Unit – 01: Introduction

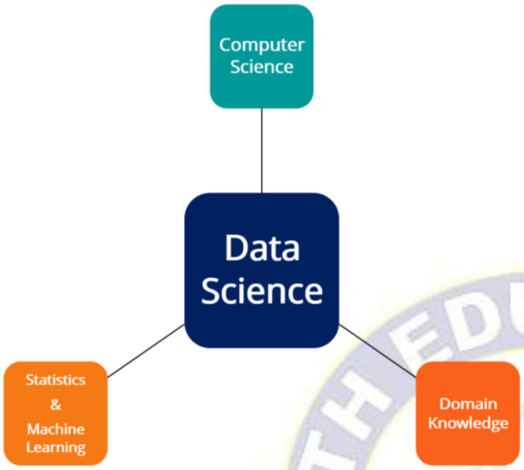
- Motivation, Importance, Definitions,
- Kind of Data, Data Mining Functionalities, Kinds of Patterns,
- Classification of Data Mining Systems,
- Data Mining Task Primitives,
- Integration of A Data Mining System, with A Database or Data Warehouse System,
- Major Issues in Data Mining,
- Types of Data Sets and, Attribute Values, Basic Statistical Descriptions of Data,
- Data Visualization, Measuring Data Similarity, PREPROCESSING: Data Quality,
- Major Tasks in Data Preprocessing, Data Reduction,
- Data Transformation, and Data Discretization, Data Cleaning and Data Integration.

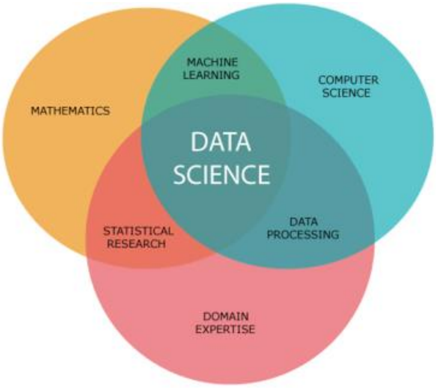
Questions to be discussed:

1. Define the term Data Warehouse. What are the advantage of Data Warehouse?
2. What is Data Mining? What are the advantage of Data Mining?
3. Differentiate between Data Warehouse and Data Mining.
4. Explain importance of data mining in various fields.
5. Define data mining and list its main functionalities.
6. Differentiate between structured and unstructured data. Provide example of each.
7. Explain the concept of measuring data similarity. Provide examples of similarity measures.
8. Describe the major tasks in data preprocessing. Why is data preprocessing essential in data mining?
9. Discuss the significance of data visualization in data analysis.

What is data science?

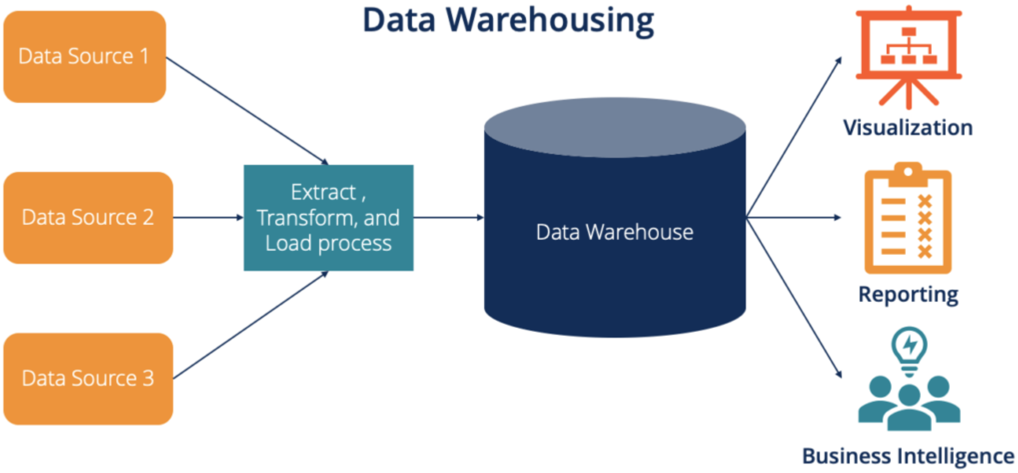
- Data Science is the study of data to find out the hidden pattern from the data.
- It is the technique to dealing with huge amounts of data to find marketing patterns.
- This pattern can be used in better decision or making good marketing strategies, in customer benefit.
- It is the study of all types of data like structured, unstructured, and semi-structured.
- This analysis helps data scientists to ask answer questions like what happened, why it happened, what will happen, and what can be done with the results.





What is Data Warehouse?

- A Data Warehouse refers to a place where data can be stored.
- In other words, we can say that it is the process of compiling & organizing data into a common database.
- It is the process of data collection and storage from various sources and managing it to provide valuable business insights.
- It is also known as electronic storage, where businesses store a large amount of data and information.
- It is built to store a huge amount of historical data and empowers fast requests over all the data, using Online Analytical Processing (OLAP).
- The idea of data warehousing was developed in the 1980.



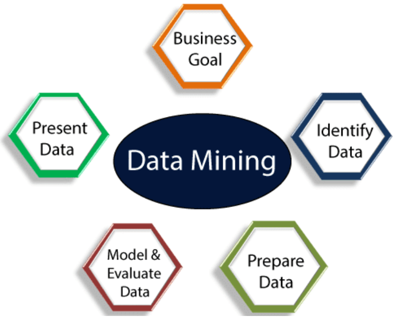
Steps in Data Warehousing:

- The following steps are involved in the process of data warehousing:

1. **Extraction of data** – A large amount of data is gathered from various sources.
2. **Cleaning of data** – Once the data is compiled, it goes through a cleaning process.
3. **Conversion of data** – After being cleaned, the format is changed from the database to a warehouse format.
4. **Storing in a warehouse** – Once converted to the warehouse format, the data stored in a warehouse.

What is Data Mining?

- Data mining refers to the analysis of data.
- We can say that processing of raw data to prepare it for some other data is known as Data Mining.
- It is also known as knowledge Discover in Database (KDD).
- Data mining is the process of searching and analyzing a large amount of raw data in order to identify patterns and extract useful information.
- The information or knowledge extracted so can be used for any of the following applications
    - ➢ Market Analysis
    - ➢ Fraud Detection
    - ➢ Customer Retention
    - ➢ Production Control
    - ➢ Science Exploration



Data Mining Techniques:

- Data mining uses algorithms and various other techniques to convert large collections of data into useful output.
- The most popular types of data mining techniques include association rules, classification, clustering, decision trees, K-Nearest Neighbor, neural networks, and predictive analysis.

Differentiate between Data Mining and Data Warehousing:

Data Mining	Data Warehousing
It is the process of determining data patterns.	It is a database system designed for analytics.
It is the process of extracting useful data from a large set of data.	It is the process of combining all the relevant data.
Data mining uses pattern recognition techniques to identify patterns.	Data warehousing is the process of extracting and storing data that allow easier reporting.
Data is analyzed regularly.	Data is stored periodically.
It is carried out by business users with the help of engineers.	It is solely carried out by engineers.

Data Mining Functionalities

- The main objective of data mining is to identify patterns, trends, or rules that explain data behavior.
- some most popular functionalities of **data mining**, such as –
  - Classification
  - Association Analysis
  - Cluster Analysis
  - Data Characterization
  - Data Discrimination
  - Prediction etc.

Why is data mining important?

- Data mining is a crucial part of any successful analytics initiative.
- Businesses can use the knowledge discovery process to increase customer trust, find new sources of revenue, and keep customers coming back.
- Below are some examples of how different industries use data mining.
  - Telecom, media, and technology
  - Banking and insurance
  - Education
  - Manufacturing
  - Retail
  - Financial services.
  - Insurance
  - Entertainment

What is Data?

- Data is a collection of information gathered by observations, measurements, research or analysis.
- They may consist of facts, numbers, names, figures or even description of things.
- Data is organized in the form of graphs, charts or tables.
- There exist data scientist who does data mining and with the help of that data analyze our world.

