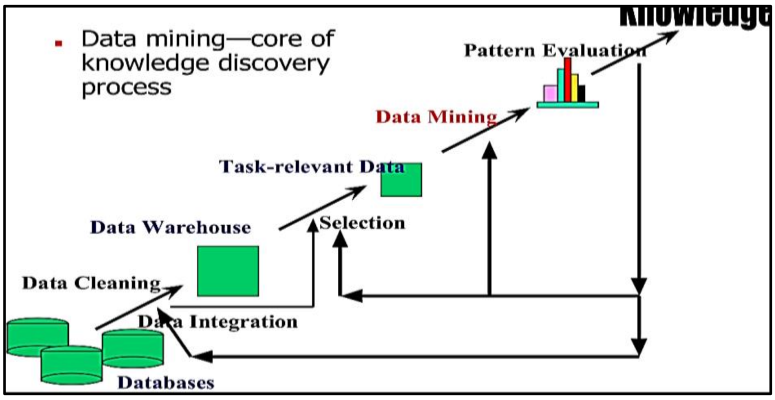
## UNIT I

### 1.EXPLAIN KDD IN DETAIL. (2013, 2014, 2017, 2018, 2018, 2019)(MID)

KDD stands for Knowledge Discovery in Databases, and it ***refers to the process of extracting meaningful, actionable knowledge from large datasets.*** Here's an overview of the steps involved in the KDD process:



**Data Mining:** Data mining is the ***process of extracting and discovering patterns in large data sets*** involving methods at the intersection of ***machine learning, statistics, and database systems.*** It is the ***process of uncovering patterns and other valuable information*** from large data sets.

**KDD Process Steps:**

1. **Data Cleaning:**
   - Removal of noisy and irrelevant data.
   - Handling ***missing values.***
   - ***Detection and correction of errors.***
2. **Data Integration:**
   - Combining ***heterogeneous data from multiple sources*** into a common source (e.g., a data warehouse).
   - Using tools like ***data migration, synchronization, and ETL (Extract-Transform-Load)*** processes.
3. **Data Selection:**
   - Deciding and retrieving data relevant to the analysis.
   - Techniques such as ***sampling, dimensionality reduction, feature, attribute, and instance selection*** can be used for data selection.
4. **Data Transformation:**
   - ***Converting data into a suitable form required by mining procedures.***
   - Involves ***data mapping, code generation, smoothening, normalization and discretization.***
5. **Data Mining:**
   - ***Applying clever techniques to extract potentially useful patterns from the transformed data.***
   - ***Transforms task-relevant data into patterns.***
   - ***Classification or characterization*** is used to determine the purpose of the model.
6. **Pattern Evaluation:** This step involves ***evaluating the patterns and information extracted from the data mining step.*** For example, we can:
   - Find the ***interestingness score*** of each pattern.
   - Use ***summarization and visualization*** to make the data and patterns understandable by the user.

7. **Knowledge Representation:**
   - Utilizing ***visualization tools to represent data mining results.***
   - ***Generating reports, tables, discriminant rules, classification rules, characterization rules, etc.***

Overall, the KDD process involves ***iterative steps*** where ***evaluation measures can be enhanced, mining techniques can be refined,*** and ***new data can be integrated and transformed to achieve different and more appropriate results.*** Pre-processing of databases typically includes both data cleaning and data integration steps to ensure the quality and usability of the data for mining purposes.

**Now, let's move on to the applications of data mining:**

1. **Healthcare**: Data mining has a lot of promise for improving healthcare systems. It ***identifies best practices for improving treatment and lowering costs*** using data and analytics.

2. **Banking and Finance**: Banks typically use data mining to ***find out their prospective customers*** who could be interested in ***credit cards, personal loans, or insurance.***

3. **Retail Industry**: Retailers use data mining to ***identify customer segments and target marketing campaigns.***

4. **Telecommunication Industry**: Telecommunication companies use data mining to predict customer churn and develop customer retention strategies.

5. **Scientific Analysis**: Scientific simulations generate bulks of data every day. Data mining techniques are capable of analysing these data.

## 2. EXPLAIN METHODS OF DATA CLEANING. (2012)

Data cleaning, also known as ***data cleansing,*** is an essential step in the data pre-processing phase of data mining. It involves ***detection and correcting errors, removal of noisy/irrelevant data, and handling missing values*** in the dataset to ensure its ***quality and reliability.*** Here's list of methods used in data cleaning:

### 1. Handling Missing Values:

1. **Ignore the tuple:** This method involves ***disregarding tuples*** with missing values. However, it's only suitable when the percentage of ***missing values per attribute is low*** and consistent. Otherwise, valuable data may be discarded.
2. **Fill in the missing value manually:** ***Time-consuming and impractical for large datasets.***
3. **Use a global constant**: Replace missing values with a constant ***(e.g., "Unknown").*** Simple but ***may bias the data and mislead analysis.***
4. **Mean/Median:** Replace missing values with the ***mean or median of the attribute.*** The choice ***depends on the distribution of the data.***
5. **Use the most probable value:** Predict missing values using ***regression, decision trees, or Bayesian inference.*** It utilizes other attributes' values to predict missing values, ***preserving relationships within the data.***

### 2. Handling Noisy Data:

1. **Regression:**
   Fits data values to a function ***(e.g., linear regression)*** to ***smooth out noise.*** Useful for identifying ***trends and patterns*** in noisy data.
2. **Outlier analysis:**

***Identifies outliers by clustering*** similar values into groups and ***considering values outside these clusters as outliers. Outliers can be indicator of noise or anomalies in the data.***

## 3. EXPLAIN TERMS. (2019)

### A. DATA INTEGRATION

Data integration ***brings information from diverse sources together,*** but this process faces several challenges. Here's a breakdown of the terms you mentioned:

**1. Entity Identification Problem**:

Imagine combining customer data from two sources. One ***identifies customers by ID, another by email.*** Matching these different representations of the same person (real-world entity) is the ***entity identification problem.***

**2. Redundancy and Correlation Analysis:**

Combining data often leads to **redundancy:** ***repetitive information across sources.*** For example, "customer address" might appear in different formats or under different labels.

***Correlation analysis helps identify and remove unnecessary redundancy.*** It ***examines how attributes relate to each other, revealing which ones can be derived from others*** (e.g., "zip code" from "full address").

***Removing redundancy reduces storage requirements and improves data quality.***

**3. Tuple Duplication:**

Tuple duplication ***refers to identical records within a single dataset.*** This can arise from ***data entry errors, inconsistencies, or different representations of the same entity.***

Duplicate tuples inflate data size and ***skew analysis results.*** Techniques like ***deduplication algorithms and fuzzy matching*** help identify and eliminate them.

### B. BINNING ***(used in data transformation)***

A ***data transformation technique*** used to ***transform continuous numeric data into discrete bins.***

**Binning Techniques:**

1. **Bin Means:** Each value within a bin is replaced by the mean (average) value of that bin. For instance, if a bin contains values 4, 8, and 15, the mean of these values (9) is used to replace each original value in the bin.
2. **Bin Medians**: Instead of using the mean, the median value of each bin is employed to replace bin values. The ***median is the middle value when the data is sorted.***
3. **Bin Boundaries:** In this method, the ***minimum and maximum values within each bin are identified as the bin boundaries.*** Each ***value in the bin*** is ***replaced by the nearest boundary value.***

### C. DATA NORMALIZATION ***(used in data transformation)***

Data normalization is a ***transformation technique*** used to ***transform numeric data into a standard scale, ensuring that all attributes contribute equally to the analysis.*** The primary goal of ***normalization is to give all attributes equal weight during data analysis.*** This is essential for algorithms where attribute scales can influence the outcome significantly.

**Methods of Data Normalization:**

1. **Min-Max Normalization:** This method performs a ***linear transformation*** on the original data to ***scale it between a specified range (e.g., [0, 1]).***
2. **Z-Score Normalization:** Also known as ***standardization,*** this method ***transforms the data to have a mean of 0 and a standard deviation of 1.*** It calculates the z-score of each value.
3. **Decimal Scaling:** In this method, ***each value is scaled by a power of 10 to ensure it falls within a specified range.*** The scaling factor depends on the maximum absolute value in the dataset.

## 4. EXPLAIN THE CONCEPTS OF DATA PREPROCESSING. DESCRIBE ITS NEEDS. HOW TO HANDLE NOISY DATA WITH PREPROCESSING?(MID)

**Data Preprocessing:** Data preprocessing is a crucial step in the data mining process. It ***transforms raw data into an understandable and efficient format.*** It involves ***Data Cleaning, Data Integration, Data Selection, Data Transformation.*** These processes ***enhance the quality of the raw data before it's fed into a machine learning algorithm.***

**Need for Data Preprocessing:** Data collected from various sources is often ***messy, inconsistent, and unreliable.*** It may contain ***errors, outliers, or missing values*** that could ***lead to inaccurate predictions or models.*** Therefore, preprocessing is needed to:

1. Improve the ***accuracy*** of the models.
2. Provide ***high-quality, reliable,*** and ***consistent data.***
3. ***Reduce the complexity*** of the data.
4. ***Transform the data into a suitable format*** for the machine learning algorithms.

**Handling Noisy Data** Noisy data is a ***meaningless data that can't provide any insight.*** It's caused by various factors like ***errors, outliers, or missing values.*** Here are some methods to handle noisy data:

**1. Regression**: Regression analysis is a ***statistical process for estimating the relationships among variables.*** It ***includes many techniques for modeling and analyzing several variables*** when the focus is on the relationship between a ***dependent variable and one or more independent variables.***

**Types of Regression:**

1. **Linear Regression:** It is ***used when the dependent variable is continuous and the independent variables are either continuous or discrete.***
2. **Logistic Regression:** It is ***used when the dependent variable is binary in nature.***
3. **Polynomial Regression:** It is a form of regression analysis in which the ***relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial.***

**2. Clustering:** Clustering is a ***technique used in machine learning 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's a method of ***unsupervised learning*** and a common technique for ***statistical data analysis*** used in many fields.

**Types of Clustering:**

1. **K-Means Clustering:** This method ***partitions the input data into K distinct clusters based on distance to the centroid of a cluster.*** The algorithm ***iteratively assigns each data point to one of the K groups.***

2. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** It's a density-based clustering algorithm, which can ***discover clusters of different shapes and sizes*** from a large amount of data, ***which is containing noise and outliers.***

3. **Hierarchical Clustering:** This algorithm builds a hierarchy of clusters where each node is a cluster consisting of the clusters of its offspring nodes.

