Name: Rushikesh Kolekar
Roll No: 282020
Batch: B3

Practical 1

Problem Statement:

Q :- Perform the following operations using R/Python on suitable data sets:

a) Read data from different formats (like csv, xls)

b) Find Shape of Data

c) Find Missing Values

d) Find data type of each column

e) Finding out Zero's

f) Indexing and selecting data, sort data,

g) Describe attributes of data, checking data types of each column,

h) counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa)

Objective:

1. This assignment introduces the Pandas library and its fundamental capabilities, such as reading different file formats, including CSV and Excel.
2. It enhances our understanding of data cleaning and preprocessing techniques.
3. It strengthens our ability to manage data in various formats, improving our analytical and manipulation skills.

Resources Used:

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

Introduction to Pandas:

Pandas is a widely-used, open-source Python library designed for efficient data manipulation and analysis. It provides intuitive data structures and functions that streamline the handling of structured datasets.

Key Components of Pandas:

- **Series:** A one-dimensional labeled array capable of storing various data types.

- **DataFrame:** A two-dimensional labeled structure where columns can contain different data types.

These structures allow users to load data from sources like CSV files, Excel sheets, and SQL databases while enabling operations such as sorting, filtering, grouping, and statistical analysis.

Basic Functions Demonstrated:

1. **pd.read_csv():** Imports data from a CSV file into a Pandas DataFrame.
2. **head():** Displays the first few rows of the DataFrame for a quick overview.
3. **sort_values():** Sorts the DataFrame based on a chosen column (e.g., arranging 'Age' in ascending order).
4. **describe():** Provides statistical insights, including count, mean, standard deviation, minimum, and maximum values for numeric columns.
5. **unique():** Extracts distinct values from a column, helping identify unique categories.

Methodology:

1) Data Collection and Exploration:

- **Data Acquisition:** Obtain a dataset for heart attack prediction containing key features like age, gender, blood pressure, cholesterol levels, etc.
- **Data Examination:** Load the dataset into a Pandas DataFrame, inspect its structure, determine feature count, data types, and check for anomalies or missing values.

2) Data Preprocessing:

- **Handling Missing Data:** Identify missing values and address them using techniques such as mean, median, or mode imputation. Alternatively, remove rows or columns with excessive missing data.
- **Data Cleaning:** Remove duplicate records, correct inconsistencies, and standardize data formats for uniformity.

3) Feature Engineering:

- **Feature Selection:** Choose essential attributes for predicting heart attacks using domain expertise, correlation analysis, or feature importance ranking.
- **Feature Encoding:** Convert categorical variables into numerical representations using methods like one-hot encoding or label encoding to prepare data for machine learning models.

Advantages:

1. **Ease of Use:** Pandas is user-friendly, making data analysis more accessible.
2. **Robust Data Structures:** Series and DataFrames simplify data handling.
3. **Extensive Functionality:** Supports a wide array of operations, from basic manipulation to advanced analysis.

Disadvantages:

1. **High Memory Consumption:** Large datasets may lead to excessive memory usage.

2. **Limited Interoperability:** Pandas' tight integration with Python may pose compatibility challenges with other programming languages.

Conclusion:

This assignment provided an introduction to the Pandas library, a vital tool for data manipulation and analysis in Python. We explored its core functions, including data import, organization, description, and handling of missing values. Through practical exercises, we developed a foundational understanding of Pandas, which will be instrumental in tackling more advanced data analysis tasks in the future.