Trend means a general tendency to change the data.

Prediction in Data :-

Eg:- Disease Prediction
Rainfall Prediction
Predicting Winners in Football

→ A prediction or forecast is about a future event or future data.

Association :

Relation among data.

Eg When we buy bread we are expected to buy jam.

Data Mining :-

It is the process of extracting knowledge or insights from large amounts of data using various statistical and computational techniques

It is the process of extracting info to identify patterns, trends and useful data that would allow the business to take the data-driven

decision from huge sets of data

Data can be structured, semistructured or unstructured and can be stored in various forms such as databases, data warehouses and data lakes

data warehouse - collection of databases
data lakes - centralized repository - (movie booking)

Extra ⇒ ERP - Enterprise Resource Planning A software which combine multiple departments

Application areas → marketing, finance, healthcare, telecommunication, etc

## Data Mining Architecture / Working of Data Mining

Components
- Data Sources
- Data Mining Engine
- Data Warehouse Server
- Pattern Evaluation
- GUI
- Knowledge Base

classmate
Date
Page

## Data Sources :-

The actual source of data is the database, data warehouse, www, text file or other documents.

Organizations store data in database or data warehouses.

Some of the data sources may be:

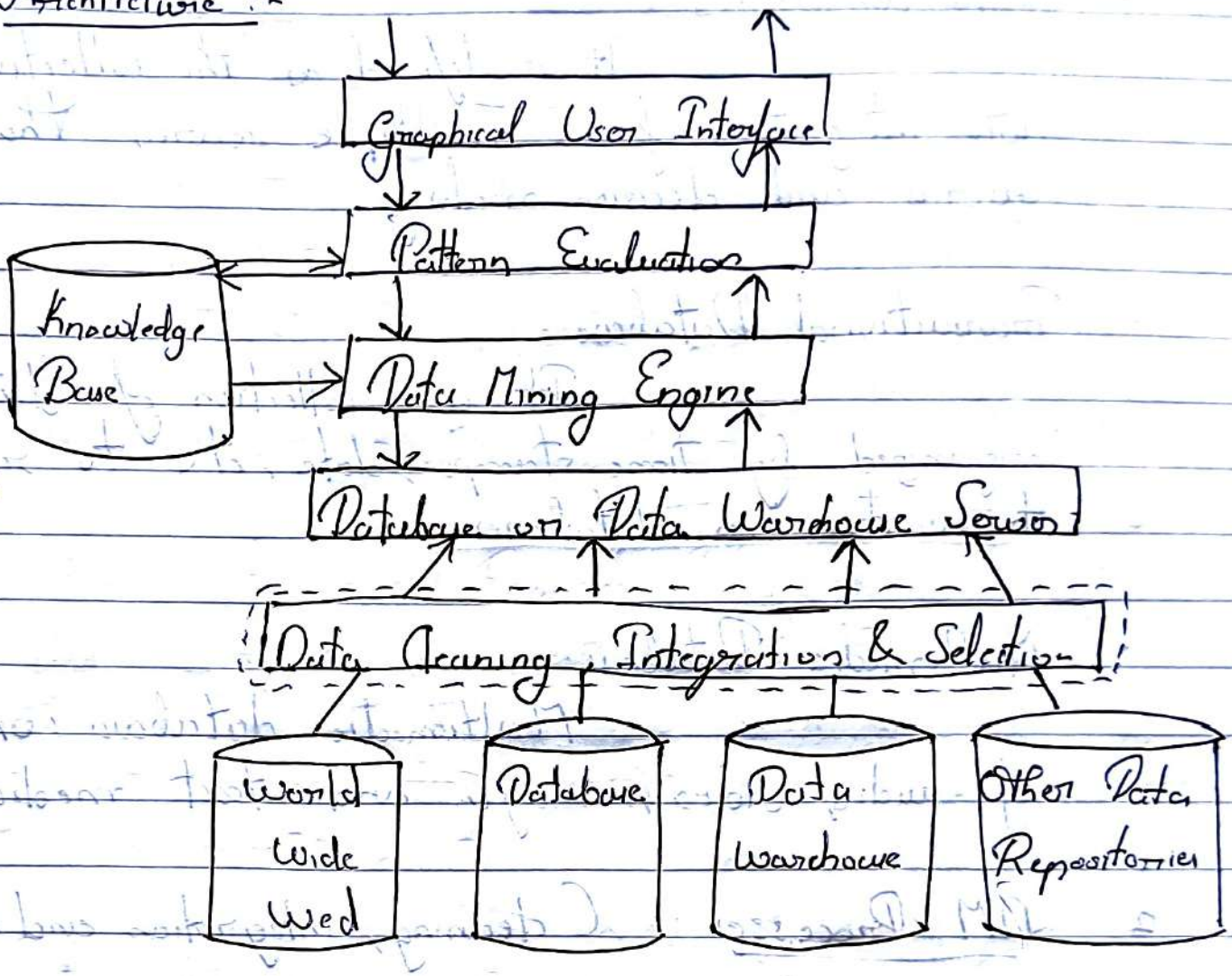
1. Flat Files :-

Data files in text or binary form with a structure that can be easily extracted by data mining algorithms.

## Architecture

Essay

## Architecture :-




## Components:-

1 Data Sources:-

. Relational Databases:-

A relational database is defined as the collection of data organized in tables with rows and columns.

- Data Warehouses :-

It is defined as the collection of data integrated from multiple source that will queries and decision making.

- Transactional Database :-

It is a collection of data organized by time stamps, date, etc to represent transaction in databases.

- Multimedia Database :-

Multimedia database consists of audio, video, image and text media.

2 DM Processes :- (cleaning, integration and selection)

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated and selected.

↳ Data Cleaning :-

As the information come from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be

- Data Warehouses :-

It is defined as the collection of data integrated from multiple sources that w queries and decision making.

- Transactional Database :-

It is a collection of data organized by time stamps, date, etc to represent transaction in databases.

- Multimedia Database :-

Multimedia database consists of audio, video, image and text media

2 DM Processes :- Cleaning, integration and selection

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated and selected.

↳ Data Cleaning :-

As the information come from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be

complete and accurate.