**3. Outlier Analysis:** Outliers are ***extreme values that deviate significantly from other observations*** in the dataset. They can be ***caused by measurement or execution error.*** The ***analysis of outlier data*** is referred to as ***outlier analysis.***

**Methods for Outlier Detection:**

1. **Z-Score:** The Z-score is a mathematical measurement that ***describes a value's relationship to the mean of a group of values.***

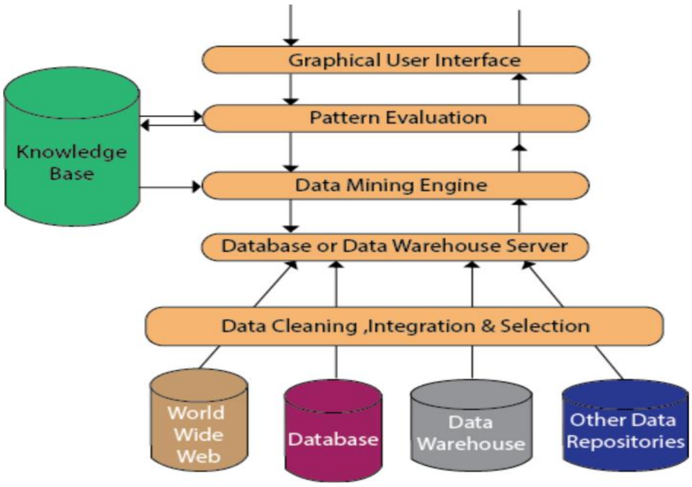
2. **IQR Score:** Any ***data point that falls below the Lower bound or above the Upper bound*** is considered an outlier.

3. **Scatter Plots:** You can start with a ***scatter plot*** of the data to ***visualize and find potential outliers.***

Remember, ***these methods are not exclusive and can be used in combination*** to preprocess your data effectively.

The goal of data preprocessing is to ***improve the data quality and make the dataset compatible with the data mining algorithms.*** It's a crucial step in the data mining process and can ***significantly impact the outcomes of the final model.***

**5. WITH BLOCK DIAGRAM EXPLAIN THE ARCHITECTURE OF TYPICAL DATA MINING SYSTEM. (2012)**



**Data mining Architecture**

Based on the architectural diagram shown, a typical data mining system consists of the following components:

1. **Data Sources:**
   - Database
   - Data Warehouse
   - World Wide Web
   - Other Data Repositories

2. **Data Cleaning, Integration, and Selection:** This component is responsible for ***extracting data from the various sources, cleaning and integrating the data,*** and ***selecting the relevant data*** for the mining process.
3. **Database or Data Warehouse Server**: The ***cleaned and integrated data is stored in a database or data warehouse server*** for efficient access and retrieval.
4. **Data Mining Engine:** This is the core component that performs the actual data mining tasks. It applies various data ***mining algorithms and techniques*** to the data stored in the database or data warehouse server to ***discover patterns, relationships, and insights.***
5. **Pattern Evaluation:** The patterns and models discovered by the data mining engine are ***evaluated and analysed in this component.*** Techniques like ***visualization, statistical analysis, and business rules*** are used to assess the ***significance and usefulness*** of the discovered patterns.
6. **Knowledge Base:** The ***validated and interpreted patterns are stored in a knowledge base*** for ***further analysis, decision-making,*** and ***use by other systems*** or ***applications.***
7. **Graphical User Interface (GUI):** This component provides a user-friendly interface for users to interact with the data mining system. It allows users to ***configure mining tasks, visualize results,*** and ***explore the discovered knowledge.***

The flow of the data mining process in this architecture can be summarized as follows:

1. Data is extracted from various sources.
2. The ***extracted data is cleaned, integrated,*** and ***relevant data is selected.***
3. The processed data is stored in a database or data warehouse server.
4. The ***data mining engine applies algorithms*** to the stored data to discover patterns.
5. The discovered patterns are evaluated for their significance and usefulness.
6. The validated patterns are stored in a knowledge base.
7. The graphical user interface allows users to interact with the system and explore the discovered knowledge.

This architecture provides a ***structured and organized approach to the data mining process,*** enabling ***efficient data management, pattern discovery***, and ***knowledge extraction*** from various data sources.

**6. Using following methods to normalize the following group of data: (2019, 2019)**

**200, 300, 400, 600, 1000**

**a. Min max normalization by setting min = 0 and max = 1**

**b. Z-score normalization using mean absolute deviation instead of standard deviation**

**c. Normalization by decimal scaling**

**7. Let a group of 12 sales price record has been sorted as follows: (2017)**

**5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215**

**Partition them into three bins by each of following methods**

**a. Equal frequency (Equal depth) b. Equal width c. Clustering**

**8. EXPLAIN HISTOGRAM AND SAMPLING FOR REDUCTION OF A DATA. (2014, 2019)**

Data reduction refers to the ***process of transforming or reducing a large dataset into a smaller, more manageable dataset*** while ***preserving the essential statistical properties*** and ***patterns***

***within the original data.*** It involves applying techniques to derive a summarized or reduced representation of the data, making it easier to ***analyse, store, and process.*** There are several data reduction techniques, and I will explain two of them: **histogram** and **sampling.**

**Histogram:** A histogram is a graphical representation that displays the distribution of data by partitioning the data into bins or intervals and ***depicting the frequency or count of data points falling within each bin.*** It ***provides a visual summary of the data's distribution,*** making it easier to identify patterns, central tendencies, and outliers.

To create a histogram, the range of data is divided into equal-sized, non-overlapping intervals (bins) along the horizontal axis. The vertical axis represents the frequency or count of data points falling within each bin. Histograms are particularly useful for understanding the shape of the data distribution, identifying modes (peaks), and detecting skewness or symmetry.

**Sampling:** Sampling is a data reduction technique that involves ***selecting a representative subset of data from the larger population or dataset.*** The ***goal is to obtain a smaller sample*** that ***accurately reflects the characteristics and patterns of the original dataset,*** allowing for analysis and inference without processing the entire dataset.

**There are various sampling techniques, including:**

1. **Simple random sampling:** Each data point in the population has an equal chance of being selected for the sample.
2. **Stratified sampling:** The population is divided into non-overlapping subgroups (strata), and samples are randomly selected from each stratum.
3. **Cluster sampling:** The population is divided into groups or clusters, and a random sample of clusters is selected.
4. **Systematic sampling:** Data points are selected at regular intervals from the population, such as every 10th item.

In summary, data reduction techniques like histograms and sampling allow researchers and analysts to work with more ***manageable and condensed representations of large datasets,*** facilitating ***analysis, visualization, and storage*** while preserving the essential patterns and characteristics of the original data.

## 2. UNIT II

### 1. WHAT DO YOU MEAN BY CLASSIFICATION? EXPLAIN K-NEAREST NEIGHBOUR ALGORITHM/ DISTANCE BASED CLASSIFICATION (2012, 2018, 2019)

Classification is a ***supervised learning technique*** in data mining and machine learning where the ***goal is to assign a class label or category to new instances*** based on a ***training dataset containing instances with known class labels.***

K-Nearest Neighbour (KNN) algorithm is a popular ***distance-based classification algorithm*** that ***assigns a class label to a new instance based on the class labels of its k-nearest neighbours*** in the training dataset. Here's how the KNN algorithm works:

1. **Distance Calculation:** For a new instance, the algorithm calculates the distance between that instance and all the instances in the training dataset. Common distance metrics used are ***Euclidean distance, Manhattan distance, and Minkowski distance.***

2. **Neighbour Selection:** After calculating the distances, the algorithm selects the k-nearest neighbours to the new instance. The ***value of k is a positive integer,*** typically small, and determined beforehand.

3. **Class Assignment:** Among the k-nearest neighbours, the algorithm identifies the classes represented by those neighbours and counts the instances of each class.

4. **Majority Vote:** The new instance is assigned the class label that is most frequent among the k-nearest neighbours. In case of a tie, various tie-breaking techniques can be employed, such as choosing the class with the nearest neighbour or using a probability model.

The KNN algorithm is a ***non-parametric method,*** meaning it ***makes no assumptions about the underlying data distribution.*** It is a ***lazy learner,*** as it ***doesn't learn a model from the training data*** but instead uses the entire training dataset for classification.

The ***choice of k and the distance metric*** can ***significantly impact the performance of the KNN algorithm.*** A ***smaller value of k can lead to overfitting,*** while a ***larger value may result in underfitting.***

KNN is ***easy to understand and implement*** and ***works well for low-dimensional data.*** However, it can be ***computationally expensive for large datasets.***

In summary, the K-NN algorithm is a straightforward yet effective method for classification tasks, relying on the principle of ***similarity between instances in the feature space to make predictions for unseen data points.***

### 2. WHAT IS A DECISION TREE? EXPLAIN DECISION TREE CLASSIFICATION. STATE AND EXPLAIN THE ALGORITHM FOR BUILDING A DECISION TREE. EXPLAIN THE ADVANTAGES AND DISADVANTAGES OF IT. (2012, 2013, 2014, 2019)

**Decision Tree:** A decision tree is a ***popular supervised learning algorithm*** used for both classification and regression tasks in data mining and machine learning. It represents a tree-like structure where each ***internal node represents a "decision"*** based on a ***feature attribute,*** each ***branch represents the outcome of the decision,*** and each ***leaf node represents the class label.***

**Decision Tree Classification:** Decision tree ***classification involves constructing a decision tree from the training data,*** where the **features and their values** are used to make decisions that lead to the classification of instances into different classes or categories.

