## Unit-I: Data Warehouse Fundamentals

### Short Answer Questions:

**Q. List characteristics of data warehouse. (Nov 23)**

Ans. A data warehouse is subject-oriented, integrated, time-variant, and non-volatile, built for query and analysis.It stores historical data to support business decisions

**Q. Define data mart. (Nov 23)**

Ans. A data mart is a subset of a data warehouse, focused on a specific business area like sales or finance.It provides faster access to data for departmental analysis.

**Q. What is metadata? Write its types.(Nov 23),(Apr 24)**

Ans. Metadata is "data about data" that describes the structure, operations, and usage of data in the warehouse.Types include business metadata, technical metadata, and operational metadata.

**Q. Define Pivot operation. (Nov 24)**

Ans. Pivot is an OLAP operation that rotates data to view it from different dimensions. It helps in reorienting the multidimensional view for better analysis.

**Q. What is the meaning of concept hierarchy? (Nov 24)**

Ans. A concept hierarchy defines levels of abstraction in data, such as city → state → country.It supports drill-down and roll-up operations in OLAP.

**Q. What is HOLAP? (Nov 24)**

Ans. HOLAP (Hybrid OLAP) combines features of ROLAP and MOLAP for efficient data storage and access.It balances scalability with performance in analytical processing.

### Long Answer Questions:

**Q. Discuss the differences between OLAP and OLTP. Provide real-world examples to illustrate these differences. (Nov 23)**

Ans. OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing) are two types of systems used in data management but serve different purposes.

**OLTP** systems are optimized for managing transaction-oriented applications. They handle a large number of short online transactions such as insert, update, and delete operations. The data is highly detailed and current. OLTP is used in day-to-day operations like banking systems, airline reservation systems, and online shopping platforms. For example, when a customer makes a purchase on Amazon, the transaction is recorded in an OLTP system.

**OLAP** systems, on the other hand, are optimized for analytical purposes. They are designed to answer complex queries and support decision-making. OLAP involves reading large volumes of data and performing complex calculations like aggregations and comparisons over time. It is used in business intelligence applications such as sales forecasting, financial reporting, and market research. For instance, a retail company analyzing the monthly sales performance across regions uses OLAP tools.

In summary, OLTP focuses on operational tasks and real-time data updates, while OLAP supports strategic analysis through historical data. OLTP requires high consistency and concurrency, whereas OLAP demands fast query performance and data summarization. Both systems are critical for a complete enterprise data solution.

**Q. Differentiate between: (Apr 24)**

**a) Operational and Informational Data Stores**

Ans. **Operational vs. Informational Data Stores (in 100 words):**

Operational Data Stores (ODS) are used to store current, real-time transactional data from various operational systems. They support day-to-day operations such as order processing or inventory tracking and are optimized for fast inserts and updates. In contrast, Informational Data Stores are used for analysis and decision-making, containing historical, aggregated, and subject-oriented data typically

used in data warehouses. These are optimized for complex queries and reporting, not frequent updates. While ODS serves operational staff for immediate business needs, informational data stores serve analysts and managers for strategic planning and business intelligence.

**Q. What is difference between data warehouse and data mining? Explain the architecture of Data warehouse in detail.(Apr 24),(Nov 24)**

Ans. A **Data Warehouse** is a centralized repository that stores integrated, historical, and subject-oriented data from multiple sources. It is used primarily for reporting and data analysis. Data warehouses enable businesses to consolidate large volumes of data for querying and decision-making purposes.

On the other hand, **Data Mining** is the process of analyzing data from different perspectives and summarizing it into useful information. It involves techniques such as clustering, classification, and association to discover hidden patterns, trends, or relationships in large datasets stored in data warehouses or databases.

**Architecture of Data Warehouse:**

The architecture of a data warehouse typically consists of three main layers:

1. **Data Source Layer:**
    - Includes various operational databases, external files, ERP/CRM systems, etc.
    - Data is extracted using ETL (Extract, Transform, Load) processes.
2. **Data Staging Area:**
    - Temporary storage where data is cleansed, transformed, and integrated.
    - Ensures data consistency and quality before loading into the warehouse.
3. **Data Storage Layer (Warehouse Database):**
    - Central repository where processed data is stored.
    - Supports multidimensional models like star or snowflake schemas.
4. **Presentation/Access Layer:**
    - Provides tools like OLAP, reporting tools, dashboards, and data mining applications for end-user access.

