEncoder()
   • Normalized the data to ensure balanced input for the model

3. Statistical Metrics Computed
   Used built-in functions like min(), max(), mean(), std(), var(), and quantile() to find:
   • Minimum and Maximum values
   • Average and Range
   • Standard Deviation and Variance
   • Percentiles (25th, 50th, 75th)

Data Visualization

1. Bar Graph: Display of chest pain type distribution
2. Histogram: Frequency distribution of patient ages
3. Scatter Plot: Correlation between age and cholesterol
4. Box Plot: Identifying outliers in cholesterol values
5. Pie Chart: Comparison between patients with and without heart disease
6. Line Graph: Cholesterol level trends across different age groups

Visualizations were created using Matplotlib and Seaborn.

Building the Classification Model

• Split the dataset using train_test_split()
• Used a Decision Tree Classifier from Scikit-learn
• Evaluated model using:

- Accuracy Score
- Confusion Matrix
- Classification Report

Strengths of Pandas and Machine Learning

1. Makes handling and modifying data straightforward
2. Allows creation of intuitive graphs for better data understanding
3. Enables automation of predictions with machine learning models

Limitations

1. High memory consumption for very large datasets
2. Requires thorough preprocessing for complex or unstructured data

Conclusion

Through this assignment, we developed a deeper grasp of structured data processing using Pandas. We explored and cleaned real-world datasets, created various types of data visualizations, and built a classification model to predict heart disease. These practices are vital for solving real-life data science and AI-related challenges.