Roll. No. 31118

Date- 10-Jan-22

Assignment 01

Title: Data Wrangling-I

Problem Statement:

Perform the following operations using Python on any open source dataset:

1> Import all required python libraries.

2> Read an opensource data from web. Provide a clear description of data & it's source

3> Load the data set into pandas data frame.

4> Data preprocessing- checking for missing values in data using pandas isnull(), describe() function to get some initial statistics.

5> Check dimensions of data frame.

6> Data formatting & data normalization- summerize the type of variable by checking the data types (i.e. character, numeric, integer, factor & logical) of the variables in the data set. If variables are not in the correct data type, apply proper conversions.

Objectives:

1> To understand data wrangling-I & it's importance in the data science cycle.

2> To understand & implement data preprocessing & data formatting functions using pandas python library.

Software & Hardware Requirements:

Processor- Intel i-5 8th gen.

64-bit windows operating system.

python 3.8 &Jupyter notebook.

Theory:

Data Wrangling:

1> It ~~involve~~ is a process which involves taking data from it's orignal state ~~to~~ to a format where we can perform meaningful analysis on it.

2> In practice there are 3 common tasks involved in the process:

- i> data cleaning
- ii> data transformation
- iii> data enrichment

3> data cleaning: It involves removing null values & storing data in correct data types.

4> data transformation: It involves changing the struture of data as per downstream analysis

Pandas:

1> Pandas is an open-source library that provides high performance data manipulation in Python.

2> Data analysis requires lot of processing such as restructuring, cleaning & merging etc.

3> There~~a~~ are different tools available for data processing such as numpy, scipy. Pandas is prefered because of it's speed, simplicity & expressiveness. compared to other tools.

Methodology:

1> Import the Required libraries

2> Read & load csv into pandas dataframe.

3> Preprocess the data. - It involves handling missing values & related tasks.

The functions at this step:

1> isnull()

2> ~~dropna()~~ dropna()

3> describe().

4> Data formatting & normalization - checks for appropriate data-types

The functions used in this step:

1> info() astype('type-name')

5> Convert categorical variables into quantitative variables

functions used. get_dummies()

Conclusion:

In this assignment lab, we has successfully implements data preprocessing, formatting & normalization using Pandas Python library.