GROUP TASK 1

PRESENTED
BY
GROUP 9

MEMBER OF GROUP

AHMAD ZIYAAD	A23CS0206
GOE JIE YING	A23CS0224
TEH RU QIAN	A23CS0191

Q1: HOW IS DATA MINING PART OF THE NATURAL EVOLUTION OF DATABASE TECHNOLOGY?

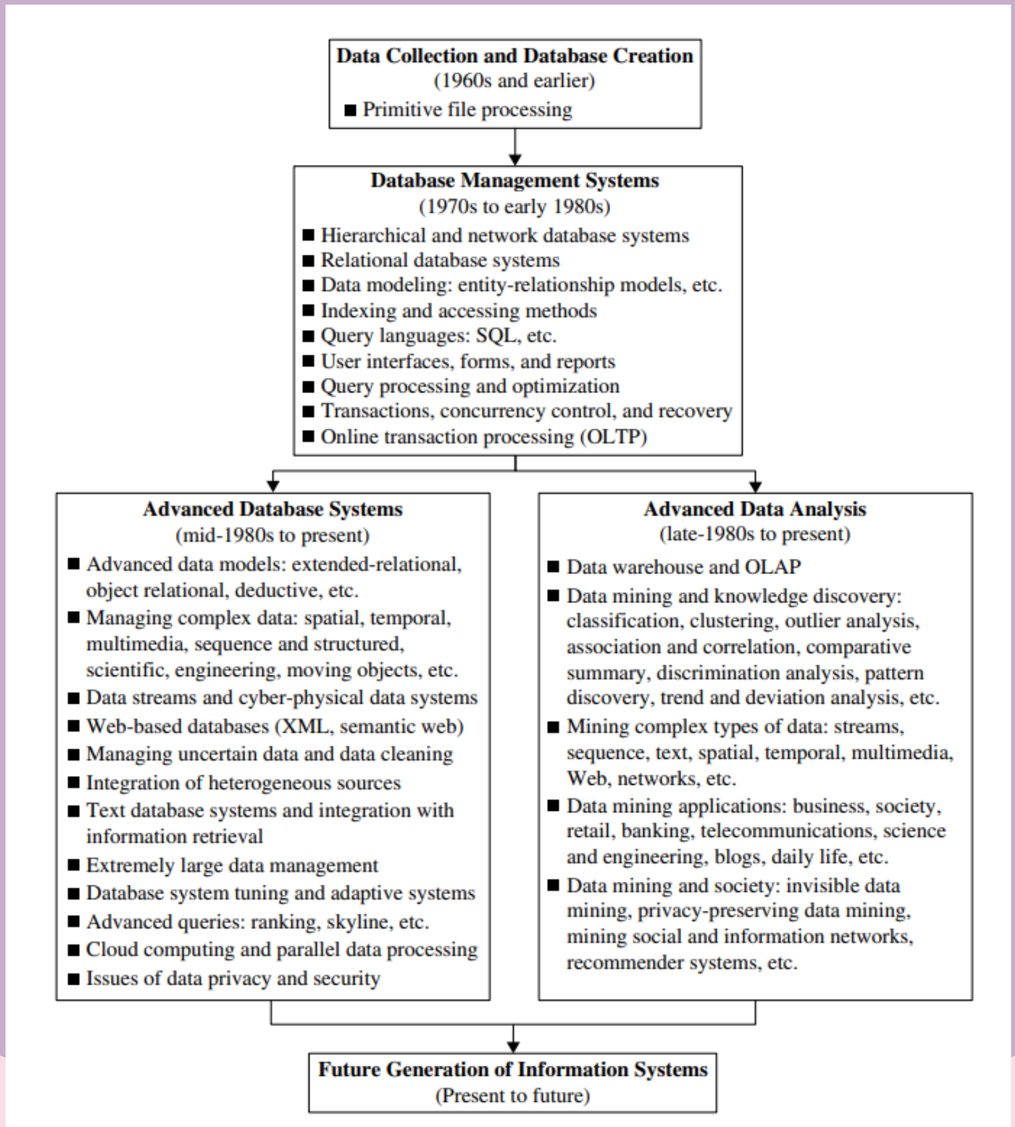


Figure 1: The evolution of database system technology

Q1: HOW IS DATA MINING PART OF THE NATURAL EVOLUTION OF DATABASE TECHNOLOGY?

Based on Figure 1,

- In data collection & database creation (Initial Phase), data is simply **stored and retrieved** from databases.
- After evolution, modern databases **introduced query processing and transaction management** to efficiently **extract useful information** from large datasets.
- As databases evolved to handle complex data, data mining became the natural next step for **discovering patterns, trends, and valuable insights** from vast amounts of data.

