Assignment 2
Name: Yash Maske
Roll No: 282007
Batch: B1

Statement

In this assignment, we aim to:
a) Generate and display summary statistics for each numerical feature (e.g., minimum, maximum, mean, range, standard deviation, variance, percentiles).
b) Visualize data distributions using histograms.
c) Carry out data cleaning, integration, transformation, and build a classification model.

Objective

1. Learn to compute descriptive statistics using Python.
2. Understand how to represent data distributions using histograms.
3. Practice essential data preparation techniques to ensure data quality.
4. Develop a classification model using a processed dataset for prediction tasks.

Resources Used

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

Introduction to Pandas and Data Analysis

Pandas is a powerful and flexible Python library designed for data analysis. It provides intuitive data structures such as Series (1D) and DataFrame (2D), enabling users to analyze and manipulate structured data efficiently.

Key Capabilities

- Reading datasets from CSV, Excel, and other formats.
- Preprocessing data by cleaning, transforming, and handling missing values.
- Generating statistical insights and visual summaries.
- Applying basic machine learning models for classification or regression.

Basic Functions Used

1. pd.read_csv() – Loads data from CSV into a DataFrame.
2. describe() – Computes count, mean, standard deviation, and other stats.

3. hist() – Plots histograms for visualizing data distribution.
4. fillna() – Replaces missing data with computed values (mean/median).
5. LabelEncoder() – Converts categorical features into numerical format.
6. train_test_split() – Divides data into training and testing subsets.
7. LogisticRegression() – Applies a classification algorithm for prediction.

Methodology

1. Data Collection and Exploration

- **Dataset Used:** A dataset related to student grades or health prediction.
- **Initial Steps:** Load the dataset and inspect columns, data types, and missing entries.

2. Data Preprocessing

- Handle missing entries using mean or median values.
- Clean the data by removing duplicates, standardizing formats, and identifying outliers.

3. Summary Statistics Computation

- Use describe() and functions like .min(), .max(), .std() to calculate key statistics for each numerical column.

4. Feature Visualization using Histograms

- Use DataFrame.hist() or sns.histplot() to generate visual plots and understand how data is distributed.

5. Data Transformation and Feature Engineering

- Convert categorical variables into numeric ones using encoders.
- Perform correlation analysis to retain meaningful features.

6. Data Integration

- If multiple datasets are involved, merge them logically using merge() or concat() while ensuring consistency.

7. Model Building (Classification)

- Split the dataset using train_test_split().
- Train a classification model like Logistic Regression.
- Evaluate performance using metrics such as accuracy score, confusion matrix, and classification report.

Advantages of Pandas and ML in Analysis

1. Pandas offers efficient tools for preprocessing and exploring datasets.
2. Visualization libraries like Matplotlib and Seaborn enhance interpretability.
3. ML models allow for pattern detection and accurate predictions.

Disadvantages

1. High memory usage for large-scale datasets.
2. Complex preprocessing steps are needed for messy or unstructured data.

Conclusion

This assignment helped explore the core functionalities of the Pandas library for performing statistical analysis, data cleaning, and visualization. It also introduced a basic machine learning workflow by implementing a classification model. These steps are crucial for anyone pursuing data science or analytics, providing the foundation for handling real-world problems with Python.