Assignment 2
Name: Srinidhi Soundarrajan
Roll No: 281064
Batch: A3

Statement

In this assignment, we aim to:
a) Compute and display summary statistics for each feature available in the dataset (e.g., minimum value, maximum value, mean, range, standard deviation, variance, and percentiles).
b) Illustrate the feature distributions using histograms.
c) Perform data cleaning, data integration, data transformation, and data model building (e.g., classification).

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.

Resources Used

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

Introduction to Pandas and Data Analysis

Pandas is an open-source Python library designed for data manipulation and analysis. It provides data structures like Series (1D) and DataFrame (2D), making it easier to analyze and process structured data efficiently.

Key Capabilities

- Importing data from various formats (CSV, Excel, SQL).
- Data preprocessing (cleaning, transforming, handling missing values).
- Statistical and analytical operations (descriptive statistics, visualization).

- Data modeling (classification, regression, clustering).

Basic Functions Used

1. pd.read_csv() – Reads data from a CSV file into a DataFrame.
2. describe() – Generates descriptive statistics for numerical columns.
3. hist() – Creates histograms to visualize data distributions.
4. fillna() – Handles missing values using strategies like mean, median, or mode imputation.
5. LabelEncoder() – Converts categorical variables into numerical representations for machine learning models.
6. train_test_split() – Splits data into training and testing sets.
7. LogisticRegression() – Builds a regression model.

Methodology

1. Data Collection and Exploration

- **Collect Data:** Obtain a dataset (e.g., heart attack prediction, customer segmentation, or student performance).
- **Explore Data:** Load the dataset into a Pandas DataFrame and examine its structure, including feature types, missing values, and unique categories.

2. Data Preprocessing

- **Handle Missing Values:** Use techniques like mean, median, or mode imputation, or remove records with excessive missing data.
- **Data Cleaning:** Remove duplicates, correct formatting inconsistencies, and handle outliers.

3. Summary Statistics Computation

- Compute the **minimum, maximum, mean, range, standard deviation, variance, and percentiles** for each numerical feature using describe() and additional statistical functions.

4. Feature Visualization using Histograms

- Use hist() and Seaborn's sns.histplot() to visualize feature distributions.

5. Data Transformation and Feature Engineering

- **Feature Encoding:** Convert categorical features into numeric using LabelEncoder() or OneHotEncoder().
- **Feature Selection:** Identify and retain relevant features based on correlation analysis.

6. Data Integration

- Merge data from multiple sources if required, ensuring consistency across datasets.

7. Model Building (Regression)

- Split the dataset into training and testing sets using train_test_split().
- Train a regression model (e.g. Logistic Regression).
- Evaluate the model's accuracy and performance using metrics like confusion matrix and classification report.

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 the use of Pandas for structured data analysis, including reading, preprocessing, and summarizing datasets. We explored feature distributions using histograms and implemented classification models to make predictions. These skills form the foundation for advanced data science projects, highlighting the efficiency of Python libraries in handling real-world datasets.