1. How does data cube aggregation help in data reduction?
Data cube aggregation helps in data reduction by summarizing data along dimensions, reducing the complexity of detailed data. By aggregating the data, less storage space is required, and analysis becomes faster, as fewer records need to be processed during queries.

2. How can you classify frequent pattern mining methods?
Frequent pattern mining methods can be classified into three main categories: Apriori-based algorithms (which generate frequent itemsets), Pattern-growth approaches (like FP-Growth), and Vertical data format-based methods. These approaches aim to efficiently discover patterns in large datasets.

3. What is the method to prune a rule set?
Rule pruning can be achieved by removing redundant or less relevant rules. Common methods include minimum support and confidence thresholds, error-based pruning, or techniques like Reduced Error Pruning (REP), which eliminates rules that contribute little to the overall accuracy.

4. Write some applications for clustering.
Clustering is used in various fields like market segmentation (grouping customers by purchasing behavior), image processing (identifying similar regions), document classification, social network analysis (detecting communities), and anomaly detection in network security.

5. Explain how unstructured data is stored.
Unstructured data, such as text, images, and videos, is stored in formats like NoSQL databases, Hadoop Distributed File Systems (HDFS), or in cloud storage systems. These storage solutions are designed to handle non-relational, schema-less data that doesn't fit into traditional databases.

6. Define a data warehouse.
A data warehouse is a centralized repository that stores large volumes of data from multiple sources. It supports business intelligence activities, like reporting and analysis, by enabling efficient querying, data mining, and decision-making processes.

7. Define FP-Tree.
An FP-Tree (Frequent Pattern Tree) is a data structure used in frequent pattern mining, specifically the FP-Growth algorithm. It compresses the dataset by grouping frequent itemsets, allowing the algorithm to mine patterns without generating candidate sets, making it more efficient than Apriori.

8. Compare the different attribute selection measures.
Attribute selection measures include information gain (which selects attributes that reduce entropy the most), gain ratio (normalizes information gain), and the Gini index (used in decision trees to select attributes based on impurity reduction). These measures help in selecting the most relevant attributes for model building.

9. How can we calculate the dissimilarity between objects described by categorical variables?
The dissimilarity between objects with categorical variables can be calculated using the simple matching coefficient or the Jaccard coefficient. These methods compare the categories

of two objects and assign a dissimilarity score based on how many categories differ between the objects.

10. Differentiate between farthest neighbor clustering algorithm and complete linkage algorithm.
The farthest neighbor algorithm (also known as complete linkage) measures cluster similarity based on the maximum distance between points in different clusters. It tends to produce more spherical clusters and avoids merging dissimilar clusters compared to single-linkage clustering (nearest neighbor), which can result in elongated clusters.

11. Mention the disadvantages of K-Means method of clustering.
The disadvantages of the K-Means method include sensitivity to the initial placement of centroids, difficulty in handling clusters of varying sizes and densities, and its assumption that clusters are spherical. It also struggles with noise and outliers, which can skew results.

12. What is a dendrogram? Explain with example.
A dendrogram is a tree-like diagram used to visualize the arrangement of clusters produced by hierarchical clustering algorithms. The vertical axis represents the distance or dissimilarity between clusters. For example, in a dendrogram of customer data, the leaves of the tree represent individual customers, and the branches represent grouped clusters.

- **Visual classification: an interactive approach to decision tree construction. Draw a flowchart to explain in detail all the main steps of visual classification**

Visual classification in decision trees involves a sequence of steps to map features to decision outcomes. The main steps include feature selection, node splitting based on certain criteria (like Gini index or information gain), branching to create child nodes, and recursively repeating the process until a stopping criterion is met. The flowchart would begin with data input, followed by the selection of the most informative feature, and branch creation based on feature values, continuing until the tree is complete.

- **Define classification and prediction**

Classification involves categorizing data into predefined classes, while prediction involves estimating the value of a continuous output variable based on input data. In classification, the outcome is discrete (e.g., yes/no), while in prediction (regression), the outcome is a continuous number (e.g., predicting a price or temperature).

- **Draw the contingency table for binary variables**

A contingency table for binary variables shows the frequency distribution of two binary variables, displaying True/False for one variable along the rows and True/False for the other along the columns. Each cell of the table represents the count of occurrences for each combination of outcomes (True-True, True-False, etc.).

- **List out and briefly explain the different interestingness measures**

Interestingness measures are used to assess the importance or relevance of discovered patterns in data mining. Examples include support (frequency of a pattern), confidence (likelihood of pattern occurrence), lift (the strength of the rule over random chance), and conviction (dependence of rules).

- **What are unstructured data? Give examples**

Unstructured data refers to data that lacks a predefined model or format, making it difficult to process and analyze with traditional tools. Examples include text documents, images, videos, social media posts, and emails, which don’t fit neatly into rows and columns like structured data.

- **Differentiate between dimension table and fact table**

Dimension tables contain descriptive attributes (dimensions) that define business objects (e.g., time, geography, product), while fact tables contain the measurable, quantitative data (facts) related to business transactions. Dimension tables help to add context to the facts stored in the fact tables in a star schema database.

- **How is Minkowski, Manhattan, and Euclidean distance related to each other?**

Minkowski distance is a generalized form of both Manhattan and Euclidean distance. When the order (p) is set to 1, it gives the Manhattan distance (sum of absolute differences between coordinates), and when $p = 2$, it gives the Euclidean distance (the straight-line distance between two points).

- **Discuss various approaches to improve the efficiency of Apriori Algorithm**

To improve the Apriori Algorithm, techniques include reducing the number of candidate itemsets by using advanced pruning techniques, incorporating hash-based methods to reduce the size of the candidate pool, and using partitioning to break the dataset into smaller, more manageable chunks. Transaction reduction and dynamic item counting also help boost efficiency.

- **What are the pre-processing steps to be done on the data before classification and prediction?**

Pre-processing steps include data cleaning (handling missing values and outliers), data normalization or standardization (to ensure features have similar scales), feature selection, and dimensionality reduction. Additionally, encoding categorical variables and splitting the data into training and testing sets are important pre-processing tasks.

- **Differentiate between single linkage clustering algorithm and complete linkage algorithm**

Single linkage clustering considers the minimum distance between clusters when merging, while complete linkage clustering uses the maximum distance between clusters. As a result, single linkage often produces elongated clusters, while complete linkage tends to create compact clusters.

- **What is sampling? Explain sampling with or without replacement**

Sampling is the process of selecting a subset of data from a larger population for analysis. In sampling with replacement, each selected data point is returned to the population, allowing it to be chosen again. In sampling without replacement, data points are not returned, so each data point can only be selected once.

- **How can we calculate the dissimilarity between objects described by categorical variables?**

The dissimilarity between objects described by categorical variables can be calculated using methods like the Hamming distance, which counts the number of positions at which two objects differ. Another approach is using a matching coefficient, which measures the ratio of mismatches to total attributes.