So the first data require to be cleaned and unified

↳ Data Integration :-

Data integration in data mining refers to the process of combing data from multiple source into a single, unified view.

↳ Data Selection :- Selection of relevant data → collection of databases

3 Database or Data Warehouse Server :-

The database or data warehouse server consists of the original data that is ready to be processed.

Hence the server is cause for retrieving data that is based on data mining as per user request. After selection, dat is stored here.

4 Data Mining Engine :-

- The data mining engine is a major component of any data mining system.
- It contains several modules for operating data mining

tasks, including association, classification, clustering, prediction, time-series analysis, etc.

- It comprise instruments and software used to obtain insights and knowledge from data collected from various data source and stored within the data warehouse.

→ It process data using input from user. i.e requirement,

5 Pattern Evaluation Module :-

D

- The pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value.

- It collaborate with the data mining engine to focus the search on exciting pattern

- The pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining technique used.

- For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining

procedure to confine the search to only fascinating patterns.

6 Graphical User Interface:

- The graphical user interface (GUI) module communicates between the data mining system and the user.
- The module helps the user to efficiently use the system without knowing the complexity of the process.
- The module cooperates with data mining system when the user specifies a query or a task and displays the results.

7 Knowledge Base:-

- The knowledge base is helpful in the entire process of data mining.
- It might be helpful to guide the search or evaluate the stake of the result patterns.

- The knowledge base may even contain user views and data from user experience that might be helpful in the data mining process.

- The data mining engine may recieve inputs from the knowledge base to make the result more accurate and reliable.

---

Data Source:- Spatial Database:-

Stores geographical information. Stores data in the form of coordinates, topology, lines, polygons, etc.

DM Process

Cleaning:-

- Fixing Structural Errors
- Managing Unwanted Outliers
- Removal of Unwanted Observations
- Handling Missing Data

After Integrating the data we need to select the relevant data. This process is called

data selection.

KDD Process (Knowledge Discovery in Database)

Why KDD?

When dealing with loads of data KDD comes into play.

Eg→ Customer Segmentation → Application Area
↳ Identifying customers based on their trend. Then grouping them who has the same trends.

→ Market Basket Analysis
↳ If you buy bread then there is a chance that you'll buy jam or butter. So the bread and jam are placed together or next to each other.

