Name: Dev Bhangale
Roll no: 281059
Batch: A3

Assignment 1

Statement:
Q. In this assignment we have to do
a) Read data from different formats
b) Indexing and selecting, sorting data
c) Describe attributes of data, checking data types of each column
d) Counting unique values of data, format of each column, converting variable data type (e.g.from long to short, vice versa)
e) Identifying missing values and fill in the missing values.

Here's a paraphrased version of your assignment:

Objective:

1. This assignment introduces the Pandas library and its fundamental functions. Pandas provides tools for reading and processing various file formats, including CSV and Excel.
2. It also covers essential data cleaning and preprocessing techniques.
3. The goal is to enhance our ability to manage and analyze data in different formats, improving overall proficiency in data manipulation.

Resources Used:

- **Software:** Google Colab
- **Library:** Pandas

Introduction to Pandas:

1. Pandas is a powerful open-source Python library widely used for data analysis and manipulation.
2. It offers user-friendly data structures and functions, making it an essential tool for structured data handling.
3. The two core data structures in Pandas are **Series** and **DataFrame**:
    - **Series**: A one-dimensional labeled array capable of holding various data types.

- **DataFrame**: A two-dimensional labeled data structure with multiple columns, each potentially containing different data types.

4. These structures enable users to perform various operations, including loading data from different sources (CSV, Excel, SQL databases), filtering, sorting, grouping, and executing statistical or analytical tasks.

Basic Pandas Functions Used in the Program:

1. **pd.read_csv()** – Reads data from a CSV file into a Pandas DataFrame.
2. **head()** – Displays the first few rows of a DataFrame, providing a quick overview of the dataset.
3. **sort_values()** – Sorts the DataFrame based on a specific column (e.g., 'Age'), arranging the data in ascending order.
4. **describe()** – Generates statistical summaries for numerical columns, including count, mean, standard deviation, minimum, and maximum values.
5. **unique()** – Returns an array of unique values in a column, helping to identify distinct categories in categorical data.

Methodology:

1. **Data Collection and Exploration:**
   - Gather a heart attack prediction dataset, ensuring it contains key attributes like age, gender, blood pressure, and cholesterol levels.
   - Load the dataset into a Pandas DataFrame and analyze its structure, including the number of records, feature types, and missing or erroneous values.
2. **Data Preprocessing:**
   - **Handling Missing Values:** Identify and address missing values using methods such as imputation (mean, median, or mode) or removing incomplete rows/columns.
   - **Data Cleaning:** Remove duplicate entries, correct errors, and maintain data consistency.
3. **Feature Engineering:**
   - **Feature Selection:** Choose relevant features for heart attack prediction based on domain knowledge and statistical methods (e.g., correlation analysis).
   - **Feature Encoding:** Convert categorical variables into numerical representations using one-hot encoding or label encoding for compatibility with machine learning algorithms.

Advantages of Pandas:

1. User-friendly and widely adopted due to its ease of use.
2. Provides powerful data structures like Series and DataFrame.
3. Offers extensive functionality for efficient data manipulation.

Disadvantages of Pandas:

1. Can consume significant memory when handling large datasets.

2. Primarily designed for Python, limiting its interoperability with other programming languages.

Conclusion:

This assignment provided a fundamental introduction to the Pandas library, a crucial tool for data manipulation in Python. We explored its essential functions, including reading and organizing data, handling missing values, and performing descriptive analysis. Through practical applications, we learned how Pandas simplifies complex data tasks, making analysis more efficient and accessible. The foundational skills acquired in this assignment will be valuable for tackling more advanced data analysis projects in the future.