This architecture supports efficient analysis and business decision-making.

**Q. How Datawarehouse and Mining has become an important process in organizations for better storage and decision making in business intelligence. (Nov 23)**

Ans. **Data Warehousing and Data Mining** have become essential components of modern organizations for enhancing **data storage** and **decision-making** within the realm of **Business Intelligence (BI)**.

A **Data Warehouse** acts as a centralized, integrated repository that consolidates data from various sources such as sales, marketing, finance, and customer service. It stores large volumes of historical and current data, structured in a way that supports quick retrieval and analysis. This enables organizations to access consistent, reliable, and well-organized data to generate reports, dashboards, and trends, aiding in strategic planning and performance monitoring.

**Data Mining**, on the other hand, extracts meaningful patterns, correlations, and insights from the data stored in warehouses. It uses advanced statistical and machine learning techniques to identify customer behavior, market trends, risk factors, and fraud detection. These insights help in **predictive decision-making**, targeted marketing, inventory optimization, and customer relationship management.

Together, data warehousing and mining support **data-driven decisions**, reduce guesswork, and provide a competitive advantage. By integrating these technologies, businesses can improve accuracy in forecasting, enhance operational efficiency, and react swiftly to market changes, making them critical tools for Business Intelligence and long-term organizational success.

Unit-II: Data Preprocessing and OLAP

Short Answer Questions:

Q. What is ETL in context to data warehouse? (Nov 23)
Ans. ETL stands for **Extract, Transform, Load**, a process used to extract data from sources, transform it for analysis, and load it into a data warehouse.

Q. How missing values are handled using data preprocessing? (Nov 23)
Ans. Missing values are handled by techniques like **deletion, mean/mode/median imputation, or prediction models** during preprocessing to maintain data integrity.

Q. What is fact table in star schema? (Nov 23)
Ans. A **fact table** stores quantitative data (facts) for analysis and is linked to **dimension tables** in a star schema structure.

Q. What is a Data Cube? (Apr 24)
Ans. A **data cube** is a multidimensional array of values used in OLAP to represent data across multiple dimensions for efficient analysis.

Q. Write some methods for data pre processing. (Apr 24)
Ans. Common preprocessing methods include **data cleaning, integration, transformation, reduction, and discretization**.

Q. Why do we need to preprocess the dataset? (Nov 24)
Ans. Data preprocessing is essential to **improve data quality, remove noise or inconsistencies**, and ensure accurate mining results.

Q. Define data characterization. (Nov 24)
Ans. Data characterization summarizes **general features of data** such as mean, min, max, and trends, often used in descriptive data mining.

Long Answer Questions:
Q. What is data preprocessing? Explain different approaches for data normalization. (Nov 23)
Ans. **Data preprocessing** is a crucial step in the data mining process, involving the transformation of raw data into a clean and usable format. Raw data often contains noise, missing values, duplicates, and inconsistencies that can negatively affect the performance of machine learning models. Preprocessing enhances data quality, ensuring more accurate, efficient, and reliable analytical outcomes.
One key aspect of preprocessing is **data normalization**, which is used to scale numerical values into a standard range. This is especially important when features have different units or scales. Common approaches to normalization include:

1. **Min-Max Normalization**: Rescales data to a fixed range, typically [0, 1]. Formula:

$X_{new}=(X-X_{min})/X_{max}-X_{min}$

2. **Z-Score Normalization (Standardization)**: Transforms data based on mean and standard deviation. Useful when data follows a Gaussian distribution. Formula:

$Z=(X-\mu)/\sigma$

3. **Decimal Scaling**: Moves the decimal point of values to scale them. Formula:

$X_{new}=X/10^{j}$

where j is chosen so that the maximum absolute value of $X_{new}$ is < 1.

Normalization ensures fair comparisons among features, improving the performance of algorithms such as KNN and SVM.