**Algorithm for Building a Decision Tree:**

1. **Selecting the Best Attribute:**
    - Choose the ***best attribute as the root of the tree.*** The best attribute is the one that ***maximizes the information gain or minimizes impurity measures*** such as ***entropy or Gini index.***
2. **Splitting the Dataset:**
    - Split the dataset into subsets ***based on the values of the selected attribute.***
3. **Recursively Building Subtrees:**
    - ***For each subset, repeat steps 1 and 2 recursively*** until one of the following conditions is met:
        - ***All instances in the subset*** belong to the ***same class (pure node).***

- No more attributes are left for splitting ***(leaf node).***
- Stopping criteria such as a ***maximum tree depth or minimum number of instances per leaf are met.***

4. **Pruning (Optional):**

   - After the tree is fully grown, prune the tree to reduce ***overfitting and improve generalization performance.*** Pruning involves ***removing nodes that do not significantly improve the tree's predictive accuracy.***

**Advantages of Decision Trees:**

- **Interpretability:** Decision trees are ***easy to interpret and visualize,*** making them useful for explaining the reasoning behind classification decisions.
- **Handles Non-linear Relationships:** Decision trees can ***capture non-linear relationships between features and target variables.***

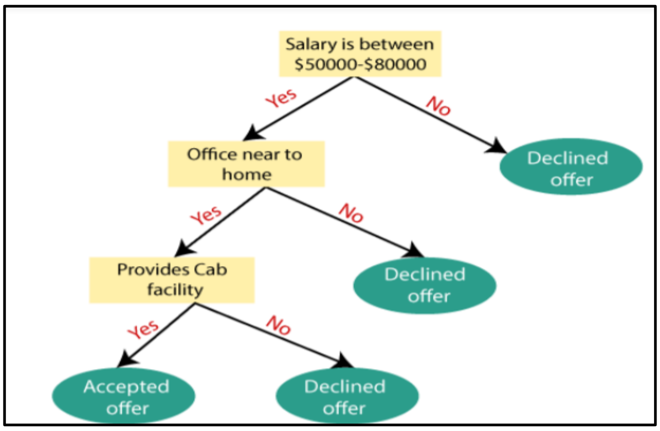
**Disadvantages of Decision Trees:**

- **Overfitting:** Decision trees are prone to overfitting, especially ***when the tree is deep*** and complex. ***Pruning techniques can help alleviate this issue.***
- **Instability:** ***Small changes in the data*** can lead to ***significant changes in the structure*** of the decision tree, resulting in instability.

In summary, decision trees are ***intuitive and powerful classification algorithms*** that offer ***interpretability and flexibility,*** but they ***require careful tuning to avoid overfitting and instability issues.***

**Example:**
Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node ***(Salary attribute by ASM).*** The root node splits further into the next decision node ***(distance from the office)*** and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node ***(Cab facility)*** and one leaf node.

Finally, the decision node splits into two leaf nodes ***(Accepted offers and Declined offer).*** Consider the below diagram:



## 3. EXPLAIN NEURAL NETWORK BASED CLASSIFICATION. EXPLAIN ADVANTAGES AND DISADVANTAGES OF IT. (2013, 2014, 2019)

Neural network-based classification is a technique used to categorize data points into predefined classes. It mimics the ***structure and function of the human brain using interconnected layers of artificial neurons.*** Here's how it works:

**1. Structure:**

- **A neural network for classification consists of three main layers:**
  - **Input layer:** Receives the raw data points (e.g., image pixels, email features).
  - **Hidden layers:** These layers ***process the data using mathematical functions inspired by how neurons work in the brain.*** There can be ***one or more hidden layers,*** and the ***complexity of the network increases with the number of layers*** and ***neurons.***
  - **Output layer:** ***Produces the classification results.*** For example, it might output the probability of an image containing a cat or a dog.

**2. Training:**

- The network is trained on a dataset where each data point has a known classification.
- The network iterates through the data, ***adjusting the connections between neurons*** (weights) to ***minimize the difference between its predicted classifications*** and the ***actual classifications.***
- This process, called backpropagation, helps the network learn the patterns that distinguish between different classes.

**3. Classification:**

- Once trained, the ***network can classify new, unseen data points.***
- The data point is fed into the input layer, processed through the hidden layers, and an output is generated by the output layer.
- The output layer typically represents the probability of the data belonging to each class. The class with the highest probability is chosen as the predicted classification.

**Example: Spam Classification**

Imagine you want to train a neural network to classify ***emails as spam or not spam.*** Each email can be represented by features like the presence of certain words or the sender's address.

- The input layer would receive these features for each email.
- The ***hidden layers would learn the complex relationships*** between these features and spam emails.
- The output layer would give a probability of the email being spam.

**Advantages of Neural Network-based Classification:**

- **High Accuracy:** Neural networks can ***achieve very high accuracy on complex classification tasks,*** especially with large amounts of training data.
- **Adaptability:** They can handle various data types like images, text, and sensor data.

- **Feature Learning:** Neural networks can automatically learn important features from the data, eliminating the need for manual feature engineering.

**Disadvantages of Neural Network-based Classification:**

- **Complexity:** Designing and training neural networks can be complex and requires significant computational power.
- **Black Box Effect:** It can be difficult to understand why a neural network makes a particular classification decision.
- **Data Dependency:** Performance heavily relies on the quality and quantity of training data.

Overall, neural network-based classification is a powerful tool for various classification tasks. However, it's important to consider its complexity and data requirements before applying it to your specific problem.

### 4. EXPLAIN REGRESSION BY DIVISION METHOD. LET A TRAINING DATA IS {1.6, 1.9, 1.88, 1.7, 1.85, 1.6, 1.7, 1.8, 1.95, 1.9, 1.8, 1.75}. WITH DIVISION METHOD CLASSIFY THE DATA INTO TWO DIFFERENT CLASSES. (2017, 2018, 2019)

Regression by division, also known as the ***mean splitting method,*** is a simple yet efficient technique for ***binary classification.*** It ***utilizes the mean of the data to create a decision boundary,*** separating the data points into two distinct classes.

**Here's how it works:**

1. **Sort the data:** Arrange the data points in either ***ascending or descending order.***
2. **Calculate the mean:** Find the average value of the sorted data.
3. **Create classes:**
   - **Class 1:** This class encompasses all data points ***less than or equal to the mean.***
   - **Class 2:** This class encompasses all data points greater than the mean.

Essentially, the ***mean acts as a threshold,*** dividing the data into two halves based on their relative values compared to the average.

**Example:**

Let's consider the following training data:

- **data =** [1.6, 1.9, 1.88, 1.7, 1.85, 1.6, 1.7, 1.8, 1.95, 1.9, 1.8, 1.75]

**Sorting the data:**

- **Sorted data =** [1.6, 1.6, 1.7, 1.7, 1.75, 1.8, 1.8, 1.85, 1.88, 1.9, 1.9, 1.95]

**Calculating the mean:**

- **mean =** ( 1.6 + 1.6 + 1.7 + 1.7 + 1.75 + 1.8 + 1.8 + 1.85 + 1.88 + 1.9 + 1.9 + 1.95 ) / 12

  **mean =** 1.78583333

**Dividing into classes:**

- **Class 1:** [1.6, 1.6, 1.7, 1.7, 1.75] ***(data points less than or equal to the mean)***

- **Class 2:** [1.8, 1.8, 1.85, 1.88, 1.9, 1.9, 1.95] ***(data points greater than the mean)***

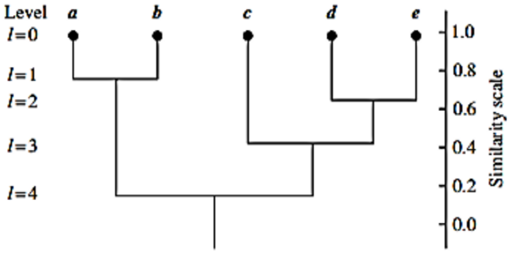
Therefore, the regression by division method classifies the data into two classes based on the ***threshold of 1.785833,*** with data points below the threshold belonging to class 1 and those above belonging to class 2.

**It's important to note that:**

- This is a simple method and may not always be the most effective for complex datasets.
- The ***choice of the threshold value can significantly impact the classification results.***
- Regression by division is often used as a ***preliminary step for more sophisticated classification algorithms.***

## 3. UNIT III

### 1. EXPLAIN DENDROGRAMS. EXPLAIN HIERARCHICAL BASED METHOD OF CLUSTERING. (2018)

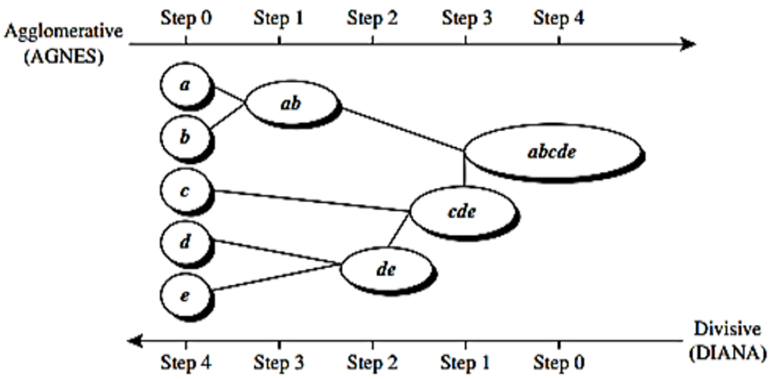


Dendrogram representation for hierarchical clustering of data objects $\{a, b, c, d, e\}$.

**Dendrograms**: A dendrogram is a ***tree-like diagram*** that represents the ***hierarchical relationships among data points or clusters*** in a ***hierarchical clustering analysis.*** It is a ***visual representation of the clustering process,*** showing how clusters are formed or divided at each step.

In a dendrogram, each ***leaf node represents an individual data point,*** and the ***branches connect the data points or clusters*** based on their ***similarity or dissimilarity.*** The ***height of the branches or the vertical axis typically represents the distance or dissimilarity*** between the clusters being merged or split.

Dendrograms are useful for understanding the ***structure and hierarchy of the clusters,*** as well as for determining the ***appropriate number of clusters*** by selecting a suitable level of the dendrogram to ***cut or truncate.***



**Hierarchical Clustering**: Hierarchical clustering is a type of ***unsupervised learning technique*** that ***builds a hierarchy of clusters*** by either ***merging or dividing data points*** based on their similarity or dissimilarity. There are two main approaches to hierarchical clustering:

1. **Agglomerative (Bottom-up) Approach**:
   - The ***agglomerative approach starts*** with ***each data point as an individual cluster.***
   - It then iteratively ***merges the closest or most similar pairs of clusters*** based on a ***similarity measure.***
   - This ***process continues until all data points are merged into a single cluster,*** forming a hierarchical structure.
   - Common ***linkage methods*** used in agglomerative clustering include ***single-linkage, complete-linkage,*** and ***average-linkage.***

The agglomerative approach is ***suitable when the number of data points is relatively small,*** as it requires ***calculating and storing the proximity matrix,*** which can be computationally expensive for large datasets.

2. **Divisive (Top-down) Approach**:
   - The ***divisive approach starts with all data points*** in a ***single cluster.***
   - It then ***iteratively splits or divides the cluster*** into smaller clusters based on a ***dissimilarity measure.***
   - The ***splitting process continues until each data point becomes its own cluster*** or until a stopping criterion is met.
   - Divisive algorithms are ***less commonly used than agglomerative algorithms*** due to their ***computational complexity*** and the ***challenges in determining the optimal splitting criteria.***

Both agglomerative and divisive approaches produce a ***hierarchical representation of the clusters,*** which can be visualized using a dendrogram. The choice between the two approaches depends on the ***size of the dataset, the desired level of granularity,*** and the ***computational resources available.***

## 2. EXPLAIN DIVISIVE APPROACH OF HIERARCHICAL METHOD OF CLASSIFICATION. (2014, 2017)

Hierarchical clustering is a way of grouping data points based on their similarity. There are two main methods: agglomerative and divisive. In this answer, we will focus on the divisive approach, which is also called top-down hierarchical clustering.