Q2: WHY IS DATA MINING IMPORTANT?

- The need for **turning huge amounts of data** into **useful information and knowledge** [3]
- **Analyze and extract pattern** [2] from enormous amounts of data, which contributing across various domains, including business strategies, knowledge management, scientific and medical research fields
- With data mining, decisions will be **based on real insights** instead of the decision maker's intuition. [1]
- **Enhance efficiency** and **drive innovation** [1]

Q3: WHAT IS DATA MINING?

- Definition 1: Data mining is **the process of discovering interesting patterns and knowledge** from large amounts of data. The data sources can **include databases, data warehouses, the Web**, other information repositories, or data that are streamed into the system dynamically. [1]
- Definition 2: Data mining is **the use of machine learning and statistical analysis** to **uncover patterns** and other valuable information from large data sets. [2]
- In summary, data mining **extracts valuable patterns and insights** from large datasets, aiding **decision-making and trend analysis**.

Q4: WHAT IS THE GENERAL ARCHITECTURE OF A DATA MINING SYSTEMS?

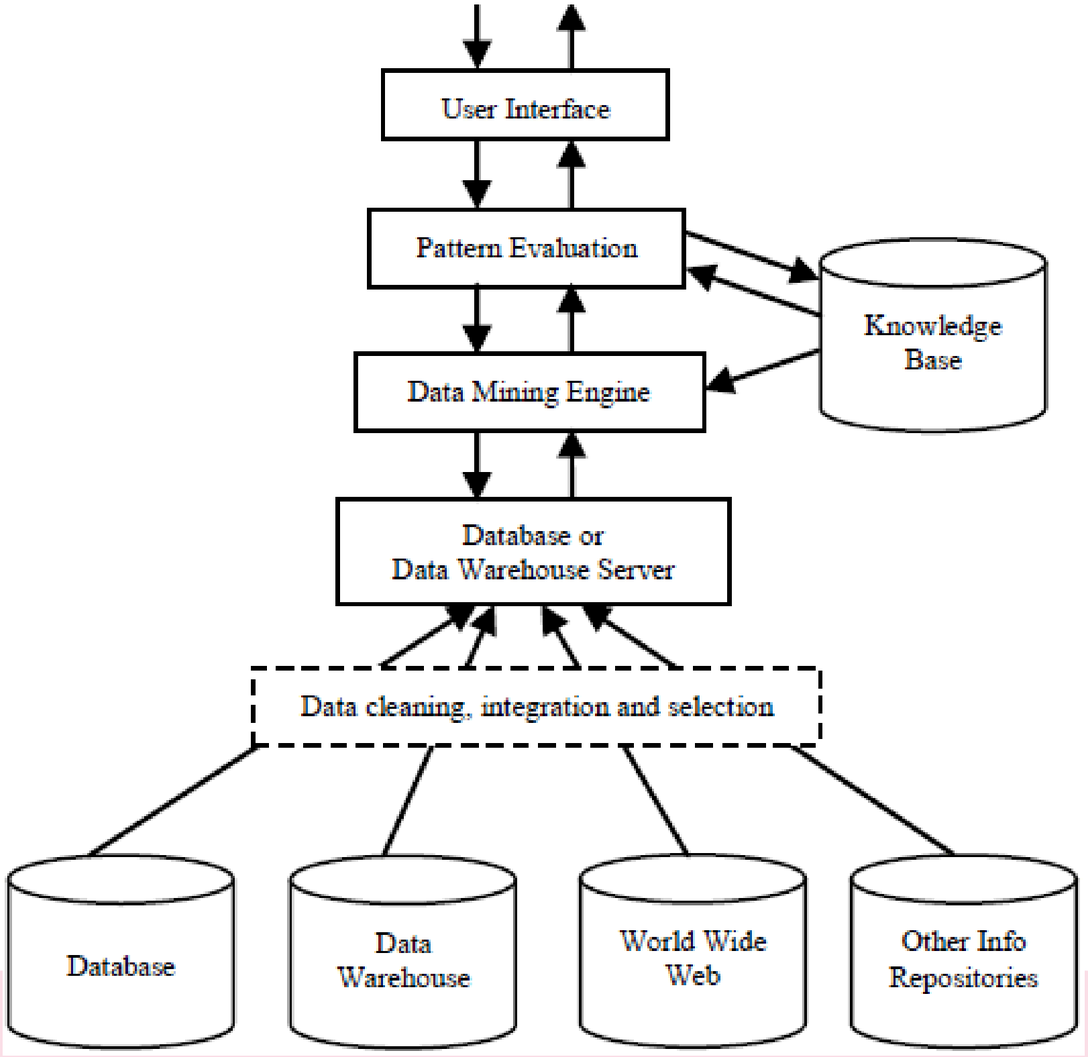


Figure 2: General Architecture of Data Mining Systems [4]

