Assignment 2

Name: Yashraj Nimbalkar

Roll No: 281039
Batch: A2

Statement

This assignment involves the following tasks:

a) Compute summary statistics for all numerical attributes (e.g., min, max, mean, range, standard deviation, variance, percentiles).
b) Visualize data using histograms to understand distribution patterns.
c) Perform data cleaning, merging, transformation, and build a classification model for prediction.

Objective

1. Learn to derive descriptive statistics using Python.
2. Understand how to interpret data distribution using visual tools like histograms.
3. Practice essential data preparation steps to improve data quality.
4. Build and evaluate a classification model using the cleaned dataset.

Resources Utilized

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

Understanding Pandas and Data Handling

Pandas is a highly capable Python package that simplifies data analysis and manipulation. It offers two core data structures:

- **Series** – One-dimensional, like a column in a table.
- **DataFrame** – A two-dimensional structure, similar to an Excel spreadsheet.

Main Features of Pandas

- Load data from formats like CSV, Excel, etc.
- Clean and format data by dealing with nulls, duplicates, or inconsistencies.
- Produce statistical summaries and generate visualizations.
- Integrate basic machine learning methods such as classification or regression.

Functions and Tools Used

1. pd.read_csv() – Imports CSV data into a DataFrame.
2. describe() – Provides detailed statistical summary of numerical columns.
3. hist() – Draws histograms for understanding distribution.
4. fillna() – Fills missing entries with a chosen value (e.g., mean or median).
5. LabelEncoder() – Transforms categorical data into numeric labels.
6. train_test_split() – Splits dataset into training and testing sets.
7. LogisticRegression() – Builds a classification model for prediction.

Approach and Methodology

1. **Loading and Exploring Data**
   • **Dataset Used**: For example, student performance or healthcare prediction dataset.
   • Inspect structure, identify datatypes, and locate missing or inconsistent values.
2. **Data Preprocessing**
   • Fill missing data using central values (mean or median).
   • Remove duplicates, standardize column formats, and handle outliers.
3. **Generating Summary Statistics**
   • Use .describe(), .min(), .max(), .std(), etc., to extract key insights.
4. **Data Visualization with Histograms**
   • Plot histograms using .hist() or sns.histplot() to view feature distributions.
5. **Transformations and Feature Engineering**
   • Encode categorical fields to numeric.
   • Run correlation checks to select impactful features.
6. **Combining Data**
   • Merge or concatenate multiple datasets using functions like merge() or concat() where required.
7. **Building the Classification Model**
   • Split dataset using train_test_split().
   • Train a logistic regression model.
   • Evaluate model using metrics such as accuracy, confusion matrix, and classification report.

Benefits of Pandas and Machine Learning

1. Simplifies data exploration and preparation.
2. Easy to generate charts and graphs for better understanding.
3. Enables predictive modeling using machine learning algorithms.

Limitations

1. May consume significant memory with large datasets.
2. Complex or unstructured data may require extra preprocessing effort.

Conclusion

This assignment helped solidify core skills in data analysis using Pandas. We worked through the full pipeline — from data inspection and cleaning to building a machine learning model. These techniques form the base of modern data science, equipping learners to tackle real-world analytics problems effectively using Python.