There are various Types of Data:

Structured Data:

- It is organized and formatted in a predefined manner and stored in relational databases.
- Structured data is represented in tables with rows and columns.
- Examples: sales records,customer information, financial data, and transaction logs.

Unstructured Data:

- Unstructured data does not have a predefined formate.
- It includes textual data, such as emails, social media posts,customer reviews, and documents like PDFs or Word files.
- Other forms of unstructured data include multimedia files (images, videos, audio recordings), sensor data, and web pages.

Semi-Structured Data:

- Semi-structured data lies between structured and unstructured data.
- It possesses some organizational structure but does notfit neatly into a traditional relational database.
- Examples of semi- structured data include XML files, JSON data, and log files.

Time-Series Data:

- Time-series data is collected and recorded over regulartime intervals.
- It represents data points or measurements taken at consecutive time points.
- Time-series data is commonly found in fields like finance (stock prices), weather (temperature recordings), energy consumption, and IoT (Internet of Things) sensor data.

Spatial Data:

- Spatial data refers to data that has a geographic or spatial component.
- It includes coordinates, maps, satellite images, and GIS (Geographic Information System) data.
- Spatial data mining techniques areused to extract patterns and relationships related to geographical locations.

Text Data:

- Textual data encompasses any form of written or typed text,including emails, documents, articles, social media posts, and online reviews.
- Text mining techniques are employed to extract valuable information from text data, such as sentiment analysis, topic modeling, and text classification.

Graph Data:

- Graph data represents entities (nodes) and their relationships(edges) in a network.
- Examples: social networks, citation networks, web graphs, and knowledge graphs.

Differentiate between Structured data and Unstructured data:

Structured data	Unstructured data
It is organized and predefined format.	It does not have a predefined format.
It is based on Relational database table	It is based on Non-Relational database table
It is represented in tables with rows and columns.	It is represented in multimedia files like images, videos, audio recordings.
It is easy to search.	Searching for unstructured data is more difficult.
Easier to organize, clean, search, and analyze.	Difficult to organize, clean, search, and analyze.
An Excel table.	A collection of video files.

Measuring Data Similarity:

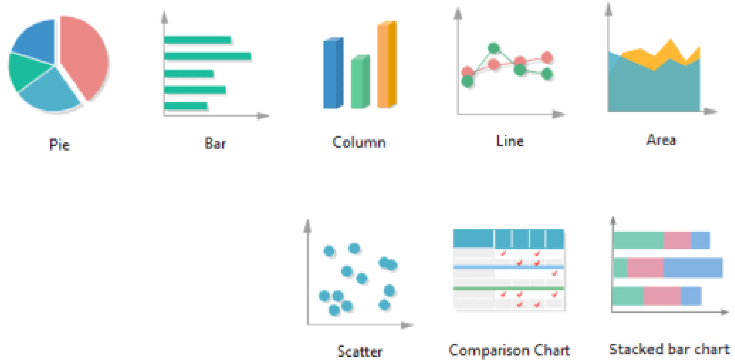
- A similarity measure is a mathematical function that quantifies the degree of similarity between two objects or data points.
- It is a numerical score measuring how alike two data points are.
- It takes two data points as input and produces a similarity score as output, typically ranging from 0 to 1.
- A similarity measure can be based on various mathematical techniques such as Cosine similarity, Jacquard similarity, and Pearson correlation coefficient.
- Similarity measures are generally used to identify duplicate records, equivalent instances, or identifying clusters.

Dissimilarity Measure:

- A dissimilarity measure is a mathematical function that quantifies the degree of dissimilarity between two objects or data points.
- It is a numerical score measuring how different two data points are.
- It takes two data points as input and produces a dissimilarity score as output, ranging from 0 (identical or perfectly similar) to 1 (completely dissimilar).
- A few dissimilarity measures also have infinity as their upper limit.
- A dissimilarity measure can be obtained by using different techniques such as Euclidean distance, Manhattan distance, and Hamming distance.
- Dissimilarity measures are often used in identifying outliers, anomalies, or clusters.

What is data visualization?

- It is the representation of data through use of common graphics, such as charts, plots, graphics etc.
- These visual displays of information is a way to understand easily complex data insights.
- It is representation of data to better understand patterns, trends, and relationships within the data set.
- Common data visualization technique include:
  - Tables
  - Pie charts
  - Line charts
  - Bar charts
  - Scatter plots
  - Heat maps:
  - Tree maps



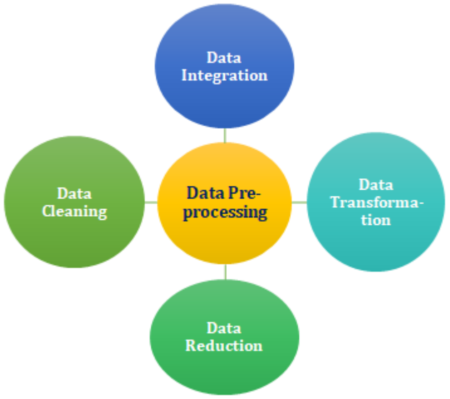
- **Tables:** This consists of rows and columns used to compare variables.
- **Pie charts:** Used to show the proportion of each category in a dataset.
- **Line charts**: Suitable for displaying trends in time series data.
- **Bar charts:** Used to compare categorical data by displaying bar of different heights.
- **Scatter plots:** Used to visualize relationships between two numerical variables.
- **Heat maps:** These graphical representation displays are helpful in visualizing behavioral data by location. This can be a location on a map, or even a webpage.
- **Tree maps,** which display hierarchical data as a set of nested shapes, typically rectangles. Tree maps are great for comparing the proportions between categories via their area size.

Data preprocessing:

- Data preprocessing is an important step in the data mining process.
- It is a data mining technique which is used to transform the raw data in a useful and efficient format.
- It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis.
- The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

- Some common steps in data preprocessing include:
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Discretization



- Data Normalization

Data Cleaning:

- This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates.
- Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

Data Integration:

- This involves combining data from multiple sources to create a unified dataset.
- Techniques such as record linkage and data fusion can be used for data integration.

Data Transformation:

- This involves converting the data into a suitable format for analysis.
- Common techniques used in data transformation include normalization, standardization, and discretization.
- Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance.
- Discretization is used to convert continuous data into discrete categories.

Data Reduction:

- This involves reducing the size of the dataset while preserving the important information.
- Data reduction can be achieved through techniques such as feature selection and feature extraction.

Data Discretization:

- This involves dividing continuous data into discrete categories or intervals.
- Discretization is often used in data mining and machine learning algorithms that require categorical data.
- Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

Data Normalization:

- This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1.
- Normalization is often used to handle data with different units and scales.
- Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Unit – 02: Data Warehousing and on-line Analytical Processing:

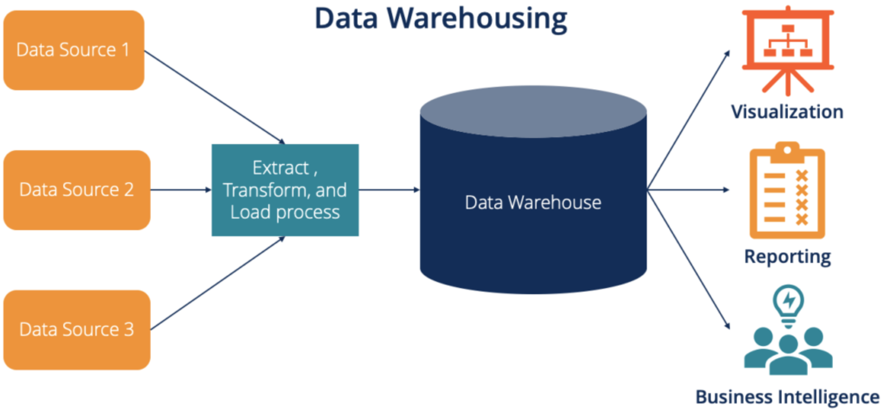
- Data Warehouse basic concepts,
- Data Warehouse Modeling - Data Cube and OLAP,
- Data Warehouse, Design and Usage,
- Data Warehouse Implementation,
- Data Generalization by Attribute-Oriented Induction,
- Data Cube Computation.