**Q. Discuss few data normalization techniques used in data mining with the help of suitable examples. (Nov 24)**

Ans. Data normalization is a preprocessing step used in data mining to standardize numerical data across features, ensuring uniformity and improving the performance of machine learning algorithms. It transforms values into a common scale without distorting their relationships.

**1. Min-Max Normalization:** This technique scales data to a specific range, usually [0,1]. It is sensitive to outliers.

**Formula:**

$X_{new}=X-X_{min}/X_{max}-X_{min}$

**Example:**

If age ranges from 18 to 60, and we want to normalize 30:

$X_{new}=(30-18)/(60-18) =12/42\approx0.286$

**2. Z-Score Normalization (Standardization):** Used when data follows a Gaussian distribution. It centers data around mean 0 and standard deviation 1.

**Formula:**

$Z= (X-\mu)/\sigma$

**Example:**

For a value X = 70, mean (μ) = 50, std deviation (σ) = 10:

$Z=70-50/10=2$

**3. Decimal Scaling:** Moves the decimal point to bring values within a standard range.

**Example:**

For a value 865, divide by 10310^3103:

$X_{new}=865/1000=0.865$

These techniques help maintain consistency and improve convergence speed in machine learning models.

**Q. Write short notes on: (Apr 24)**

**a) Multi dimensional data and its schemas**

**b) OLAP Query Processing.**

Ans. **a) Multi-dimensional Data and Its Schemas (100 words):** Multi-dimensional data refers to the structured organization of data into multiple dimensions, allowing users to analyze it from various perspectives (e.g., time, location, product). This is commonly used in data warehousing and OLAP systems. The most common schemas for organizing multi-dimensional data are:

- **Star Schema**: A central fact table linked to dimension tables.
- **Snowflake Schema**: A normalized form of the star schema with hierarchies in dimension tables.
- **Fact Constellation Schema**: Also known as galaxy schema, it involves multiple fact tables sharing dimension tables.

  These schemas enhance analytical efficiency and enable detailed business insights.

**b) OLAP Query Processing (100 words):** OLAP (Online Analytical Processing) query processing enables fast and interactive analysis of multi-dimensional data. OLAP queries typically involve operations like slicing, dicing, drilling down/up, and pivoting to explore data from different viewpoints. OLAP systems can be categorized as:

- **ROLAP** (Relational OLAP): Uses relational databases and SQL queries.
- **MOLAP** (Multidimensional OLAP): Uses specialized multidimensional databases for fast processing.
- **HOLAP** (Hybrid OLAP): Combines the strengths of both ROLAP and MOLAP. Efficient OLAP query processing ensures rapid retrieval of aggregated and summarized data, helping users make timely and informed decisions.

Unit-III: Data Mining Techniques
Short Answer Questions:

Q. What is formula of calculating confidence? (Nov 23)
Ans. Confidence is calculated as:
Confidence(A → B) = Support(A ∩ B) / Support(A); it indicates how often B occurs when A occurs.

Q. What are frequent patterns? (Nov 24)
Ans. Frequent patterns are itemsets or subsequences that appear frequently in a dataset, useful in association rule mining.

Q. What is the meaning of confidence used in Association Rule Mining? (Nov 24)
Ans. Confidence measures the reliability of an inference made by a rule; it reflects the likelihood of B given A.

Q. Differentiate classification and clustering. (Apr 24)
Ans. Classification is a supervised technique that assigns predefined labels, whereas clustering is unsupervised and groups data based on similarity.

Q. What is Information Gain? (Apr 24)
Ans. Information gain measures the reduction in entropy after splitting a dataset, helping in building efficient decision trees.

Long Answer Questions:

Q. Explain the data mining process. What are the various issues in data mining? (Apr 24)
Ans. Data mining is the process of extracting meaningful patterns, trends, and insights from large datasets using statistical, machine learning, and database techniques. It helps organizations in decision-making, forecasting, and discovering hidden relationships.
The data mining process typically involves the following steps:

1. Data Cleaning – Removes noise and inconsistent data.
2. Data Integration – Combines data from multiple sources.
3. Data Selection – Selects relevant data for mining.
4. Data Transformation – Converts data into suitable formats for mining.
5. Data Mining – Applies algorithms to extract patterns.
6. Pattern Evaluation – Identifies truly interesting patterns based on measures like support, confidence, and lift.
7. Knowledge Presentation – Presents mined knowledge in visual or report formats for easy understanding.

