Dr. SNS RAJALAKSHMI COLLEGE OF ARTS AND SCIENCE

(AUTONOMOUS)

Accredited by NAAC (Cycle- III) with 'A+' Grade

DEPARTMENT OF B.SC CS (GCD)

22UDA501 – INTRODUCTION TO DATA ANALYTICS

UNIT- II

Data Preprocessing

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data.

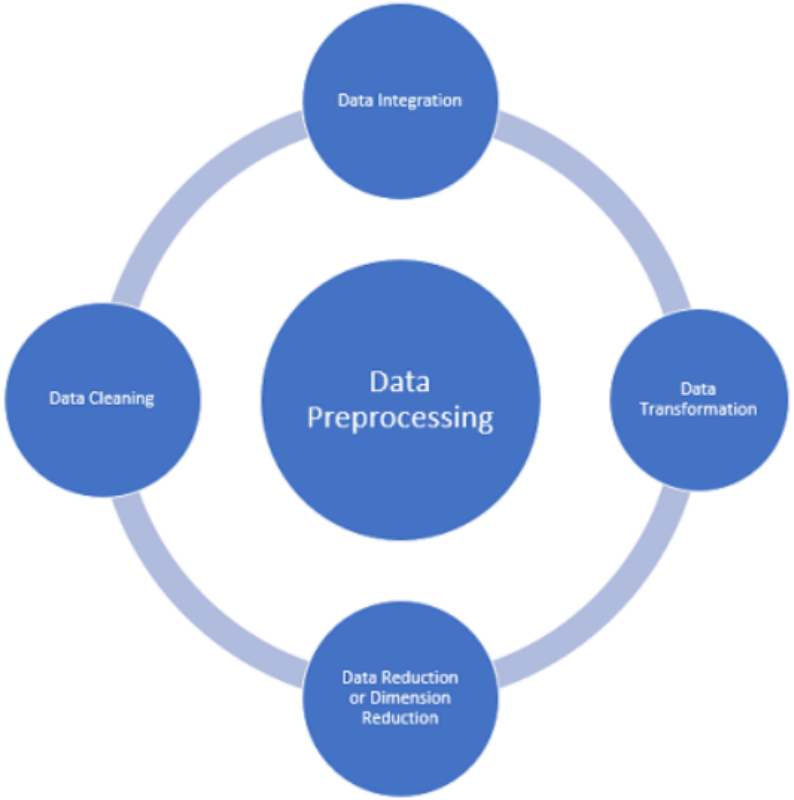
The quality of the data should be checked before applying machine learning or data mining algorithms.

Important of Data Preprocessing

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.
- **Consistency:** To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.
- **Interpretability**: The understandability of the data.

Major Tasks in Data Preprocessing

There are 4 major tasks in data preprocessing – Data cleaning, Data integration. Data reduction. and Data transformation



Handling Missing Values

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually, but it is not recommended when that dataset is big.
- While using regression or decision tree algorithms, the missing value can be replaced by the most probable value.

Handling Noisy Data

- Noisy generally means random error or containing unnecessary data points.
- **Binning:** This method is to smooth or handle noisy data. First, the data is sorted then, and then the sorted values are separated and stored in the form of bins.
- **Regression:** This is used to smooth the data and will help to handle data when unnecessary data is present.
- **Clustering:** This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

Data Integration

The process of combining multiple sources into a single dataset.

- **Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.
- **Entity identification problem:** Identifying entities from multiple databases.
- **Detecting and resolving data value concepts**: The data taken from different databases while merging may differ.

Data Reduction

This process helps in the reduction of the volume of the data, which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space.

- **Dimensionality reduction:** This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced.
- **Numerosity Reduction**: In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.
- **Data compression**: The compressed form of data is called data compression. This compression can be lossless or lossy.

Data Transformation

The change made in the format or the structure of the data is called data transformation.

- **Smoothing:** With the help of algorithms, we can remove noise from the dataset, which helps in knowing the important features of the dataset.
- **Aggregation:** In this method, the data is stored and presented in the form of a summary.
- **Discretization:** The continuous data here is split into intervals. Discretization reduces the data size.
- **Normalization:** It is the method of scaling the data so that it can be represented in a smaller range.