Questions to be discussed:

1. Describe the concept of data cube and online analytical processing.
2. Discuss the integration of data mining system with a database or database warehouse system.
3. Identify and elaborate on the major issue in data mining.

What is Data Warehouse?

- A Data Warehouse refers to a place where data can be stored.
- In other words, we can say that it is the process of compiling & organizing data into a common database.
- It is the process of data collection and storage from various sources and managing it to provide valuable business insights.
- It is also known as electronic storage, where businesses store a large amount of data and information.
- It is built to store a huge amount of historical data and empowers fast requests over all the data, using Online Analytical Processing (OLAP).
- The idea of data warehousing was developed in the 1980.



Types of Data Warehouse:

There are three main types of data warehouse.

1. Enterprise Data Warehouse (EDW)
2. Operational Data Store (ODS)
3. Data Mart

Enterprise Data Warehouse (EDW):

- This type of warehouse serves as central database that facilitates decision-support services.
- The advantage to this type of warehouse is that it provides access to cross-organizational information.

Operational Data Store (ODS):

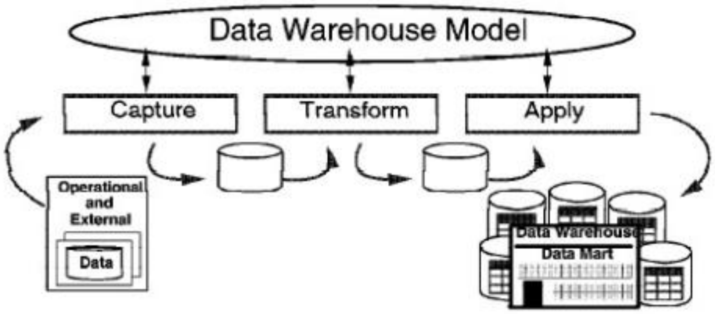
- This type of data warehouse refreshes in real-time.
- It is often preferred for routine activities like storing employee records.

Data Mart:

- It is a subset of a data warehouse built to maintain a particular department, region, or business unit.
- Every department of a business has a central repository or data mart to store data.
- The data from the data mart is stored in the ODS periodically.
- The ODS then sends the data to the EDW, where it is stored and used.

Data Warehouse Modeling:

- It is the process of designing the schemas in details and summarized information of the data warehouse.
- The goal of data warehouse modeling is to develop a schema describing the reality.
- In contrast, data modeling in operational database systems targets efficiently supporting simple transactions in the database such as retrieving, inserting, deleting, and changing data.

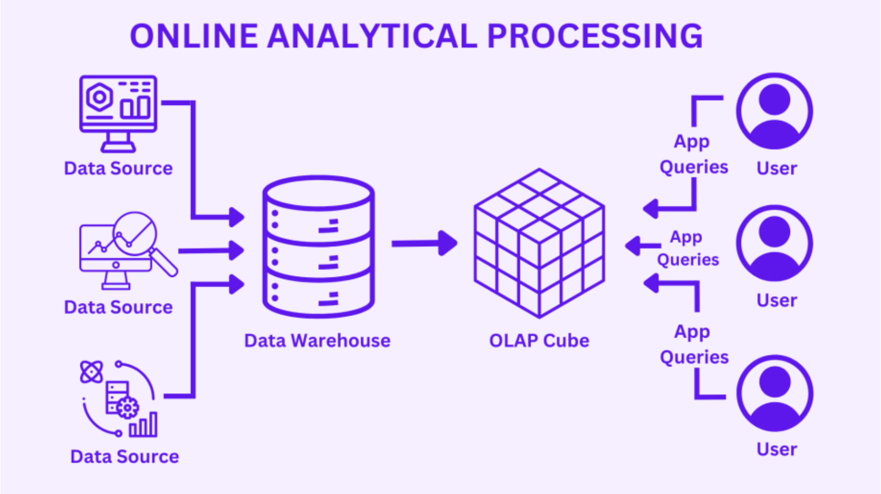


What is OLAP?

- OLAP stands for Online Analytical Processing.
- It is software technology which is use to analyze business data from different points of view
- It provides interactive access to large amounts of data and supports complex calculations.
- OLAP is used to support business intelligence and decision-making processes.

OLAP cubes:

- Grouping of data in a multidimensional matrix is called data cubes.
- In Data ware housing, the data will be represented by multiple dimensions and multiple attributes.
- This multidimensional data is represented in the data cube.
- The Data cube pictorially shows how different attributes of data are arranged in the data model.
- Below is the diagram of a general data cube.



Difference between OLAP and OLTP:

OLAP (Online Analytical Processing)	OLTP (Online Transaction Processing)
It is well-known as an online database query management system.	It is well-known as an online database modifying system.
Consists of historical data from various Databases.	Consists of only operational current data.
It makes use of a data warehouse.	It makes use of a standard DBMS.
In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
Relatively slow as the amount of data involved is large. Queries may take hours.	Very Fast as the queries operate on 5% of the data.
Only read and rarely write operations.	Both read and write operations.
The process is focused on the customer.	The process is focused on the market.

Discuss the integration of data mining system with a database:

- Data integration is the process of combining data from multiple sources into a single source.
- This can involve cleaning and transforming the data, as well as resolving any inconsistencies.
- The goal of data integration is to make the data more useful and meaningful for the purposes of analysis and decision making.
- Techniques used in data integration include data warehousing, ETL (extract, transform, load) processes, and data federation.
- The data integration approaches are formally defined as triple <G, S, M> where,

  G stand for the global schema,
  S stands for the heterogeneous source of schema,
  M stands for mapping between the queries of source and global schema.

What is Data Mining?

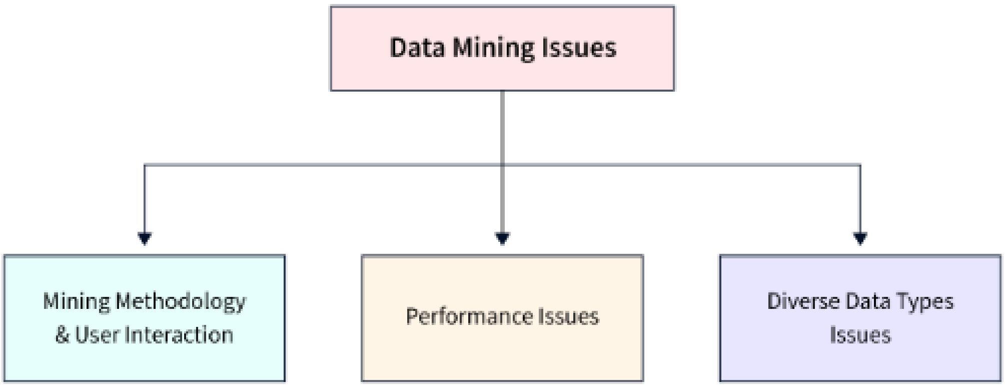
- Data mining is the process of extracting knowledge and patterns from large volumes of data.
- It involves using techniques from statistics, machine learning, and database management to uncover hidden insights, relationships, and trends within datasets.
- This invaluable discipline aids decision-making, prediction, and knowledge discovery across various fields and industries.

Purposes of Data Mining:

- Data mining serves several critical purposes, including identifying patterns and trends, predicting future outcomes, and improving decision-making.
- It enables businesses to gain a competitive edge by optimizing marketing strategies, understanding customer behaviour, and detecting anomalies or fraud.
- In healthcare, data mining aids in disease prediction and treatment optimization.
- It's also instrumental in scientific research, helping to extract meaningful insights from complex datasets.
- Essentially, data mining transforms raw data into actionable knowledge, driving innovation and informed choices across diverse domains.