Q4: WHAT IS THE GENERAL ARCHITECTURE OF A DATA MINING SYSTEMS? [3]

DATABASE, DATA WAREHOUSE, OR OTHER INFORMATION REPOSITORY

- Store raw data in databases, data warehouses, spread sheets, or other kinds of information repositories.
- Perform data cleaning and data integration

DATABASE OR DATA WAREHOUSE SERVER

- Responsible for retrieving relevant data based on the user's data mining request.

KNOWLEDGE BASE

- Stores domain knowledge to guide data mining and evaluate pattern relevance.
- Includes metadata and predefined constraints to refine results.

Q4: WHAT IS THE GENERAL ARCHITECTURE OF A DATA MINING SYSTEMS? [3]

DATA MINING ENGINE

- characterization
- association analysis
- classification evolution
- deviation analysis

PATTERN EVALUATION

- Uses interestingness measures to filter and refine patterns.
- Works with the data mining engine to focus on meaningful insights.

GRAPHICAL USER INTERFACE (GUI)

- Enables user interaction, task specification, and result visualization.

Q5 WHAT TYPES OF DATA CAN BE MINED?

Structured Data	Unstructured Data
Easily analyzable	Requires advanced techniques
Processed with SQL, machine learning models	Often in text, image or video formats
Easier to mine using traditional data mining techniques	Difficult to analyze using traditional methods
Example: • Relational databases (tables, rows, columns) • Spreadsheets (Excel, CSV files) [1]	**Example**: • Text data (emails, articles) • Multimedia (images, videos, audio) [1]

Q6 WHAT ARE THE MAIN STEPS/PROCESSES INVOLVED IN DATA MINING?

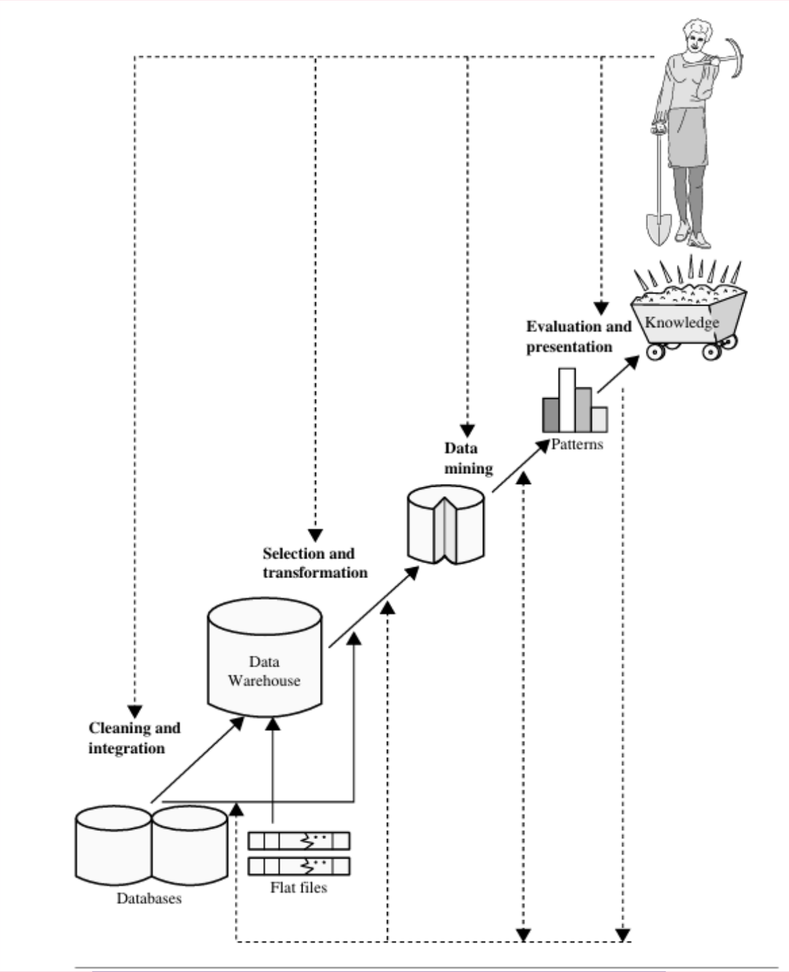


Figure 2: Data mining as a step in the process of knowledge discovery [1]

Q6 WHAT ARE THE MAIN STEPS/PROCESSES INVOLVED IN DATA MINING?

