Assignment 1
Name: Sakshi Vyavahare
Roll No: 281023
Batch: A1

Statement

In this assignment, we focus on the following tasks:
a) Import data from different file formats.
b) Apply indexing, data selection, and sorting techniques.
c) Analyze data attributes, determine data types, and count distinct values.
d) Modify and reformat columns, and convert data types when needed.
e) Detect and manage missing data efficiently.

Objective

1. Introduce the **Pandas** library and its powerful tools for handling structured data, including reading files like CSV and Excel.
2. Learn essential data cleaning and transformation methods.
3. Gain hands-on experience in handling and processing real-world datasets to build a solid foundation in data analysis.

Resources Used

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

Introduction to Pandas

Pandas is a powerful open-source Python library used for data manipulation and analysis. It provides flexible data structures and functions that make it easy to work with structured datasets.

1. Core Data Structures

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

2. Key Features

Pandas supports operations like:

- Reading data from multiple sources (CSV, Excel, SQL).
- Data filtering, grouping, reshaping, and sorting.

- Conducting descriptive and statistical analysis.

Basic Functions Used

1. pd.read_csv() – Loads data from a CSV file into a DataFrame.
2. head() – Displays the first few entries in the dataset.
3. sort_values() – Sorts the dataset based on column values.
4. describe() – Provides summary statistics for numerical data.
5. unique() – Returns unique values in a specific column.

Methodology

1. Data Import and Overview

- **Dataset Used:** A sample dataset (e.g., diabetes prediction, patient health records) with features such as age, glucose level, BMI, etc.
- **Load and Explore:** Import the dataset using Pandas and examine its shape, structure, column types, and presence of missing values.

2. Data Cleaning and Preparation

- **Handling Null Values:** Replace missing data using imputation (mean/median/mode) or remove them if necessary.
- **Cleaning:** Remove duplicates, fix incorrect values, and standardize formats.

3. Feature Engineering

- **Selection:** Choose important features for analysis based on correlation or domain knowledge.
- **Encoding:** Convert categorical data into numerical form using methods like one-hot or label encoding.

Advantages of Pandas

1. Easy-to-learn and efficient for data manipulation.
2. Robust structures like DataFrames and Series.
3. Broad range of features for data analysis tasks.

Disadvantages of Pandas

1. Can be slow and memory-intensive with large datasets.
2. Primarily works within the Python ecosystem; limited support for other languages.

Conclusion

This assignment provided an introduction to the Pandas library for data handling in Python. We practiced reading, organizing, cleaning, and summarizing data through hands-on tasks. These foundational concepts will be crucial for future projects involving data science, enabling efficient data analysis using Python.