Data Mining Issues:

There are three key data mining issues, as mentioned below -


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues

Mining Methodology Issues:

- Methodology-related data mining issues encompass challenges related to the choice and application of mining algorithms and techniques.
- Selecting the right method for a specific dataset and problem can be daunting.

Performance Issues:

- Performance-related issues revolve around scalability, efficiency, and handling large datasets.
- As data volumes continue to grow exponentially, it becomes essential to develop algorithms and infrastructure capable of processing and analyzing data promptly.

Diverse Data Types Issue:

- The diverse data types issues highlight the complexity of dealing with heterogeneous data sources.
- It involves integrating data from various formats, such as text, images, and structured databases.

Data Warehouse Implementation:

- Data Warehouse Implementation is a series of activities that are essential to create a fully functioning Data Warehouse, after classifying, analyzing and designing the Data Warehouse with respect to the requirements provided by the client.
- The various phases of Data Warehouse Implementation are ‘Planning’, ‘Data Gathering’, ‘Data Analysis’ and ‘Business Actions’.
- Every Data Warehouse needs a few important components, that needs to be defined while designing the implementation of the system, such as Data Marts, OLTP/ OLAP, ETL, Metadata, etc.

Data Cube Computation:

- Data cubes can be computed on top of a data warehouse, which allows for fast querying of the data.
- Data warehouses can be expensive to set up and maintain, and may not be suitable for all organizations.
- In this approach, the data cube is computed using a distributed computing system, such as Hadoop or Spark.
- The advantage of this approach is that it allows for the data cube to be computed on a large dataset, which may not fit on a single machine.

Unit – 03: Patterns, Associations and Correlations

- Mining Frequent Patterns, Associations and Correlations: Basic Concepts,
- Efficient and Scalable Frequent, Item set Mining Methods,
- Pattern Evaluation Methods, Applications of frequent pattern and associations,
- Frequent Patterns and Association Mining: A Road Map, Mining Various Kinds of Association Rules, Constraint-Based Frequent Pattern Mining, Extended Applications of Frequent Patterns.

Questions to be discussed:

1. Describe the term Frequent Patterns, Associations and Correlations in brief.
2. What is association rule? What are the application of association rule?
3. Describe pattern evaluation methods and their importance in association mining.
4. Explain the different kinds of pattern that can be mined from data.
5. Discuss Frequent Patterns and Association Mining with its application.

Describe the term Frequent Patterns, Associations and Correlations in brief.

Frequent Patterns:

- Frequent patterns are the sets of items that frequently co-occur together in a dataset.
- These patterns can be used to identify association & relationship between different items/attributes.

Association Rules:

- Association rules are logical statements that describe relationships between items in a dataset.
- They consist of an antecedent (a set of items) and a consequent (another set of items), with a measure of support and confidence indicating the strength of the rule.

Correlations:

- Correlations measure the statistical relationship between two or more variables in a dataset.
- They help to identify patterns of co-occurrence or dependency between variables.

Frequent Item set Mining Methods:

- It is a market basket analysis methodology that helps to find patterns.
- In this methods we can find shopping behaviors of users across different shopping platforms.
- These relationships are represented in the form of association rules.
- Frequent element pattern mining is used due to its wide applications in pattern mining & correlations.

Association Rules:

- Association Rules search for frequent patterns, associations, correlations, or causal structures between sets of items or objects in transaction databases, relational databases, and other available information repositories.

Applications

- Analysis of banking data
- Cross-marketing
- Catalog design

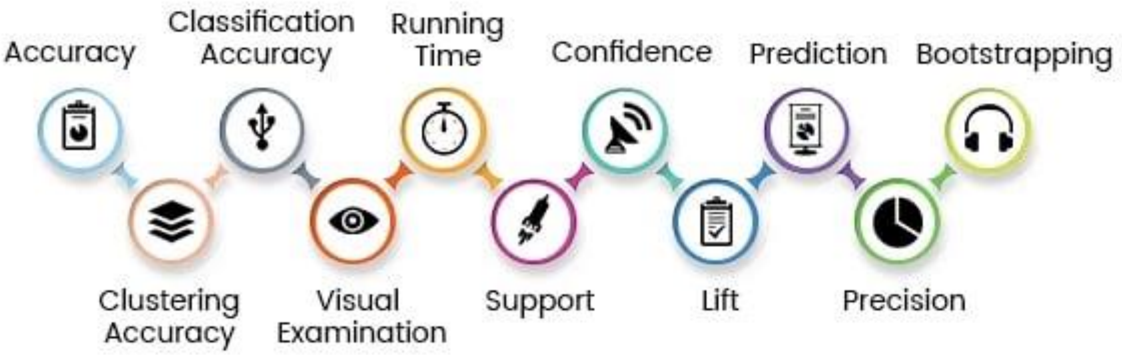
Association rules help to predict the occurrence of one item based on the occurrences of other items in a set of transactions.

Examples

- People who buy bread will also buy milk
- People who buy milk will also buy eggs
- People who bought soda will also buy potato chips
- People who buy bread will also buy jam

Pattern Evaluation Methods:

- In the field of data mining, the objective is to draw useful information and from vast amounts of data.
- Finding patterns, trends, and correlations in data allows for the discovery of hidden information.
- That hidden information can help with decision–making and problem–solving.
- An essential step in this process is pattern evaluation.
- The process of finding the useful patterns from large amount of data is known as pattern evaluation.
- There are various pattern evaluation methods/algorithms are available which are given below:

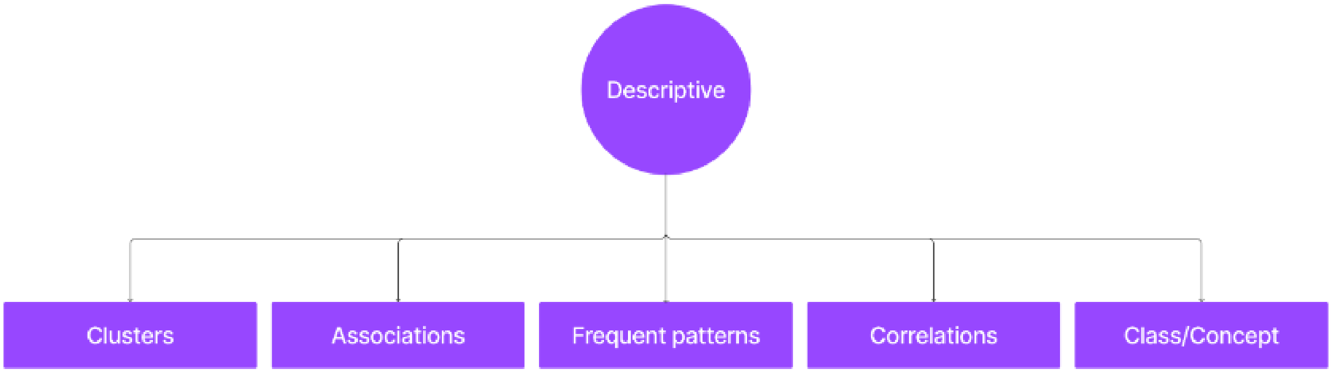


Explain the different kinds of pattern that can be mined from data.