Issues in data mining include:

- Data Quality – Incomplete, noisy, or irrelevant data can reduce accuracy.
- Scalability – Mining large datasets demands high computational resources.
- Privacy and Security – Sensitive data must be protected during analysis.
- Algorithm Selection – Choosing the right algorithm affects performance and results.
- Integration with Existing Systems – Data mining tools must be compatible with current databases and applications.

Addressing these issues is vital to ensure successful data mining and useful knowledge discovery.

Q. What is ARM? Explain Apriori algorithm, and how is it used in data mining? (Nov 23)
Ans. Association Rule Mining (ARM) is a key technique in data mining used to uncover interesting relationships or associations among items in large datasets. ARM is commonly applied in market basket analysis to identify products frequently bought together. For example, if a customer buys bread, they are also likely to buy butter. The two primary metrics used in ARM are support (how

frequently items appear together in the dataset) and **confidence** (how often the rule has been found to be true).

The **Apriori Algorithm** is a classic algorithm used for mining frequent itemsets and deriving association rules. It works on the principle that all subsets of a frequent itemset must also be frequent. The algorithm uses a breadth-first search and generates candidate itemsets of increasing lengths, pruning those that do not meet the minimum support threshold.

**Steps in Apriori Algorithm:**

1. Identify frequent individual items (1-itemsets) in the transaction database.
2. Generate candidate itemsets of length k from frequent (k−1)-itemsets.
3. Prune itemsets that do not meet minimum support.
4. Repeat until no more frequent itemsets are found.
5. Use the frequent itemsets to generate strong association rules that meet minimum confidence.

Apriori is widely used due to its simplicity and efficiency in handling large transactional data.

**Q. Explain the Apriori algorithm. Consider the given dataset with given transactions. (Apr 24)**

| TID | Items bought |
|---|---|
| 1 | (B, C, E, J} |
| 2 | {B, C, J} |
| 3 | {B, M,Y} |
| 4 | {B, J, M} |
| 5 | {C, J,M} |

**Generate rules using Apriori algorithm with given support 50% and 75% confidence.**

Ans. To solve this using the **Apriori algorithm**, we will follow these steps:

**Step 1: List all transactions**

| TID | Items bought |
|---|---|
| 1 | B, C, E, J |
| 2 | B, C, J |
| 3 | B, M, Y |
| 4 | B, J, M |
| 5 | C, J, M |

**Step 2: Minimum thresholds**

- **Minimum Support** = 50% → at least 3 transactions (out of 5)
- **Minimum Confidence** = 75%

**Step 3: Find frequent itemsets**

**Frequent 1-itemsets (count in transactions ≥ 3)**

| Item | Count | Support |
|---|---|---|
| B | 4 | 80% ✅ |
| C | 3 | 60% ✅ |
| J | 4 | 80% ✅ |
| M | 3 | 60% ✅ |
| E | 1 | 20% ❌ |
| Y | 1 | 20% ❌ |

✅ Keep: B, C, J, M

**Step 4: Generate candidate 2-itemsets and check support**

| Pair | Transactions | Count | Support | |
|---|---|---|---|---|
| B,C | 1, 2 | 2 | 40% | ✗ |
| B,J | 1, 2, 4 | 3 | 60% | ✓ |
| B,M | 3, 4 | 2 | 40% | ✗ |
| C,J | 1, 2, 5 | 3 | 60% | ✓ |
| C,M | 5 | 1 | 20% | ✗ |
| J,M | 4, 5 | 2 | 40% | ✗ |

✓ Keep: B,J and C,J

**Step 5: Generate rules from frequent itemsets**
**From {B, J} (support = 60%)**

- Rule: **B → J**, confidence = 3/4 = 75% ✓
- Rule: **J → B**, confidence = 3/4 = 75% ✓

**From {C, J} (support = 60%)**

- Rule: **C → J**, confidence = 3/3 = 100% ✓
- Rule: **J → C**, confidence = 3/4 = 75% ✓

✓ **Final Association Rules:**

1. **B → J** (support: 60%, confidence: 75%)
2. **J → B** (support: 60%, confidence: 75%)
3. **C → J** (support: 60%, confidence: 100%)
4. **J → C** (support: 60%, confidence: 75%)

These are the **valid rules** generated using the Apriori algorithm for the given dataset with **≥50% support** and **≥75% confidence**.