STEP 1

Data cleaning (to remove noise and inconsistent data)
Data integration (where multiple data sources may be combined) [1]

STEP 2

Data selection (where data relevant to the analysis task are retrieved from the database)
Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations) [1]

Q6 WHAT ARE THE MAIN STEPS/PROCESSES INVOLVED IN DATA MINING?

STEP 3

Data mining (an essential process where intelligent methods are applied to extract data patterns) [1]

STEP 4

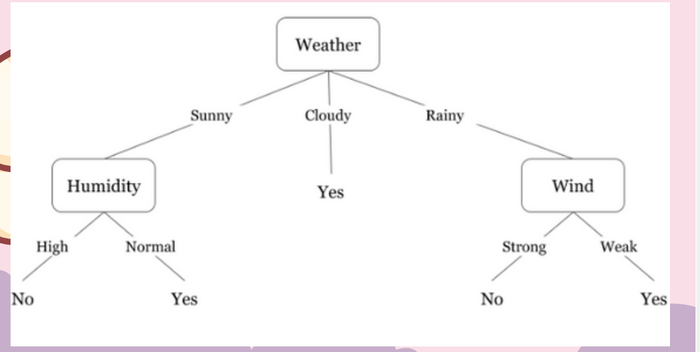
Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)

Knowledge presentation (where visualization and knowledge representation tech niques are used to present mined knowledge to users) [1]

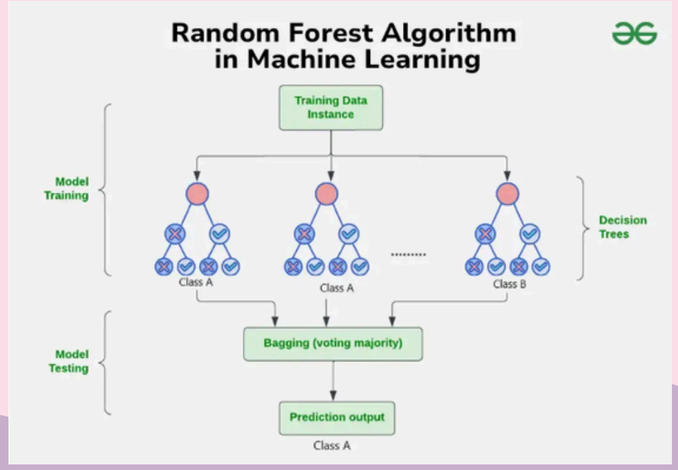
Q7 WHAT ARE THE COMMON DATA MINING TASKS/TECHNIQUES AND THEIR ASSOCIATED ALGORITHMS?

Classification:

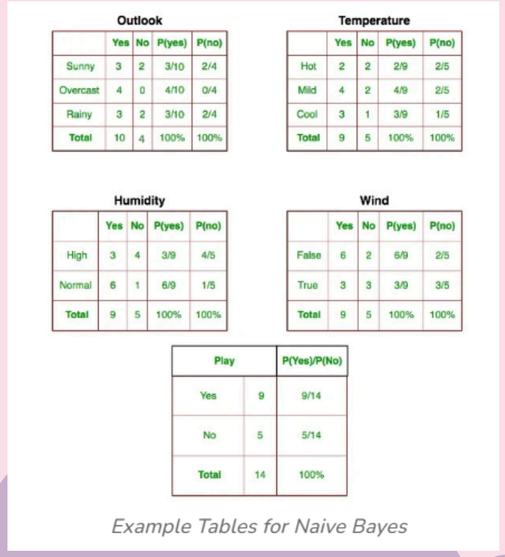
1. Assign data into predefined categories based on input features.
2. Applications: Spam detection, customer churn prediction, image recognition, credit scoring.
3. Algorithms:
    - Decision Trees [5]
    - Random Forest [6]
    - Naive Bayes [7]
    - Support Vector Machines (SVM) [8]



Decision Trees



Random Forest



Outlook

	Yes	No	P(yes)	P(no)
Sunny	3	2	3/10	2/4
Overcast	4	0	4/10	0/4
Rainy	3	2	3/10	2/4
Total	10	4	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

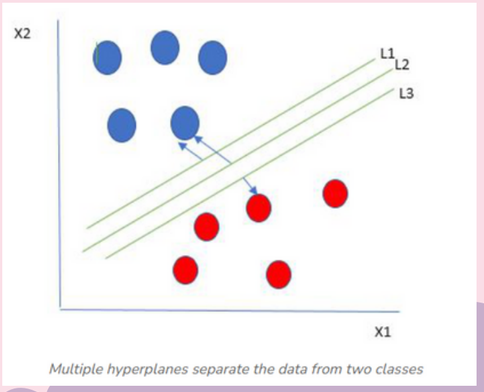
Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Example Tables for Naive Bayes

