DSBDAL

Assignment - 1

Omkar Gaikwad
31126

Page No.
Date

Title : Data wrangling - I

Problem Statement :

perform following operations using python on any open source dataset.

1. Import all required libraries.
2. Locate an open source data from the web Provide a clear description of the data and its source
3. Load the dataset into Panda's library dataframe
4. Data processing
   Check for missing values in data use pandas describe() function to get some initial statistics. Provide variable descriptions. Types of variables. Check dimension of dataframe
5. Data formatting and data Normalization.
   summarize the types of variables by checking the datatypes of the variables in dataset
   If variables are not in correct datatype apply proper conversions.
6. Turn categorical values into quantitative variables

learning Objectives :

- To learn and understand data wrangling using Pandas
- To perform data preprocessing, formatting and normalization
- To perform one hot encoding on categorical variable.

Learning Outcomes:

Students will be able to

- Perform basic data preprocessing, data formatting and dd data normalization
- perform encoding for conversion

S/W Requirements:

Windows 10 - 64bit, 8GB RAM, 256GB SSD intel i5-8265U, vs code, Python 3.8

Theory:

When working with tabular data such as data stored in spreadsheets or databases, pandas is the right tool. Pandas helps to explore, clean & restore data. In pandas datatable is called as dataframe Importing data from each of these data sources is provided by function with prefix read_* Similarly to_* methods are used to store data. When selecting a single column of pandas Dataframe, the result is pandas service To select column use label of column in [].

A pandas series have no column labels but have row labels The describe method provides a quick overview of the numerical data in dataframe.

Pandas represents missing value with speaid float value NaN series is na() & series not na() can be used to filter rows. dropna() is used to drop rows with missing values Fillna can be used to fill rows with missing data.

method "ffil" for forward fill from previous rows categorical variable takes on a limited usally fixed number of possible values. They might

have an order.

df.shape returns a tuple of the shape of underlying data. df.size returns number of elements underlying data.

df.astype(dtype) converts / casts the type of an object to the specified datatype.

Analysis/Methods:

The given dataset contained 3580 rows × 21 columns with some missing values in some columns that was filled with default '0'. Some column data cells that didn't satisfy dtype were converted to numerical variables by the use of get_dummies(). The end result were printed on console & dataframe was save in file.

Conclusion: successfully performed the mentioned operations on the given dataset.