UNIT-I

INTRODUCTION TO DATA MINING

1. Introduction
2. What is data Mining
3. Definition
4. KDD
5. challenges
6. Data Mining tasks (Pending)
7. Data Preprocessing
8. Data cleaning
9. Missing Data
10. Dimensionality Reduction
11. Feature Subset selection
12. Discretization and Binarization
13. Data Transformation
14. Measures of Similarity and Dissimilarity - Basics.

Introduction :-

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words we can say that data mining is mining knowledge from data.

→ There is a huge amount of data available in the information industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

→ Extraction of information is not only the process we need to perform; Data mining also involves other processes such as data cleaning, Data integration, Data transformation, Pattern Evaluation and Data presentation. Once all these processes are over, we would be able to use this information in many applications such as

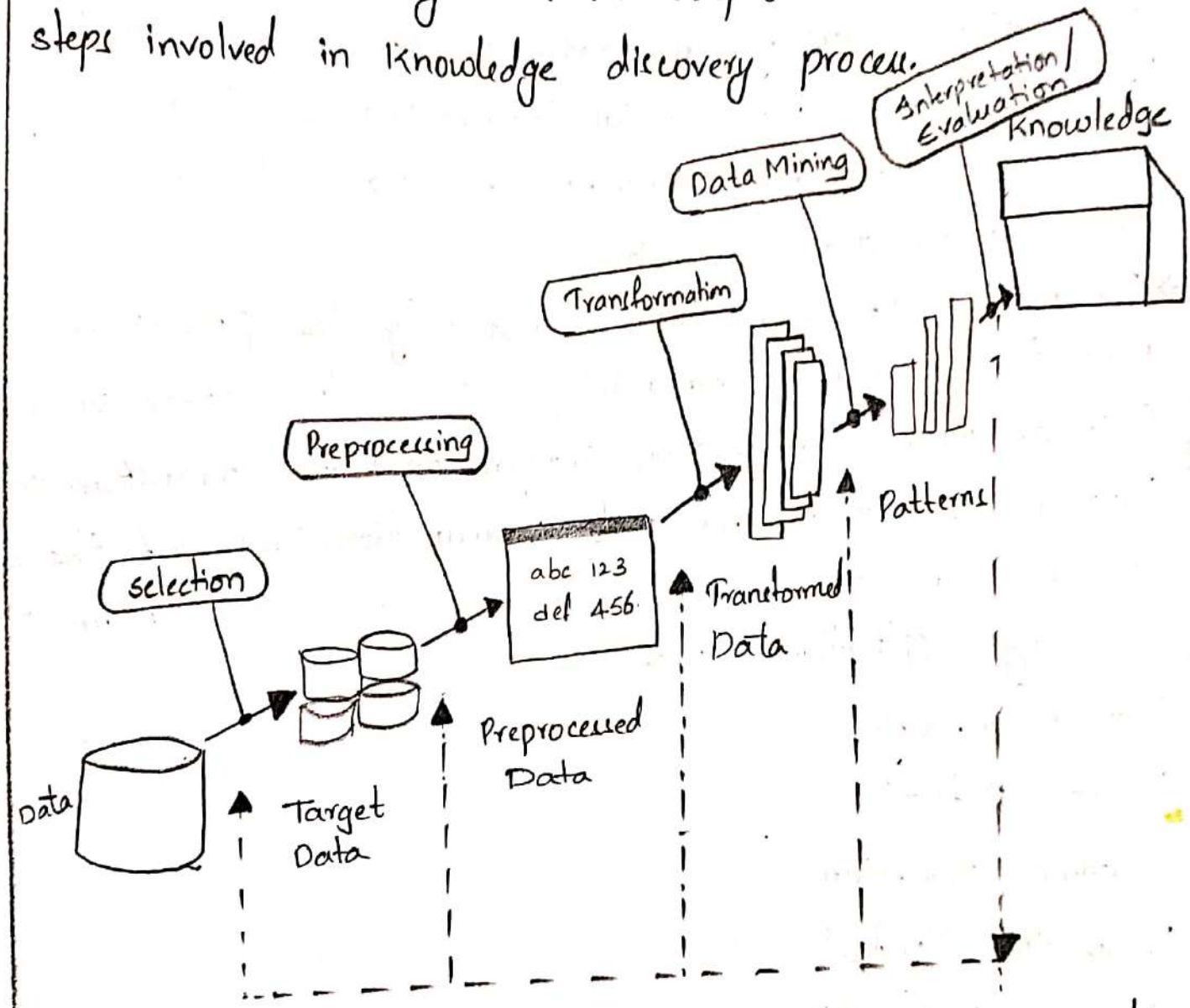
Market Analysis (M)
Fraud Detection (F)
Customer Retention (C)
Production Control (P)
Science Exploration. (S)

* WHAT IS DATA MINING :-

Data Mining is defined as extracting information from huge sets of data. The information or knowledge extracted can be used for many applications.