Naive Bayes



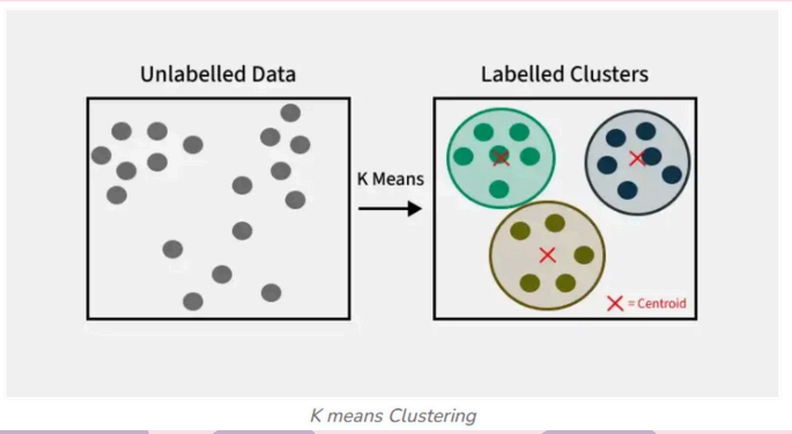
Multiple hyperplanes separate the data from two classes

SVM

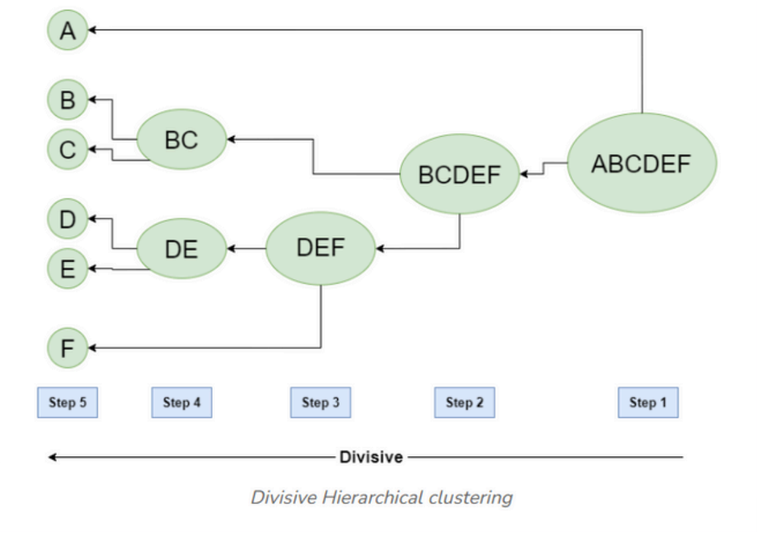
Q7 WHAT ARE THE COMMON DATA MINING TASKS/TECHNIQUES AND THEIR ASSOCIATED ALGORITHMS?

Clustering:

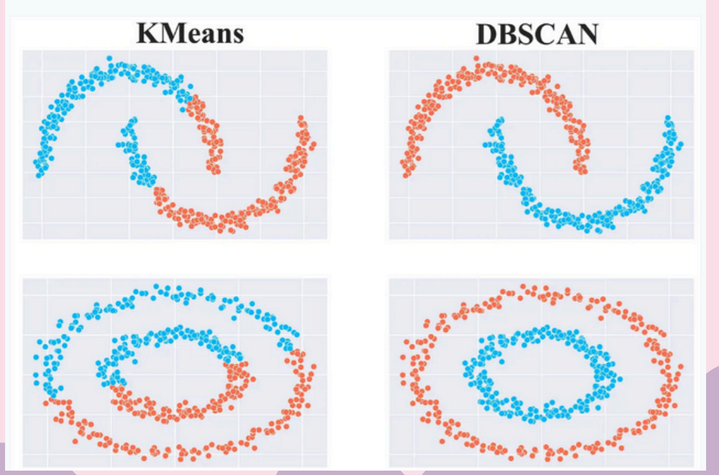
1. Group data points into clusters based on similarity without predefined labels.
2. Applications: Market segmentation, image compression, anomaly detection.
3. Algorithms:
    - K-Means Clustering [9]
    - Hierarchical Clustering [10]
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [11]



K means Clustering



Divisive Hierarchical clustering



Q7 WHAT ARE THE COMMON DATA MINING TASKS/TECHNIQUES AND THEIR ASSOCIATED ALGORITHMS?

