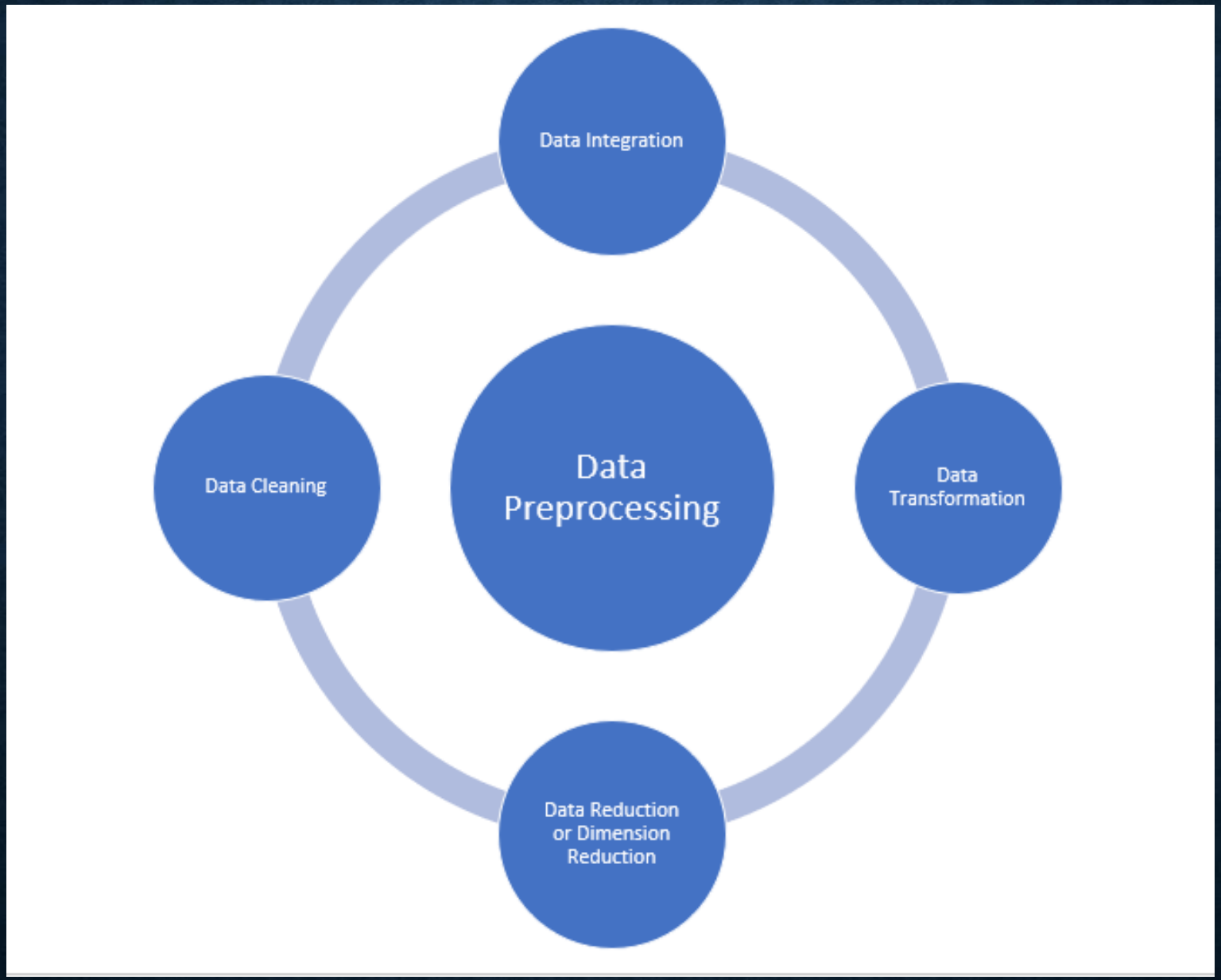
Introduction To **Data Mining**


Data Integration
Data Cleaning
Data Preprocessing
Data Transformation
Data Reduction or Dimension Reduction

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.

Why Pre processing of data is important

•Preprocessing of data is mainly to check the data quality. The quality can be checked by the following

•**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.

•**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 :

- **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.

- **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).

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

Data integration:

The process of combining multiple sources into a single dataset. The Data integration process is one of the main components in data management. There are some problems to be considered during data integration.

•**Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.

•**Entity identification problem:** Identifying entities from multiple databases. For example, the system or the use should know student _id of one database and student name of another database belongs to the same entity.

•**Detecting and resolving data value concepts**: The data taken from different databases while merging may differ. Like the attribute values from one database may differ from another database. For example, the date format may differ like “MM/DD/YYYY” or “DD/MM/YYYY”.

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).