* KDD:-

KDD stands for Knowledge Discovery in Database. Data mining is an essential step in the process of knowledge discovery. There are seven different stages in KDD process. This process takes raw data as input and provides useful information desired by users as output. Here is the list of steps involved in knowledge discovery process.



- Preprocessing of database consists of Data cleaning and Data integration.
- KDD is an iterative process.

DATA MINING CHALLENGES:-

The challenges are

1. Complex Heterogeneous data.
2. Distributed data.
3. Scalability
4. Non-Traditional Analysis
5. High Dimensionality

1. **Complex Heterogeneous data:-**

The growth in various fields such as Science, medical and finance produced large complex heterogeneous and non-traditional data. Some of such data includes semi-structured text, unstructured text and multimedia. This type of data cannot be handled by classical data analysis techniques.

2. **Distributed data:-**

Data needed for analysis in certain circumstances does not belong to single owner or stored in single geographic location. This distributed data analysis needs new techniques.

(i) Techniques to minimize resources needed for distributed computing.

(ii) Integration of data mining results from heterogeneous sources.

(iii) Handling Data Security.

3. Scalability:-

Data mining algorithms must be capable to handle and incorporate huge volumes of data. These algorithms can be made more scalable by

(i) Sampling data

(ii) Implementing data structure

(iii) Developing distributed and parallel algorithms.

4. Non-Traditional Analysis:-

This analysis methods are based on hypothesis and testing. This method needs high volumes of resources which becomes difficult in present data mining scenario.

5. High Dimensionality:-

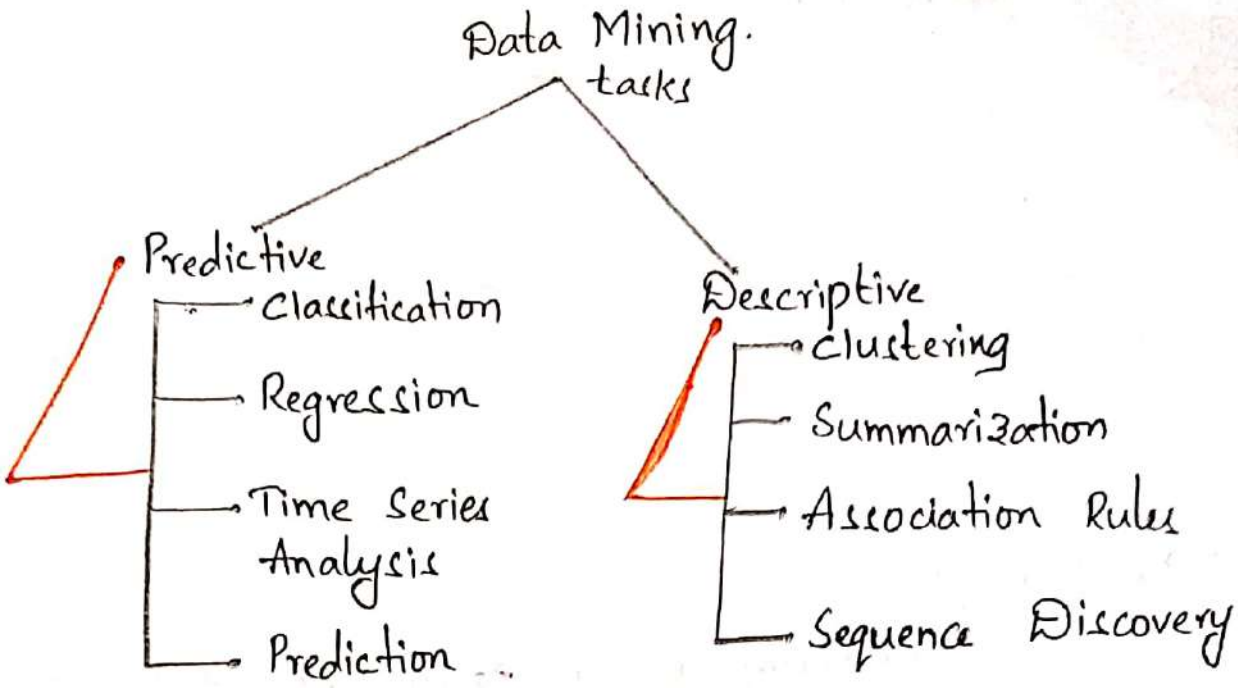
Data sets of present era has several hundreds of attributes. There are number of dimensions that grow with technology advancement.

* DATA MINING TASKS:-

The data mining tasks can be classified generally into two types. Those two categories are

① Predictive Tasks

② Descriptive tasks.



Predictive:- It makes prediction about values of data using known results from different data or based on historical data.

Descriptive:- It identifies patterns or relationship in data, it serves as a way to explore properties of data.

→ Classification:- discovery of a function that classifies a data item into one of several predefined classes.

Given a collection of records.