**The divisive approach works by following these steps:**

- **Step 1:** Start with one cluster that contains all the data points.
- **Step 2:** Find the most dissimilar sub-cluster within the current cluster and split it into a new cluster.
- **Step 3:** Repeat step 2 until you reach a stopping criterion, such as a desired number of clusters, a minimum similarity level, or a maximum dissimilarity level.

**Things to remember about the divisive approach:**

- It starts with one cluster and splits it into smaller ones, unlike the agglomerative approach, which starts with many clusters and merges them into larger ones.
- It splits based on dissimilarity, which means it tries to find the most different sub-clusters within each cluster.

- It is a top-down approach, which means it works from the general to the specific.
- It can be useful for finding natural groupings in the data, especially when you don't know how many clusters there are.

**Drawbacks of the divisive approach:**

- It is ***less intuitive than the agglomerative approach,*** which is easier to visualize and understand.
- It can be potentially suboptimal, because the initial split can affect the final clusters, and there is no way to go back or change it.
- It can be ***computationally expensive,*** because it requires a lot of calculations to measure the dissimilarity between data points.

The divisive approach is one of the two main methods in hierarchical clustering. It offers a unique perspective for grouping data points based on their dissimilarity. However, it also has some limitations that need to be considered. It is important to compare it to the agglomerative approach and choose the best method for your data and analysis goals.

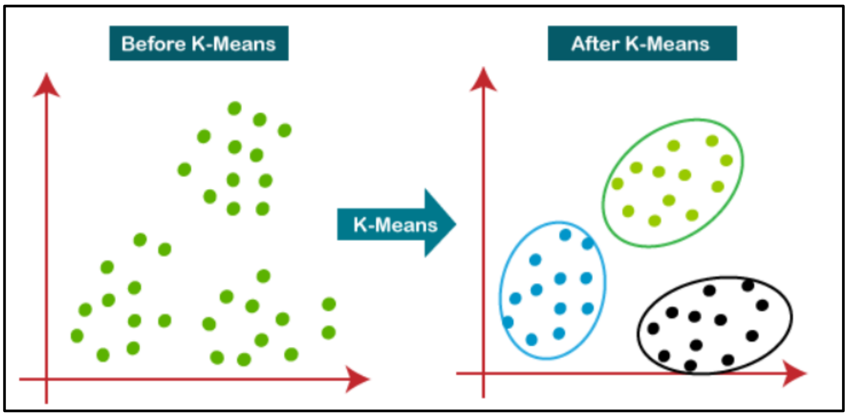
## 3. ILLUSTRATE K-MEANS CLUSTERING WITH AN EXAMPLE. (2014, 2017, 2019)

**What is clustering:**
Clustering is a way of organizing data into groups, or "clusters," based on similarities between data points. It's like sorting a bag of marbles into different piles based on their colours or sizes. In clustering:

1. Data points that are similar to each other are grouped into the same cluster.
2. Clustering ***helps find patterns and structures in data,*** making it easier to understand and analyse large datasets.
3. It's an ***unsupervised learning technique,*** meaning the algorithm learns from the data itself without any predefined labels or categories.
4. Common clustering algorithms include ***K-means, hierarchical clustering, and DBSCAN.*** Each algorithm has its own way of defining similarity and grouping data points into clusters.

**K-Means Clustering:**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here ***K defines the number of pre-defined clusters that need to be created*** in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



**Suppose we have a dataset containing the following points in a two-dimensional space:**

(2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4), (1, 2), (4, 9)

Our goal is to cluster these points into k clusters using the k-means algorithm.

**Step 1: Initialization**

First, we randomly select k points from the dataset as the initial centroids. Let's say we choose k=3 and select the following initial centroids:

Centroid 1: (2,10)

Centroid 2: (5,8)

Centroid 3: (1,2)

**Step 2: Assignment**

Next, we assign each point to the nearest centroid. We calculate the Euclidean distance between each point and each centroid and assign each point to the centroid with the smallest distance.

After the assignment, our clusters might look like this:

Cluster 1: (2,10), (2,5), (1,2)
Cluster 2: (8,4), (7,5), (6,4)
Cluster 3: (5,8), (4,9)

**Step 3: Update**

Now, we recalculate the centroids based on the mean of all points assigned to each cluster.

New Centroid 1: (2+2+1 / 3, 10+5+2 / 3) = (1.67, 5.67)

New Centroid 2: (8+7+6 / 3, 4+5+4 / 3) = (7, 4.33)

New Centroid 3: (5+4 / 2, 8+9 / 2) = (4.5,8.5)

**Step 4: Repeat**

We repeat the assignment and update steps until the centroids no longer change significantly or until a maximum number of iterations is reached.

## 4. WHAT ARE OUTLIERS? ILLUSTRATE OUTLIER CLUSTERING PROBLEM. HOW ARE THEY EFFECTIVE IN CLUSTERING? (2012, 2014)

1. **Outliers:** Outliers are ***data points that significantly deviate from the rest of the dataset.*** They are observations that lie far away from other observations in a dataset. Outliers can occur due to various reasons such as ***measurement errors, experimental errors,*** or they may ***indicate important but rare events*** or ***phenomena in the data.***

2. **Illustration of Outlier Clustering Problem**: Consider a dataset containing information about the salaries of employees in a company. Most of the salaries fall within a certain range corresponding to the salary structure of the company. However, there might be one or a few data points representing salaries that are exceptionally high or low compared to the rest of the salaries. These extreme values could be outliers.

   In a clustering problem, outliers pose a challenge because they can ***distort the natural grouping or clustering of data points.*** For instance, if we're trying to cluster employees based on their salaries into different income brackets, the presence of outliers might result in clusters that are not truly representative of the underlying patterns in the data.

3. **Effectiveness of Outliers in Clustering:** Outliers can actually be effective in clustering in certain scenarios:
   - **Identifying Anomalies:** ***Outliers can represent rare or anomalous patterns*** in the data ***that might be of interest.*** By identifying and clustering outliers separately, ***we can gain insights into these unusual phenomena.***
   - **Improved Cluster Separation:** In some cases, outliers may help in better separating clusters by ***acting as natural boundaries*** between different groups of data points.
   - **Robustness Testing:** Outliers can be ***used to test the robustness of clustering algorithms.*** By ***intentionally introducing outliers into the dataset,*** we can assess how well a clustering algorithm performs.

In conclusion, ***while outliers present challenges*** in clustering by p**otentially distorting the natural grouping of data points,** they can als***o offer valuable insights and enhance the effectiveness of clustering algorithms*** in certain scenarios. Proper handling and interpretation of outliers are essential for accurate and meaningful clustering results.

## 5. EXPLAIN PRIORITY-BASED METHOD OF CLUSTERING. (2017)

Priority-based clustering involves ***organizing data points into clusters*** based on ***predefined priorities or criteria.*** In the context of coronavirus symptoms, we can use priority-based clustering to group patients based on the ***severity of their symptoms.***

**Here's a straightforward example:**

Imagine we have a dataset of patients who have reported symptoms related to ***COVID-19.*** Each patient is described by various symptoms such as ***fever, cough, shortness of breath, fatigue,*** and ***loss of taste or smell.***

1. **Priority Criteria**:
   - We define the priority criteria based on the severity of symptoms. For example:
     - **High Priority:** Fever, shortness of breath
     - **Medium Priority:** Cough, fatigue
     - **Low Priority:** Loss of taste or smell
2. **Clustering Process**:
   - We start by assigning each patient to a cluster based on their highest priority symptom.
   - ***Patients with fever or shortness of breath*** are grouped into the ***high priority cluster.***
   - ***Patients with cough or fatigue*** but without fever or shortness of breath are grouped into the ***medium priority cluster.***
   - ***Patients with only loss of taste or smell,*** without other symptoms, are grouped into the ***low priority cluster.***

3. **Example**:
    - Patient A reports fever and cough. They would be assigned to the high priority cluster.
    - Patient B reports only fatigue. They would be assigned to the medium priority cluster.
    - Patient C reports loss of taste or smell. They would be assigned to the low priority cluster.
4. **Result**:
    - After clustering, we have separate groups of patients ***based on the severity of their symptoms.***
    - This ***allows healthcare professionals to prioritize resources and interventions*** based on the severity of symptoms. For instance, ***patients in the high priority cluster may require immediate medical attention,*** while those in the low priority cluster may need less urgent care.

Priority-based clustering ***simplifies the process of organizing and prioritizing patients*** based on the severity of their symptoms, which is ***crucial in managing healthcare resources efficiently*** during a ***pandemic like COVID-19.***

## 4. UNIT IV

### 1. WHAT IS THE ASSOCIATION RULE? EXPLAIN WITH ITS APPLICATION. (2012, 2017)

An association rule is a form of ***pattern discovery and data mining technique*** used to ***identify relationships or associations between different items or attributes in a dataset.*** It is represented as an ***implication*** of the form ***X ⇒ Y,*** where X and Y are disjoint itemsets (sets of items).

The association rule X ⇒ Y can be interpreted as ***"if X occurs, then Y is likely to occur as well."*** In other words, it suggests that the ***presence of itemset X in a transaction implies a high probability of finding itemset Y in the same transaction.***

Association rules are typically **characterized by two measures:**

1. **Support**: The support of an association rule X ⇒ Y is the ***proportion of transactions in the dataset that contain both X and Y.*** It represents the ***frequency or prevalence*** of the rule in the data. ***Support(X ⇒ Y) = P(X ∪ Y)***
2. **Confidence**: The confidence of an association rule X ⇒ Y is the ***conditional probability of finding Y in a transaction, given that X is present.*** It ***measures the reliability or strength of the rule. Confidence(X ⇒ Y) = P(Y | X) = Support(X ∪ Y) / Support(X)***

Association rules are considered interesting and useful ***if they satisfy user-defined minimum support and confidence thresholds.***

**Applications of Association Rule Mining:**

Association rule mining has numerous applications across various domains, including:

1. **Market Basket Analysis**: ***Identifying relationships*** between ***products purchased together*** by customers in retail stores. This information can be used for ***product placement, cross-selling, promotions, and marketing strategies***.
2. **Web Usage Mining**: ***Analysing patterns in web browsing data*** to understand user behaviour, ***optimize website design,*** and provide ***personalized recommendations.***

3. **Financial Analysis**: Identifying relationships between ***financial indicators, market trends, and investment patterns*** to support investment decisions and risk management.
4. **Fraud Detection**: ***Detecting patterns of fraudulent behaviour in transactions,*** such as ***credit card fraud or insurance claims,*** by identifying associations between various attributes.
5. **Telecommunication Analysis**: Analysing calling patterns, service usage, and customer behaviour to improve network optimization, customer segmentation, and churn prediction.

