**What is the main difference between supervised and unsupervised learning?**

**Answer: A) Supervised learning has labeled data, unsupervised learning does not.**

Supervised learning involves training a model on labeled data, where each input has a corresponding output. This allows the model to learn a direct mapping from inputs to outputs, making it suitable for tasks like classification and regression. In contrast, unsupervised learning works with unlabeled data, identifying patterns and structures without explicit labels, making it useful for clustering and anomaly detection.

**Which of the following is an example of an unsupervised learning task?**

**Answer: C) Customer segmentation for marketing**

Unsupervised learning is used when there are no predefined labels, and the goal is to discover hidden patterns in the data. Customer segmentation involves grouping customers based on shared characteristics, such as purchasing behavior or demographics, without prior labels. This helps businesses tailor marketing strategies and improve customer engagement.

**How can you measure the success or value of a machine learning project?**

**Answer: B) By evaluating improvements in key business metrics**

While achieving high accuracy is important, the real success of a machine learning project is measured by its impact on business objectives, such as increased revenue, reduced costs, or improved customer satisfaction. A model that enhances key business metrics provides tangible value, making it more beneficial than one that simply performs well on test data.

**How can you compute the Return on Investment (ROI) of a machine learning project?**

**Answer: A) (Total Revenue Generated - Model Development Cost) / Model Development Cost**

ROI measures the financial return of a machine learning project relative to its cost. It is calculated by subtracting the total cost from the revenue generated by the model and dividing by the cost. A high ROI indicates that the project has delivered significant value compared to the investment required.

**In a company with low data maturity, which type of project is more beneficial to start with?**

**Answer: B) BI project, to improve data organization, reporting, and governance**

A company with low data maturity typically lacks structured data and robust data management practices. Business Intelligence (BI) projects help improve data collection, organization, and reporting, creating a solid foundation for future AI initiatives. Jumping directly into AI without proper data governance can lead to unreliable results and project failures.

**What is a common source of bias in open-source generative AI models?**

**Answer: B) Bias in the training data used to develop the model**

AI models learn patterns from the data they are trained on, and if that data contains biases, the model will likely reflect those biases. This can result in unfair or skewed outputs, reinforcing existing stereotypes or systemic inequalities. While model architecture influences performance, the primary source of bias is often the dataset itself.

**When should a company internalize a data project instead of outsourcing it?**

**Answer: A) When the company has the necessary in-house expertise and wants to build long-term capabilities**

Internalizing a data project makes sense when a company has the technical expertise and aims to develop long-term capabilities. This approach provides greater control, customization, and scalability over time. However, if expertise is lacking or speed is a priority, outsourcing may be a better option.

**What is the main objective of clustering in machine learning?**

**Answer: A) To group similar data points together**

Clustering is an unsupervised learning technique used to identify natural groupings within a dataset. It helps uncover hidden patterns, such as customer segments or anomaly detection, by grouping data points that share similar characteristics. Unlike classification, clustering does not require predefined labels.

**Which of the following is not a clustering algorithm?**

**Answer: C) Decision Trees**

Decision trees are used for supervised learning tasks such as classification and regression, where labeled data is available. Clustering algorithms like K-Means, DBSCAN, and hierarchical clustering are used for grouping data points based on similarity without prior labels.

**In K-Means clustering, what does the "K" represent?**

**Answer: A) The number of clusters & D) The number of centroids at the beginning**

The "K" in K-Means refers to the predefined number of clusters that the algorithm aims to identify. The algorithm assigns each data point to one of these K clusters based on similarity, iteratively adjusting the cluster centroids to optimize grouping.

**Which of the following metrics is commonly used to evaluate clustering quality?**

**Answer: B) Silhouette Score**

The Silhouette Score measures how well data points fit within their assigned clusters. A high score indicates that clusters are well-separated and distinct, while a low score suggests overlap or poor clustering.

**What is the purpose of Principal Component Analysis (PCA)?**

**Answer: B) To transform high-dimensional data into a lower-dimensional space**

PCA is a dimensionality reduction technique that transforms a dataset with many correlated features into a smaller set of uncorrelated variables called principal components. This helps reduce complexity while preserving as much variance as possible, improving computational efficiency and visualization.

**How does PCA achieve dimensionality reduction?**

**Answer: B) By creating new uncorrelated variables (principal components)**

PCA transforms the original features into a new set of orthogonal (uncorrelated) variables, ordered by the amount of variance they capture. The first few components retain most of the original information, allowing for dimensionality reduction without significant information loss.

**What is a principal component in PCA?**

**Answer: B) A new axis that captures the maximum variance in the data**

A principal component is a linear combination of the original features, chosen to maximize variance. The first principal component captures the most variance, while subsequent components capture the remaining variance in descending order, ensuring that the most informative directions are prioritized.

**If the first two principal components explain 95% of the variance in the dataset, what does that mean?**

**Answer: A) The remaining components contain very little useful information**

When two principal components explain 95% of the variance, it means that most of the important information in the dataset is captured by just these two dimensions. The remaining components contribute little additional variance, allowing for a significant reduction in dimensionality while preserving the dataset’s key structure.