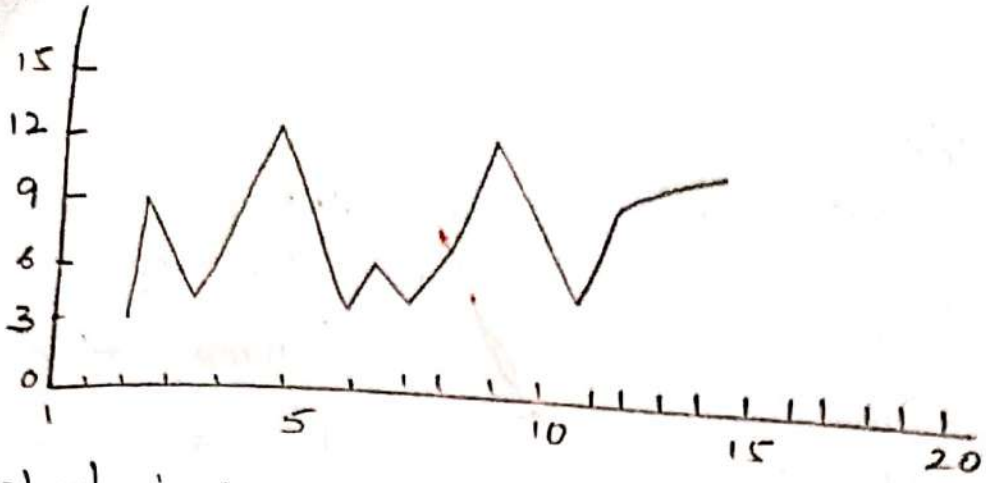
Each record contains a set of attributes, one of the attributes is class.

Ex:- Pattern Recognition.

→ Time series analysis:-

- The value of attribute is examined as it varies over time.
- A time series plot is used to visualize time series.

Ex:- Stock Exchange.



→ Clustering:-

clustering is a task of segmenting a diverse group into a number of similar subgroups or clusters. Most similar data are grouped in clusters

Ex:- Bank Customer.

DATA PREPROCESSING:-

Data Preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent and is likely to contain many errors. Data Preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

* Data preprocessing is used database driven application such as customer relationship management and rule-based applications.

* Data preprocessing is required because

⟹ Real world data are generally

- Incomplete:- Missing attribute values.
- Noisy:- Containing errors or outliers.
- Inconsistent:- Containing discrepencies in codes or names

Need for preprocessing the data:-

1) Attributes of interest may not be available always.
2) Relevant data may not be recorded due to misunderstanding or because of equipment malfunctions.
3) Data that is inconsistent with other recorded data might be deleted.
4) Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.
5) The data collection instruments used may be faulty

6) There may be human or computer errors occuring at entry.

7) Errors in data transmissions can also occur.

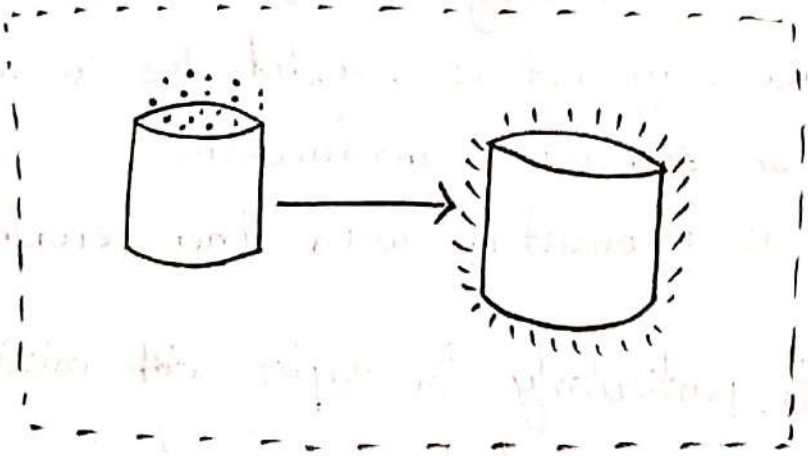
8) There may be technical limitations such as limited buffer size for coordinality synchronized data transfer and consumption.

* To overcome the above problems the following data preprocessing techniques are required.

