Unit 5 -

Data Mining

- A procedure of extracting information from huge sets of data.
- also defined as mining knowledge of data.
- Data mining domain has mining as fundamental process involved in every application.
- It is a key step in KDD process.

Data Mining Tasks:

① Descriptive Tasks ② Predictive Tasks

- aim to summarize / describe the patterns, relationships, or structures within a dataset.
- These tasks provide insights to the data and help in understanding its underlying properties.

Types :-

(i) Clustering - grouping of similar data points based on patterns, or some attribute.
Eg. Grouping customers based on age, habits or preferences.

(ii) Association Rule Mining - finding relationships in dataset
Eg. Who buy milk often buy bread.

(iii) Anomaly Detection - identify rare data points
Eg. Detecting fraud transaction in Banking systems.

② Predictive Tasks

- focus on historical data to predict future outcomes.
- These tasks are essential for decision-making & forecasting.

Types :-

(i) Classification - classify data Eg. fraud / not fraud
(ii) Regression - predict continuous numeric values
Eg. prediching house price
(iii) Decision Trees -

	Descriptive Tasks	Predictive Tasks
Objective	Describe pattern & relationship in data.	Predict future outcomes or unknown values.
focus	Understanding past/present	forecasting & decision-making.
Tech used	Clustering, Association, Anomaly Detection.	Classification, Regression, time-series
Example.	Identifying customer segments.	Predicting whether customer will churn.

Issues in Data mining Tasks :-

Data Related issues:

Data Quality - inconsistent, incomplete can lead to inaccurate results

High Dimentionality - high dimentions increase complexity

Imbalance Data - if one class dominates, biased

Algorithm issues:

Scalability - hard to handle large dataset

Overfit & Underfit

Parameter Selection -

Privacy & Security.

Cost & Resources - costly

Applications of Data Mining :-

① Market & Retail -

- analyse customer behaviour, preference & pattern of purchase.
- Retailers use data mining for improving sales strategy.

② Banking & finance -

- detect fraud / unusual pattern
- risk management, safer decision-making.

③ Healthcare & Medicine.
- predict disease outbreaks, improve care.
- predict based on patients history

④ Education
- monitor student performance & factor affecting success
- personized learning, predict risk of dropout.

⑤ Manufacturing.
- optimizes production proces, quality control & equipment maintainance.
- predicts machine failure, reduce downtime & improve product quality by analysis.

⑥ Telecommunication.
- use to understand costumer pattern, improve customer satisfaction.
- fraud detect, target maketing for telecom service,
- predict customer churn.

⑦ Crime Prevention.
- analyse crime pattern & predict future activities.
- identifies high risk areas

⑧ E-commerse.
- like amazon, use to analyze data of customer purchase history & browsing behaviour.
- recommendations, such as He will byy this, if he buyed that.

⑨ finance
- use to access credit card risks by analysis of person's history & behaviour.
- decision making.

⑩ Social Media
- recommendations.

Benefits of Data Mining

① Enhance Decision making
② CRM improve
③ fraud Detect & manage
④ Cost Efficiency
⑤ Predichive analysis
⑥ Scalable & speed

⑦ Support Medical Research
⑧ Real-time insights

KDD (Knowledge Discovery in Database)

① Data Selection.
- select relevant data for analysis task.
- identify & collect useful data.
  gathering data from different source/system.

for eg. Retail business, collecting from customer purchase, browse history, etc.

- challenge is to select relevant data, as irrelvent can cause inconsistency / poor mining results.

② Data Preprocessing.
- data is cleaned, transformed & prepared for analysis.
- involves handling missing values, remove noise, deal with duplicates, resolve inconsistency.
- transforms data into suitable format by normalization.
- goal is to improve quality of data & make it ready for mining.
- challenge - handling large data is time consuming.