Unit-IV: Classification and Clustering
Short Answer Questions:

**Q. What is difference between precision and recall? (Nov 23)**
Ans. Precision is the ratio of correctly predicted positive observations to the total predicted positives, while recall is the ratio of correctly predicted positives to all actual positives.

**Q. Let us suppose that there are 200 pages available on Internet for machine learning.**
The search on this term returns total 210 pages, out of which 190 belongs to Machine Learning, calculate precision and recall for our algorithm. (Apr 24)
Ans. Precision = 190/210 ≈ 0.9047; Recall = 190/200 = 0.95; these metrics evaluate the effectiveness of the search algorithm.

**Q. What are advantages of interactive data visualization? (Nov 23)**
Ans. Interactive visualizations help in better **data understanding** and decision-making by allowing users to explore and analyze data dynamically.

**Q. List few real-world applications of clustering algorithms. (Nov 23)**
Ans. Clustering is used in **market segmentation**, **image compression**, **social network analysis**, and **document classification**.

**Q. Name some attribute selection measures. (Apr 24)**
Ans. Common measures include **Information Gain**, **Gain Ratio**, **Gini Index**, and **Chi-Square**, used in feature selection for classification.

**Q. Name the steps in data classification. (Nov 24)**
Ans. Key steps are: **Data collection**, **data preprocessing**, **model building**, **evaluation**, and **deployment**.

**Q. What is the false positive rate? (Nov 24)**
Ans. It is the proportion of actual negatives that are incorrectly classified as positives; calculated as **FP / (FP + TN)**.

**Q. How do we increase the accuracy of a classifier? (Apr 24)**
Ans. Accuracy can be improved using **feature engineering**, **model tuning**, **ensemble methods**, and **balanced datasets**.

**Q. What is the significance of high false negative results? (Apr 24)**
Ans. High false negatives can be **dangerous**, especially in critical applications like **medical diagnosis**, as they miss identifying actual positives.

Long Answer Questions:

**Q. What is data visualization? Why data visualization is important? Discuss few tools and and techniques for data visualization. (Apr 24),(Nov 24),(Nov 24)**
Ans. **Data Visualization** is the graphical or visual representation of data using charts, graphs, maps, and other tools. It helps transform raw data into a visual context, such as a bar chart or pie chart, to make data easier to understand and interpret. The goal is to communicate information clearly and efficiently to users, enabling quicker decision-making and better insight discovery.
Importance of Data Visualization:

1. **Simplifies complex data** and highlights trends and patterns.
2. **Enhances decision-making** by making data easily interpretable.
3. **Identifies outliers or anomalies** quickly.
4. **Improves communication** of data-driven insights across departments or teams.
5. Makes presentations and reports more engaging and interactive.

Tools for Data Visualization:

- **Tableau:** Popular for its user-friendly interface and real-time data analytics.
- **Power BI:** Microsoft's tool integrated with Excel and cloud services.
- **Google Data Studio:** Free and easy-to-use tool for creating dashboards.
- **Matplotlib & Seaborn (Python):** Libraries used in programming for customizable visualizations.

Techniques:

- **Bar charts and histograms:** To compare quantities.
- **Line graphs:** To show trends over time.
- **Pie charts:** To show part-to-whole relationships.
- **Heatmaps and scatter plots:** To show relationships and density.

In conclusion, data visualization bridges the gap between complex datasets and actionable insights.

**Q. What is supervised learning in context to classification? Explain the concept of decision trees used for classification processes. (Nov 23)**

Ans. **Supervised learning** is a type of machine learning where the model is trained on a labeled dataset, meaning the input data is paired with the correct output. In the context of **classification**, supervised learning involves teaching the model to categorize data into predefined classes. For example, identifying whether an email is "spam" or "not spam" based on historical examples.