1. Data cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction.

1. Data cleaning:-

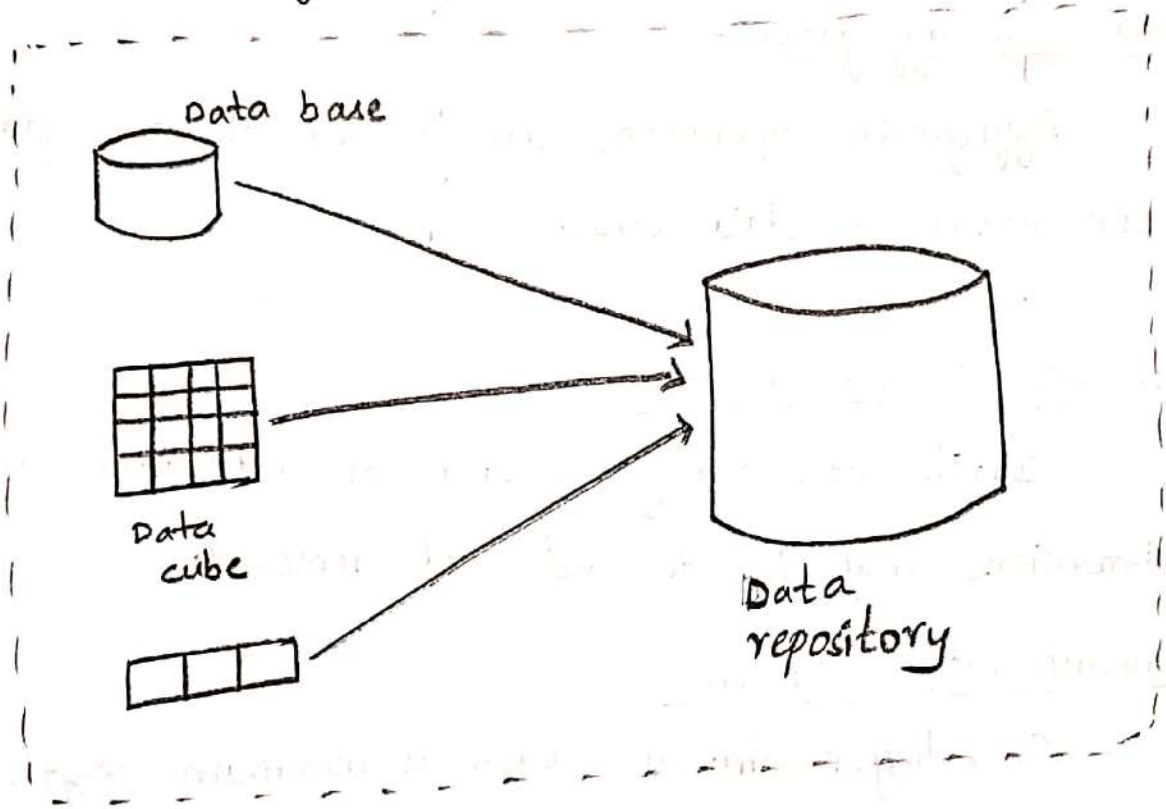
This routine work is to 'clean' the data by filling in missing values, Smoothing noisy data, identifying or removing outliers and resolving inconsistencies.



Data cleaning.

Data Integration:-

This is the process of integrating multiple databases, data cubes, or files. Yet some attributes representing a given concept may have different names in different databases, causing inconsistencies and redundancies.



3) Data Transformation:-

This is a kind of operation in which we use normalization and aggregation.

- 4, 29, 100, 40, 80 → -0.04, 0.29, 1.00, 0.40, 0.80
- 1, 20, 50, 100 → 0.01, 0.20, 0.50, 1.0

4) Data reduction :-

This is the reduced representation of the data set that is much smaller in volume, yet produces the same analytical results.

* Following are different data reduction strategies:

(a) Data cube aggregation:-

Aggregation operations are applied to the data in the construction of data cube.

→ sum, count, add, Min, Max

(b) Attribute subset selection:-

Irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.

(c) Dimensionality reduction:-

Encoding mechanisms such as minimum length encoding or wavelets are used to reduce the data set size.

(d) Numerosity reduction:-

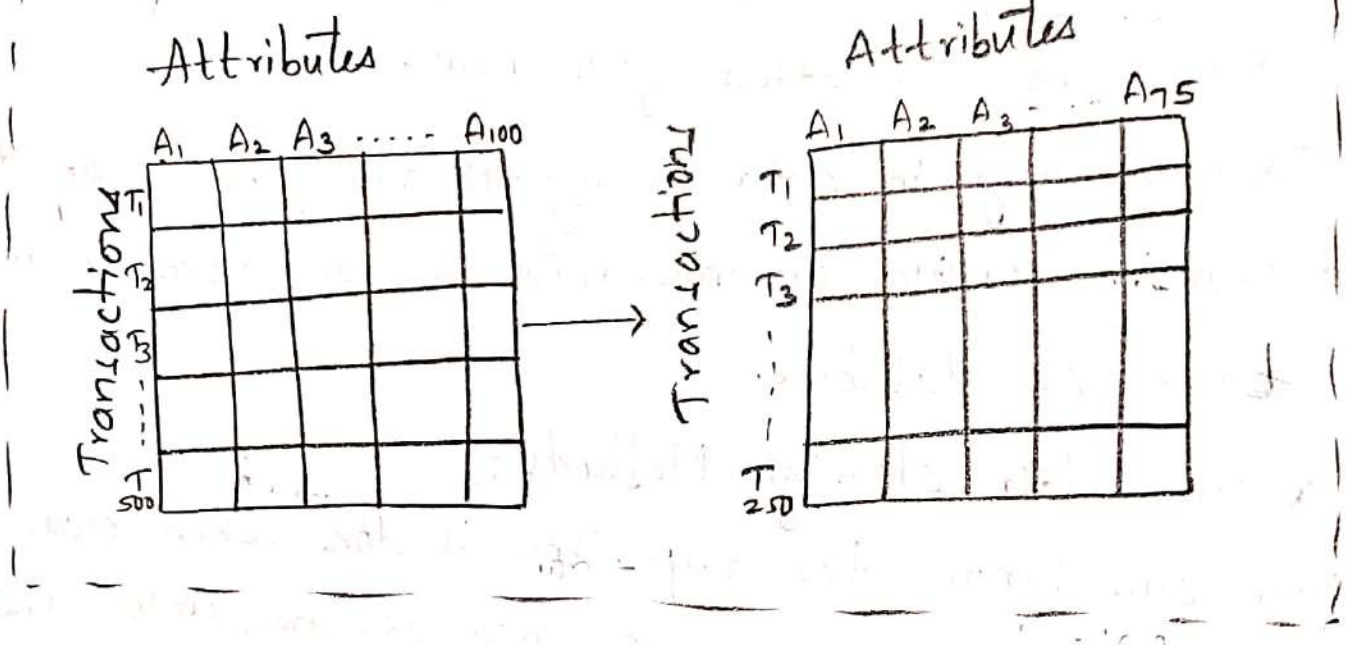
The data is replaced or estimated by alternative, smaller data representations such as clusters or parameters models, histograms, Sampling.

