Data Preprocessing in Data Mining :-

Data preprocessing is an important step in the data mining process. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Data preprocessing also ensures that there are not any incorrect or missing value due to human error or bugs.

Factors Contributing to Data Quality :-

Before looking at how data is preprocessed, let's discuss at some factors contributing to data quality —

1) **Accuracy :-** Accuracy means that the information is correct. Outdated information, typos (small mistake in a text made when it was typed or printed) and redundancies can affect a dataset accuracy.

2) Consistency :- The data should have no contradictions. Inconsistent data may give you different answers to the same questions.

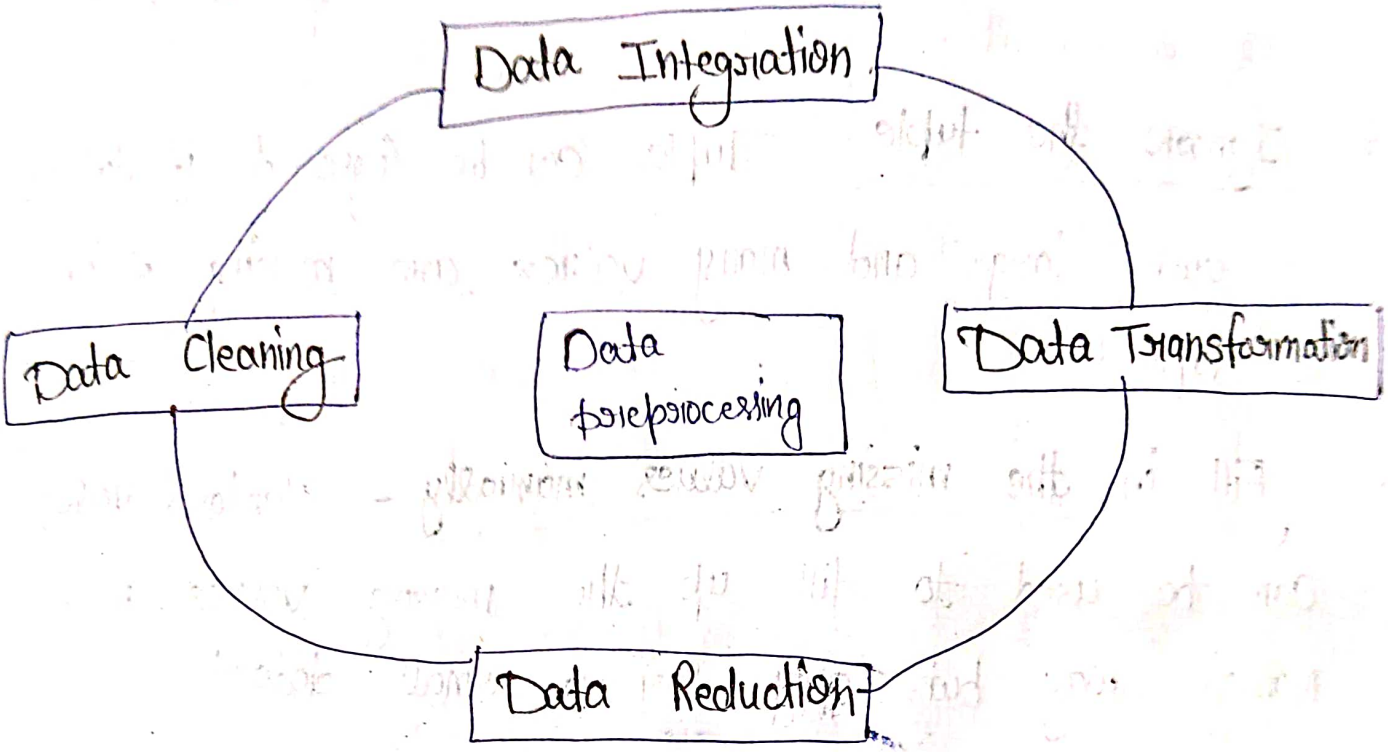
3) Completness :- The dataset should not have imcomplete field or lack empty fields.

This characteristic allows data scientists to perform accurate analyses as they have access to a complete picture of situation the data describes.

4) Validity – A dataset is considered valid if the data samples appear in the correct format, are within a specified range, and are of the right type. Invalid datasets are hard to organize and analyze.

5) Timeliness - Data should be collected as soon as the event it represents occurs. As time passes, every dataset becomes less accurate and useful as it doesnot represent the current reality.

Data Preprocessing Stages - There are four stages of data processing :-



1) Data Cleaning - Data cleaning or cleansing is the process of cleaning datasets by removing outliers, replacing missing values, smoothing noisy data and correcting inconsistent data. Many techniques are used to perform each of these tasks, where each technique is specific to a user's preference or problem set.

Some techniques used in data cleaning are-

Handling Missing Values - This type of scenario occurs when some data is missing. The following methods execute missing value over a number of attributes -

a) Ignore the tuple - Tuples can be ignored if dataset is quite large and many values are missing within a tuple.

b) Fill in the missing values manually - Standard values can be used to fill up the missing values in a manual way but only for a small dataset.

c) Use global constant to fill in the missing value - Replace all missing attributes with the same constant, such as a "unknown" or ∞.

d) Attribute's mean and median values can be used to replace the missing value in normal and non-normal distribution of data.

Noisy Data - A large amount of meaningful data is called noise.

Ex If you need to predict whether a person can drive, information about their hair color, weight will be irrelevant.

Following are the methods for solving the problem of noisy data -

1) Binning - This method handle noisy data to make it smooth. Data gets divided equally and stored in the forms of bins and then methods are applied to smoothing or completing the tasks.

2) Regression - Regression function is used to smoothen the data. Regression can be linear (consist of one independent variable) or multiple (multiple independent variable)

3) Clustering - The process of ~~Combining data of~~ grouping the similar data in groups or clusters and used to finding outliers.

4) Data Integration - The process of combining data from multiple sources (databases, spreadsheets, text files) into a single dataset. Single and consistent view of data is created in this process.

Major problems during data integration are schema integration (Integrates set of data collected from various sources), Entity identification and detecting and resolving data values concept.

Data Transformation - In this part, change in format or structure of data in order to transform the data suitable for mining process. Methods for data transformation are -

a) Normalization - Method of scaling data to represent it in a specific smaller range

b) Discretization - It helps reduce the data size and make continuous data divided into intervals.

c) Attribute Selection - To help the mining process, new attributes are derived from the given attributes.

d) Hierachy Generation - In this, the attributes are changed from lower level to higher level in hierachy.

e) Aggregation - A summary of data gets stored which depends upon quality and quantity of data to make the result more optimal.

Data Reduction - The last stage of data preprocessing is data reduction. It helps in increasing storage efficiency and reducing data storage to make the analysis easier by producing almost the same results.