Unit - II

Data Mining :-

Data mining refers to extract (or) mining knowledge (or) patterns from large amounts of data.
Similar type of data

Uses:-

Data mining is used to extracting significant, implicit previously unknown and potentially useful data.

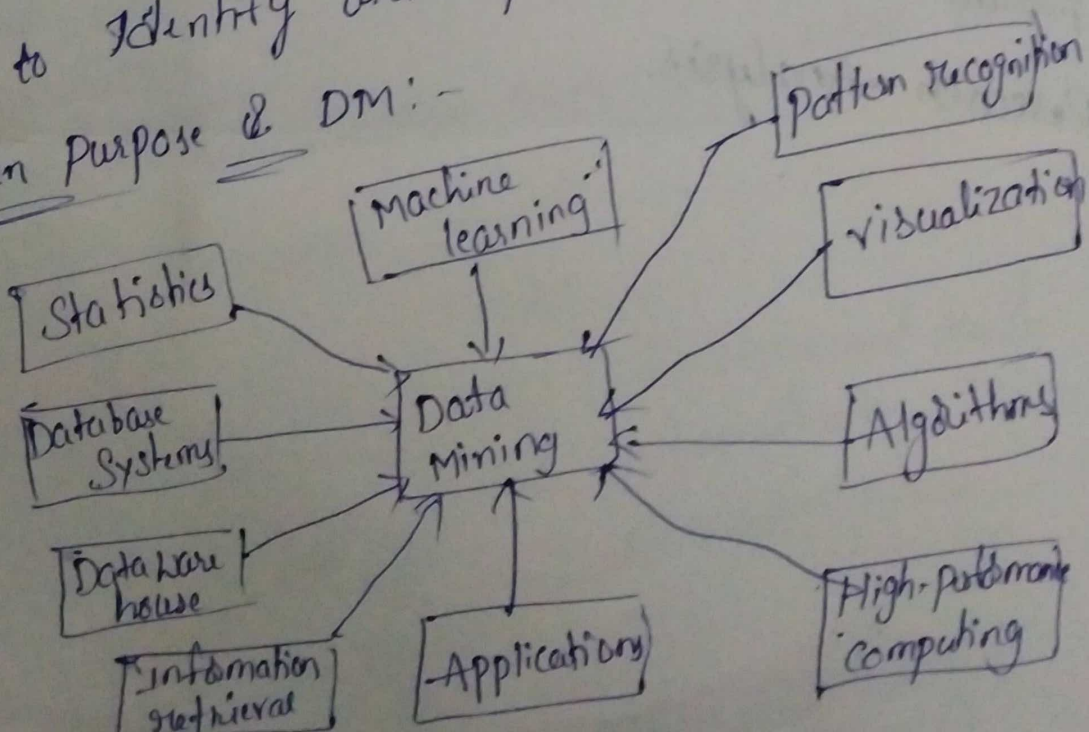
It is also known as knowledge discovery in data (KDD) (or) knowledge extraction.

Process:-

Mining process starts with giving certain input as the data to data mining tools that uses some statistics and algorithms to show reports (or) patterns.

Various techniques such as classification, clustering, regression and outlier analysis etc. are applied to data to identify useful patterns.

Main Purpose of DM:-



Basically, Data mining has been integrated with many other techniques from other domains such as Statistics, machine learning etc.

The whole process of Data mining consists of three main phases:

1. Data pre-processing - Data cleaning, integration, selection, and transformation takes place.
2. Data Extraction - occurence of exact data mining.
3. Data Evaluation and presentation - Analyzing and presenting results. (Diagram)

Applications of DM:-

1. Financial Data Analysis: → Banking services include loans, investments, credit, debit etc.
2. Biological Analysis: Biological data deals with Genomics, proteomics (protein) etc. → medical pharmacy, cancer therapies
3. Scientific Analysis → Scientific domains (Geosciences, Astronomy, Meteorology) chemical engineering etc.
4. Intrusion Detection → Viruses attack.
5. Fraud Detection
6. Research Analysis.