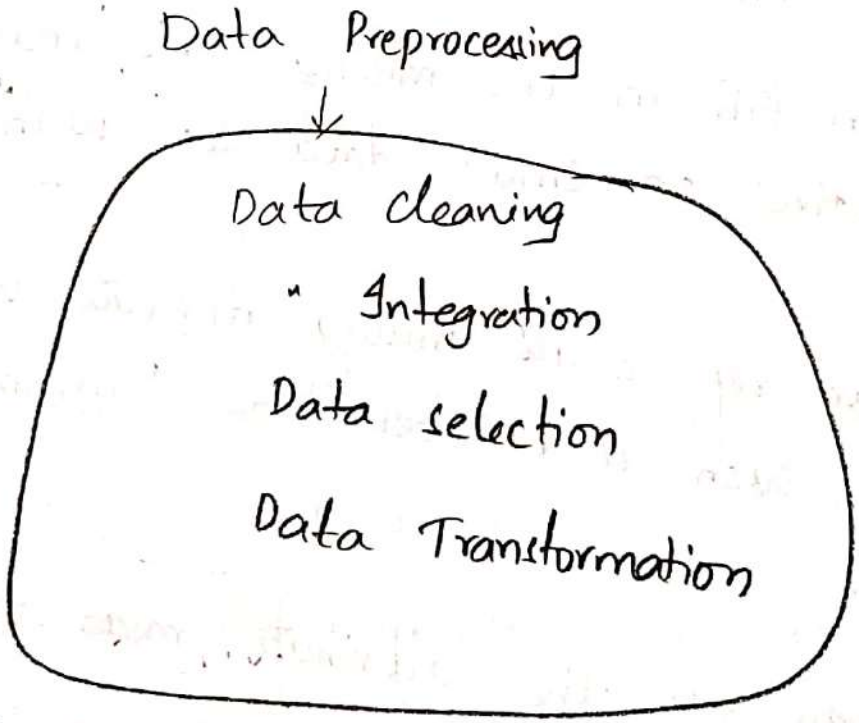
(e) Data discretization and concept hierarchy generation:-

Generalization:-

Raw data values for attributes are replaced by ranges or higher-level concepts. For Example, raw values for age may be replaced by higher-level concepts, such as youth

adult or senior. Automatic generation of concept hierarchies from numerical data.





DATA CLEANING :-

Quality of your data is critical in getting to final analysis. Any data which tend to be incomplete, noisy and inconsistent can effect your result.

* Data cleaning in data mining is the process of detecting and removing corrupt or inaccurate records from a record set, table or database.

