Data Preprocessing

Data preprocessing is the data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent and it may contain too many errors in it. Data preprocessing is a proven method to solve such issues. Data preprocessing prepare raw data for further processing.

Why we need data preprocessing?

Data preprocessing is required because:

- Real-world data are generally **Incomplete** (missing attribute values, having only aggregate data)
- Real-world data are **Noisy** (containing error or outliers)
- Real-world data are **Inconsistent** (Containing discrepancies in codes or names)

Steps for Data Preprocessing

The following are the steps that we used to preprocess the raw data into useful material.

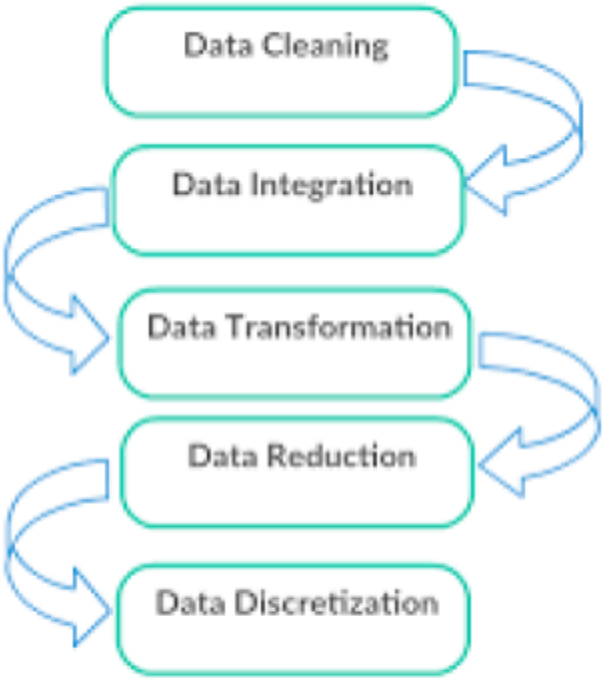


***Figure:** Steps of Data Pre-processing*

Data Cleaning:

Data cleaning also called data **cleansing** or **scrubbing**.

- Fill in missing values, smooth noisy data, identify or remove the outliers, and resolve inconsistencies.
- Data cleaning is required because source systems contain "**dirty data**" that must be cleaned.

Data Integration:

It compromises the merging of data from multiple sources.

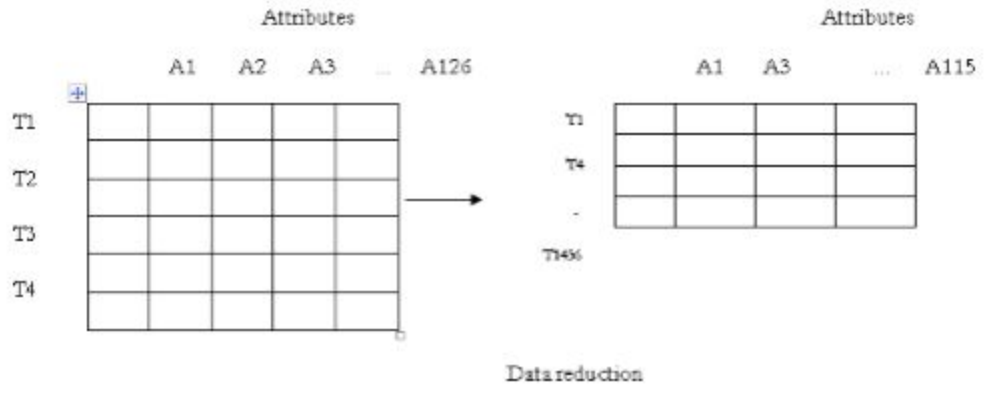
- This process must be carefully implemented in order to avoid redundancies and inconsistencies in the resulting data set.
- Combines data from multiple sources into a coherent data store e.g. data warehouse.

Data Transformation:

Data is normalized, aggregated and generalized. The data is converted or consolidated so that the mining process result could be applied or maybe more efficient

Data Reduction:

- It obtains reduced representation in volume but produces the same or similar analytical results.
- Need for data reduction:



Data Discretization:

Reduce the number of values for a given continuous attribute by divide the range of a continuous attribute into intervals. Interval labels can then be used to replace actual data values.