One of the most commonly used algorithms for classification in supervised learning is the **Decision Tree**. A decision tree is a flowchart-like structure where each internal node represents a decision or a test on an attribute (e.g., "Is age > 30?"), each branch represents the outcome of the test, and each leaf node represents a class label (e.g., "Approved" or "Denied").

The process of building a decision tree involves:

- Selecting the best attribute to split the data using measures like **Gini Index**, **Information Gain**, or **Gain Ratio**.
- Recursively splitting the dataset into subsets until all data points in a subset belong to the same class or a stopping condition is met.

Decision trees are popular because they are easy to understand, interpret, and visualize. However, they can overfit on training data, which can be managed by pruning or using ensemble methods like **Random Forest**.

**Q. Define classification. Explain K-Nearest-Neighbour Classifiers with the help of a suitable example. (Nov 24)**

Ans. **Classification** is a supervised learning technique in machine learning that involves categorizing data into predefined classes or labels based on input features. It is used in numerous real-world applications such as spam detection, disease diagnosis, and sentiment analysis.

One popular classification algorithm is the **K-Nearest Neighbour (K-NN)** classifier. It is a simple, instance-based learning method that classifies new data points based on the majority label of their 'k' nearest neighbors in the feature space.

How K-NN works:

1. Choose the number of neighbors**K**.
2. Calculate the **distance** (commonly Euclidean) between the new data point and all other points in the dataset.
3. Select the **K closest points**.
4. Assign the new point to the **most common class** among those K neighbors.

Example:

Suppose we want to classify a fruit based on its weight and color. Given a dataset of labeled fruits (e.g., apples, oranges, bananas), if a new fruit with unknown type has a weight of 150g and is red, the K-NN algorithm will:

- Measure distances to all known fruits.
- Identify the K nearest ones (say K=3).
- If two of them are apples and one is an orange, the new fruit is classified as an **apple**.

K-NN is simple and effective, but can be computationally expensive for large datasets.

**Q. How do we evaluate classifier's accuracy? Explain different techniques for accuracy estimation. (Apr 24)**

Ans. Evaluating a classifier's accuracy is crucial to understand its performance and reliability in predicting outcomes. **Accuracy** refers to the proportion of correctly classified instances out of the total instances. However, accuracy alone might not be sufficient, especially with imbalanced datasets. Therefore, multiple evaluation techniques are used.

**1. Confusion Matrix:**

A table used to describe the performance of a classification model by comparing actual vs. predicted classes. It includes:

- **True Positives (TP)** – Correct positive predictions.
- **True Negatives (TN)** – Correct negative predictions.
- **False Positives (FP)** – Incorrectly predicted as positive.
- **False Negatives (FN)** – Incorrectly predicted as negative.

From this, metrics like:

- **Precision** = TP / (TP + FP)
- **Recall (Sensitivity)** = TP / (TP + FN)
- **F1-Score** = 2 * (Precision * Recall) / (Precision + Recall)

**2. Cross-Validation:**

Data is split into K parts (folds). The model is trained on K parts and tested on the remaining part. This process is repeated K times, and the average accuracy is taken.

**3. Holdout Method:**

Data is split into a training set and a test set. The classifier is trained on the training set and evaluated on the test set.

These methods ensure that the classifier generalizes well to unseen data.

**Q. What is clustering? Explain different types of clusters. Explain k-means clustering method by taking a suitable example. (Apr 24)**

Ans. **Clustering** is a type of unsupervised learning technique used in data mining to group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups. It helps in identifying inherent patterns or structures in data without pre-labeled responses.

**Types of Clusters:**

1. **Exclusive (or Hard) Clustering:** Each data point belongs to exactly one cluster (e.g., k-means).
2. **Overlapping Clustering:** Data points can belong to multiple clusters (e.g., fuzzy c-means).
3. **Hierarchical Clustering:** Builds a hierarchy of clusters either bottom-up (agglomerative) or top-down (divisive).
4. **Density-Based Clustering:** Forms clusters based on dense regions of data points (e.g., DBSCAN).

**K-Means Clustering:**

K-Means is a popular exclusive clustering algorithm. It partitions the dataset into *k* clusters where each data point belongs to the cluster with the nearest mean.