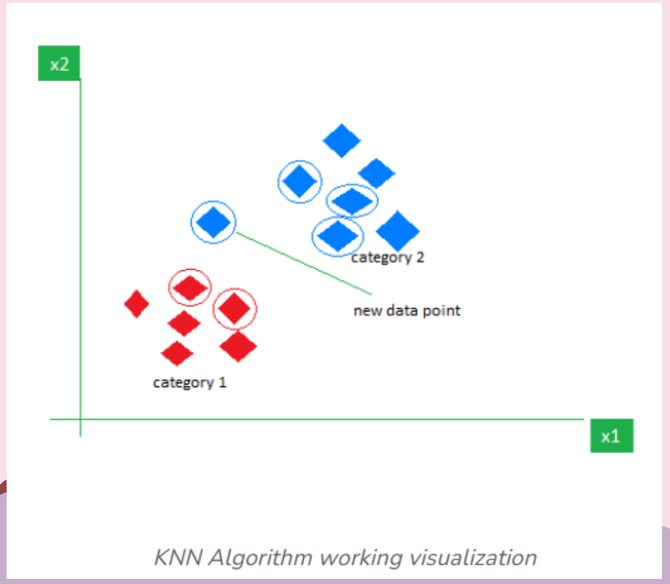
Association Rule Mining:

1. Discover relationships or associations between different variables in large datasets (often used in marketing analysis)
2. Applications: Retail product bundling, cross-selling, website link analysis.
3. Algorithms:
    - Apriori Algorithm - find frequent item sets in transactional databases
    - FP-growth Algorithm - builds a compact tree structure to discover frequent item sets efficiently
    - Eclat Algorithm - Uses a vertical data format and depth-first search to find frequent itemsets

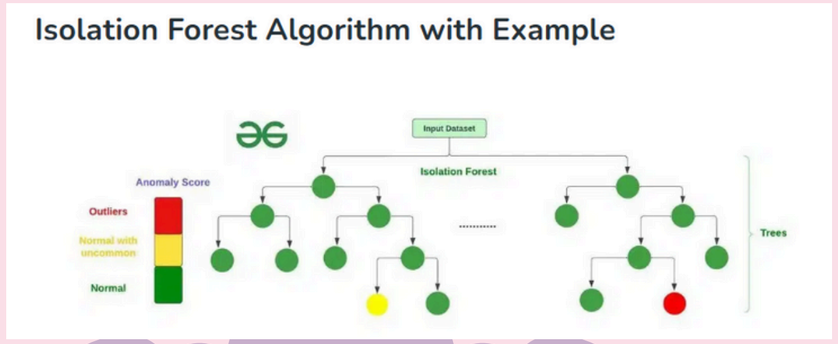
Q7 WHAT ARE THE COMMON DATA MINING TASKS/TECHNIQUES AND THEIR ASSOCIATED ALGORITHMS?

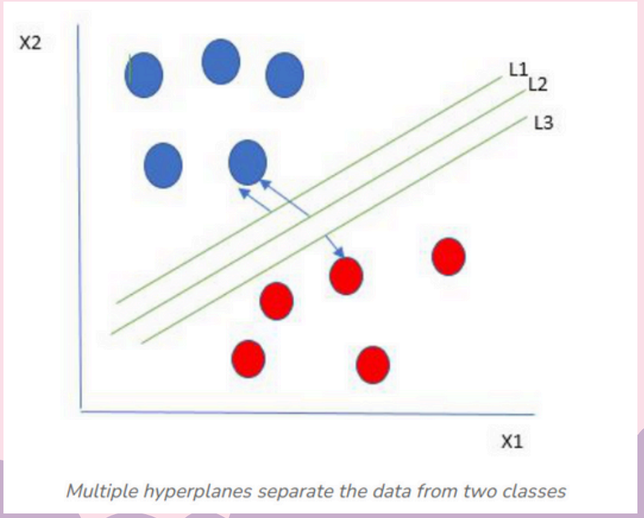
Outlier Analysis:

1. Detect data points that significantly different from the expected patterns
2. Applications: Fraud detection, network security, quality control.
3. Algorithms:
    - k-Nearest Neighbours (k-NN) [12]
    - Isolation Forest [13]
    - Support Vector Machines (SVM) [8]



