1) What is the difference between supervised and unsupervised learning?

- Supervised learning involves training a model on labeled data, where the outcome variable is known, to make predictions or classifications. Unsupervised learning, on the other hand, deals with unlabeled data and is used to find patterns or structures within the data, like clustering or association.

2) Explain overfitting and how can you prevent it?

- Overfitting occurs when a model learns the training data too well, including the noise and outliers, and performs poorly on new, unseen data. It can be prevented by methods like cross-validation, regularization, reducing model complexity, and gathering more data.

3) How would you evaluate a machine learning model's performance?

- A model's performance is typically evaluated using metrics like accuracy, precision, recall, F1 score for classification tasks, and mean squared error or mean absolute error for regression.

4) What are the assumptions required for linear regression?

- Linear regression assumes a linear relationship between the independent and dependent variables, independence of errors, and normal distribution of error terms.

5) What is the difference between classification and regression?

- Classification is used when the output variable is a category, such as 'spam' or 'not spam', while regression is used for predicting a continuous value like house prices.

6) How do decision trees prevent overfitting?

- To prevent overfitting, decision trees use strategies like pruning (removing parts of the tree that provide little power to classify instances) and setting a maximum depth for the tree

7) What is feature scaling and when do you need it?

- Feature scaling involves standardizing the range of independent variables or features of data. It is essential when using algorithms that calculate distances or assume normality, such as KNN or logistic regression.

8) How do you handle missing or corrupted data in a dataset?

- Missing or corrupted data can be handled by imputation (replacing missing values with statistical measures like mean, median), deletion (removing rows or columns with missing values), or using algorithms that can handle missing values.

9) What are the different types of machine learning algorithms?

- The primary types are supervised learning (e.g., linear regression, decision trees), unsupervised learning (e.g., clustering, principal component analysis), and reinforcement learning, where an agent learns to make decisions by performing actions and receiving feedback.

10) Can you explain the concept of a training set and a test set?

- In machine learning, data is split into a training set used to train the model and a test set used to evaluate its performance. The training set helps the model learn the patterns, while the test set assesses its generalization to new, unseen data.

11) Can you explain what ensemble learning is?

- Ensemble learning is a technique where multiple models (often called "weak learners") are trained and combined to improve the overall model's accuracy. Methods like bagging, boosting, and stacking are common approaches in ensemble learning.

12) What are the differences between clustering and classification?

- Clustering is an unsupervised learning technique where you group similar data points together, whereas classification is a supervised learning technique where you categorize data points into predefined classes.

13) Explain how a Random Forest algorithm works.

- A Random Forest is an ensemble learning method that constructs multiple decision trees at training time and outputs the mode of the classes for classification or mean prediction for regression of the individual trees.