**Steps:**

1. Choose *k* initial centroids randomly.
2. Assign each data point to the nearest centroid.
3. Update centroids by calculating the mean of assigned points.
4. Repeat steps 2 and 3 until convergence.

**Example:**

Given data points: (1,1), (2,2), (8,8), (9,9) and $k = 2$, initial centroids could be (1,1) and (8,8). After several iterations, two clusters will form: one around (1,1)-(2,2) and another around (8,8)-(9,9).

**Q. Why clustering is used in data mining? Explain the different types of clustering methods used in data mining. (Nov 24)**

Ans. **Clustering** is used in data mining to identify natural groupings or patterns within datasets where data points share similar characteristics. Unlike classification, clustering doesn't rely on pre-labeled

data, making it ideal for exploratory data analysis, pattern recognition, image processing, and customer segmentation.

Why Clustering is Used:

- **Discover patterns:** Helps uncover hidden structures in data.
- **Data compression:** Groups similar data points for simplified analysis.
- **Outlier detection:** Identifies unusual data points that don't belong to any group.
- **Customer segmentation:** Groups customers based on behavior or preferences.
- **Image and document classification:** Automatically groups similar files.

Types of Clustering Methods:

1. **Partitioning Methods (e.g., K-Means):** Divides data into non-overlapping subsets where each data point belongs to exactly one cluster. Simple and efficient for large datasets.
2. **Hierarchical Methods (e.g., Agglomerative, Divisive):** Builds nested clusters in a tree-like structure (dendrogram), useful for detailed analysis.
3. **Density-Based Methods (e.g., DBSCAN):** Forms clusters based on dense regions in the data, effective for irregularly shaped clusters and noise detection.
4. **Grid-Based Methods (e.g., STING):** Divides the data space into a finite number of cells and forms clusters based on density within these cells.

Each method suits different data structures and clustering goals, offering flexibility for diverse mining applications.

**Q. Clustering's role in unsupervised learning. Explain k-means clustering algorithm with the help of example. (Nov 23)**

Ans. **Clustering** plays a vital role in **unsupervised learning**, where the goal is to discover hidden patterns or groupings in data without predefined labels. It helps in identifying structures within data by grouping similar instances into clusters based on feature similarity. Clustering is widely used in customer segmentation, document categorization, image analysis, and anomaly detection.

K-Means Clustering Algorithm:

K-Means is one of the most popular partitioning clustering algorithms. It divides the dataset into **K distinct, non-overlapping clusters** where each data point belongs to the cluster with the nearest mean (centroid).

Steps:

1. Choose the number of clusters **K**.
2. Initialize **K** centroids randomly.
3. Assign each data point to the nearest centroid.
4. Update centroids by calculating the mean of all data points in each cluster.
5. Repeat steps 3 and 4 until centroids no longer change significantly (convergence).

Example:

Consider data points: (1,2), (1,4), (3,4), (5,6), (6,5), (7,7) and **K = 2**. Initially, two centroids are selected randomly. Based on distance, points are assigned to the nearest centroid. Centroids are recalculated and reassigned until clusters stabilize. This results in two well-defined groups representing natural divisions in the dataset.

K-Means is efficient and simple, but requires specifying K and is sensitive to outliers.

**Q. Write a short note on: (Nov 24)**

**a) Decision Tree Induction**

**b) Data Cube Computation.**

Ans. **a) Decision Tree Induction:** Decision Tree Induction is a popular supervised learning technique used for classification and prediction. It involves breaking down a dataset into smaller subsets while developing an associated decision tree incrementally. The final result is a tree with decision nodes and leaf nodes. Attributes are selected based on measures like information gain or Gini index to split the data. It is easy to understand and interpret, making it widely used in data mining for decision-making tasks.

**b) Data Cube Computation:** Data Cube Computation is a core operation in Online Analytical Processing (OLAP) systems that organizes data into multidimensional structures for analysis. It enables users to

view and analyze data from multiple perspectives, such as sales by region, product, and time. A data cube computes group-by aggregations across combinations of dimensions. Techniques like iceberg cubes and materialized views are used to optimize storage and speed up querying, making it crucial for efficient data analysis in data warehousing.