Assignment I

Data Mining And Big Data Analystics

NAME : SARATH BHARATHI B

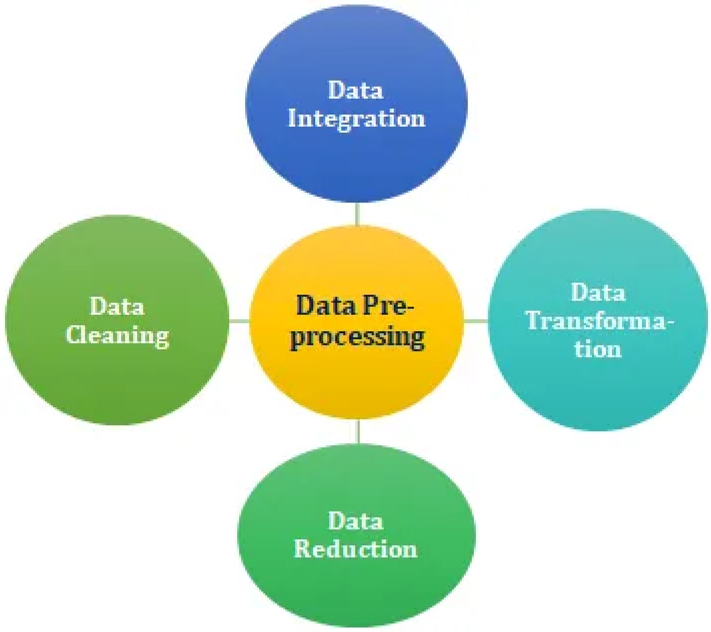
CLASS : I - MCA - A

TOPIC : Preprocessing in Data Mining

DATE : 20/02/2023

Preprocessing in Data Mining

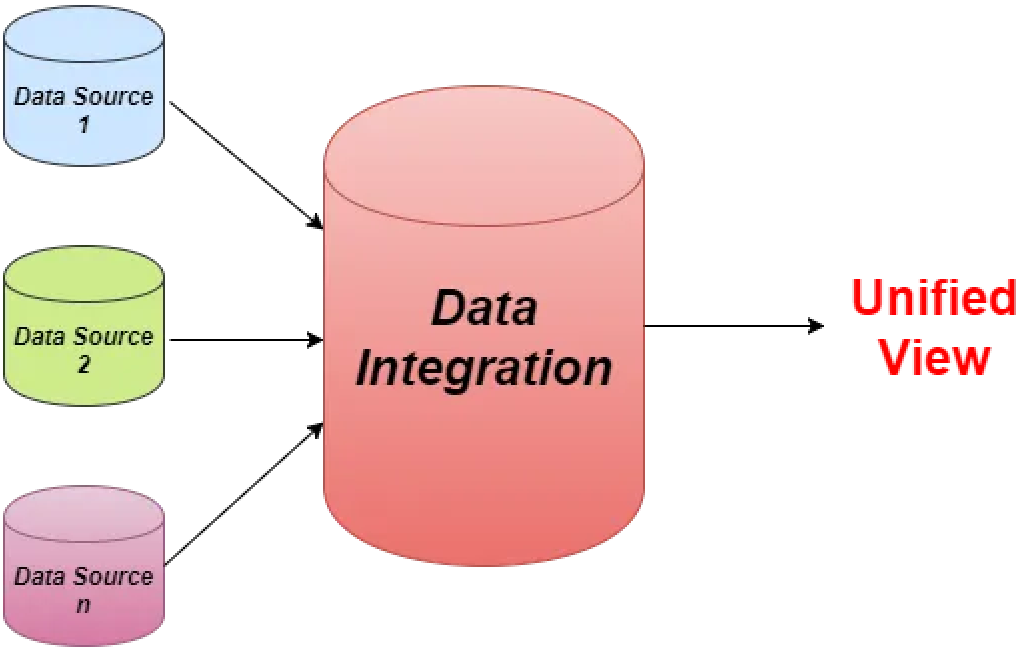
Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.



Some common steps in data preprocessing include:

- **Data cleaning:** this step involves identifying and removing missing, inconsistent, or irrelevant data. This can include removing duplicate records, filling in missing values, and handling outliers.

- **Data integration:** this step involves combining data from multiple sources, such as databases, spreadsheets, and text files. The goal of integration is to create a single, consistent view of the data.



- **Data transformation:** this step involves converting the data into a format that is more suitable for the data mining task. This can include normalizing numerical data, creating dummy variables, and encoding categorical data.
- **Data reduction:** this step is used to select a subset of the data that is relevant to the data mining task. This can include feature selection (selecting a subset of the variables) or feature extraction (extracting new variables from the data).
- **Data discretization:** this step is used to convert continuous numerical data into categorical data, which can be used for decision tree and other categorical data mining techniques.

> **Note :** By performing these steps, the data mining process becomes more efficient and the results become more accurate.