- Different types of data can be mined in data mining.
- However, the data should have a pattern to get helpful information.
- Based on the data functionalities, patterns can be further classified into two categories.
  1. Descriptive
  2. Predictive

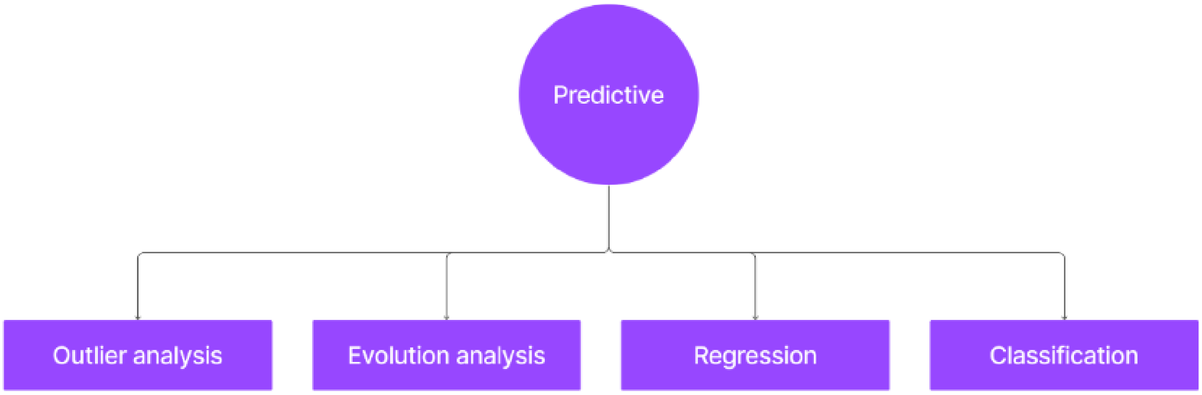
Descriptive patterns:

- It deals with the general characteristics and converts them into relevant and helpful information.
- Descriptive patterns can be divided into the following patterns:



Predictive patterns:

- It predicts future values by analyzing the data patterns and their outcomes based on the previous data.
- It also helps us find missing values in the data.
- Predictive patterns can be categorized into the following patterns.



Frequent Patterns and Association Mining:

- It is the process of identifying patterns or associations within a dataset that occur frequently.
- This is done by analysing large datasets to find items that appear together frequently.
- Frequent patterns are patterns that appear frequently in a data set.
- For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent item set.
- A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.

Applications of Frequent Pattern Mining:

Here are some applications of frequent pattern mining in bullet points -

- Market basket analysis
- Recommendation systems
- Fraud detection
- Social network analysis
- Healthcare
- Quality control etc.

Constraint-Based Frequent Pattern Mining:

- Constraint-based frequent pattern mining involves incorporating user-defined constraints or interestingness measures during the mining process.
- This allows formore focused and targeted mining based on specific requirements.
- Constraints can be used to guide the mining process by specifying certain item sets or patterns of interest, minimum thresholds, or domain-specific rules.

Extended Applications of Frequent Patterns:

- Frequent pattern mining has found applications beyond traditional associationrule mining.
- Some of the extended applications include:
  1. Text Mining
  2. Spatial Data Mining
  3. Social Network Analysis

Text Mining:

- Frequent pattern mining techniques can be applied to textdata for tasks such as document clustering, text classification, and information retrieval.
- Frequent term sets or n-grams can be extracted to identify common patterns in text documents.

Spatial Data Mining:

- Frequent pattern mining can be extended to spatial data sets for analyzing spatial relationships and patterns.
- It helps in identifying spatial associations, hotspots, and spatial dependencies in geographic datasets.

Social Network Analysis:

- Frequent pattern mining is useful for analyzingsocial networks and identifying patterns of connections or interactions between individuals or entities.
- It helps in understanding community structures, detecting influential nodes, and predicting network behaviors.

Unit – 01: Classification

- Basic Concepts,
  - Decision Tree Induction,
  - Bayesian Classification Methods,
  - Rule Based Classification,
  - Model Evaluation and Selection,
- Techniques to Improve Classification Accuracy:
  - Ensemble Methods,
  - Handling Different Kinds of Cases in Classification,
  - Classification by Neural Networks, Support Vector.
  - Machines, Pattern-Based Classification, Lazy Learners (or Learning from Your Neighbors).

Questions to be discussed:

1. Describe the classification of data mining system based on functionalities.
2. Define classification and its significance in data mining.
3. How do neural networks contribute to classification? Provide an example.
4. Discuss Bayesian classification methods and their application in data analysis.
5. Discuss the concept of lazy learners in classification. How do they differ from eager learners?

What is classification in data mining?

- It is a method used by data scientists to classify data into a given number of classes.
- This system can be used on structured or unstructured data.
- Its main purpose is to determine which category or class a new data set belongs to.
- In data science, four classification algorithms are commonly utilized.

Uses of Classification in Data Science:

- Classification algorithms can be used in different places.
  - Spam Mail Detection
  - Speech Recognition
  - Identifies Cancer tumor cells.
  - Classification of Drugs
  - Biometric Identification, etc.

Types of classification in Data Science:

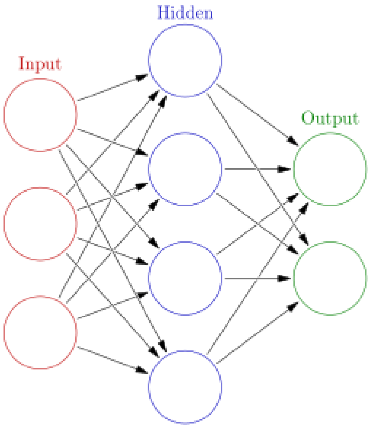
1. Neural Network
2. K- Nearest Neighbours
3. Decision Tree
4. Random Forest

Neural Network:

- It is a method that teaches computers to process data in a way that is inspired by the human brain.
- It is a type of machine learning that uses interconnected nodes or neurons in a layered structure that resembles the human brain.
- It is a model inspired by the structure and function of biological neural networks in animal brains.
- Neural networks are used in data science to help cluster and classify complex relationships.
- Neural networks are also called artificial neural networks (ANN) or simulated neural networks (SNN).
- They are a subset of machine learning, and at the heart of deep learning models.

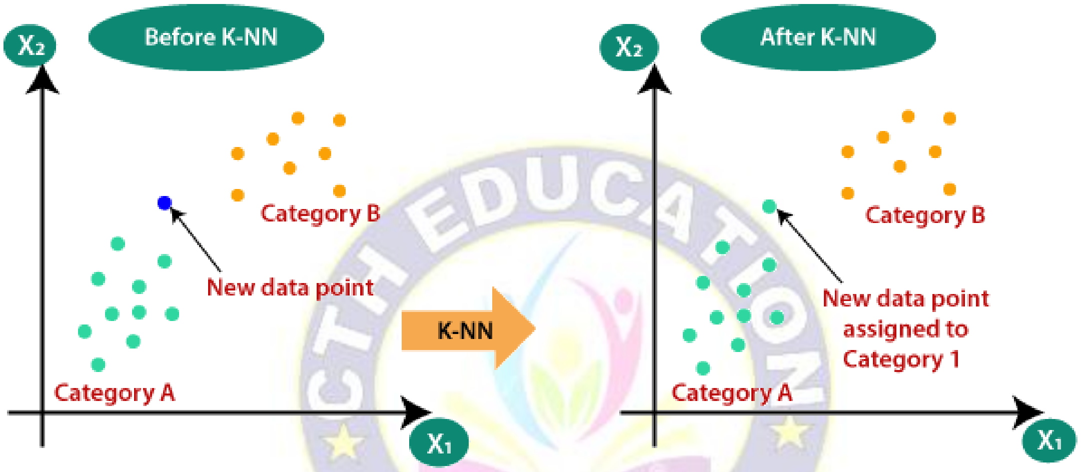
Application of Neural Networks:

- Neural networks are used, with applications for financial operations, enterprise planning, trading, business analytics, and product maintenance.
- Neural networks have also used in business applications such as forecasting and marketing research solutions, fraud detection, and risk assessment.



K- Nearest Neighbours:

- KNN stands for K-Nearest Neighbours.
- It is one of several algorithms used in data mining and machine learning.
- KNN is a classifier technique based on the similarity of data (a vector) to others.
- It is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

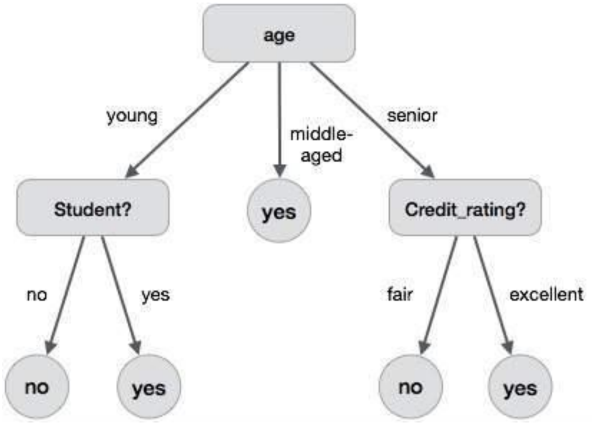
Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Decision Tree:

- In supervised learning methods, the decision tree algorithm is used.
- It is a tree that helps us in decision-making purposes.
- It is a structure that includes a root node, branches, and leaf nodes.
- Each internal node denotes a test and each branch denotes the outcome of a test, and each leaf node holds a class label.
- The topmost node in the tree is the root node.