If you buy a drawing then crayons might be bought

→ Demand Prediction
↳ Predicting prices like petrol prices in the future.

Def. KDD is a process that involve the extraction of useful, previously unknown and potentially valuable information from large datasets.

- It is a field of interest to researchers in various fields like:
  - AI
  - Machine learning
  - Pattern recognition
  - Databases
  - Knowledge Acquisition for expert systems,
  - Data Visualization.

Essay KDD Process

KDD process is an iterative process and it requires multiple iterations of the various steps to extract accurate knowledge from the data.

i Data Cleaning :-

Aka data cleansing or data scrubbing. It is process of identifying and correcting errors, inconsistencies and inaccuracies within a dataset

ERP - Enterprise Resource Planning
LOB - Last Objects
CRM - Customer Relationship Management

It can be done by filling the missing values (meaningless data) and smoothing noisy data, analyzing and removing outliers, and removing inconsistencies in the data.

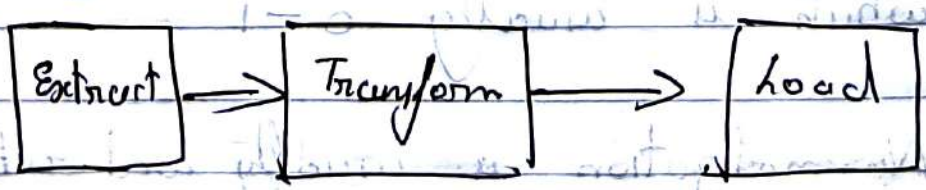
- Inconsistencies can be recorded in various transactions, during data entry, or arising from integrating information from multiple databases.

ii Data Integration :

- It is defined as the process of combining data from multiple sources to create a unified dataset.

- The heterogenous data from multiple sources are combined in to a common source.

- It is also known as ETL (Extract-Transform-load) process



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 the data into appropriate form required by mining procedure

Common methods:
- Discretization
- Standardization
- Normalization

↳ Normalization :-

Normalization is used to rescale the features to a standard range of values which is usually 0 - 1

Normalization is usually used when diffrent features have different range of values and some feature might contribute more to the

model learning process, normalization helps in equilizing the range of the features and makes sure that the features contribute equally to the learning process. algorithm.

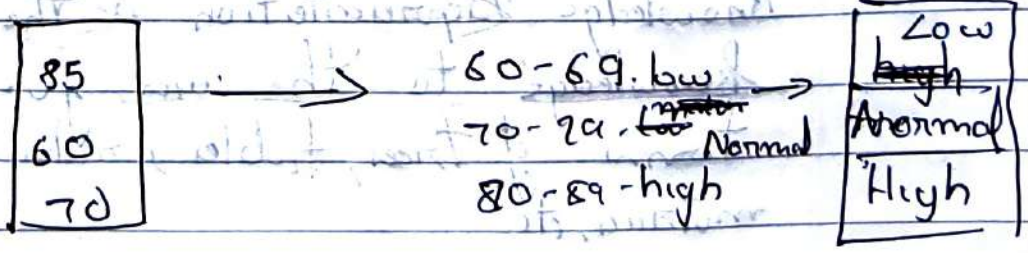
Minimum and maximum value are used for scaling

↳ Standardization:-

- It is used to transform the data into a standardized format
- Mean and standard deviation is used for scaling.

↳ Discretization

- Discretization is used to convert a huge number of data values into smaller ones so that the evaluation and management becomes easy.



## Data Mining :

→ Data mining is defined as techniques that are applied to extract patterns potentially useful.

→ It transforms task relevant data into patterns, and decides purpose of model using classification or characterization.

## Pattern Evaluation :

- It is the process of assessing the quality of discovered pattern.
- This process is important in order to determine whether the patterns are suitable for users or not. ie (if they are useful and if they can be trusted).

## Knowledge Representation

- Knowledge Representation is the presentation of knowledge to the user for visualization in terms of trees, tables, rules graphs, charts, matrices, etc.

