DATA MINING

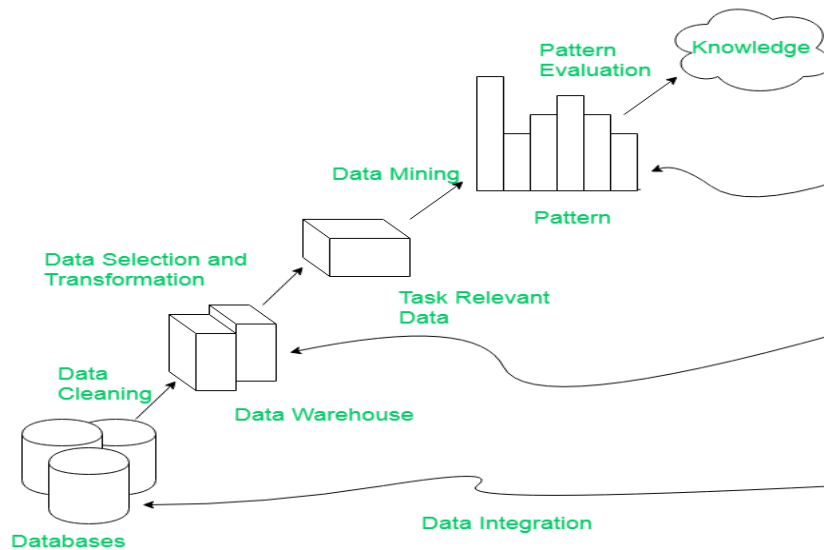**KDD Process**

Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.
Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
Data Transformation is a two step process:

Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.s
attern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
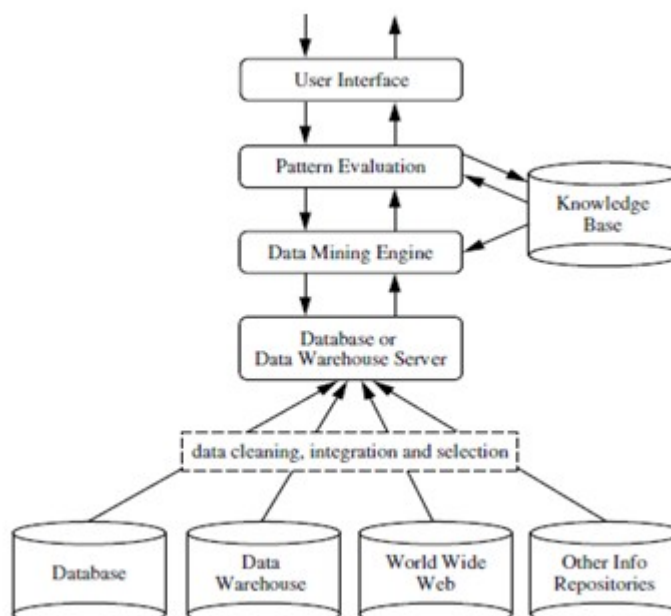
Classification:Catagorization of new data with the help of current data or old data.
           EG:Grouping of patients according to there medical records.
Prediction:Predicting the missing or unknown value based past or current data.
           Predictiing the correct treatment for the patient based on there medical condition.

Data mining system:

Nominal Attribute:
Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Sex of the Person.

Ordinal Attribute:
The ordinal attribute value provides sufficient information to order the objects and not checking the quantity. Such as Rankings, Grades, Height.

Interval Scaled attribute:
It is measured on a scale of equal size units,these attributes allows us to compare such as temperature in C or F and thus values of attributes have order.

Ratio Scaled attribute:
Ratio scale is a type of variable measurement scale which is quantitative nature.It allows any researcher to compare the intervals or differences.

Data Preprocessing in Data Mining:

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:

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

2.**Fill the Missing values**: You can choose to fill the missing values manually, by attribute mean or the most probable value.It is time consuming.

3.**Global constant:**You can replace all missing attribute values with global constant, such as a label like "Unknown" or minus infinity.

4.**attribute mean:**Replace missing values of an attribute with the mean (or median if its discrete) value for that attribute in the database.For example, in a database of US family incomes, if the average income of a US family is X you can use that value to replace missing income values.

5.**The most probable value:**The value can be determined using regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms.

•**(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 reconstuction 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).