Association rule mining provides valuable insights from ***patterns and relationships*** present in large datasets, ***enabling organizations to make data-driven decisions, optimize processes, and improve customer experiences.*** However, it is essential to carefully interpret and validate the discovered rules, as they may not necessarily imply causality or have practical significance in all cases.

## 2. DESCRIBE THE CONCEPT OF FREQUENT ITEMS AND ASSOCIATION RULE.

**Frequent Items:**

- Frequent items ***refer to individual elements or items that appear frequently*** in a dataset, such as ***products in a retail store*** or ***words in a document.***
- In the context of association rule mining, the ***Apriori algorithm is commonly used to identify frequent items.*** This algorithm works by ***iteratively discovering item sets*** (combinations of items) that ***meet a minimum support threshold.***
- The support of an itemset is the proportion of transactions in which the itemset appears. ***Items that have a support value above the minimum threshold*** are considered frequent items.

**Association Rule:** refer to the same answer of question no 01.

## 3. EXPLAIN APRIORI BASED ALGORITHM FOR FINDING ASSOCIATION RULE. (2013, 2018, 2019, 2019)

The Apriori algorithm is a classic data mining technique ***used to discover frequent itemsets and association rules in transactional datasets.*** It is ***widely employed in market basket analysis,*** where the goal is to identify patterns of co-occurring items in transactions. Here's an explanation of how the Apriori algorithm works:

1. **Support Measure**:
    - The ***Apriori algorithm uses a support measure to identify frequent itemsets.*** Support measures the ***frequency of occurrence of an itemset in the dataset.*** For example, if an ***itemset {A, B} occurs in 100 out of 1000 transactions,*** its support is **100/1000 = 0.1.**
2. **Apriori Principle**:
    - The key insight behind the Apriori algorithm is the Apriori principle, which states that ***if an itemset is frequent, then all of its subsets must also be frequent.*** This ***principle allows the algorithm to prune the search space*** and ***avoid considering candidate itemsets that cannot possibly be frequent.***
3. **Algorithm Steps**:

    a. **Generate Candidate Itemsets**:

    - Start by identifying all individual items (1-itemsets) and their support.
    - Then, generate candidate 2-itemsets by joining pairs of frequent 1-itemsets.

- Next, generate candidate 3-itemsets by joining pairs of frequent 2-itemsets, and so on, ***until no new candidate itemsets can be formed.***

b. **Prune Infrequent Itemsets**:

- After generating candidate itemsets, calculate their support and ***discard those that do not meet a minimum support threshold.*** This step ***eliminates infrequent itemsets and reduces the search space.***

c. **Generate Association Rules**:

- Once frequent itemsets are identified, association rules are generated from them.
- Association rules have the form ***{A} => {B},*** where A and B are itemsets, and the rule indicates that ***if A occurs, then B is likely to occur as well.***
- ***For each frequent itemset, generate all possible non-empty subsets*** and ***calculate the confidence of association rules*** formed by these subsets.
- Confidence ***measures the likelihood of occurrence*** of the ***consequent (B) given the antecedent (A).*** It is calculated as ***support({A, B}) / support({A}).***

d. **Filter Rules by Confidence and Support**:

- Filter the generated association rules based on ***user-defined thresholds*** for ***confidence and support.***
- ***Association rules that meet the specified thresholds*** are considered ***interesting and retained for further analysis.***

4. **Iterative Process**:

   - The Apriori algorithm iterates through the steps described above ***until no new frequent itemsets can be discovered*** or ***until a specified maximum itemset size*** is reached.

In summary, the Apriori algorithm efficiently discovers frequent itemsets and association rules by leveraging the Apriori principle to ***prune the search space and focus on promising candidates.*** It is a ***fundamental technique in data mining for uncovering meaningful patterns and insights in transactional datasets.***

**3. EXPLAIN FOLLOWING TERMS W.R.T. ASSOCIATION RULE MINING: (2019)**

**A. ASSOCIATION RULE, B. SUPPORT, C. CONFIDENCE, D. FREQUENT ITEM SET, E. LARGE ITEM SET.**

**a. Association rule:**

- An association rule is a relationship or pattern discovered in a dataset that describes the correlation between items. It typically takes the form of "if {X} then {Y}", where X and Y are sets of items.
- These rules are used to uncover interesting relationships between different items in a dataset, such as "customers who buy product X are also likely to buy product Y."

**b. Support:**

- Support is a measure used to evaluate the frequency of occurrence of an itemset in a dataset.
- It indicates how often a particular itemset appears in the dataset relative to the total number of transactions.

- High support values suggest that the itemset occurs frequently, making it more significant in association rule mining.

**c. Confidence:**

- Confidence is a measure used to assess the reliability or strength of an association rule.
- It indicates the likelihood that the consequent (Y) of a rule will be present in a transaction given that the antecedent (X) is present.
- Confidence is calculated as the ratio of the support of the itemset containing both X and Y to the support of the itemset containing only X.
- Higher confidence values indicate a stronger association between the antecedent and consequent of the rule.

**d. Frequent itemset:**

- A frequent itemset is a set of items that occurs frequently in a dataset, with a support value greater than or equal to a predefined minimum support threshold.
- In association rule mining, frequent item sets are essential as they form the basis for generating meaningful association rules.
- Identifying frequent item sets helps to focus on the most relevant patterns in the dataset.

**e. Large itemset:**

- A large itemset, also known as a frequent itemset, is a set of items that meets the minimum support threshold in a dataset.
- Large item sets are identified during the process of frequent itemset mining and are used as the building blocks for generating association rules.
- These item sets represent patterns or combinations of items that occur frequently enough to be considered interesting or significant in the dataset.

## 5. UNIT V

### 1. DIFFERENTIATE IN BETWEEN OLTP AND OLAP. EXPLAIN SCHEMA INTEGRATION AND DATA CLEANING WITH DATA WAREHOUSE. (2019)

**OLTP (Online Transaction Processing):** A ***system that manages a high volume of short, concurrent transactions*** typically encountered in ***operational databases.*** These transactions involve tasks like ***updating product inventory, processing customer orders,*** or ***managing financial accounts.***

**OLAP (Online Analytical Processing):** A ***system that supports complex data analysis and retrieval for business intelligence purposes.*** It ***focuses on analysing large datasets from various sources*** to ***identify trends, patterns, and insights*** for informed decision making.

**Differentiation between OLTP and OLAP**

The following table summarizes the key differences between OLTP and OLAP systems:

| Feature | OLTP | OLAP |
|---|---|---|

| | | |
|---|---|---|
| **Focus** | Transaction processing (***CRUD operations***) | ***Data analysis and reporting*** |
| **Data Volume** | small, ***current operational data*** | Large, ***historical data*** from multiple sources |
| **Data Update Frequency** | High | Low (***periodic updates***) |
| **Transactions** | Short, **frequent,** concurrent | Complex, ***infrequent,*** user-driven |
| **Queries** | Simple, ***focused on specific data retrieval*** | Complex, involving ***aggregations and calculations*** |
| **Response Time** | ***Critical,*** requires fast response times | Less critical, but ***accuracy*** is essential |
| **Schema** | ***Normalized*** for ***efficient updates*** | ***Denormalized*** for ***faster analysis*** |
| **Example Applications** | Point-of-sale systems, ***Banking transactions*** | ***Market analysis, Sales forecasting, Customer segmentation*** |

**Schema Integration and Data Cleaning in Data Warehouses**

A ***Data Warehouse serves as a central repository for historical data*** from various operational systems. Here's how schema integration and data cleaning play crucial roles:

- **Schema Integration:** Combines data from different sources that may have different structures and formats. This involves ***resolving naming conflicts, defining common data types, and ensuring data consistency*** across the warehouse.
- **Data Cleaning:** Identifies and ***corrects errors and inconsistencies within the data.*** This includes handling ***missing values, outliers, duplicate entries,*** and ***formatting*** inconsistencies. Cleaning ensures the ***accuracy and reliability of data for analysis*** in the Data Warehouse.

Effective schema integration and data cleaning are essential for building a reliable and trustworthy Data Warehouse that supports insightful business analysis.

## 2. LIST AND ELABORATE OLAP OPERATIONS WITH AN EXAMPLE. (2013, 2017, 2018, 2019)

**OLAP Operations:**

**Roll-up:**

- **Definition:** Roll-up, also known as drill-up, involves aggregating data along a concept hierarchy by ***climbing up to higher levels of summarization.***
- **Example:** Consider a data cube containing sales data aggregated by city. Performing a roll-up operation on the "location" dimension by climbing up the hierarchy from city to country would aggregate the data by country instead.

**Drill-down:**

- **Definition:** Drill-down ***navigates from less detailed data to more detailed data by stepping down*** a concept hierarchy or introducing additional dimensions.

- **Example:** In a data cube summarized by quarters, ***drilling down on the "time" dimension from quarter to month*** would provide more detailed sales data broken down by month.

**Slice and dice:**

- **Slice:** Selects a sub cube by performing a ***selection on one dimension.***
- **Dice:** Defines a sub cube by performing a ***selection on two or more dimensions.***
- **Example:** Slicing the data cube by ***selecting only the sales data for the first quarter*** (***time = "Q1"***). Dicing the cube by ***selecting sales data for specific combinations of dimensions,*** such as (***location = "Toronto" or "Vancouver"***) and (***time = "Q1" or "Q2"***) and (***item = "home entertainment" or "computer"***).

**Pivot (Rotate):**

- **Definition:** Pivot (or rotate) operation ***rearranges the axes of a cube*** to provide an ***alternative data presentation.***
- **Example:** ***Rotating the axes of a 2-D slice of the cube,*** such as swapping the item and location axes to view the data from a different perspective.

**Other OLAP Operations:**

- **Drill-across:** Executes queries ***involving more than one fact table.***
- **Drill-through:** Uses relational SQL facilities to drill down to the underlying relational tables.

