Name: Prachi Shedge
Roll no: 281050
Batch: A2
Assignment 4

Statement:

In this assignment, we have to perform:
a) Read data from different formats.
b) Indexing, selecting, and sorting data.
c) Describing attributes of data and checking data types of each column.
d) Counting unique values of data, formatting each column, and converting variable data types.
e) Identifying missing values and filling them appropriately.

Objective:

1. This assignment aims to introduce you to the Pandas library and its basic functions. The library provides functionality for reading different file formats such as CSV and Excel.
2. Additionally, it familiarizes users with data cleaning and preprocessing techniques.
3. Enhances skills in handling and analyzing data using Python.

Resources Used:

- **Software used:** Visual Studio Code
- **Libraries used:** Pandas, NumPy, Scikit-learn

Introduction to Pandas:

Pandas is a powerful and widely-used open-source Python library for data manipulation and analysis. It provides easy-to-use data structures and functions, making it an essential tool for working with structured data. The main data structures in Pandas are:

- **Series:** A one-dimensional labeled array capable of holding any data type.
- **DataFrame:** A two-dimensional labeled data structure with columns of potentially different types.

Some Basic Functions Used in the Program:

1. **pd.read_csv():** Reads data from a CSV file into a DataFrame.
2. **head():** Displays the first few rows of the DataFrame.
3. **sort_values():** Sorts the DataFrame by the values of a specified column.
4. **describe():** Generates descriptive statistics for numerical columns.
5. **unique():** Returns an array of unique values in a column.

Methodology:

1. **Data Collection and Exploration:**
    - Load the dataset into a pandas DataFrame and explore its structure.
    - Identify missing or erroneous values.
2. **Data Preprocessing:**
    - Handle missing values appropriately.
    - Perform data cleaning tasks such as removing duplicates and correcting erroneous entries.
3. **Feature Engineering:**
    - Encode categorical variables into numerical format.
    - Standardize numerical features.
4. **Model Training and Evaluation:**
    - Split data into training and testing sets.
    - Train a **Random Forest Classifier** and evaluate performance using metrics such as accuracy, precision, recall, and F1-score.

Program Implementation:

1. Importing Libraries:

2. Loading the Dataset:

3. Checking for Missing Values:

4. Feature Engineering:

5. Encoding Categorical Variables:

6. Data Preprocessing:

7. Splitting the Data:

8. Model Training:

9. Predictions and Evaluation:

10. Printing the Results:

Advantages:

1. Pandas provides powerful data structures like Series and DataFrame.
2. Scikit-learn simplifies machine learning model implementation.
3. Random Forest is robust against overfitting and handles missing data effectively.

Disadvantages:

1. Pandas may consume significant memory when handling large datasets.
2. Model training with a large dataset may require substantial computational power.

Conclusion:

This assignment introduced us to data handling using Pandas and machine learning techniques using Scikit-learn. We explored data cleaning, preprocessing, feature engineering, and model training with Random Forest. This hands-on experience provided a solid foundation for further data science and machine learning projects.