KDD:- Knowledge Discovery in Databases:-

Data cleaning:- unnecessary & noisy data

Data Integration:- multiple sources data → single data base
transfer

Data Selection:- Selecting relevant data.

Data Transformation:- Raw data converts Required format. (convert)

Patterns-Evaluation & presentation :- Identifying the relevant patterns.

Knowledge:- using the Visualization tools presentation knowledge.

Generate reports
tables
rules (discriminant rule
classification "
characterization "

Steps / Tasks & data processing / Data preprocessing in DM :-

Databases have noisy, missing and inconsistent data due to their huge size.

→ Low quality data leads to low quality data mining.

→ Data prepsocessing is used to improve the quality & data and minining result.

various techniques like data cleaning, data integration, data reduction and data transformation are used in data preprocessing.

Fig :-

1. Data cleaning :- It is applied to remove noise and correct inconsistant & data fills missing values, smooth out noise while identifying outliers.

2. Data Integration :- merges data from multiple sources into a single data source such as datawarehouse which helps to reduce redundent data.

3. Data reduction :- Reduces the size & the data by using aggregation, clustering methods (or) by eliminating redundent data.

4. Data Transformation :- Data is scalled to fall with in a smaller range like 1.0 → 0.0.

Data pre-processing :-

Data mining technique used to transform the raw data into organized format (knowledge).

Real-world data is often incomplete, inconsistent and is, lacking in certain behaviours (or) trends, and is likely to contain many errors.

Why do we need Data preprocessing :-

There are many factors comprising data quality, including accuracy, completeness (with all the required fields), consistency, timeliness (update time to time), believability (reflects how much the data are trusted by users), and interpretability (reflects how easy the data are understood).

The data you wish to analyze by data mining techniques are

* Incomplete (lacking attribute values (or) certain attributes of interest, (or) containing only aggregate data).
* Inaccurate (or) noisy (containing errors, (or) values that deviate from the expected); and.
* Inconsistent (e.g containing discrepancies in the departme-nt codes used to categorize items).

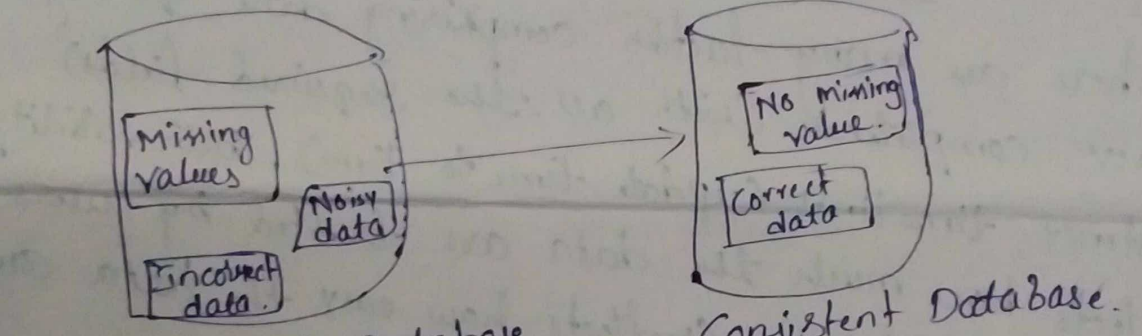
Data preprocessing improves the quality of the data in data warehouse.

Major Tasks & Data preprocessing.

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

1. Data cleaning :-

Data cleaning (or data cleaning) routines attempt to fill in missing values, smooth out noise. while identifying outliers, and correct inconsistencies in the data.



Inconsistent Database Consistent Database.

Approaches in Data cleaning :-

1. Missing values
2. Noisy data.

1. Missing values :-

1) Ignoring tuples with missing data values.
2) Fill in the missing values manually
3) Use a global constant to fill in the missing values e.g(NA)
4) Use a measure of central tendency for the attribute (e.g. the mean (or) median) to fill in the missing value.
5) Use the most probable value to fill in the missing value. (e.g., using decision tree).

