Interview Questions:

Q1.How does K-Means clustering work?
Ans.K-Means clustering is an unsupervised machine learning algorithm that partitions data into K distinct groups (clusters) based on similarity. It starts by randomly selecting K centroids, then assigns each data point to the nearest centroid based on distance (usually Euclidean). The centroids are then updated to the mean of their assigned points, and the process repeats until the centroids stabilize. The goal is to minimize intra-cluster variance (how far points are from their cluster center), creating compact and well-separated groups in the data.

Q 2. What is the Elbow method?
Ans.The Elbow method is a technique used to determine the optimal number of clusters (K) in K-Means clustering. It involves plotting the within-cluster sum of squares (inertia) against different values of K. As K increases, inertia decreases, but after a certain point, the rate of improvement drops sharply—forming an “elbow” shape in the plot. The K at this elbow point is considered optimal, as it balances clustering quality and simplicity.

Q 3. What are the limitations of K-Means?
Ans.K-Means has several limitations that can affect its effectiveness in certain scenarios. One major limitation is that it assumes clusters are spherical and equally sized, which may not hold true for real-world data. As a result, it can struggle with identifying clusters that have irregular shapes or varying densities.

Another drawback is its sensitivity to the initial placement of centroids. Poor initialization can lead to suboptimal clustering results or convergence to a local minimum. Although techniques like K-Means++ help mitigate this, randomness can still affect outcomes.

K-Means also requires you to predefine the number of clusters (K), which may not be intuitive and can lead to either overfitting or underfitting. Additionally, it is sensitive to outliers, as extreme values can distort the centroids and skew the clustering.

Lastly, K-Means works best with numeric data and Euclidean distance; it may not perform well with categorical features or non-Euclidean similarity measures.

Q4.How does initialization affect results?
Ans. Initialization in K-Means significantly affects the final clustering results because the algorithm converges to a local minimum based on the starting positions of the centroids. Poor initialization can lead to suboptimal clusters with higher intra-cluster variance, while good initialization helps the algorithm find more stable and accurate groupings. To reduce this sensitivity, methods like K-Means++ are used, which choose initial centroids more strategically to improve convergence and clustering quality.

Q 5.What is inertia in K-Means?
Ans. Inertia in K-Means is a metric that measures the sum of squared distances between each data point and the centroid of the cluster it belongs to. It reflects how tightly the data points are grouped around the centroids—lower inertia means clusters are more compact and better defined. Inertia is often used to evaluate and compare the quality of clustering results.

Q 6.What is Silhouette Score?
Ans. Silhouette Score is a metric that evaluates how well clusters are separated and how compact they are in clustering analysis. It ranges from -1 to 1, where a higher score indicates that data points are well matched to their own cluster and poorly matched to neighboring clusters, signifying clear and distinct clustering. A low or negative score suggests overlapping or poorly defined clusters.

Q7.How do you choose the right number of clusters?
Ans. Choosing the right number of clusters involves balancing cluster compactness and separation. Common methods include the Elbow Method, which looks for a point where adding more clusters yields diminishing returns in reducing inertia, and the Silhouette Score, which identifies the number of clusters that maximizes cluster quality. Domain knowledge and practical interpretability also guide the choice to ensure meaningful and actionable clusters.

Q8. What's the difference between clustering and classification?
Ans. Clustering is an unsupervised learning technique that groups data points based on similarity without predefined labels, aiming to discover natural patterns or structures in the data. Classification, on the other hand, is a supervised learning method that assigns predefined labels to new data based on a trained model using labeled examples. Essentially, clustering finds groups without prior knowledge, while classification predicts known categories.