③ Data Transformation.
- transform data into appropriate form
- Data mapping - map the base to destination
  Code generation - create actual program
- challenges - transforming incorrectly can lead loss of imp features

④ Data Mining:
- core step in KDD
- apply techniques & algorithms to extract patterns & knowledge

Challenge - selection of right algorithm

⑤ Pattern Evaluation -

- identify most interesting & meaningful patterns.
- Evaluate patterns based on predefined measures (eg. accuracy, support, etc)
- Discard patterns that are irrelevant

Challenge - irrelevant are selected

⑥ Knowledge Presentation -

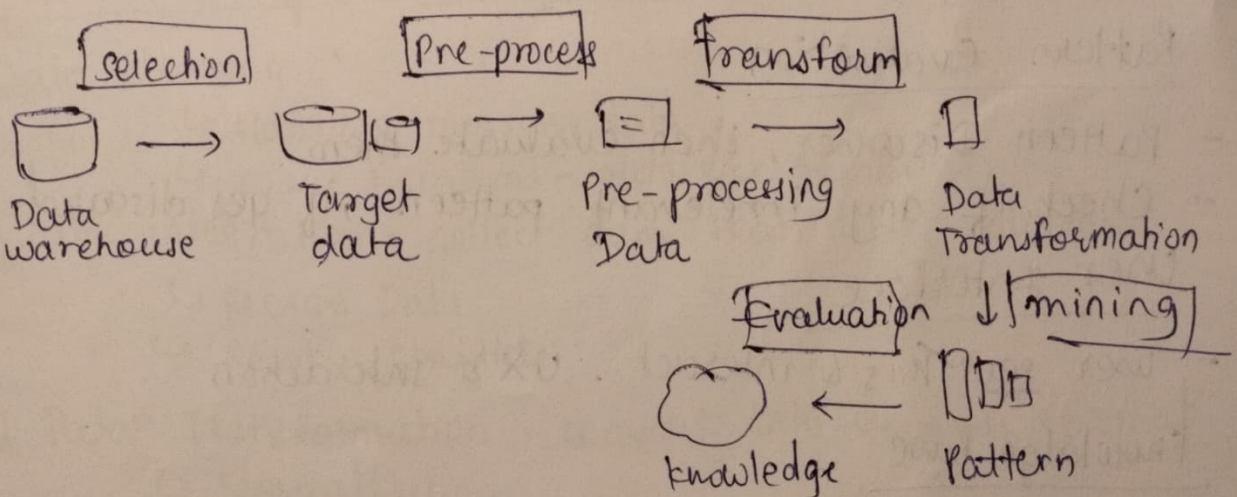
- present discovered knowledge in understandable format
- visualizing patterns / knowledge that are discovered by graphs, charts, reports or decision trees.

Challenge - Complex patterns.

⑦ Knowledge Deployment -

- deploy to Improve decision making
- use insights for better marketing strategies

Challenge - Deployment into real world may require integration with existing processes.



Data Mining Architecture

step 1: Data source, clean & Integration.

- gather data from diff. places like db, spreadsheets and online sources.
- data is often messy, with missing values.
- so clean & integration layer, to fix errors.
- and written in uniform format

Data Warehouse server

- After cleaning, data comes here. This is like large storage area.
- Data is kept is orgaized way, for easy acces & retrive.

Data mining Engine

- core of the system.
- various mining techniques are applied to find patterns, trends or relationships.
- clustering, classifn, Regression, etc

Pattern Evaluation

- Pattern Discover, then evaluate them.
- Check if any irrelevant pattern, if yes discard

