Assignment 1

Name: Srinidhi Soundarrajan

Roll No: 281064

Batch: A3

Statement

In this assignment, we aim to:
a) Read data from different formats.
b) Perform indexing, selecting, and sorting data.
c) Describe data attributes, check data types of columns, and count unique values.
d) Format columns, convert variable data types (e.g., from long to short, and vice versa).
e) Identify and fill missing values.

Objective

1. Introduce the Pandas library and its core functionality for working with structured data, including reading file formats like CSV and Excel.
2. Familiarize users with data cleaning and preprocessing techniques.
3. Develop skills for handling data in various formats, enhancing proficiency in data analysis and manipulation.

Resources Used

- **Software:** Visual Studio Code
- **Library:** Pandas

Introduction to Pandas

Pandas is a versatile and widely used open-source Python library for data manipulation and analysis. It simplifies handling structured data through its easy-to-use data structures and functions.

1. Key Data Structures

- **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.

2. Capabilities

Pandas enables operations such as:

- Loading data from various file formats (CSV, Excel, SQL databases).
- Manipulating data (sorting, filtering, grouping).
- Performing statistical and analytical tasks.

Basic Functions Used

1. **pd.read_csv()** – Reads data from a CSV file into a DataFrame.
2. **head()** – Displays the first few rows of the DataFrame to provide an overview of the data.
3. **sort_values()** – Sorts the DataFrame by the values of a specified column (e.g., 'Age').
4. **describe()** – Generates descriptive statistics for numerical columns (e.g., count, mean, std dev, min, max).
5. **unique()** – Returns an array of unique values in a column, useful for identifying distinct categories in categorical data.

Methodology

1. Data Collection and Exploration

- **Collect Data:** Obtain a dataset (e.g., heart attack prediction) with relevant features such as age, gender, blood pressure, cholesterol levels, etc.
- **Explore Data:** Load the dataset into a Pandas DataFrame and examine its structure, including the number of samples, features, data types, and missing/erroneous values.

2. Data Preprocessing

- **Handle Missing Values:** Use strategies like imputation (mean, median, mode) or remove rows/columns with excessive missing data.
- **Data Cleaning:** Remove duplicates, correct erroneous entries, and ensure consistency in formatting.

3. Feature Engineering

- **Feature Selection:** Identify relevant features for prediction tasks using domain knowledge or techniques like correlation analysis.
- **Feature Encoding:** Convert categorical variables into numerical formats using one-hot encoding or label encoding to prepare them for machine learning algorithms.

Advantages of Pandas

1. Simple and user-friendly, making it popular among data analysts.
2. Provides powerful data structures like Series and DataFrame.
3. Offers extensive functionality for data manipulation and analysis.

Disadvantages of Pandas

1. May consume significant memory when working with large datasets.
2. Limited interoperability with programming languages other than Python.

Conclusion

This assignment introduced the Pandas library, an essential tool for data manipulation and analysis in Python. Through practical exercises, we explored basic functions such as reading data from various formats, organizing and describing data, and handling missing values. These foundational skills will serve as a strong starting point for more advanced data analysis projects, demonstrating the efficiency and accessibility of Pandas for simplifying complex data tasks.