Assignment 2

Name : Aditya Deshmukh
PRN : 22310607
Roll No : 281024
Batch : A2

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 build a data model (e.g., classification).

Objective

1. Utilize Python and Pandas to analyse and preprocess structured data.
2. Develop skills in exploratory data analysis, statistical computation, and data visualization.
3. Implement data transformation techniques to prepare datasets for machine learning.
4. Train and evaluate a classification model using machine learning algorithms.

Resources Used

- **Software:** VS Code
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

---

Introduction to Pandas and Machine Learning

Pandas is a widely-used Python library for data manipulation and analysis, providing easy-to-use data structures and functions. It simplifies handling structured data and integrates well with other libraries like NumPy and Matplotlib.

Key Functionalities Used:

1. **Data Handling with Pandas**
    - pd.read_csv(): Reads data from a CSV file into a DataFrame.
    - describe(): Computes summary statistics for numerical columns.
    - drop(): Removes specified columns or rows from a DataFrame.
2. **Data Visualization with Matplotlib and Seaborn**
    - sns.histplot(): Generates histograms to illustrate feature distributions.
    - plt.show(): Displays plotted graphs.
3. **Data Preprocessing & Model Building**
    - train_test_split(): Splits the dataset into training and testing sets.
    - StandardScaler(): Standardizes features by removing the mean and scaling to unit variance.
    - RandomForestClassifier(): A machine learning model used for classification tasks.
    - confusion_matrix(), accuracy_score(), precision_score(), recall_score(), f1_score(): Metrics to evaluate model performance.

Methodology

1. Data Collection and Exploration

- **Dataset Used:** *admission.csv*
- **Features:** Various attributes related to student admission predictions.
- **Initial Steps:**
    - Loaded the dataset using Pandas.
    - Displayed the first few rows to understand the structure.
    - Dropped irrelevant columns (e.g., Serial No.).

2. Data Cleaning and Preprocessing

- **Handled Missing Values:** Checked for and addressed missing data.
- **Performed Data Transformations:** Standardized numerical features using StandardScaler().
- **Visualized Data Distributions:** Used histograms to analyze feature distributions.

3. Data Model Building

- **Split the Data:** Divided into training and testing sets using an 80-20 ratio.
- **Built a Classification Model:** Used RandomForestClassifier() for classification.
- **Evaluated Performance:** Computed accuracy, precision, recall, F1-score, and confusion matrix.

Advantages of Pandas & Machine Learning

1. Simplifies data handling and manipulation.
2. Provides extensive statistical and analytical functions.
3. Enables efficient data visualization.
4. Machine learning enhances predictive capabilities.

Disadvantages

1. Memory-intensive when handling large datasets.
2. Training machine learning models can be computationally expensive.

Conclusion

In this assignment, we performed structured data analysis, preprocessing, and classification using Pandas and Scikit-learn. We computed summary statistics, visualized feature distributions, cleaned the data, and built a classification model. This assignment strengthened our understanding of data analysis and machine learning, providing a foundation for more advanced predictive modeling tasks.