OLAP systems may offer operations for ***ranking top or bottom items, computing moving averages, growth rates, interests, internal return rates, depreciation, currency conversions,*** and ***statistical functions.***

## 3. EXPLAIN THE CONCEPT OF DATA WAREHOUSE. EXPLAIN DATA WAREHOUSE MODELLING AS OLAP. (2014)

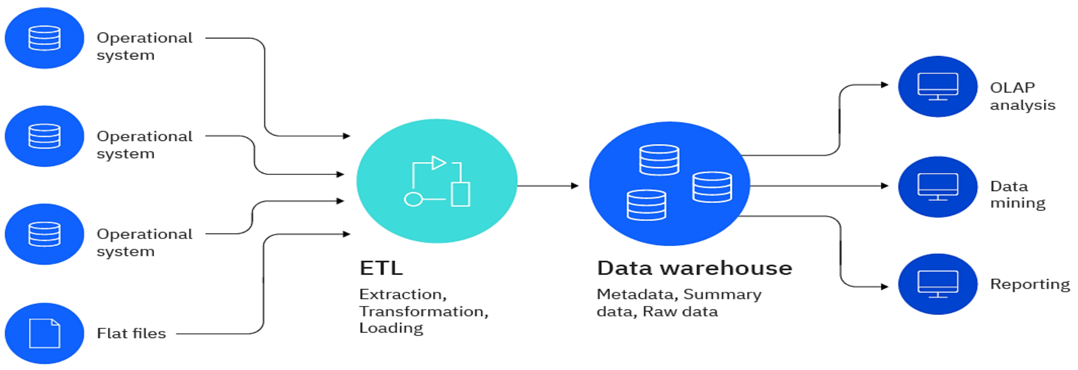
A data warehouse, or enterprise data warehouse (***EDW***), is a system that ***aggregates data from different sources into a single, central, consistent data store*** to support ***data analysis, data mining,*** artificial intelligence (AI) and machine learning. Key features of a data warehouse include:

- **Integration:** Data from disparate sources such as ***operational databases, spreadsheets, and external systems*** are ***consolidated and transformed*** to ensure consistency and uniformity.
- **Subject-oriented:** Data in a ***warehouse is organized around key subject areas relevant to the organization's business processes,*** such as ***sales, finance***.
- **Time-variant:** Data in a warehouse includes historical records, ***allowing analysts to analyse trends and patterns over time.***
- **Non-volatile:** Once data is loaded into the warehouse, it is typically ***not updated or deleted.*** Instead, ***changes are tracked*** through incremental updates.
- **Designed for analysis:** Data in a warehouse is ***structured and optimized*** for ***querying and analysis,*** enabling ***complex analytical queries*** and ***reporting.***

**Data warehouse architecture**

Generally speaking, ***data warehouses have a three-tier architecture,*** which consists of a:

**Bottom tier**: The bottom tier consists of a ***data warehouse server,*** usually a ***relational database system,*** which ***collects, cleanses, and transforms data*** from ***multiple data sources*** through a process known as Extract, Transform, and Load (ETL). For most organizations that use ETL, the ***process relies on automation,*** and is ***efficient, well-defined,*** and ***batch-driven.***



**Architecture(standardized by IBM)**

- **Middle tier**: The middle tier consists of an OLAP (online analytical processing) server which ***enables fast query speeds.*** Three types of OLAP models can be used in this tier, which are known as ***ROLAP, MOLAP and HOLAP.*** The type of OLAP model used is ***dependent on the type of database system that exists.***

- **Top tier**: The top tier is represented by some kind of ***front-end user interface or reporting tool,*** which enables end users to conduct data analysis on their business data.

**Data Warehouse Modelling as OLAP:**

Online Analytical Processing (OLAP) involves ***modelling data in a way that facilitates multidimensional analysis and reporting.*** Data warehouse modelling, therefore, aligns closely with OLAP principles to support efficient querying and analysis of data. Common data warehouse modelling techniques include:

**Dimensional Modelling:**

In dimensional modelling, ***data is organized into two types of tables:***

**fact tables and dimension tables.**

- Fact tables ***contain quantitative data (facts)*** that are ***typically numeric and represent business metrics,*** such as ***sales revenue or quantity sold.***
- Dimension tables ***contain descriptive attributes that provide context for the facts***, such as ***time, product, location.***

**Star Schema and Snowflake Schema:**

- ***Star schema is a simple dimensional model*** where a ***central fact table is surrounded by dimension tables, forming a star-like structure.***
- ***Snowflake schema extends the star schema by normalizing dimension tables,*** resulting in a ***more complex but normalized structure.***

***Both schemas are widely used in data warehousing to organize data for OLAP analysis,*** with ***star schema being more commonly used due to its simplicity*** and ***query performance*** benefits.

Overall, ***data warehouse modelling focuses on creating a structure that optimizes analytical querying and reporting,*** enabling users to explore and analyse data across multiple dimensions efficiently.

## 4. ILLUSTRATE WITH SUITABLE EXAMPLE, STAR SCHEMA FOR MULTIDIMENSIONAL DATA MODEL/ ILLUSTRATE WITH SUITABLE EXAMPLE MULTIDIMENSIONAL DATA MODEL. (2012)

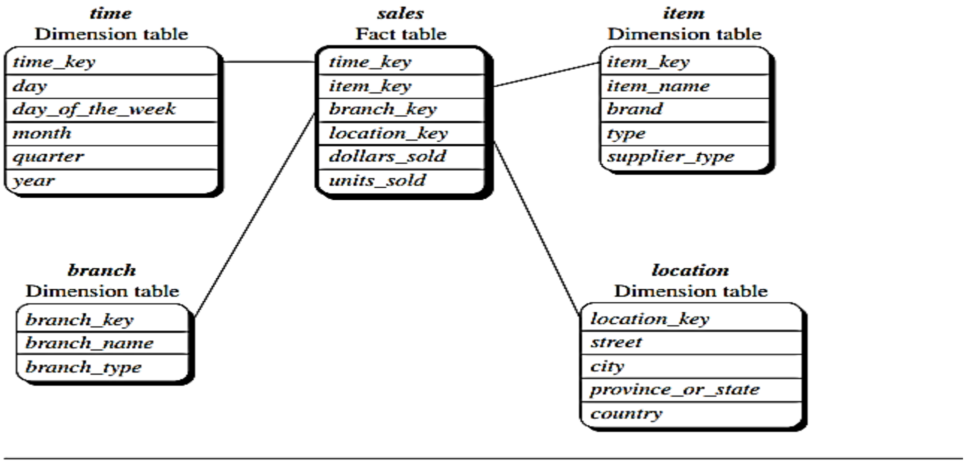


**Figure 4.6** Star schema of *sales* data warehouse.

**Example of Star Schema for All Electronics Sales:**

In the All-Electronics company's data warehouse, sales data is organized using a star schema. Let's illustrate this schema with an example:

**Fact Table: Sales**

- The central fact table, named "Sales," contains the bulk of the data related to sales transactions.
- It ***includes keys to each of the four dimensions***: time key, item key, branch key, and location key.
- Additionally, it contains measures such as ***dollars sold*** and ***units sold.***

**Dimension Tables:**

**Time Dimension:**

- Contains attributes related to time, such as year, quarter, month, and day.
- **Example attributes:** time key, year, quarter, month, day.

**Item Dimension:**

- Contains ***attributes related to the items sold, such as item ID, name, category, and brand.***
- **Example attributes:** item key, item ID, name, category, brand.

**Branch Dimension:**

- Contains ***attributes related to branches or stores where the sales occurred,*** such as branch ID, name, and address.
- **Example attributes:** branch key, branch ID, name, address.

**Location Dimension:**

- Contains attributes related to geographical locations, such as street, city, province/state, and country.
- **Example attributes:** location key, street, city, state, country.

In this example, ***each row in the fact table represents a sales transaction,*** with keys to the respective dimensions and measures of ***dollars sold and units sold.***

## 5. EXPLAIN THE CONCEPT OF DATA WAREHOUSE. EXPLAIN DATA WAREHOUSE MODELLING AS DATA CUBE. (2013, 2018)

**What is data warehouse?** -> Refer to question no 03 of this chapter.

**Data Warehouse Modelling as Data Cube:**

Data warehouse modelling as a data cube is a ***multidimensional approach*** to ***organizing and structuring data in a way that facilitates efficient analysis and reporting from multiple perspectives.*** The data ***cube model is an extension of the dimensional modelling technique,*** which is commonly used in data warehouses for OLAP (Online Analytical Processing) applications.

In a data cube model, ***data is organized into three main components***:

1. **Dimensions**: Dimensions represent the ***different perspectives or entities*** by which the data can be analysed. Examples of dimensions include ***time, product, and location.*** Each ***dimension consists of hierarchical levels,*** such as ***year, quarter, month, and day for the time dimension.***
2. **Measures**: Measures are the ***numerical values or facts*** that are being ***analysed and aggregated.*** Examples of measures include ***quantity sold and sales revenue.*** These ***measures are typically stored in a fact table.***
3. **Fact Table**: The fact table is the central table in the data cube model, which ***stores the measures*** and ***foreign keys linking to the dimension tables.***

The data cube model represents the multidimensional structure by combining the dimensions and measures, creating a cube-like structure. ***Each cell in the data cube contains the aggregate values of the measures for the specific combination of dimension values.***

**Example:** Explain it with an ***"All-Electronic Sells"*** example given in syllabus

## 6. DISCUSS DIFFERENT SCHEMAS FOR DATA WAREHOUSING. (2012)

In data warehousing, ***schemas define the logical structure and organization of data within the data warehouse.*** Different schemas are ***used to model relationships between tables*** and optimize ***data retrieval and analysis.*** Here are some common schemas used in data warehousing:

**Star Schema:**

- Star schema is the ***most common modelling paradigm*** in data warehousing.
- It consists of a ***central fact table*** surrounded by ***dimension tables,*** forming a ***star-like structure.***
- The fact table contains ***quantitative data (facts)*** such as ***sales revenue or quantity sold.***
- Dimension tables ***contain descriptive attributes that provide context for the facts,*** such as ***time, product, and location***.