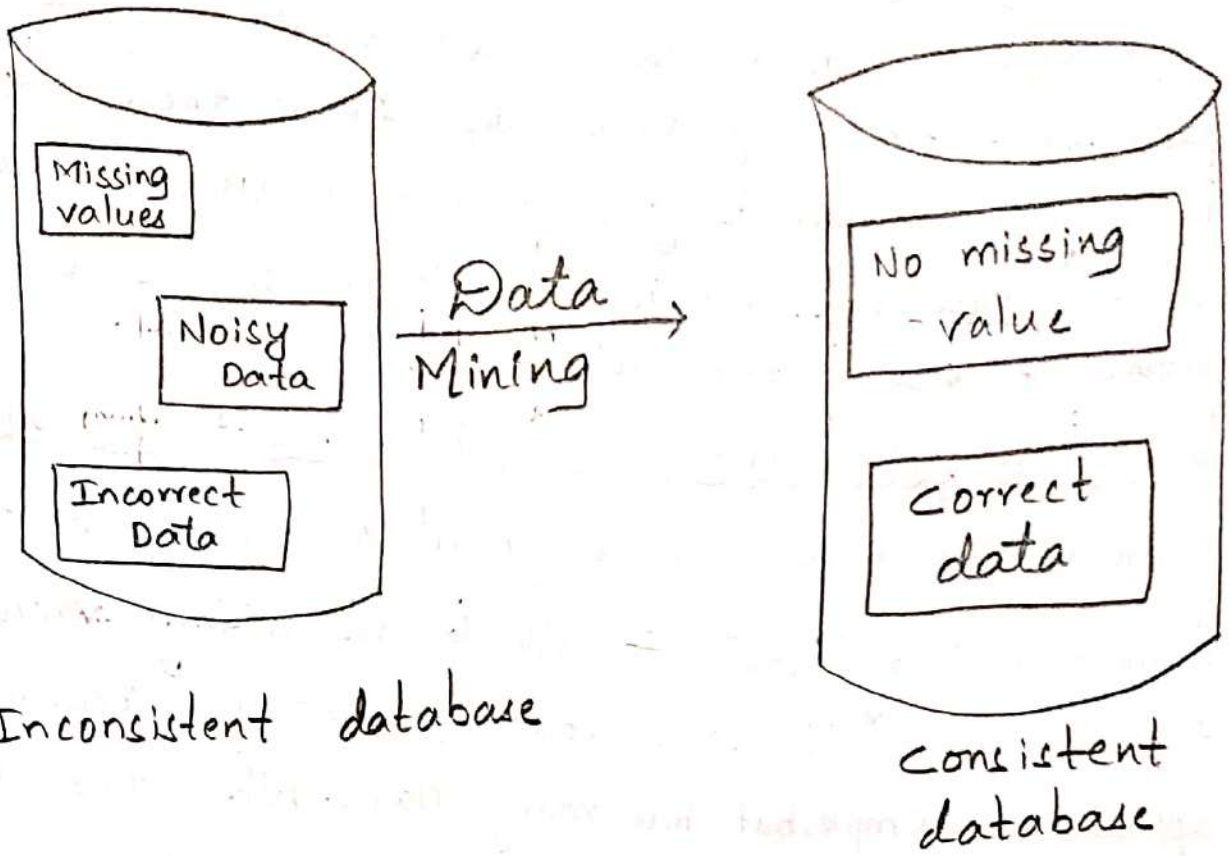
* Some data cleaning Methods:-

1) You can ignore the tuple. This is done when class label is missing. This method is not very effective, unless the tuple contains several attributes with missing values.

2) You can fill in the missing value manually. This approach is effective on small data set with some missing values.

3) You can replace all missing attribute values with global constant, such as a label like 'unknown' or minus infinity.

4) You can use the attribute mean to fill in the missing value.



* MISSING DATA:-

In data warehouses data is stored in relational format i.e in the form of rows and columns. If any of the attribute value is mistakenly not recorded for any tuple, then it will lead to inaccurate results. In order to handle these missing values, the following methods are employed by data cleaning process.

① Ignoring the tuple with missing data values:-

This approach is beneficial when more number of attribute values are missing within the same tuple. when the number of missing values varies significally.

2) Manually filling the missing data values:-

In this method, the user himself should try to find out the tuples with missing values and fill in those tuples manually. This method is generally not advantageous as it consumes more time and is not suitable to use when massive volume of data contains missing values.

3) Using a global constant to fill in the missing value:-

This method makes use of global constant such as "unknown" or "∞" to fill in the missing value. All tuples with missing value contain identical constant. This approach is simple, but has many flaws. When data analysis is being performed these values are unintentionally treated as special value by the data mining process → inaccurate result.

4) Using attribute average value:-

For Example in student database the missing value in marks attribute can be filled by calculating the average marks of all the students i.e all missing value tuples will contain identical mean value.

5) Use the most probable value to fill in the missing value:-

Eg:- Using the other customer attributes in your data set you may construct a decision tree to predict the missing values for income.

* Missing value may not always result in an Error.

DIMENSIONALITY REDUCTION:-

Dimensionality Reduction is the process of reducing the number of random variables or attributes under consideration by using different encoding schemas.

* Dimensionality Reduction methods include

a) Wavelet transforms

b) Principal Component analysis. (PCA)

a) Wavelet transforms:-

The discrete wavelet transform (DWT) is a linear signal processing technique. It transforms a vector into a numerically different vector. (D to D') of wavelet coefficients

→When applying this technique we consider each tuple as an n-dimensional data vector depicting 'n' measurements made on the tuple from 'n' different database attributes.

→Wavelet transform data can be truncated

→A small compressed approximation of the data can be retained by storing only a small ~~function~~ fraction of the strongest wavelength co-efficient.

→Removes noisy without smoothing out the main feature of data making effective for data cleaning.

→Wavelet transforms can be applied to multidimensional data such as data cubes. wavelet transforms have many real world applications including the compression of finger print images, computer vision and analysis of time-series data.

Principle Component Analysis:- (PCA)

(Korhonen – leone - KI method)

Data to be reduced consists of tuples or data vectors described by n-attributes or dimensions.

→ PCA searches for k-dimensional orthogonal vectors that can be best used to represent the data where c<=k.

