PICT, PUNE

DSBDAL

Shrikrushna S Tirape
B1284

Assignment - 01

DOP: 7/11/2022 DOS: 9/11/2022

Problem Statement:-
Data Wrangling

Perform the following operations using python or any other open source dataset

1) Import all the required libraries.
2) locate the open source dataset from web. provide a clear description of data and it's source.
3) Load the dataset into pandas dataframe
4) Data preprocessing: check for missing value in the data using isnull(). describe function to get some initial statistics. provide a variable description Types of variable etc. check for the dimension of dataframe.
5) Data formatting and Data Normalization. Summerize the types of variable by checking the data types of variable in the dataset. If variable not in correct type apply proper type conversion
6) Turn the categorical values into quantitative variable in python.

Aaikrushna G Zirape
B1284

PICT, PUNE

Learning objective:-
To learn and understand data wrangling using pandas.
To perform data preprocessing formatting and normalization
To perform one hot encoding on categorical values.

Learning outcomes:-
- Students will be able to
- perform basic data preprocessing, data formatting and data normalization.
- perform encoding for conversion.

SW/HW requirements:-
Windows 10 OS, 64-bit OS, 8GB RAM, 1TB HDD, Intel is 8th gen, Jupyter notebook, etc.

Theory:-

While working with tabular data stored in excel sheet or in a dataframe Panda is the best tool helps to explore and process data.

In pandas a dataset is called dataframe. pandas supports integration with many file formats (csv, excel, sql, json). Importing data from each of these data source is provided by function with prefix read_

Shatkrushna G Zirape
B1284

PICT, PUNE

similarly to method are used to store data. When selecting a single column of pandas data frame we use column name label in [].

The describe() method gives quick overview of numerical data in dataframe.

Pandas represents missing data with a special float value NaN. Series.isNA() and Series.notNA() can be used to filter rows. .dropNA() is used to drop rows with missing values.

.fillna() can be used to fill rows with missing values.

method = 'ffill' for forward fill from previous rows. Categorical variable takes on a limited, usually fixed, number of previous values. They might have an order.

df.shape() returns a tuple of the shape of underlying data.

df.size() returns number of elements in the underlying data.

df.astype(dtype) Converts/Casts the type of object to the specified datatype.

PICT, PUNE

Shaikrushna S Zirape

Analysis /method :-

The given dataset contains 13580 rows and 21 columns with missing value in some columns that was filled with default '0' some column's that didn't satisfy dtype are type casted to appropriate datatype one of the categorical variable type was converted to numerical variable. by the use of get-dummies The end result were printed on console and the dataframe was saved in file.

Conclusion

Successfully performed the mentioned operation on the given dataset.