Q: Noisy Data :-

Noise is a random error (or) variance in a measured ~~values~~ variable.

Approaches in Noisy Data :-

i. Binning
ii. Regression.
iii. outlier analysis.

i. Binning :-

a) partition into equal frequency bins
b) Smoothing by bin means.
c) Smoothing by bin boundaries.

Ex :- 6, 10, 17, 22, 22, 25, 27, 30, 36.

a. partition into equal frequency bins.

Bin 1 : 6, 10, 17
Bin 2 : 22, 22, 25,
Bin 3 : 27, 30, 36.

b. Smoothing by by bin means.

Bin1 : 11, 11, 11
Bin2 : 23, 23, 23
Bin3 : 31, 31, 31.

c. Smoothing by bin boundaries

Bin 1: 6, 6, 17

Bin 2: 22, 22, 25

Bin 3: 27, 27, 36.

ii) Regression :- 2 methods.

linear Regression

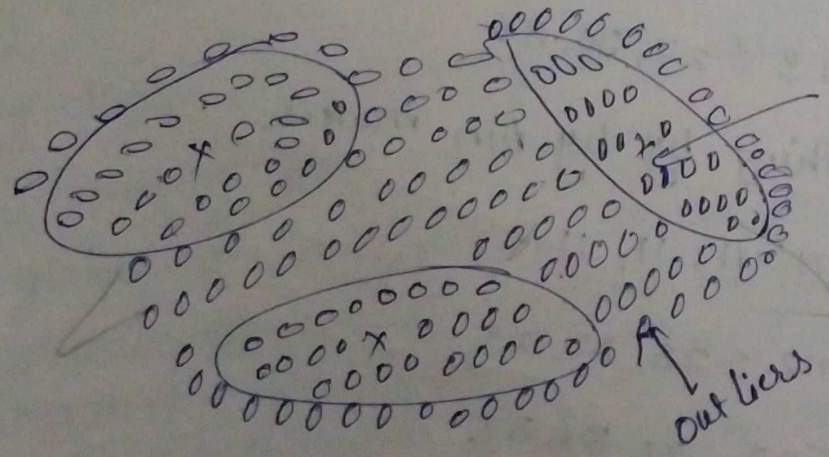
multiple linear regression.

Linear Regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.

Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multiple dimensional surface.

iii) Outlier analysis :-

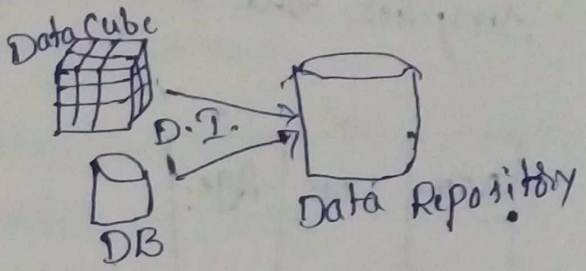
Outliers may be detected by clustering, for example, where similar values are organized into groups (or) "clusters". Intuitively, values that fall outside of the set of clusters may be considered outliers.



Data Integration and ~~transformation~~:-

Data Integration:-

Merging of data collected from multiple sources. Careful Integration can help reduce redundancies and inconstancies in the resulting dataset.



~~Approaches~~ in Data Integration:-

1. Entity Identification problem.
2. Redundancy and corrulation analysis.
3. Tuple Duplication.
4. Data value conflict Detection and Resolution.

1. Entity Identification problem:-

Schema integrateion and object matching are very important issues in Data integration.

Schema Integration - Mismatch in Attribute names.

~~Eg~~: cust_id, customer_id, cust_no etc.

handling blank, zero, null values.

object Matching:- Mismatch in Structure of the data.

~~Eg~~:- Discount issues.

currency type.