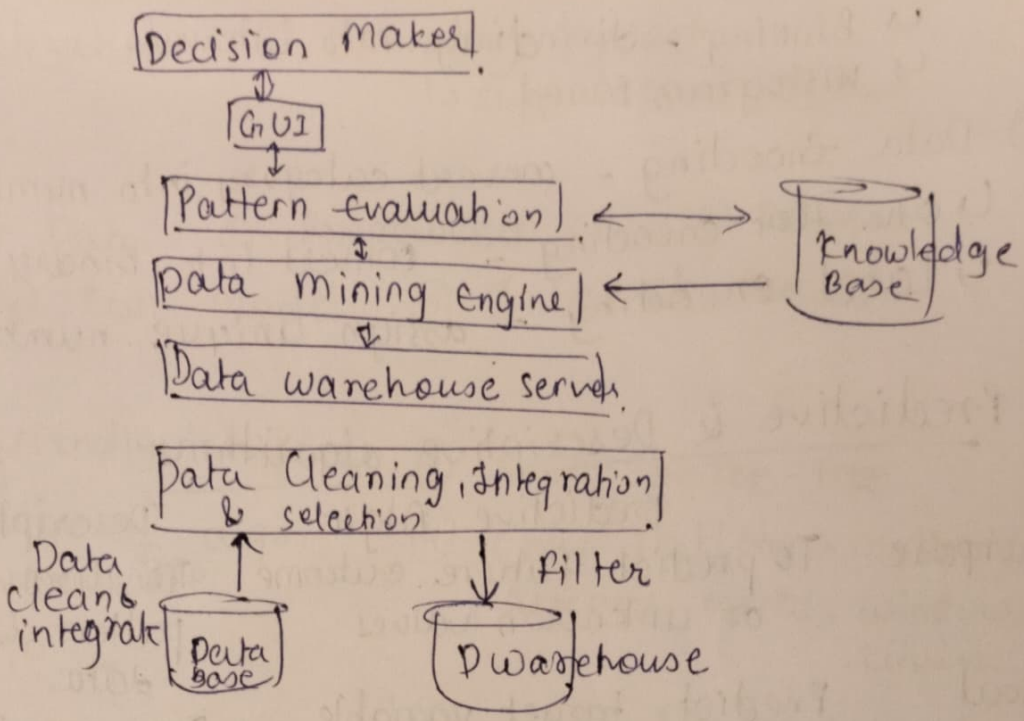
User Interface

- user sees this & interact. UX & interaction

Knowledge Base

- store info.
- refine searching & more accurate results.

overall, convert large data into useful insights.



Data Preprocessing

- process of preparing & cleaning raw data before it is used in data mining models.
- raw data consists of inconsistency, missing values, errors, etc.

Techniques of Data Preprocessing

1. Data cleaning
   - ↳ Handling missing data
   - ↳ Noise Removal - outliers removed
2. Data integration - collect data from diff. sources.
   - ↳ Merge Data
   - ↳ Resolve conflicts
3. Data Transformation - convert into uniform format
   - ↳ Normalization
   - ↳ Aggregation - summing up into one
4. Data Reduction - reduce volm
   - ↳ Dimention Reduction - using PCA
   - ↳ Data Sampling - use smaller subsets
5. Data Discretization - convert continuous data to discreate

↳ Binning - dividing into intervals
↳ Histogram based

⑤ Data encoding - convert category into number format
↳ One Hot encoding - convert into Binary values
↳ Label encoding - assign unique number.

Predictive & Descriptive algorithms

	Predictive Algo.	Descriptive Algo.
Purpose	To predict future outcome or unknown values	To discover pattern & understand data.
Goal	Predict target variable	Summarize data & identify relations
Type of learning	Supervised	Unsupervised
Output	Predictions	rules, pattern, clusters
Eg	Classifn, Regression	Clustering, Association Rule mining
Applⁿ	fraud detect, sales forecasting, disease prediction.	Customer segmentⁿ, market analysis

Explain any data mining tool:

WEKA (Waikato Environment for Knowledge Analysis)
- open source mining tool developed by University of Waikato in New Zealand.
- provides collection of ML algo & tools for processing, classifying, etc.

Features
① Data preprocessing
② Classification & regression
③ Clustering
④ Association mining rule

⑤ Visualization ⑥ scripting support.
↳ support Java / Python

Steps

① Import Data ② Preprocess Data ③ Choose Algo ④ Run model ⑤ Interpret Results.

Adv:-

① User friendly interface
② Wide range of algo. available
③ Open source - free
④ Cross Platform - multiple platforms MacOS, Windows, Linux

~~DIS~~ Applⁿ:-

① Customer segmentation
② fraud detect
③ Marked Basket Analysis
④ Healthcare

Disadv:-

① Scalability - .
② Algo complenity
③ Limited Real-Time support.