- This algorithm could be used to deconstruct regression and classification problems.
- A decision tree is a tree structure that is used to develop classification or regression models.
- J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser).

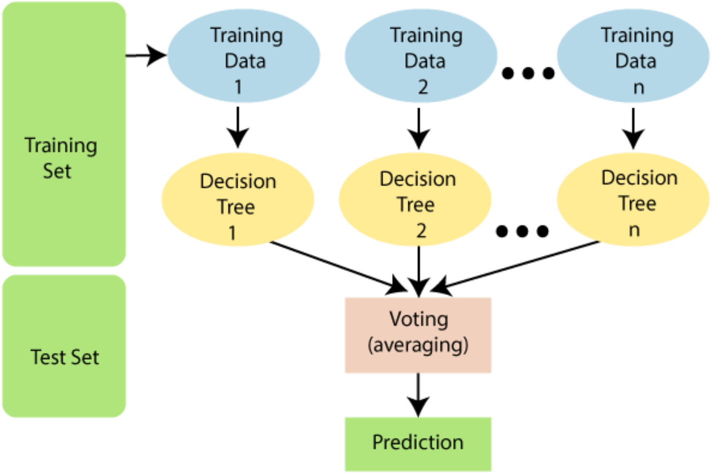


Benefits of decision tree:

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

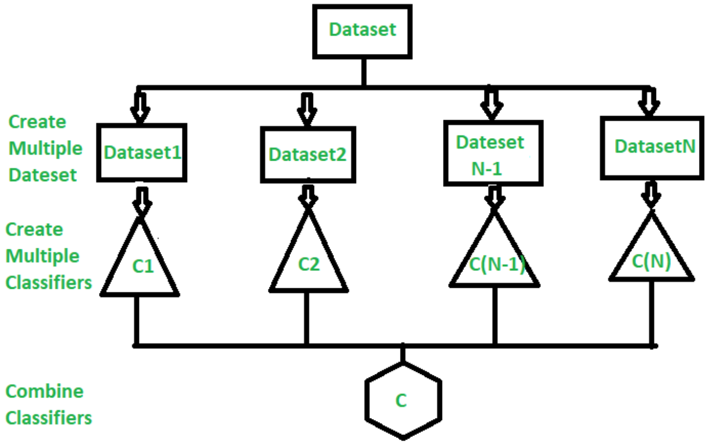
What is the Random Forest Algorithm?

- Random Forest algorithm is a powerful tree learning technique in Machine Learning.
- It works by creating a number of Decision Trees during the training phase.
- It can be used for both Classification and Regression problems in ML.
- It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- The greater number of trees in the forest leads to higher accuracy & prevents the problem of overfitting.



What are Ensemble Learning models?

- Ensemble learning helps improve machine learning results by combining several models.
- This approach allows the production of better predictive performance compared to a single model.
- Basic idea is to learn a set of classifiers (experts) and to allow them to vote.
- Ensemble learning models work just like a group of diverse experts teaming up to make decisions – think of them as a bunch of friends with different strengths tackling a problem together.



Advantage : Improvement in predictive accuracy.

Disadvantage : It is difficult to understand an ensemble of classifiers.

Bayesian Classification Methods:

- Bayesian classification uses Bayes theorem to predict the occurrence of any event.
- Bayesian classifiers are the statistical classifiers with the Bayesian probability understandings.
- The theory expresses how a level of belief, expressed as a probability.
- Bayes theorem came into existence after Thomas Bayes, who first utilized conditional probability to provide an algorithm that uses evidence to calculate limits on an unknown parameter.
- Bayesian classification is based on Bayes' Theorem.
- Bayesian classifiers are the statistical classifiers.
- It can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem

- Bayes' Theorem is named after Thomas Bayes.
- There are two types of probabilities –

1. Posterior Probability [P(H/X)]
2. Prior Probability [P(H)]

where X is data tuple and H is some hypothesis. According to Bayes' Theorem,

$$P(H/X)= P(X/H)P(H) / P(X)$$

Types of Learners in Classification:

We have two types of learners in respective to classification problems –

1. Lazy Learners
2. Eager Learners

Lazy Learners:

- As the name suggests, such kind of learners waits for the testing data to be appeared after storing the training data.
- Classification is done only after getting the testing data.
- They spend less time on training but more time on predicting.
- Examples of lazy learners are
  - K-nearest neighbor and
  - case-based reasoning.

Eager Learners:

- As opposite to lazy learners, eager learners construct classification model without waiting for the testing data to be appeared after storing the training data.
- They spend more time on training but less time on predicting.
- Examples of eager learners are
  - Decision Trees,
  - Naïve Bayes and
  - Artificial Neural Networks (ANN).

How do they differ from eager learners?

Lazy Learning	Eager Learning
The model is built during prediction.	The model is built before prediction.
Fast, store the data while traning.	Slow, try to learn from data while training.
Faster during training, but slower during prediction.	Slower during training, but faster during prediction.
k-Nearest Neighbors (KNN)	Decision Trees, Support Vector Machines (SVM), Neural Networks
Less memory usage during training, but more during prediction.	More memory usage during training, but less during prediction.

Unit – 05: Cluster Analysis

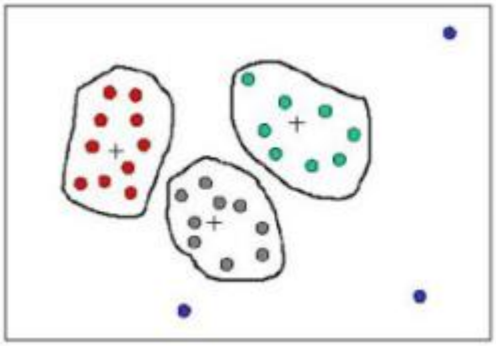
- Basic Concepts of Cluster Analysis, Clustering Structures, Major Clustering Approaches,
  - Partitioning Methods,
  - Hierarchical Methods,
  - Density-Based Methods,
  - Model-Based Clustering,
- Why outlier analysis? Identifying and handling of outliers, Outlier Detection Techniques,
- WEB MINING: Basic concepts of web mining, Different types of web mining,
- PAGE RANK Algorithm, HITS Algorithm.

Questions to be discussed:

1. Describe the term Cluster. What do you mean by cluster analysis?
2. Discuss different algorithms used for cluster analysis.
3. Explain the different clustering structures and approaches in cluster analysis.
4. What is Outlier Analysis in Data Mining? Also explain its type.
5. Differentiate between data mining and web mining.
6. Explain the page rank algorithm and its role in web mining.
7. Describe the HITS algorithm and its significance in web mining.

What is a Cluster?

- Cluster is a group of objects that belongs to the same class.
- In other words, similar objects are grouped in one cluster & dissimilar objects are in another cluster.
- We can say that a cluster is a subset of similar objects
- A collection of data object that are similar to one another and that can be treated as one group.

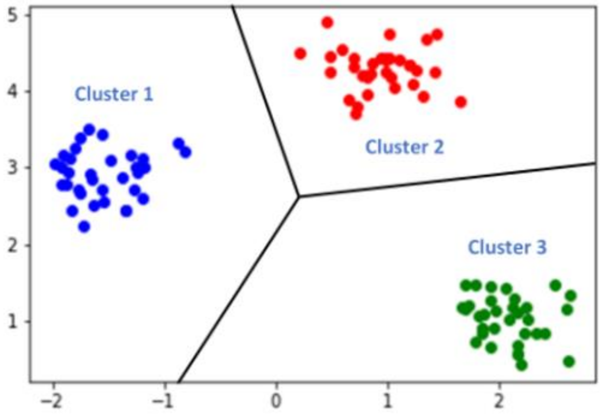


What is Clustering in Data Mining?

