Assignment No.2

Name - Kuldharon Sumit Dattatraya
class - TE
Div - 4
Subject - DSBDAL

Problem Statement -

Create an "Academic Performance" dataset of students and perform the following operations using Python.

1. Scan all variables for missing values and inconsitencies. If there missing values and/or inconsitencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outliers. If there are outliers use any of the suitable techniques to deal with them.
3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decreases the skeconess and convert the distribution into a normal distribution.

Reason and document your approoch properly

Theory -

1) Explain the methods to detect the outlier.

→ ① High Dimensional outlier Detection -

Real coord data sets are mostly very high dimensional. In many applications, data sets may contain thousands of features. The traditional outlier detection approaches such as PCA & LOF will not be effective. High contrast subspoces for

Density-Based outlier Ranking (HICS) method explained in this paper as an effective method to find outliers in high dimensional data sets. Using methods such as LOF, which are based on the nearest neighborhood, for high dimensional data sets will lead to outlier scores which are close to each other.

② Proximity Method -

once you have explored the simpler extreme value methods, consider moving onto proximity based methods.

i) Use clustering methods to identify the natural clusters in the data (such as the k-means algorithm).

ii) Identify and mark the cluster centroids.

iii) Identify data instances that are a fixed distance or percentage distances from cluster centroids.

iv) Filter out the outliers candidate from training dataset and assess the model's.

③ Projection Method -

Projection methods are relatively simple to apply and quickly highlight extraneous values.

i) Use projection methods to summarize your data to two dimensions (such as PCA, SOM or Sammon's mapping)

ii) Visualize the mapping and identify outliers by hand.

iii) Use proximity measures from projected values or codebook vectors to identify outliers.

iv) Filter out the outliers candidate from training dataset & assess the model's performance.

2) Explain data transformation methods.

→ ① Aggregation -

Data aggregation is the method where raw data is gathered and expressed in a summary form for statistical analysis.

Types of data aggregation - ① time aggregation
② Spatial aggregation

② Attribute Construction -

This method helps create an efficient data mining process. In attribute construction or feature construction of data transformation, new attributes are constructed & added from the given set of attributes to help the mining process.

③ Discretisation -

Data discretisation is the process of converting continous data attribute values into a finite set of intervals and associating with each interval some specific data value.

④ Generalisation -

Data generalisation is method of generating successive layers of summary data in an evaluational database to get a more comprehensive view of a problem or situation.

⑤ Integration -

Data integration is a crucial step in data preprocessing that involves combining data residing in different sources & providing users with a unified view of these data.

Approaches - ① tight coupling approach
② loose coupling approach

⑥ Manipulation -

Data manipulation is the process of changing or altering data to make it more readable & organised

⑦ Normalisation -

Data normalisation is a method to convert the source data into another format for effective processing.

⑧ Smoothing -

Data smoothing is a technique for detecting trends in noisy data where the shape of the trend is unknocon.

③ Write the algorithm to display the statistics of Null values present in the dataset.

→

5) What is use of Seaborn library?

→ Seaborn is a library for making statistical graphics in python. It builds on top of matplotlib & integrates closely with pandas data structures. Seaborn helps you explore & understand your data

6) What is matplotlib?

→ Matplotlib is a cross platform, data visualization and graphical plotting library for python and its numerical extension Numpy. As such, it offers a viable opensource alternative to MATLAB.