what is data preprocessing and why preprocess, what forms of data preprocessing

Data mining is the process of transforming raw data into meaningful and understandable data.

It is also a crucial step in a data mining process (knowledge discovery process) where the quality of data should be checked before it is implemented in the models and the data mining algorithms.

The data which are found in the real world tend to be incomplete, noisy, dirty, and inconsistent. The data preprocessing method can help to improve the quality of the data which ensures the accuracy and efficiency of the subsequent mining process. In other words, data preprocessing is a mandatory phase because through this we can check missing values, noisy data, and other inconsistencies before executing it to the algorithms.

Forms of Data processing:

1. Data Cleaning/Cleansing

Real-world data tend to be incomplete, noisy, and inconsistent. Data Cleaning/Cleansing routines to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Data can be noisy and have incorrect attribute values. Owing to the following, the data collection instruments used may be at fault. Maybe human or computer errors occurred during data entry. Errors in data transmission can also occur.

“Dirty” data can cause confusion in the mining procedure. Although most mining routines have some procedures, they deal with incomplete or noisy data, which are not always robust. Therefore, a useful Data Preprocessing step is to run the data through some Data Cleaning/Cleansing routines.

1. Data Integration

Data Integration is involved in data analysis task which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files.

For example, how can a data analyst be sure that customer\_id in one database and cust\_number in another reference to the same entity? The answer is metadata. Databases and data warehouses typically have metadata. Simply, metadata is data about data.

Another most important issue is redundancy. The attribute which is derived from another table may be replicated.

Inconsistencies in attributes and names can also cause redundancy in the dataset.

## Data Transformation

## This step is taken in order to transform the data into appropriate forms suitable for the mining process. This involves the following ways:

## Normalization:

## It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

1. Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

1. Discretization:

Interval levels or conceptual levels are used to replace the raw values of numeric attributes.

1. Concept Hierarchy Generation:

In this case, attributes are being transformed from a lower to a higher level in the hierarchy. The characteristic "city," for example, can be changed to "country."

1. Data Reduction:

Data mining is a strategy for dealing with large amounts of data. When dealing with large amounts of data, analysis becomes more difficult. We employ a data reduction technique to get rid of this. Its goal is to improve storage efficiency while lowering data storage and analysis expenses.