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:

1. **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(A) Missing Data:** This situation arises when some data is missing in the data. It can be handled in various ways. **Some of them are:**

    - **Ignore the tuples:** This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

    - **Fill the Missing values:** There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

  (B) Noisy Data: Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

    1. **Binning Method:** This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

    2. **Regression:** Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

    3. **Clustering:** This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. **Data Transformation:** This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

    1. **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)

    2. **Attribute Selection:** In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

    3. **Discretization:** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

    4. **Concept Hierarchy Generation:** Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. **Data Reduction:** Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

    ### The various steps to data reduction are:

    1. **Data Cube Aggregation:** Aggregation operation is applied to data for the construction of the data cube.

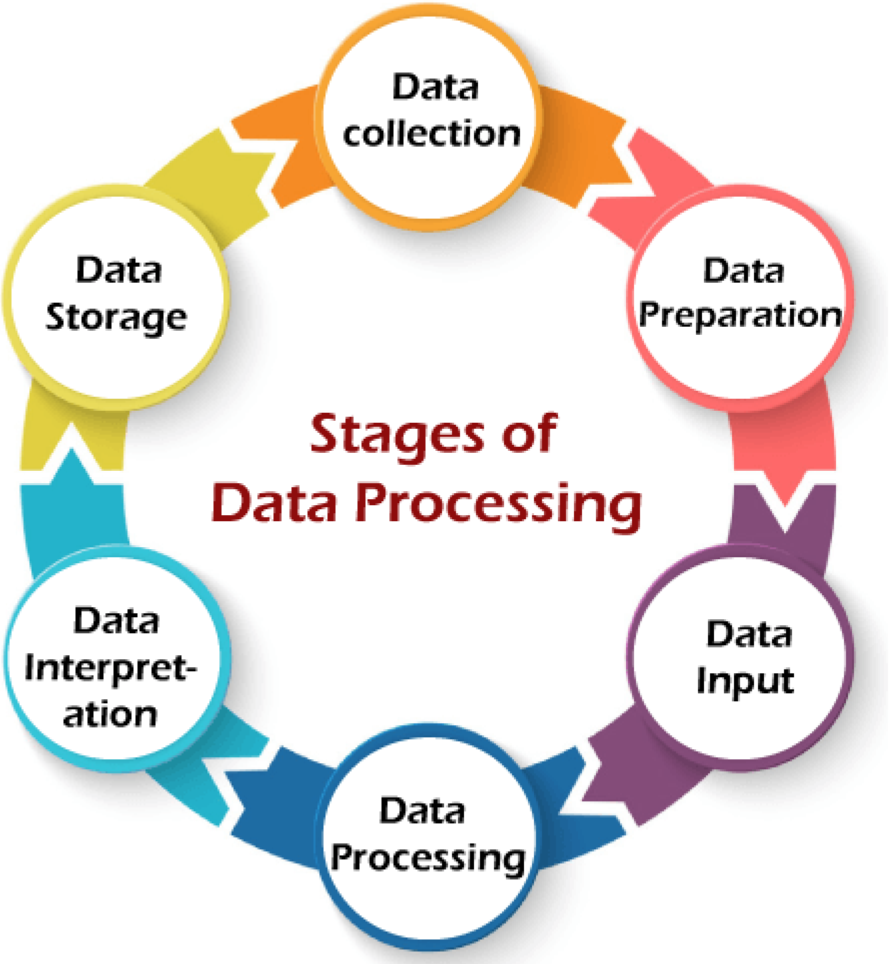
    2. **Attribute Subset Selection:** The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

3. **Numerosity Reduction:** This enable to store the model of data instead of whole data, for example: Regression Models.

4. **Dimensionality Reduction:** This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

Stages of Data Processing

The data processing consists of the following six stages.



Examples of Data Processing

Data processing occurs in our daily lives whether we may be aware of it or not. Here are some real-life examples of data processing, such as:

- Stock trading software that converts millions of stock data into a simple graph.
- An e-commerce company uses the search history of customers to recommend similar products.
- A digital marketing company uses demographic data of people to strategize location-specific campaigns.
- A self-driving car uses real-time data from sensors to detect if there are pedestrians and other cars on the road. Importance of Data Processing in Data Mining In today's world, data has a significant bearing

on researchers, institutions, commercial organizations, and each individual user. Data is often imperfect, noisy, and incompatible, and then it requires additional processing. After gathering, the question arises of how to store, sort, filter, analyze and present data. Here data mining comes into play.

- The complexity of this process is subject to the scope of data collection and the complexity of the required results. Whether this process is time-consuming depends on steps, which need to be made with the collected data and the type of output file desired to be received. This issue becomes actual when the need for processing a big amount of data arises. Therefore, data mining is widely used nowadays.

- When data is gathered, there is a need to store it. The data can be stored in physical form using paper-based documents, laptops and desktop computers, or other data storage devices. With the rise and rapid development of such things as data mining and big data, the process of data collection becomes more complicated and time-consuming. It is necessary to carry out many operations to conduct thorough data analysis.

- At present, data is stored in a digital form for the most part. It allows processing data faster and converting it into different formats. The user has the possibility to choose the most suitable output.

Reference :

- geeksforgeeks.org
- javatpoint.com
- medium.com