**Snowflake Schema:**

- Snowflake schema extends the star schema by ***normalizing dimension tables,*** resulting in a ***more normalized but complex structure.***
- Dimension tables are ***normalized into multiple related tables,*** forming a ***snowflake-like structure.***
- ***Normalization reduces data redundancy and improves data integrity.***

**Fact Constellation Schema:**

- Fact constellation schema, also known as ***galaxy schema,*** consists of ***multiple interconnected star schemas.***
- It ***includes multiple fact tables sharing common dimension tables,*** forming a ***constellation-like structure.***

**Normalized Schema:**

- Normalized schema follows ***traditional database normalization principles*** to minimize data redundancy and improve data integrity.
- It ***decomposes data into multiple related tables,*** each representing a distinct entity or relationship.
- Normalized schemas are ***suitable for transactional databases*** where ***data integrity and update efficiency are paramount.***

**7. DESCRIBE SNOWFLAKE SCHEMA. LET FOR UNIVERSITY DATABASE CONSIST OF FOUR DIMENSIONS AS STUDENT, COURSE, SEMESTER AND INSTRUCTOR. THE MEASURING ATTRIBUTE IS AVERAGE GRADE, WHERE AVERAGE GRADE STORES THE ACTUAL COURSE GRADE OF THE STUDENT. DRAW SNOWFLAKE SCHEMA FOR THE ABOVE-MENTIONED DATA. (2013, 2014, 2019)**

A snowflake schema is an extension of the star schema, which is a ***dimensional modelling approach*** used in data warehousing and OLAP (Online Analytical Processing) systems. In a snowflake schema, some ***dimension tables are further normalized into multiple related tables,*** resulting in a more complex structure with additional dimension tables.

The main difference between a star schema and a snowflake schema is that in a star schema, all the dimension attributes are stored in a single dimension table, while in a snowflake schema, some dimension tables are further normalized into multiple tables, creating a more extended hierarchy.

Here's an illustration of a snowflake schema for the University database you described, with **four dimensions:** ***student, course, semester, and instructor,*** and the measure attribute being ***average grade:***

**In this snowflake schema:**

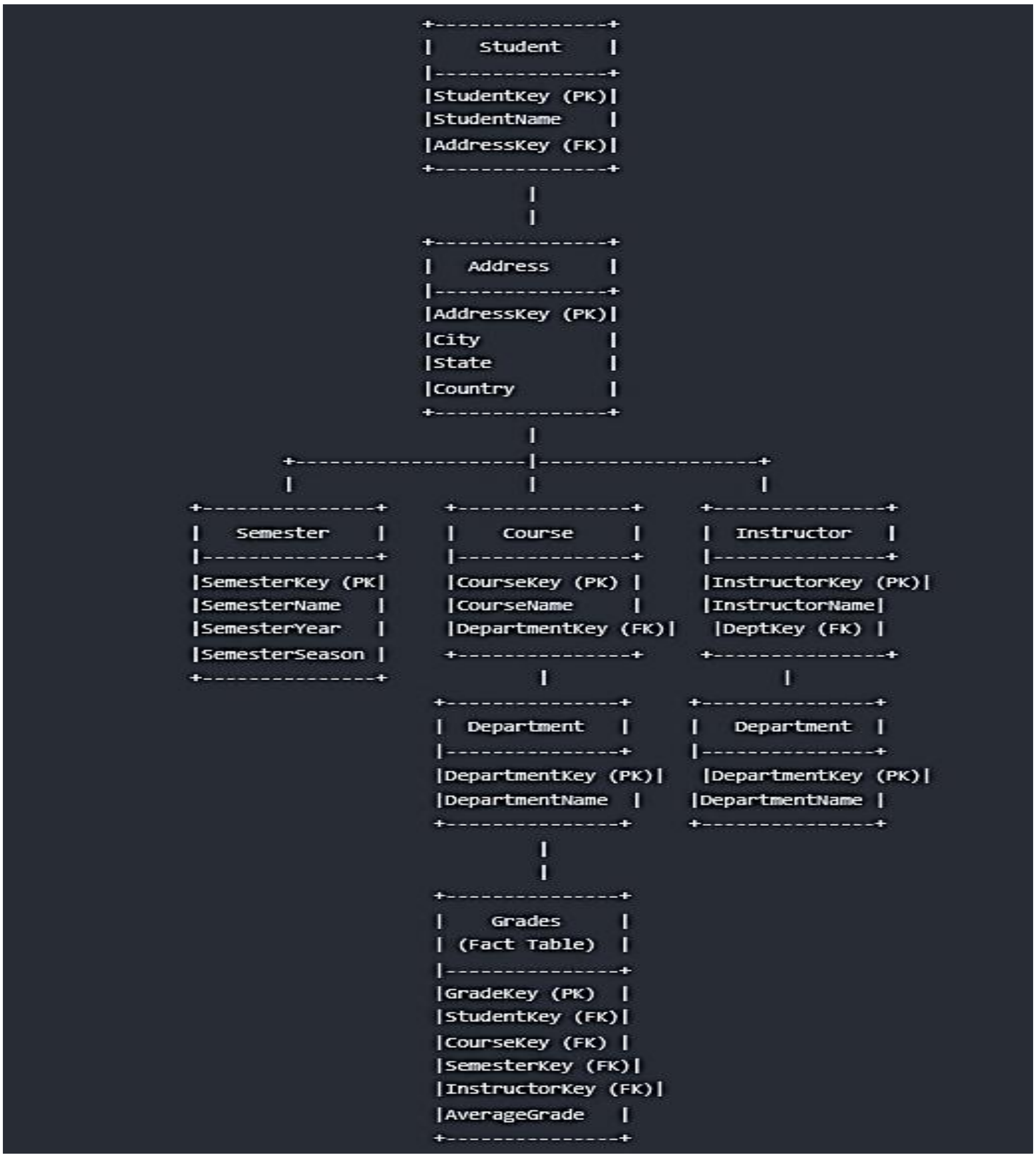
1. The central table is the ***"Grades" fact table,*** which contains the measure attribute ***"Average Grade"*** and ***foreign keys linking to the respective dimension tables.***

2. The ***"Student" dimension table is normalized further into the "Address" table,*** which stores the address details of the student.

3. The ***"Course" dimension table is normalized further into the "Department" table,*** which stores the department information for each course.

4. The ***"Instructor" dimension table is also normalized into the "Department" table,*** as instructors belong to specific departments.

5. The ***"Semester" dimension table remains unnormalized,*** containing all the relevant semester attributes.



```
                    +---------------+
                    |    Student    |
                    |---------------+
                    |StudentKey (PK)|
                    |StudentName    |
                    |AddressKey (FK)|
                    +---------------+
                              |
                              |
                    +---------------+
                    |   Address     |
                    |---------------+
                    |AddressKey (PK)|
                    |City           |
                    |State          |
                    |Country        |
                    +---------------+
                              |
        +---------------------|--------------------+
        |                     |                    |
+---------------+      +---------------+      +---------------+
|   Semester    |      |    Course     |      |  Instructor   |
|---------------+      |---------------+      |---------------+
|SemesterKey (PK|      |CourseKey (PK) |      |InstructorKey (PK)|
|SemesterName   |      |CourseName     |      |InstructorName|
|SemesterYear   |      |DepartmentKey (FK)|    |DeptKey (FK) |
|SemesterSeason |      +---------------+      +---------------+
+---------------+               |                    |
                     +---------------+      +---------------+
                     |  Department   |      |   Department  |
                     |---------------+      |---------------+
                     |DepartmentKey (PK)|    |DepartmentKey (PK)|
                     |DepartmentName  |     |DepartmentName |
                     +---------------+      +---------------+
                               |
                               |
                     +---------------+
                     |    Grades     |
                     | (Fact Table)  |
                     |---------------+
                     |GradeKey (PK)  |
                     |StudentKey (FK)|
                     |CourseKey (FK) |
                     |SemesterKey (FK)|
                     |InstructorKey (FK)|
                     |AverageGrade   |
                     +---------------+
```

The ***snowflake schema provides a more normalized structure compared to the star schema,*** which can help ***reduce redundancy and data integrity.*** However, it may introduce additional joins and complexity when querying the data, potentially impacting query performance.

## 6. Unit VI

### 1. WHAT ARE SOCIAL IMPLICATIONS OF DATA MINING? (2012, 2017, 2018, 2019, 2019)

The social implications of data mining are multifaceted and deserve careful consideration. Here are some key points:

- **Privacy Concerns:**
  Data mining often ***involves the collection and analysis of personal information.*** This ***raises concerns about individuals' privacy rights*** and the ***potential for misuse of their data. Unauthorized access or breaches*** can lead to ***identity theft, discrimination,*** and other forms of harm.
- **Surveillance and Control:**
  The widespread use of data mining can ***contribute to a surveillance culture*** where ***individuals feel constantly monitored and controlled.*** This can ***erode trust in institutions*** and ***foster a climate of suspicion.***

- **Discrimination and Bias:**
  ***Biases in data can lead to biased algorithms.*** Imagine an algorithm trained on data that reflects ***past discrimination*** in ***societal inequalities.*** It might continue those ***unfair practices*** in areas like **hiring or loan approvals.**
- **Loss of Autonomy:**
  As data mining gets better, ***people might lose control over how their data is used.*** This can ***make individuals feel powerless*** and ***increase the gap in power between them and organizations.***
- **Transparency and Accountability:**
  There's a need for greater ***transparency and accountability*** in data mining practices to ***ensure that decision-making processes are fair and understandable.*** Individuals should know about ***how their data is being collected, analysed, and used.***
- **Digital Divide:**
  Data mining technologies may exacerbate existing disparities in access to information and resources. Those who lack access to technology or who are unable to understand its implications may be further marginalized in a data-driven society.

In conclusion, while data mining offers numerous benefits, including insights into human behaviour and societal trends, it also ***poses significant social challenges*** that must be ***addressed through careful regulation, ethical guidelines.***

## 2. DESCRIBE THE CONCEPT OF TEXT MINING IN DETAIL. (2019)

***Text mining is an interdisciplinary field*** that ***harnesses techniques*** from ***data mining, machine learning, statistics,*** and ***computational linguistics*** to extract valuable insights from large volumes of text data. Given that a significant portion of information is stored as text in various forms such as ***news articles, technical papers, emails, blogs, and web pages,*** text mining has become increasingly vital in today's data-driven world.

The ***primary goal of text mining is to derive high-quality information from text.*** This ***involves discovering patterns and trends within textual data*** through methods such as ***statistical pattern learning, topic modelling,*** and ***statistical language modelling.*** To achieve this, text mining typically involves several steps:

- **Text Structuring:** The input text is structured, which may involve ***parsing, removing irrelevant information,*** and then ***inserting the structured data into a database.*** This ***step is crucial for preparing the text for analysis.***
- **Pattern Discovery:** Once the text is structured, ***patterns and relationships are derived from the data using various analytical techniques.*** These patterns may include ***associations between words, sentiment analysis, and entity relationships.***
- **Evaluation and Interpretation:** The output from text mining is evaluated based on criteria such as ***relevance, novelty, and interestingness.*** Interpretation involves ***understanding the discovered patterns and their implications.***

**Some typical tasks in text mining include:**

- **Text Categorization:** ***Assigning predefined categories to text documents*** based on their content.
- **Text Clustering:** ***Grouping similar documents*** together based on their content.
- **Document Summarization:** Generating ***concise summaries of lengthy*** documents.
- **Sentiment Analysis:** ***Determining the sentiment or emotional tone*** expressed in text.

