1. What is the difference between supervised and unsupervised learning? Give some examples to illustrate your point.

>>>Supervised Learning: This type of learning involves training a model on a labeled dataset, where the algorithm learns from input-output pairs. The goal is to learn a mapping function that can make predictions on new, unseen data. Examples: Image classification, spam email detection.

Unsupervised Learning: Here, the algorithm works with unlabeled data, aiming to find patterns, structures, or relationships in the data.

2. Mention a few unsupervised learning applications.

>>>>Clustering: Grouping similar data points, like customer segmentation for marketing.

Dimensionality Reduction: Reducing the number of features while retaining important information.

Anomaly Detection: Identifying unusual patterns in data, such as fraud detection.

Topic Modeling: Identifying topics in text data, helpful for document categorization.

3. What are the three main types of clustering methods? Briefly describe the characteristics of each.

>>>Hierarchical Clustering: Builds a hierarchy of clusters, either bottom-up (agglomerative) or top-down (divisive).

Partitioning Clustering: Divides data into non-overlapping clusters, like the k-means algorithm.

Density-Based Clustering: Identifies clusters based on dense regions separated by less dense regions, as seen in DBSCAN.

4. Explain how the k-means algorithm determines the consistency of clustering.

>>>The k-means algorithm aims to minimize the within-cluster sum of squared distances (WCSS), which is the sum of squared distances between data points and their assigned cluster center. Consistency is determined by observing how much the WCSS changes as clusters are formed and recalculated.

5. With a simple illustration, explain the key difference between the k-means and k-medoids algorithms.

>>>In k-means, cluster centers are the mean of data points in the cluster, whereas in k-medoids, cluster centers are actual data points.

6. What is a dendrogram, and how does it work? Explain how to do it.

>>>A dendrogram is a tree-like diagram used in hierarchical clustering to represent the arrangement of the clusters. It shows how data points are merged or divided as the algorithm progresses.

7. What exactly is SSE? What role does it play in the k-means algorithm?

>>>SSE is a measure of the total variance within a cluster. It's calculated by summing the squared distances between each data point and the center of its assigned cluster. In k-means, SSE helps assess the quality of the clustering solution; lower SSE indicates more compact clusters.

8. With a step-by-step algorithm, explain the k-means procedure.

>>>Initialize: Choose k initial cluster centers.

Assign: Assign each data point to the nearest cluster center.

Update: Recalculate the cluster centers as the mean of assigned data points.

Repeat: Iteratively assign and update until convergence (center positions stabilize).

9. In the sense of hierarchical clustering, define the terms single link and complete link.

>>>Single Link: Measures the distance between the closest points of two clusters. It's sensitive to outliers and can result in long, chain-like clusters.

Complete Link: Measures the distance between the farthest points of two clusters. It's less sensitive to outliers but can create tight, spherical clusters.

10. How does the apriori concept aid in the reduction of measurement overhead in a business basket analysis? Give an example to demonstrate your point.

>>>Apriori is a rule-based algorithm used in market basket analysis to find associations between items frequently purchased together. It reduces measurement overhead by setting a minimum support threshold, so only significant associations are considered.