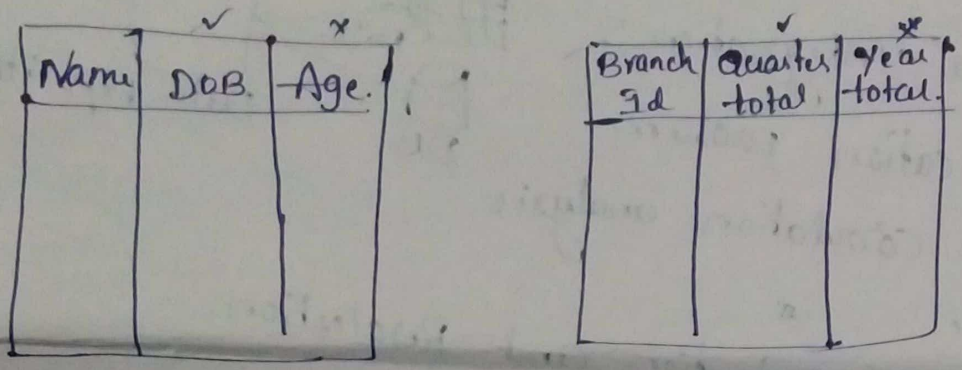
2. Redundancy and correlation analysis:-

Redundancy - An attribute may be redundant if it can be "derived" from another attribute (or) set of attributes.

Eg:- DOB, Age

Quarter Sales, year Sales.

Correlation analysis:- Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.



3. Tuple duplication:-

The use of denormalized tables (often done to improve performance by avoiding joins) is another source of data redundancy. Inconsistencies often arise b/w various duplicates due to inaccurate data entry (or) updating some but not all data occurences.

Name	~~Age~~ ~~DOB~~	Branch	occupation	Address.
A	25	TPG	Govt	TPG
B	30	Tnk	Govt	Rjy
A	25	TPG	private	TPG
C	30	tnk	~~private~~ Govt.	Rjy.

4. Data value conflict Detection and Resolution:-

-Attribute values from different sources may differ. This may be due to differences in representation, scaling (or) encoding.

-Eg:- Weight (metric (or) british empherial units)
School curricalum (grading system)

-Attributes may also differ on the abstraction level, where an attribute in one system is recorded at, say, a lower abstraction level than the "same" attribute in another.

-Eg:- monthly total sales in a store.
& monthly total sales from all stores in that region.

---

3. Data Transformation:-

The data are transformed (or) consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

Data Transformation Strategies include:

1. ~~Smoothing~~ Normalization
2. Attribute selection.
3. Discretization

4B. concept hierarchy generation.

1. Normalization: The attribute data are scaled so the with in a smaller range, such as.

-1.0 to 1.0

(or) 0.0 to 1.0

Data Transformation by normalization:-

The measurement unit used can affect the data Analysis. To help avoid dependence on the choice of measurement units, the data should be normalized (or) Standardized.

Normalizing the data attempts to give all attributes an equal weight.

Methods for data normalization:-

1. min-max normalization.
2. Z-score normalization
3. normalization by decimal scaling.

let A be a numeric attribute with n observed values,

$V1, V2, \ldots, Vn$.

min-max normalization performs a linear transformation the Oginal data. Suppose that minA and maxA are the minimum and maximum values of an attribute, A.

$$Vi' = \frac{Vi - \min A}{\max A - \min A} (new\_\max A - new\_\min A) + new\_\min A$$

Let the minimum and maximum values are 10 and 100 respectively Then min-max normalization for value 75 is.

$$Vi' = \frac{Vi - \min A}{\max A - \min A} (new\_\max A - new\_\min A) + new\_\min A$$

$$Vi' = \frac{75-10}{100-10} (1-0) + 0$$

$$= \frac{65}{90} * 1 = 0.722$$

1. Smoothing:-

To remove noise from the data, techniques include Binning, Regression, and clustering (data cleaning)

2. Attribute construction:-

New Attributes are constructed and added from the given set of attributes. where new attributes can be created and added to given set of attributes to simplify mining process more efficient.

3. Aggregation:-

Summary and Aggregation functions can be applied on the data for constructing data cube for data analysis (data reduction).