- Clustering is the process of making a group of abstract objects into classes of similar objects.
- In clustering, a group of different data objects is classified as similar objects.
- One group means a cluster of data.
- Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data.
- After the classification of data into various groups, a label is assigned to the group.
- It helps in adapting to the changes by doing the classification.

Cluster Analysis:

- Cluster Analysis is the process to find similar groups of objects in order to form clusters.
- It is an unsupervised machine learning-based algorithm that acts on un-labelled data.
- Cluster analysis is also known as clustering,
- It is a method of data mining that groups similar data points together.



Applications of Cluster Analysis:

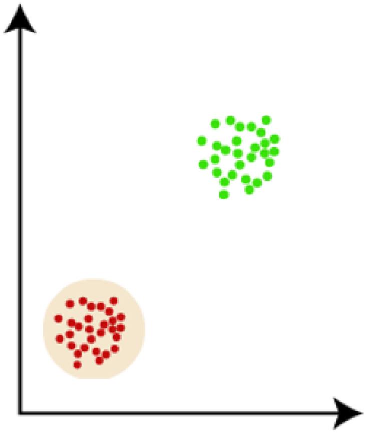
- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base.
- It can be used in the field of biology.
- It also helps in information discovery by classifying documents on the web.

Algorithms used for cluster analysis:

- There are many different algorithms used for cluster analysis, such as:
  1. Partitioning Method
  2. Hierarchical Method
  3. Density-based Method
  4. Model-Based Method etc.

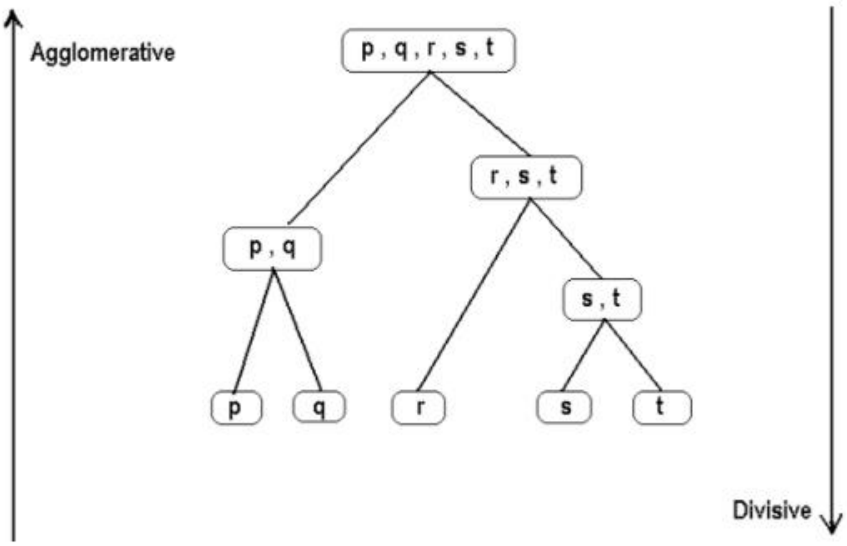
Partitioning Method:

- It is used to make partitions on the data in order to form clusters.
- If "n" partitions are done on "p" objects of the database, then each partition is represented by a cluster and n < p.
- The two conditions which need to be satisfied with this Partitioning Clustering Method are:
  - One objective should only belong to only one group.
  - There should be no group without even a single purpose.



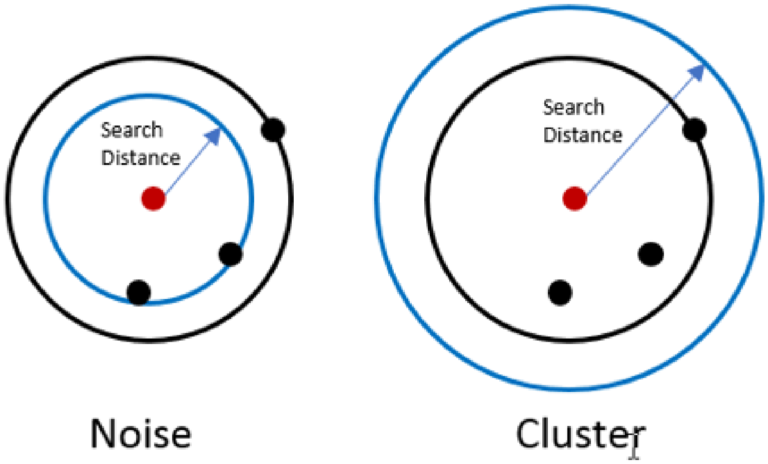
Hierarchical Method:

- In this method, a hierarchical decomposition of the given set of data objects is created.
- We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed.
- There are two types of approaches for the creation of hierarchical decomposition, they are:
  1. Agglomerative Approach
  2. Divisive Approach



Density-Based Method:

- The density-based method mainly focuses on density.
- In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster.
- The radius of a given cluster has to contain at least a minimum number of points.

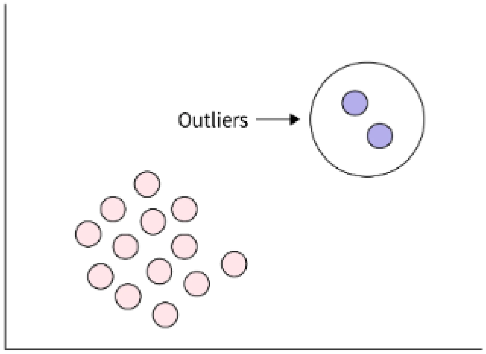


Model-Based Method:

- Here, all the clusters are hypothesized in order to find the data which is best suited for the model.
- The clustering of the density function is used to locate the clusters for a given model.
- It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account.
- Therefore, it yields robust clustering methods.

What is Outlier Analysis in Data Mining?

- It is the process of identifying & examining data points that significantly differ from rest of the dataset.
- Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner.
- They can be caused by measurement or execution errors.
- The analysis of outlier data is referred to as outlier analysis or outlier mining.
- An outlier cannot be termed as a noise or error.



Applications of Outlier Analysis:

Outlier analysis has many applications in various fields:

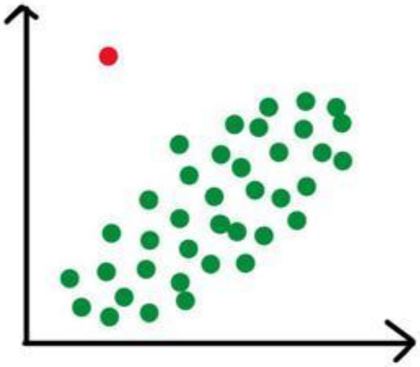
- Finance –
- Healthcare
- Manufacturing
- Marketing
- Environmental science
- Cybersecurity etc.

Types of Outliers in Data Mining:

- Outliers are of three types, namely –
  1. Global (or Point) Outliers
  2. Collective Outliers
  3. Contextual (or Conditional) Outliers

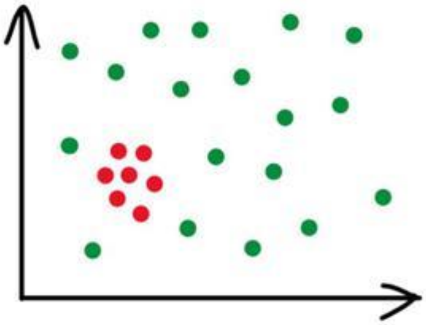
Global Outliers:

- Global outliers are data points that deviate significantly from the overall distribution of a dataset.
- Errors in data collection, measurement errors, or truly unusual events can result in global outliers.
- Global outliers can distort data analysis results and affect machine learning model performance.