→ The original data are thus projected onto a much smaller space

→ The basic procedure is as follows:-

i) The input data are normalized, so that each attribute falls within the same range. This step helps to ensure that attributes with larger domains will not dominate attributes with smaller domains.

ii) PCA computes orthonormal vectors to provide a basis for the normalized input data. These are unit vectors that point in a direction perpendicular to others. These vectors are referred to as the principal components.

→ Principal components may be used as input to multiple regression and cluster analysis.

→ PCA handles better the sparse data.

→ PCA is computationally inexpensive and it can be ordered or unordered.

FEATURE SUBSET SELECTION :-

Dimensions can be reduced ~~to~~ through feature subset Selection.

→ A database consists of massive volumes of data set which inturn are the collection of the records. Each record consists of numerous attributes. Out of these attributes set many of the attributes are duplicate, inconsistent and irrelevant. It is very time consuming if data analysis is performed on all these attributes.

→ It is difficult for a data analyst to select the relevant attributes when the characteristics of the data is unknown. The selection of irrelevant attribute can lead to poor quality pattern, confusion and degradation in performance of mining process.

→ In order to eliminate the usage of irrelevant attributes a strategy called "Attribute Subset Selection" is used. This strategy compresses the actual size of data set by deleting those attributes that are redundant and irrelevant. The advantage of using this strategy is that, the discovered patterns can be easily understood.

Feature subset selection has three approaches.

a) Embedded
b) Filter
c) Wrapper

a) Embedded:-

Data algorithms during their operations decide whether to apply a particular attribute or not. Decision tree classifier algorithm operates on this principle.

b) Filter:-

characteristics of data set are selected prior to applying datamining algorithms through an independent technique.

c) Wrapper:-

Target datamining are applied on black box to search for best subset of attributes. The following steps are followed to select subset features.

(i) Meaures to evaluate a subset:

The current features subsets must be compared against new features generated. This comparision needs an evaluation criteria to know subsets attributes for data mining operations such as clustering or classification. In wrapper method subset evaluates data mining results whereas filter method (applies) attempts to determine the performance of data mining algorithm on a set of attributes.

ii) Strategy to control new subset Generation:

In this method, all possible feature subsets are searched to select features. These can be selected using various search strategies. however the selected strategy should find optimal or near optimal feature set.

iii) Criteria to stop Feature Generation:-

The subsets can be generated in huge volume which cannot be examined individually therefore a stopping criteria is needed. This criteria is based on certain conditions such as

- The number of iterations.
- The value that measures subset whether it is met with the optimal level or not.
- Whether any improvement can be made using search strategy.
- Whether creation size of subset is obtained or not.
- Whethe evaluation criteria and simultaneous size is obtained or not.

iv) Validation Procedure:-

The result produced by data mining algorithm on target subsets need to be validated. This can be done by running the algorithm with entire features and then comparing the results against the result obtained from feature subset.

* DISCRETIZATION AND BINARYZATION - DATA DISCRETIZATION:

Data discretization converts a large number of data values into smaller ones, so that data evaluation and data management becomes very easy.

(or)

Discretization is the process of putting values into buc so that there are a limited no. of possible states. The buckets themselves are treated as ordered and discrete values.

→ There are several methods that you can use to discrete data. If your data mining solution uses relational data, you can control the number of buckets for grouping data by setting the value of the Discretization Bucket Count property. The default no. of buckets is 5.

Example:-

We have an attribute age with following values:

Age	10, 11, 12, 13, 14, 17, 19, 30, 31, 32, 38, 40, 42, 70, 72, 73, 75

Table : Before discretization

Age	10, 11, 13, 14, 17, 19	30, 31, 32, 38, 40, 42
	70, 72, 73, 75	

Table: How to discretization

Young	old	Mature

Age	Young	Mature

Table: After discretization

*There are different methods which are used for performing data discretization.

a) Supervised Discretization:

It data is discretized using class information then it is refered as supervised or organized discretization.

b) Unsupervised Discretization:

It data values are reduced by substituting them by limited interval description but without using class Information then it is refered to as unsupervised discretization

c) Top - Down Discretization:

It the process starts by first finding one or a few points to split the entire attribute range, and then repeats this recursively on the resulting intervals, then it is called top-down discretization or splitting.

d) Bottom-up discretization:

If the process starts by considering all of the continuous values as potential split-points, removes some by merging neighbour-hood values to form intervals then it is called bottom-up discretization or merging.

Techniques of Data Discretization:

1. Histogram analysis
2. Binning
3. correlation analysis

4) Clustering analysis
5) Decision tree analysis
6) Equal width partitioning
7) Equal depth partitioning
8) Entropy based discretization

1) Histogram analysis:-

Histogram analysis does not use class information so it is an unsupervised discretization technique. Histograms partition the values for an attribute into disjoint ranges called buckets.