## 3. DISCUSS SPATIAL MINING WITH ITS APPLICATION. (2013, 2017, 2019)

Spatial mining, also known as spatial data mining, is a specialized area of data mining that focuses on ***discovering patterns, trends, and knowledge*** from spatial data. ***Spatial data typically refers to geospatial data, which is data related to the Earth's surface and its features.*** This data can include information such as ***geographic coordinates, land use, elevation,*** and other attributes relevant to a specific location.

Spatial data can be stored in various formats, ***including vector or raster formats,*** as well as ***imagery and geo-referenced multimedia.*** With the advent of ***geographic information systems*** (GIS) and ***spatial databases,*** large repositories of spatial data have been constructed, integrating data from diverse sources such as ***satellite imagery, GPS data, and census information.***

***Spatial mining operates on these spatial data repositories*** to extract valuable insights and knowledge. Some key concepts and applications of spatial mining include:

- **Spatial Data Warehouses:** These are repositories that integrate ***thematic and geographically referenced data*** from multiple sources. They ***enable the construction of spatial data cubes,*** which contain ***spatial dimensions and measures*** for ***multidimensional spatial analysis. Spatial OLAP*** (Online Analytical Processing) techniques can then be applied to analyse spatial data cubes.
- **Spatial Association Mining:** This involves discovering relationships and associations between spatially referenced objects or phenomena. For example, identifying co-occurrences of features within a specific geographic area, such as the association between certain types of businesses and their proximity to transportation hubs.
- **Spatial Clustering:** Spatial clustering aims to group spatially proximate objects or locations based on similarity in their attributes or spatial characteristics. It can help identify spatially coherent regions or patterns within the data, such as identifying hotspots of crime activity or clustering of similar land cover types.
- **Spatial Classification:** In spatial classification, machine learning algorithms are applied to assign spatially referenced objects to predefined categories or classes based on their attributes and spatial relationships. For example, classifying land cover types from satellite imagery or predicting land use changes over time.
- **Spatial Trend and Outlier Analysis:** This involves identifying trends and anomalies in spatial data over time or space. For example, ***detecting areas experiencing rapid urbanization*** or ***identifying outliers in environmental monitoring data.***

Overall, spatial mining plays a crucial role in various domains such as ***urban planning, environmental management, public health, transportation,*** and ***natural resource management,*** enabling informed decision-making and facilitating better understanding of spatially referenced phenomena.

## 4. DISCUSS TEMPORAL MINING WITH ITS APPLICATION. (2012, 2012, 2014, 2018, 2019)

Temporal mining, also known as ***time series data mining or temporal data mining,*** focuses on discovering patterns, trends, and knowledge from ***data that vary over time.*** Temporal mining is a process in data mining that focuses on ***analysing data with respect to time.*** Let me give you a simple example to illustrate temporal mining:

Imagine you own a coffee shop and you have ***sales data for the past year.*** This data includes the ***date and time*** of each sale, what was purchased, and the amount spent. You want to understand ***customer behaviour over time*** to make ***better business decisions.***

***Using temporal mining, you could analyse this data*** to discover patterns such as:

- **Time-based Patterns**: You might find that coffee ***sales increase significantly*** between ***7 AM and 9 AM on weekdays,*** suggesting a rush of customers before work.
- **Seasonal Trends**: There could be an ***uptick in sales of cold beverages*** during the ***summer*** months and ***hot beverages*** in the ***winter.***
- **Event-driven Spikes**: You may notice a spike in sales around ***local events or holidays,*** indicating that these events influence customer purchases.

By ***identifying these temporal patterns,*** you can adjust ***your inventory, staffing,*** and ***marketing strategies*** accordingly. For instance, you might decide to run special promotions during slow periods or stock up on certain items before expected busy times.

Temporal mining helps you ***make sense of time-stamped data to uncover valuable insights*** that are not immediately obvious, enabling you to predict future trends and make informed decisions based on historical patterns.

**Applications:**

1. **Stock Market Analysis**: Temporal mining is used to predict stock prices and market trends based on historical data.
2. **Customer Behaviour Analysis**: It helps businesses understand customer purchase patterns over time and improve marketing strategies.
3. **Healthcare Monitoring**: In healthcare, temporal mining can track patient health trends and predict outcomes.
4. **Social Media Trend Analysis**: It's used to analyse social media data to identify trending topics and user behaviour over time.
5. **Environmental Monitoring**: Temporal mining assists in studying climate patterns and predicting weather-related events.

### 5. WHAT ARE THE APPLICATIONS OF DATA MINING? WRITE A SHORT NOTE ON WEB MINING. (2013, 2019)

Data mining is a powerful tool used across various domains to extract valuable insights from large datasets. Here are some of the key applications of data mining:

1. **Healthcare**: Data mining helps in predicting disease trends, understanding patient behaviour, and improving healthcare services.
2. **Banking and Finance**: It's used for credit scoring, fraud detection, customer segmentation, and risk management.
3. **Retail**: Retailers use data mining for ***market basket analysis, inventory management,*** and customer relationship management.
4. **Telecommunications**: Telecom companies apply data mining for churn analysis, fraud detection, and customer retention strategies.
5. **Education**: Educational institutions use it for student performance analysis and to enhance learning experiences.

**Web Mining**: ***Web mining is the application of data mining techniques*** to ***discover patterns from the World Wide Web.*** It involves ***extracting and analysing information from web resources*** to gain insights into user behaviour, preferences, and web structure. Web mining can be ***categorized into three types:***

- **Web Content Mining**: This focuses on the ***extraction of useful information from web page contents,*** which can include ***text, images, and videos.***
- **Web Structure Mining**: This analyses the structure of the web, ***understanding the link hierarchy*** and the ***relationship between web pages.***
- **Web Usage Mining**: This ***deals with user interaction data,*** such as clickstream data, to understand user behaviour and preferences.

Applications of web mining include ***improving website navigation, targeted marketing, search engine optimization,*** and ***competitive intelligence.*** By analysing web data, businesses can tailor their services to better meet the needs of their customers and improve their online presence.

## 6. DESCRIBE PAGE RANK ALGORITHM WITH ITS APPLICATIONS. (2014, 2019, 2019)

The PageRank algorithm is a foundational algorithm used in web search engines to rank web pages based on their importance and relevance. Originally developed by Larry Page and Sergey Brin, the co-founders of Google, PageRank forms the basis of Google's search engine algorithm and remains a fundamental component of web search technology.

Here's a detailed description of the PageRank algorithm and its applications:

**Algorithm Overview:**

- PageRank operates on the ***principle of link analysis,*** evaluating the importance of a web page based on the quantity and quality of incoming links from other pages.
- The algorithm views the web as a graph, with web pages as nodes and hyperlinks between pages as edges.
- PageRank assigns each web page a numerical score, known as its PageRank score, which represents the probability that a random surfer (user) will arrive at that page by following links from other pages.
- The PageRank score of a page is influenced by the PageRank scores of the pages linking to it, with higher scores from more authoritative pages carrying greater weight.
- PageRank is calculated iteratively through an algorithmic process that distributes PageRank scores across the web graph until convergence is achieved.

**Applications:**

- **Web Search:** The primary application of PageRank is in web search engines, where it forms the basis for ranking search results. Pages with higher PageRank scores are typically considered more relevant and are ranked higher in search engine results pages (SERPs). This helps users discover authoritative and trustworthy content on the web.
- **Link Analysis:** PageRank is used in link analysis tasks to identify important nodes (web pages) in large networks. Beyond web search, PageRank has applications in social network analysis, citation analysis in academic literature, and recommendation systems, where it helps identify influential nodes or entities.
- **Spam Detection:** PageRank can be used as a component in spam detection algorithms to identify and penalize manipulative or low-quality web pages that attempt to artificially inflate their importance through link schemes or other deceptive practices. Pages with suspicious link

patterns or unusually high PageRank scores relative to their content may be flagged as potential spam.

- **Page Importance Metrics:** PageRank serves as a metric for assessing the importance and authority of web pages. Website owners and marketers use PageRank scores as a measure of their site's visibility and influence on the web. Improving PageRank can be a goal of search engine optimization (SEO) efforts, leading to higher rankings and increased organic traffic.

### 7. EXPLAIN IN DETAIL WEB STRUCTURE MINING. (2012, 2014, 2017, 2018)

Web structure mining is a process that utilizes graph and network mining techniques to analyse the nodes and connections within the structure of the World Wide Web. This approach focuses on extracting patterns from hyperlinks, which serve as structural components connecting web pages to each other. Additionally, web structure mining may involve analysing the document structure within individual web pages, such as the hierarchical structure described by HTML or XML tags. By examining these structural elements, web structure mining aims to gain insights into web content and transform it into relatively structured datasets.

**Here's a detailed overview of web structure mining:**

- **Graph Representation:** In web structure mining, the World Wide Web is represented as a graph, where web pages are nodes and hyperlinks between pages are edges. This graph structure captures the interconnected nature of web pages and allows for the application of graph mining techniques to analyse relationships and patterns within the web.
- **Hyperlink Analysis:** Hyperlink analysis is a primary focus of web structure mining. By examining the patterns of hyperlinks between web pages, it is possible to uncover valuable insights about the organization, connectivity, and relevance of web content. Common techniques used in hyperlink analysis include link analysis algorithms like PageRank and HITS (Hyperlink-Induced Topic Search), which assess the importance and authority of web pages based on their link structure.
- **Document Structure Analysis:**
  Web structure mining may also involve analysing the document structure within individual web pages. This includes examining the hierarchical organization of elements within a page, such as HTML or XML tags
- **Pattern Extraction:**
  The primary goal of web structure mining is to extract meaningful patterns and relationships from the structure of the web. This involves identifying recurring patterns in hyperlink connectivity, such as clusters of closely connected pages or hubs of pages with many outgoing links
- **Understanding Web Content:**
  Through the analysis of web structure, web structure mining helps to understand the content and organization of information on the web. By uncovering patterns in hyperlink connectivity and document structure.
- **Data Transformation:**
  Web structure mining can also facilitate the transformation of unstructured web content into relatively structured datasets. By identifying patterns in hyperlink connectivity and document structure, web content can be organized and represented in a more structured format, making