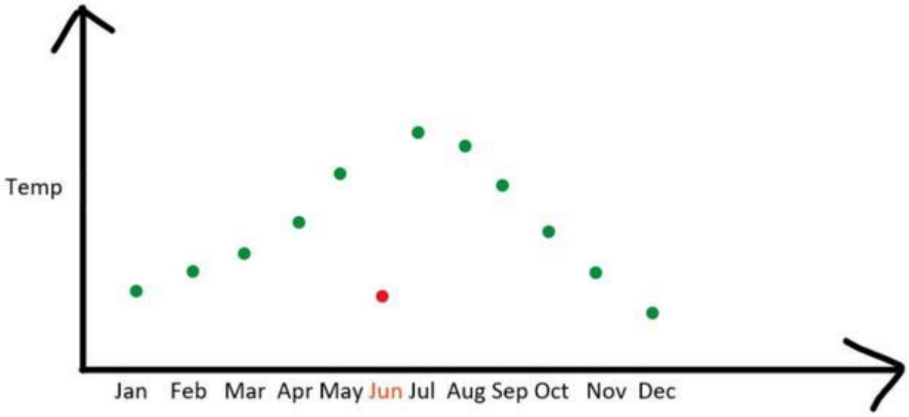
Collective Outliers:

- Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.
- Techniques for detecting collective outliers include clustering algorithms, density-based methods etc.
- Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.



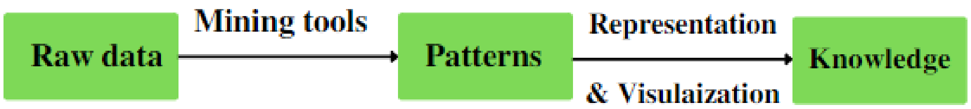
Contextual Outliers:

- Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup.
- Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.
- Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.



What is web mining?

- Web mining is the process of discovering patterns, structures, and relationships in web data.
- It involves using data mining techniques to analyze web data and extract valuable insights.
- It is the process of Data Mining techniques to automatically discover and extract information from Web documents and services.
- The main purpose of web mining is to discover useful information from the World Wide Web.
- The process of web mining is given below:

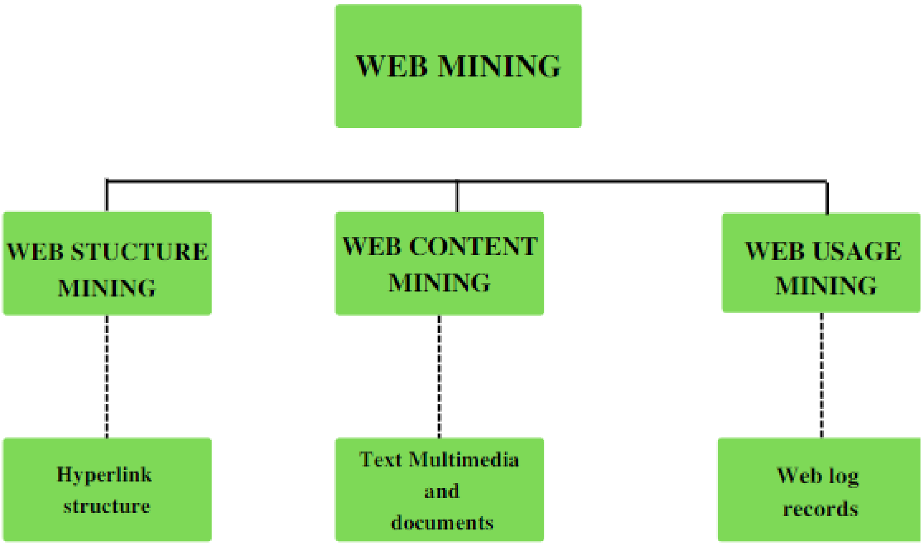


Applications of Web Mining:

- The applications of web mining are wide-ranging and include:
  - Personalized marketing
  - E-commerce
  - Search engine optimization
  - Fraud detection
  - Sentiment analysis
  - Web content analysis
  - Customer service
  - Healthcare

Types of Web mining:

- Web mining can be broadly divided into three different types of techniques of mining:
  1. Web Content Mining,
  2. Web Structure Mining
  3. Web Usage Mining.

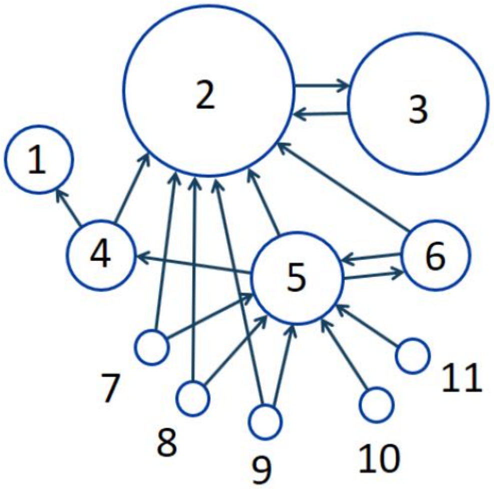


Differentiate between Data Mining and Web Mining:

Data Mining	Web Mining
It is the process that attempts to discover pattern and hidden knowledge in large data sets in any system.	It is the process to automatically discover & extract information from web documents.
Data Mining is very useful for web page analysis.	Web Mining is very useful for a particular website and e-service.
Uses Data scientist and data engineers.	Uses Data scientists along with data analysts.
It includes tools like machine learning algorithms.	Special tools for web mining are Scrapy, PageRank and Apache logs.
Suffer from Clustering, classification, regression, prediction, optimization and control.	It suffer from Web content mining, Web structure mining.

PageRank Algorithm (PR Algorithm):

- It is an algorithm used by Google Search to rank websites in their search engine results.
- It was introduced by Larry Page in the late 1990s.
- PageRank was named after Larry Page, one of the founders of Google.
- PageRank is a way of measuring the importance of website pages.
- According to Google, PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.
- The underlying assumption is that more important websites are likely to receive more links from other websites.



HITS Algorithm:

- Hits stands for Hyperlink Induced Topic Search.
- It is a Link Analysis Algorithm that rates webpages.
- Hits algorithm developed by Jon Kleinberg.
- It is used to web link-structures to discover & rank the webpages relevant for a particular search.
- HITS uses hubs and authorities to define a recursive relationship between webpages.
- The algorithm performs a series of iterations, each consisting of two basic steps:
  - Authority update and
  - Hub update

Authority update:

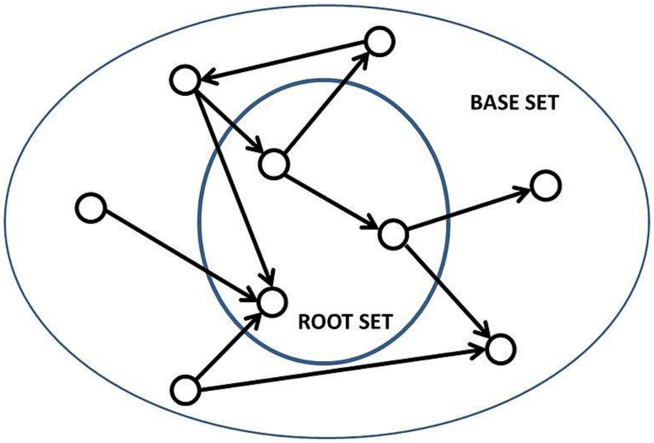
- Update each node's authority score to be equal to the sum of the hub scores of each node.
- That is, a node is given a high authority score by being linked from pages that are recognized as Hubs for information.

Hub update:

- Update each node's hub score to be equal to the sum of the authority scores of each node.
- That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

The Hub score and Authority score for a node is calculated with the following algorithm:

1. Start with each node having a hub score and authority score of 1.
2. Run the authority update rule
3. Run the hub update rule
4. Normalize the values by dividing each Hub score by square root of the sum of the squares of all Hub scores & dividing each Authority score by square root of the sum of the squares of all Authority scores.
5. Repeat from the second step as necessary.


BASE SET
ROOT SET