2) Binning:

Binning is a top-down splitting technique based on a specified number of bins. Binning is an unsupervised discretization technique.

a) Equal-width binning
b) Equal-depth binning

a) Equal-width binning:-

Given a range of values [min, max] we divide in intervals of approximately same width; either we set the width arbitarily to w, or we set the desired number of bins to n, in this case w is calculated as.

$$w = max - min / n$$

Eg:-

If the range is [0,100] and we want 4 bins, each bin will have a width of

$$100 - 0/4$$
$$= 25$$

the bins will be [0,24], [25,49], [50,74], [75,100]

b) Equal-depth binning:-

Given a range of values [min,max], we place approximately the same number of instances in each bin by dividing the total number of samples nb by the desired number of samples in each bin(depth) d, in that case number of bins are calculated as:

$$n = nb/d$$

Eg:-

If the range is [0,100] for 100 samples of different values (for eg 99 is missing), we want 20 samples in each bin, the no.of bins will be

$$100/20$$
$$= 5$$

the bins will be [0,19], [20,39], [40,59], [60,79], [80,100]

Advantage:-

*Equal width binning is more simple however very sensitive to outliers in the data.

* Equal-depth binning scales well by keeping the distribution of data however the bin values may be more difficult to interpret.

3) correlation analysis:-

correlation is often used as a preliminary technique to discover relationships between variables. More precisely, the correlation is a measure of the linear relationship between two variables.

4) clustering analysis:-

clustering is the process of making a group of abstract objects into classes of similar objects

→ cluster analysis is a popular data discretization method. A clustering algorithm can be applied to discrete a numerical attribute of A by partitioning the values of A into clusters or groups.

→ Each initial cluster or partition may be further decomposed into several subcultures, forming a lower level of the hierarchy.

5) Decision tree analysis:-

A decision tree is a structure that includes a root node, branches and leaf nodes. Each internal node denotes a test on a attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is root node.

6) Entropy based discretization:-

Here entropy means lack of order or prediction.

→ Uses the concept called information gain.

→ It is supervised top-down splitting. Each value of 'A' can be considered as a potential interval boundary or splitting point to partition the range 'A' i.e a splitting for 'A' can partition the tuple in 'B', into 2 subsets satisfying $A \leq$ splitting point or $A >$ splitting point respectively. Hereby creating a binning discretization.

tuples:

$$\text{Info}_A(D) = \frac{|D_1|}{|D|} \text{Entropy}|D_1| + \frac{|D_2|}{|D|} \text{Entropy}|D_2|$$

Classes:

$$\text{Entropy}(D_1) = \sum_{i=1}^{m} P_i \log_2(P_i);$$

* DATA TRANSFORMATION:

In data transformation process data are transformed from one format to other format, that is more appropriate for data mining.

In normalization we have 3 methods

1) Min-max normalization
2) Z-score normalization
3) Decimal scaling normalization.

1) Min-Max normalization:-

Performs a linear transformation on the original data. Suppose that min A and max A are the minimum and maximum values of an attribute 'A'.

min-max normalization maps a value v of A to v' in the range [new-min A, new-max A]

$$v' = \frac{v - \min A}{\max A - \min A} (new\_max_A - new\_min_A) + new\_min_A$$

Example:-

Q) suppose the minimum & maximum values for an attribute income are 12,000 and 98,000 respectively. we would like to map income range [0,1] by min-max normalization a value of 73,600 for income transformed.

Sol:- $$v' = \frac{v - \min_A}{\max_A - \min_A} (new\_max_A - new\_min_A) + new\_min_A$$

$$v' = \frac{73,600 - 12000}{98000 - 12000} (1-0) + 0$$

$$= 0.716$$

* MEASURES OF SIMILARITY AND DISSIMILARITY

Distance or Similarity measures are essential to solve many pattern recognition problems such as classification and clustering. various distance/similarity measures are available in literature to compare two data distributions. As the names suggest, a similarity measures how close two distributions are

1. Similarity measure
2. Dissimilarity measure.

Similarity measure:-

The Similarity between the two objects is a numerical measure of the degree to which the two objects are alike. Consequently similarities are higher for pairs of objects that are more alike. Similarities are usually non-negative and are often between [0, 1]

Dissimilarity Measure:-

The dissimilarity between two objects is the numerical measure of the degree to which the two objects are different. Dissimilarity is lower for more similar pair object.

2) Z-score normalization:-

This is also known as zero-score normalization. The values for an attribute A are normalized based on the mean and standard deviation of 'A'. A value of 'A' is normalized to v' by computing.

$$v' = \frac{v - \bar{A}}{\sigma A}$$

Example:-

a) Suppose the mean and standard deviation of the value for the attribute income are 54,000 & 16,000 respectively. with z-score normalization a value of 73,600 is transformed to

$$v' = \frac{73,600 - 54,000}{16,000}$$

$$= 1.225$$

3) Decimal scaling normalization:-

By moving the decimal point of values of attributes 'A' the number of decimal point moved depends on maximum obsolute value of 'A'. A value 'v' of attribute 'A' is normalized to v' by computing

$$v' = \frac{v}{10^I}$$

where I is smallest integer such that $\max(|v'|) < 1$