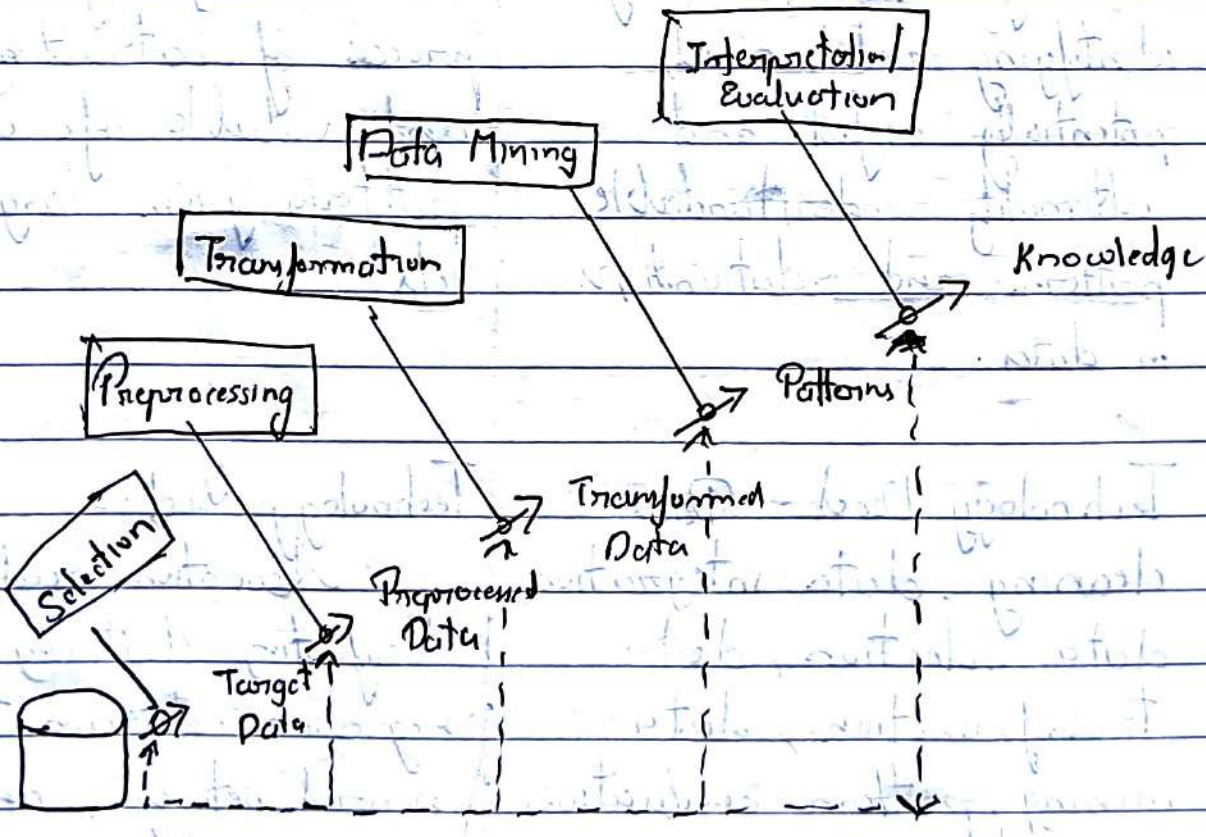
- It is a place where visualization and knowledge representation techniques are used to present mined knowledge to users

Note: Data mining is a step in KDD process

- I/p of data mining - pattern, association or insights.
  O/p

| KDD | Data Mining |
|---|---|
| • KDD refers to a process of identifying valid, novel, potentially useful and ultimately understandable patterns and relationships in data | Data Mining refers to a process of extracting useful and valuable info or patterns from large data sets |
| • Technology Used:- Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation and visualization | Technology Used :- Association rules, classification, clustering, regression, decision trees, neural network, and dimensionality reduction |

| | |
|---|---|
| Output: Structured info such as rules and models that can be used to make decisions or predictions | Output: Patterns, associations or insights that can be used to improve decision making or understanding |
| Focus is on the discovery of useful knowledge, rather than simply finding patterns in data | Data mining focus is on the discovery of patterns or relationships in data |



Essay

# Data Pre-Processing

→ It is the process of converting raw data into an understandable format.

## Steps

→ Data Cleaning
→ Data Integration
→ Data Transformation
→ Data Reduction → (Similar to Selection)

Step 2

## Data Cleaning:-

Involves identifying and correcting errors or inconsistencies in the data such as missing value, outliers, duplicates.