KNN Algorithm working visualization



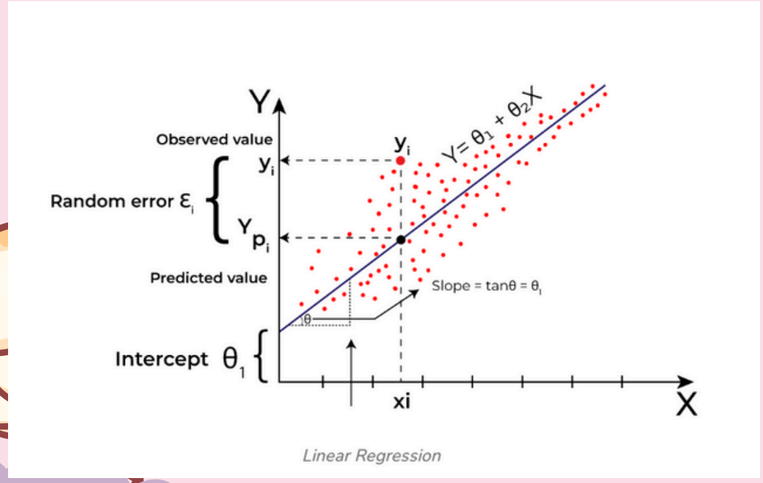


Multiple hyperplanes separate the data from two classes

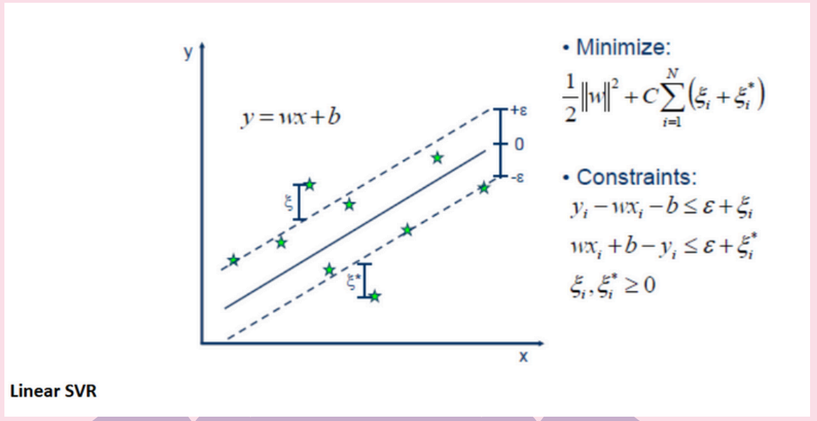
Q7 WHAT ARE THE COMMON DATA MINING TASKS/TECHNIQUES AND THEIR ASSOCIATED ALGORITHMS?

Regression:

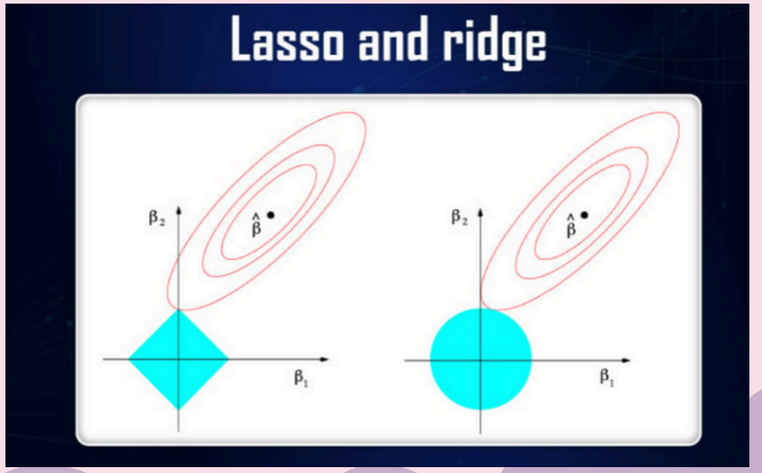
1. Predict continuous numerical values based on input data.
2. Applications: Stock price prediction, real estate price forecasting, sales prediction.
3. Algorithms:
    - Linear Regression [14]
    - Support Vector Regression(SVR) [15]
    - Ridge and Lasso Regression [16]



Linear Regression



Linear SVR



Q8 – WHAT ISSUES OR CHALLENGES ARE COMMONLY ENCOUNTERED IN DATA MINING?

- **User Interaction Issues**: Ensuring that data mining systems are user-friendly and provide actionable insights requires addressing issues related to human-computer interaction and result interpretation.
- **Efficiency and Scalability**: With the rapid growth of data, creating algorithms that can handle large-scale data efficiently is crucial.
- **Diverse Data Types**: Handling Various Data Formats: Extracting insights from different types of data, including multimedia, spatial, and temporal data, requires specialized approaches.
- **Ensuring Data Privacy and Security**: Safeguarding confidential information during data mining is crucial for maintaining user confidence and adhering to legal requirements.
- **System Integration Challenges**: Effectively embedding data mining insights into current systems and workflows can be a challenging task.

