Q1. What is the difference between supervised and unsupervised learning? Give some examples to illustrate your point.

Supervised Learning: In supervised learning, the algorithm is provided with a labeled dataset where each data point is associated with a corresponding target or output. The goal is to learn a mapping from inputs to outputs based on the provided examples. The algorithm learns to make predictions or classifications for new, unseen data based on patterns it learned from the labeled data. Examples include linear regression, classification (like logistic regression), and support vector machines (SVMs).

Unsupervised Learning: In unsupervised learning, the algorithm is provided with an unlabeled dataset where there are no explicit target values. The goal is to find patterns, structures, or relationships within the data without any predefined categories. Unsupervised learning aims to uncover hidden insights or groupings in the data. Examples include clustering, dimensionality reduction (like PCA), and anomaly detection.

Examples:

* Supervised Learning: Predicting house prices based on features like square footage and location. The dataset includes historical prices for training.
* Unsupervised Learning: Grouping customer purchasing behaviors without knowing the categories in advance. The algorithm identifies patterns in the purchasing data without labeled categories.

Q2. Mention a few unsupervised learning applications.

Unsupervised Learning Applications:

* Clustering: Grouping similar items together, like customer segmentation, image segmentation, and document clustering.
* Dimensionality Reduction: Reducing the number of features while retaining the most relevant information, useful for visualization and feature selection.
* Anomaly Detection: Identifying unusual or unexpected patterns in data, applicable in fraud detection or network security.
* Topic Modeling: Extracting underlying themes or topics from a collection of text documents.
* Recommendation Systems: Suggesting relevant items or content to users based on their preferences and behaviors.

Q3. What are the three main types of clustering methods? Briefly describe the characteristics of each.

Types of Clustering Methods:

* Hierarchical Clustering: Builds a tree-like structure of clusters, where each data point starts as a separate cluster and is gradually merged into larger clusters based on similarity.
* Partitioning Clustering: Divides data into non-overlapping subsets or clusters.
* Density-Based Clustering: Groups data points that are dense in some region of the feature space, separating areas with low density.

Q4. Explain how the k-means algorithm determines the consistency of clustering.

**K-means Algorithm and Clustering Consistency:** The k-means algorithm determines the consistency of clustering by minimizing the sum of squared distances between data points and the centroids of their respective clusters. It iteratively updates the cluster assignments and centroids until convergence. A consistent clustering minimizes the within-cluster variance and maximizes the between-cluster variance.

Q5. With a simple illustration, explain the key difference between the k-means and k-medoids algorithms.

**Difference between K-means and K-medoids:**

In k-means, centroids are the mean values of the data points within a cluster.

In k-medoids, medoids are actual data points within the cluster that represent the center.

K-medoids is less sensitive to outliers compared to k-means since medoids are less affected by extreme values.

Q6. What is a dendrogram, and how does it work? Explain how to do it.

A dendrogram is a tree-like diagram that shows the arrangement of the clusters in hierarchical clustering. It illustrates the order in which clusters are merged. The height of the branches represents the similarity between clusters. Closer branches indicate higher similarity.

Q7. What exactly is SSE? What role does it play in the k-means algorithm?

**SSE (Sum of Squared Errors):** SSE is a measure of how far data points within a cluster are from the centroid of that cluster. In the k-means algorithm, SSE is minimized by adjusting the cluster assignments and centroids. Lower SSE indicates tighter, more compact clusters.

Q8. With a step-by-step algorithm, explain the k-means procedure.

K-means Algorithm Step-by-Step:

1. Choose the number of clusters (k) and initialize centroids.
2. Assign each data point to the nearest centroid, forming k clusters.
3. Recalculate the centroids as the mean of data points within each cluster.
4. Repeat steps 2 and 3 until centroids stabilize or a maximum number of iterations is reached.

Q9. In the sense of hierarchical clustering, define the terms single link and complete link.

Single Link and Complete Link in Hierarchical Clustering:

* Single Link: Measures similarity between clusters based on the minimum pairwise distance between their data points. Tends to create long, straggly clusters.
* Complete Link: Measures similarity based on the maximum pairwise distance between data points in different clusters. Tends to create compact, spherical clusters.

Q10. 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 Concept and Measurement Overhead Reduction:** The Apriori algorithm is used in market basket analysis to find associations between items frequently bought together. It reduces measurement overhead by focusing only on item sets that satisfy a minimum support threshold. For example, if the support threshold is set to 5%, the algorithm identifies item sets that appear in at least 5% of all transactions, filtering out infrequent item combinations and reducing computational complexity.