## I Methods for handling missing data

i Ignore the tuples:-

↳ This approach is suitable when the dataset is large and multiple values are missing in a tuple

ii Fill the Missing Values:-

→ Manually or by most probable value.

iii Use a global constant to fill the missing value

→ Replace all missing attribute by same constant

iv Use a measure of central tendency for the attribute to fill the missing value.

→ In this replace all the missing attribute by central tendency values such as mean or median

II Methods for handling Noisy data.

- Noisy data is a meaningless data that can't be interpreted by machines

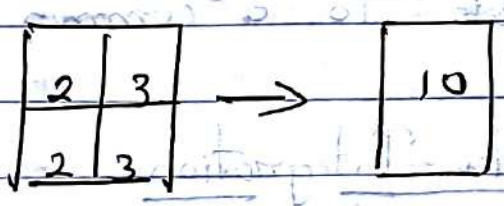
- It can be generated due to faulty data collection, data entry errors, etc

Methods → Binning
↳ Regression
↳ Clustering.

Binning :

→ The data is divided into several segments of equal size.

After that, the different methods are executed to complete the task




Regression

Data can be made smooth by applying a regression function. It can be linear (having 1 independent variable) or multiple (having multiple independent variable).

Clustering

→ Clustering is also known as Outlier Analysis.
→ Groups the data in a cluster.
→ Then the outliers are detected and can be removed

Step 2 Data Integration

- It combines data from multiple sources and provides a unified view of data (heterogenous data)
- It identifies and access different data sources, mapping data to a common format

Issues in Data Integration

i Data Redundancy
ii Duplicate data attributes - (columns)
iii Irrelevant attributes
iv Entity Identification Problem

Data Redundancy :-

→ This happens when we merge data from multiple databases

→ If the redundant data is not removed, incorrect results will be obtained during data analysis.

## Duplicate Data Attribute

Identifying them and removing them.

## Irrelevant Attributes

Some data are not important and they are not considered while performing the data mining tasks.

## Entity Identification Problem:

- The integration of data from multiple resources, some data resources match each other and they will become redundant if they are integrated.
- For example: A.cust-id = B.cust-number. Here A, B are two different database table. cust-id is the attribute of table A, cust-number is the attribute of table B.
- Here cust-id and cust-number are attributes of different tables and there is no relationship between these tables but the cust-id attributes and cust-number attributes are taking the same ~~number~~ values

- It helps in detecting and resolving data value conflicts.

## Data Transformation:

- This involves converting the data into a suitable format for analysis
- Common techniques used in data transformation include.

The different methods are:
- Normalization
- Standardization
- Discretization

## Methods for Data Normalization

- Decimal Scaling
- Min-Max Normalization
- Z-Score Normalization (Zero-mean Normalization)

## Data Reduction:-

Data reduction is a crucial

step in the data mining process that involves reducing the size of the dataset while preserving the important information.

- This is done to improve the efficiency of data analysis and to avoid overfitting of the model.

- Data cube aggregation
- Attribute subset selection
- Dimensionality Reduction

## Data Cube Aggregation:

- The technique is used to aggregate data in a simpler form.

- Data Cube Aggregation is a multi dimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction.

## Attribute Subset Selection:

- The large data set has many attributes, some of which are irrelevant to data mining.

- The core attribute subset selection reduces the data volume by eliminating redundant and irrelevant attributes.

- In each iteration, one attribute is added to a reduced set.

4. Stepwise Backward Elimination

- Here all the attributes are considered in the initial set of attributes.

- In each iteration, one considered in the initial is eliminated from the set of attributes whose p-value is higher than significance level.

Combination of Forward Selection and Backward Elimination

- The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently.

- This is the most common technique which is

generally used for attribute selection.

## Dimensionality Reduction

- The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.

- It is a way of converting the higher dimensions dataset into lower dimension dataset ensuring that it provides similar information.

## Problems

## Measures of Central Tendency

- Mean
- Median
- Mode
- Mid Range
- Range

→ Mean :- $(\bar{x})$

$$\text{Mean} = \frac{\text{Sum of Observation}}{\text{No of Observations}}$$