DSBDAL - Assignment 2

Page No.
Date

Name :- Om Khedkar
Roll no :- 31325
Batch :- L-3
Date :- 03/02/2022

* Title :- Data Wrangling II

* Problem statement :-
Perform the following operation using Python on any open source dataset.
1. Scan all variables for missing values & inconsistencies. If there are missing values & inconsistencies, use any of the suitable technique to deal with them.
2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable technique to deal with them.
3. Apply data transformation on at least one of the variables.

* Learning objective :-
1. To deal with inconsistent data.
2. To deal with 'outliers' in data.
3. Learn & apply data transformation on underlying dataset.

* Learning outcome :-
1. Understood the importance of data cleansing for preprocessing data set.
2. Lean about inconsistencies & the ways to deal with missing values.
3. Learned the types of outliers / anomaly.

31325

Page No.
Date

* Theory :-

Preprocessing of data is to mainly check data quality. The data quality can be checked by the following :-

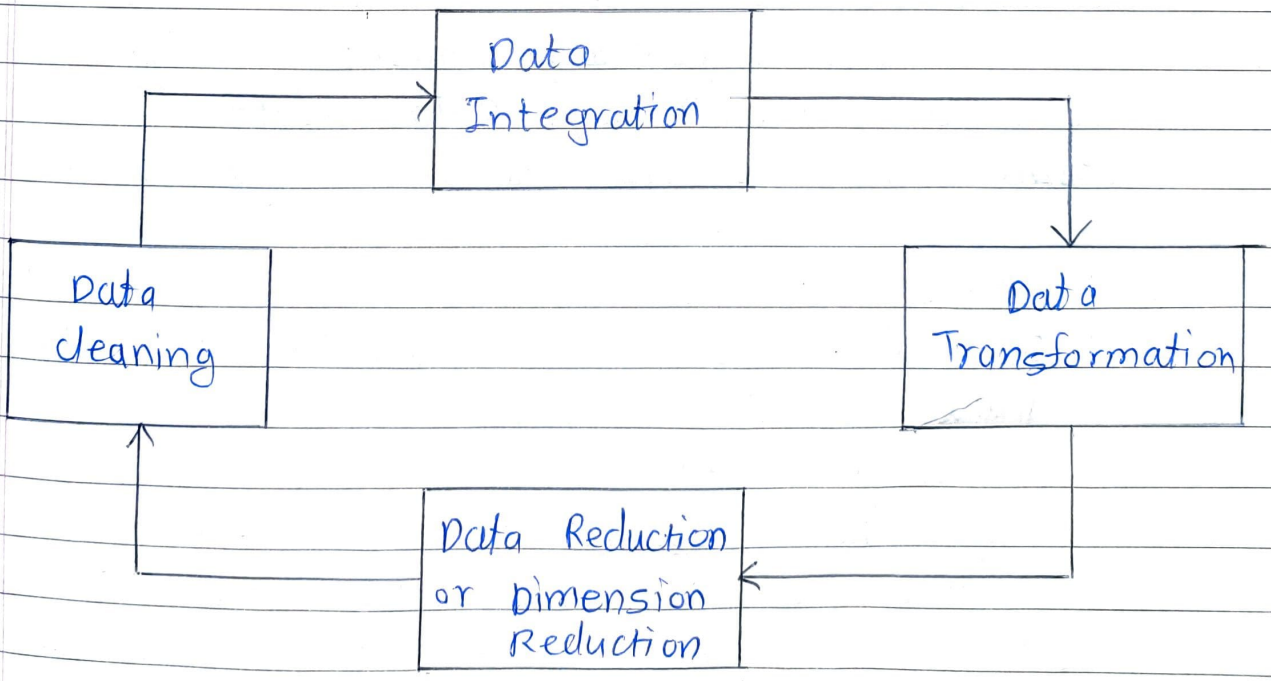
1. Accuracy :- To check whether data entered is correct or not.

2. Completness :- To check whether the data is available or not.

3. Consistency :- To check whether some data is present.

4. Timeliness :- The data should be updated correctly.

Major task in data properocessing :-



1. Data cleaning :-

It is the process to remove incorrect data, incomplete data & inaccurate data from the dataset & it also replaces the missing values.
There are some technique in data cleaning.:

a) Handling missing values :
1. Standard values like 'Not available' or 'NA' can be used to replace null values.
2. Mean value of attribute can be inserted when data is normally distributed.
3. In case of non normally distribution meadin values of attribute can be used.
4. while using regression or decision tree algorithm, the missing value can be replaced by most preferrable value.

b) Noise :
Noise generally means random error or containing unnecessary data points. There are some methods to handle noisy data.

c) Binning :
1. Smoothing by bin mean method, here the mean value of bin.
2. By bin median.
3. By bin boundry - using minimum & maximum values of the bin values are taken & the values are replaced by the closed boundry value.

31325

d) Regression :

For the analysis purpose regression helps to decide the ~~suitable~~ variable which is suitable for analysis.

e) clustering :

This is used for finding outliers & also in grouping the data clustering is generally used ~~to~~ in unsupervised learning.

2. Data integration :-

The process of combining multiple source into single data set.

a) Schema integration :

Integer meta data from different sources.

b) Entity Identification problem :

Identifying entities of multiple dataset for eg:- stud_id of one database & stud_name of another database belongs to same entity.

c) Detection & resolution of data value concepts :

The data taken from different database while merging may differ like the attribute value from one database may differ from another database.

31325

3) Data Reduction :-

This process reduces volume of data which makes the analysis easier yet produces the same or almost same results. There are some techniques in data reduction they are as follows :-

a) Dimensionality Reduction :

This process is necessary for real world problems as the data size is big. In this process the reduction of random variables or attribute is done so that the dimensionality of data set can be reduced. Combining & merging the attribute of data without losing its original characteristic

b) Numerosity reduction :

In this method the representation of data is made smaller by reducing the volume. There will not be any loss of data in this reduction.

c) Data compression :

The compressed form of data is called data compression. This compression can be lossless or lossy.

4) Data Transformation :-

The change made in the format or structure of data is called the data transformation. This can be simple or complex.

31325

Page No.
Date

Methods in data transformation :

a) Smoothing :
with the help of algorithms we can remove noise from the dataset & helps in knowing the important feature of dataset.

b) Aggregation :
In this method, the data is stored & presented in the form of summary.

c) Discretization :
The continuos data is split into intervals. It reduces data size.

d) Normalization :
It is the method of scaling the data so that it can be represented in smaller range.

* Conclusion :-
Understodd the concept of data processing in terms of way to do it & achieved good quality of dataset.