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

illustrate your point.

The main difference between supervised and unsupervised learning: Labeled data. The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.

2. Mention a few unsupervised learning applications.

Products Segmentation.

Customer Segmentation.

Similarity Detection.

Recommendation Systems.

Labelling unlabelled datasets.

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

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.

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

It is an iterative process of assigning each data point to the groups and slowly data points get clustered based on similar features. The objective is to minimize the sum of distances between the data points and the cluster centroid, to identify the correct group each data point should belong to.

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

algorithms.

K -means attempts to minimize the total squared error, while k -medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers (medoids or exemplars).

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

A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities. Each branch is called a clade. on. There is no limit to the number of leaves in a clade.

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

The SSE is defined as the sum of the squared Euclidean distances of each point to its closest centroid. Since this is a measure of error, the objective of k-means is to try to minimize this value. The purpose of this figure is to show that the initialization of the centroids is an important step.

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

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. ...

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

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

In Single Linkage, the distance between two clusters is the minimum distance between members of the two clusters. In Complete Linkage, the distance between two clusters is the maximum distance between members of the two cluster

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.

The Apriori algorithm is commonly cited by data scientists in research articles about market basket analysis. It identifies frequent items in the database and then evaluates their frequency as the datasets are expanded to larger sizes.