Q9 – WHAT ARE SOME EXAMPLES OF DATA MINING APPLICATIONS AND TOOLS?

Field/Area	Application / Tools
Business Management	WEKA: An open-source suite of machine learning software written in Java, offering a collection of algorithms for data mining tasks.
Financial	Fraud detection, risk assessment using clustering, association rule mining, anomaly detection. (SAS Enterprise Miner, Apache Mahout)
Education	Provides insights into past behaviors and trends, predicting student performance. (WEKA, Orange)
Manufacturing	Predictive maintenance to prevent equipment failure and optimize production. (IBM SPSS Modeler)
Medical / Health Care	Disease prediction, personalized medicine, identifying customer segments with similar purchasing habits. (IBM Watson Health)

Q10 – WHAT ARE THE DIFFERENCES BETWEEN DESCRIPTIVE AND PREDICTIVE DATA MINING?

Aspect	Descriptive Data Mining	Predictive Data Mining
Objective	Summarizes and interprets existing data to identify patterns and relationships.	Uses existing data to predict future outcomes or trends.
Techniques Used	Clustering, association rule mining, anomaly detection.	Classification, regression analysis, time-series analysis.
Outcome	Provides insights into past behaviors and trends.	Forecasts future events or behaviors based on historical data.
Example	Identifying customer segments with similar purchasing habits.	Predicting which customers are likely to respond to a marketing campaign.

REFERENCES

[1] jiawei han, micheline kamber, and jian pei, “Data Mining Third Edition,” 2011. Available: https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[2] J. Holdsworth, “Data mining,” Ibm.com, Jun. 28, 2024. https://www.ibm.com/think/topics/data-mining

[3] SECP2753 Data Mining Lecture Slide Module 1a_2024 Introduction to Data Mining_Part1

[4] Architecture of a typical data mining system | download scientific diagram, https://www.researchgate.net/figure/Architecture-of-a-Typical-Data-Mining-System-1_fig2_290212859 (accessed Mar. 26, 2025).

[5] GeeksforGeeks, “Decision Tree,” GeeksforGeeks, Oct. 16, 2017. https://www.geeksforgeeks.org/decision-tree/?ref=header_outind

[6] GeeksforGeeks, “Random forest algorithm in machine learning,” GeeksforGeeks, Jul. 12, 2024. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

REFERENCES

[7]GeeksforGeeks, “Naive Bayes Classifiers,” GeeksforGeeks, Mar. 03, 2017. https://www.geeksforgeeks.org/naive-bayes-classifiers/?ref=header_outind

[8]GeeksforGeeks, “Support Vector Machine (SVM) Algorithm,” GeeksforGeeks, Jan. 20, 2021. https://www.geeksforgeeks.org/support-vector-machine-algorithm/?ref=header_outind

[9]GeeksforGeeks, “K means Clustering Introduction,” GeeksforGeeks, May 02, 2017. https://www.geeksforgeeks.org/k-means-clustering-introduction/?ref=header_outind

[10]GeeksforGeeks, “Hierarchical Clustering in Data Mining,” GeeksforGeeks, Feb. 05, 2020. https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/?ref=header_outind (accessed Mar. 26, 2025).

[11]A. Chawla, “The Limitations of DBSCAN Clustering Which Many Often Overlook,” Dailydoseofds.com, Oct. 04, 2023. https://blog.dailydoseofds.com/p/the-limitations-of-dbscan-clustering

REFERENCES

[12] GeeksforGeeks, “KNearest Neighbor(KNN) Algorithm,” GeeksforGeeks, Apr. 14, 2017. https://www.geeksforgeeks.org/k-nearest-neighbours/?ref=header_outind

[13]GeeksforGeeks, “What is Isolation Forest?,” GeeksforGeeks, Apr. 02, 2024. https://www.geeksforgeeks.org/what-is-isolation-forest/?ref=header_outind (accessed Mar. 26, 2025).

[14]GeeksforGeeks, “Linear Regression in Machine learning,” GeeksforGeeks, Sep. 13, 2018. https://www.geeksforgeeks.org/ml-linear-regression/?ref=header_outind

[15] “Support Vector Regression,” www.saedsayad.com. https://www.saedsayad.com/support_vector_machine_reg.htm

[16] P. Bose, “Guide to Lasso and Ridge Regression Techniques with Use Cases,” Blogs & Updates on Data Science, Business Analytics, AI Machine Learning, Aug. 25, 2023. https://www.analytixlabs.co.in/blog/lasso-and-ridge-regression/


THANK
YOU!