4. Discretization:- Numeric values are replaced by Interval labels (or) conceptual labels. The labels in turn can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numerical attributes.

Eg:- 0-10, 11-20 etc.
Youth, adult, Senior.

5. Concept hierarchy generation for nominal data:-

- Attributes of lower level concepts can be generalized to higher-level concepts.

- Eg1:- Street A, Street B, Street C can be generalized to town (or) city.

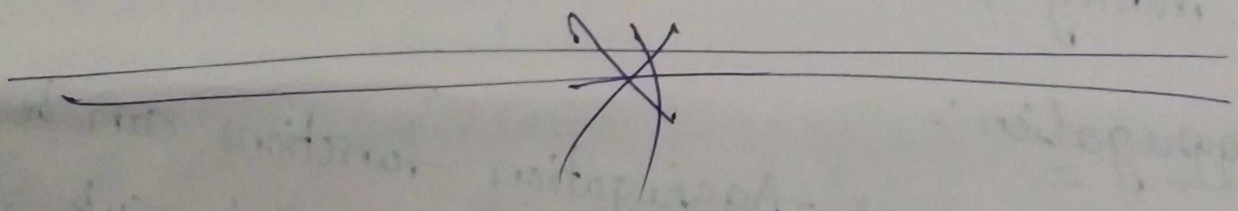
- Eg2:- lower level groups into higher level concepts.

6. Normalization:- The attribute data are scaled so as to with in a smaller range, such as

-1.0 to 1.0

(or) 0.0 to 1.0.

Data



2. Z-Score normalization :- Suppose the mean and Standard deviation values are 64 and 8.

Then z-score normalization for value 75 is.

$$v_i' = \frac{75-64}{8}$$

$$= \frac{11}{8} = 1.375.$$

3. Decimal Scaling - Suppose that the recorded values of A range from -845 to 945. To normalize by decimal Scaling, we therefore divide each value by 1000 (range from -1 to 1).

So that -845 normalizes to -0.845
and 945 normalizes to 0.945.

Data Reduction:-

Data Reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume. Minning on the reduced dataset should be more efficient yet produce the same (or almost the same analytical results

Methods for Data Reduction:-

1. Dimensionality Reduction
2. Numerosity
3. Data compression.
4. Data cube aggregation.
5. Attribute Subset Selection.
6. ~~Discreti~~

1. Dimensionality Reduction:-

It eliminates the redundant attributes which are weakly important across the data.

i. Stepwise forward Selection.
ii. Stepwise backward elimination.
iii. Decision tree induction.

i. Stepwise forward Selection:

Eg:- Initial / original attribute Set,
$\{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8\}$.

1: $\{\}$ → Initial reduced Set.
2: $\{P_2\}$
3: $\{P_2, P_4\}$

4: {P2, P4, P6}.

5: {P2, P4, P6, P8} ⇒ Result reduced set of attribute

(ii). Stepwise backward elimination:

Eg:- Initial attribute set {P1, P2, P3, P4, P5, P6, P7, P8}

Initial reduced set {

Step1: {P1, P2, P3, P4, P5, P6, P7, P8}

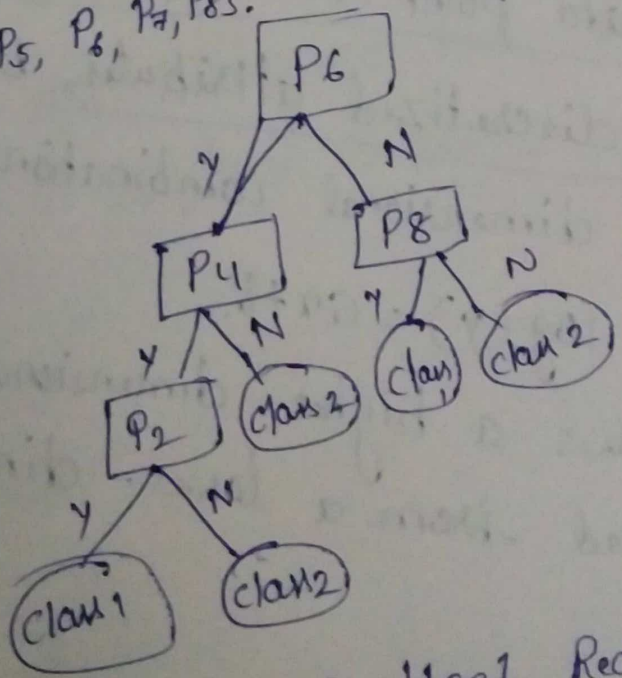
2: {P2, P4, P5, P6, P7, P8}

3: {P2, P4, P6, P7, P8}

4: {P2, P4, P6, P8} → Resultant reduced set of attributes.

(iii). Decision tree induction:-

{P1, P2, P3, P4, P5, P6, P7, P8}.



{P2, P4, P6, P8} → Resultant Reduced set of attributes.

2. Numerosity Reduction:-

Replaces the original data with small form of data representation. There are 2 methods parametric and Non-parametric reduction.

i. Parametric method: Used to estimate the data. So that only parameters of data are required to be stored, instead of actual data.

a) Regression:-

Simple linear Regression (to fit a straight line)

$(Y = ax + b)$

multiple linear Regression.

(with 2 or more predictor variables)

b). Log-linear model: Used to estimate the probability of each data point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.

$\log(y) = ax + b.$

This allows a higher-dimensional data space to be constructed from a lower-dimensional attributes.

ii. Non-Parametric method: Used to store reduced representation of data. It includes.

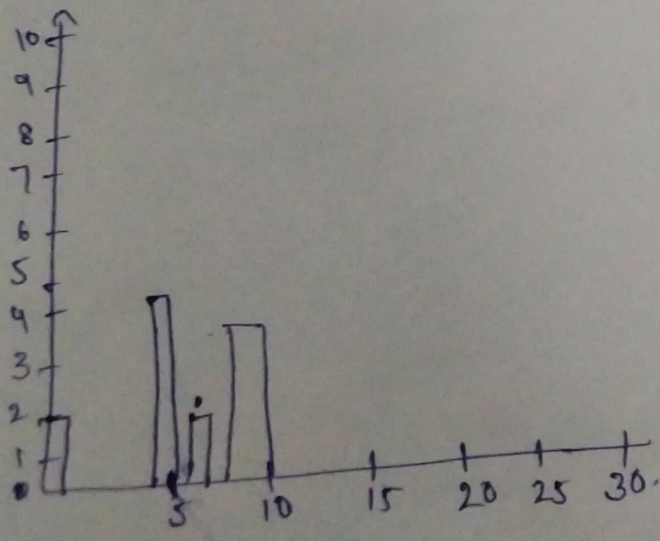
a. Histrograms.
b. Clustering
c. Sampling
d. Data cube aggregation.

a. Histrograms :- Binning to approximate data distributions.

1. Equal-width Histrogram
2. Equal-frequency Histrogram

① According to this histrogram rule, data set is divided into buckets of constant width.

② According to this histrogram rule, data set is divided into buckets of constant frequency



b) Clustering:-

Partitions the whole data into different clusters. Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid.

c) Sampling:-

A large data set to be represented by a much smaller random data sample.

fig:-

d) Data cube aggregation:- The data can be aggregated so that the resulting data summarize the total sale per year.

fig:-

3. Data compression:-

Reduces the size of the files using different encoding mechanisms. There are 2 types.

i. Lossless compression.

Without any loss after compression.

ii. Lossy compression:

The decompressed data may differ to the original data but are useful enough to retrieve information from them. They are

a) Discrete wavelet Transforms.
b) principal component Analysis.

a. Discrete wavelet Transforms:-

Transforms pixels of the images into wavelets, those will be used for wavelet based compression and coding.

Eg:- An image of size 10MB compressed to 100kB.

b) pr

b) Principal component Analysis:

Used to reduce data size using 'k' orthogonal vectors. Unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal component.