1. What is a Decision Tree, and how does it work?

Answer: A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It works by splitting the dataset into subsets based on the value of input features. The splits are determined to maximize the separation of the target classes (for classification) or to minimize variance (for regression). This process continues recursively, creating a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label or continuous value.

2. What does the `criterion` parameter in `DecisionTreeClassifier` specify?

Answer: The `criterion` parameter specifies the function to measure the quality of a split. For classification, common options are: - **"gini":** Uses the Gini impurity to evaluate the splits. - **"entropy":** Uses the information gain (entropy) to evaluate the splits.

3. What is the purpose of the `max_depth` parameter in `DecisionTreeClassifier`?

Answer: The `max_depth` parameter controls the maximum depth of the tree. It helps to prevent overfitting by limiting the number of levels in the tree. A smaller `max_depth` results in a simpler model that may underfit, while a larger `max_depth` can lead to overfitting by capturing more details of the training data.

4. How do you interpret the accuracy score of a model?

Answer: The accuracy score represents the proportion of correctly classified instances out of the total instances in the test set. It is calculated as: Accuracy = Number of Correct Predictions/Total Number of Predictions A higher accuracy indicates better model performance, but it may not always reflect the model's ability to handle imbalanced datasets.

5. What information can you obtain from a confusion matrix?

Answer: A confusion matrix provides a summary of the model's prediction results. It shows: - **True Positives (TP):** Correctly predicted positive cases. - **True Negatives (TN):** Correctly predicted negative cases. - **False Positives (FP):** Incorrectly predicted positive cases. - **False Negatives (FN):** Incorrectly predicted negative cases. It helps to assess the model's performance in terms of precision, recall, and F1-score.

6. How is a classification report useful in evaluating model performance?

Answer: A classification report provides detailed metrics for each class, including: - **Precision:** The proportion of true positives among the predicted positives. - **Recall:** The proportion of true positives among the actual positives.

- **F1-Score:** The harmonic mean of precision and recall. - **Support:** The number of actual occurrences of each class in the test set. These metrics help to understand how well the model performs across different classes.

7. What is the purpose of visualizing a Decision Tree?

Answer: Visualizing a Decision Tree helps to: - Understand the model's decision-making process. - Interpret the splits and decision rules. - Identify which features are most important in making predictions. - Communicate the model's logic to stakeholders.

8. Why might you use the `entropy` criterion instead of the default `gini`?

Answer: The `entropy` criterion uses information gain, which can sometimes provide better performance by considering the amount of information gained from each split. However, the choice between `entropy` and `gini` often depends on the specific dataset and problem. Both criteria generally perform similarly, and the choice may come down to personal preference or computational efficiency.

9. What does the `random_state` parameter control in the `train_test_split` function?

Answer: The `random_state` parameter controls the random number generator used for shuffling and splitting the data. Setting a fixed `random_state` ensures that the split is reproducible, allowing for consistent results across different runs of the code.

10. How do you address overfitting in a Decision Tree model?

Answer: To address overfitting in a Decision Tree model, you can: - **Limit the tree depth** using the `max_depth` parameter. - **Set a minimum number of samples per leaf** using `min_samples_leaf`. - **Set a minimum number of samples per split** using `min_samples_split`. - **Prune the tree** by removing nodes that provide little predictive power.

Additional Viva Questions

1. **What is the role of feature importance in Decision Trees?**
   - **Answer:** Feature importance in Decision Trees measures the contribution of each feature in predicting the target variable. It helps in understanding which features have the most influence on the model's decisions and can be used for feature selection.
2. **How do you handle categorical features in Decision Trees?**
   - **Answer:** Categorical features should be encoded into numerical values before using them in Decision Trees. This can be done using techniques like one-hot encoding or label encoding.

3. **What are some advantages and disadvantages of using Decision Trees?**
   - **Answer:**
     - **Advantages:** Easy to interpret, no need for feature scaling, can handle both numerical and categorical data.
     - **Disadvantages:** Prone to overfitting, can be biased towards features with more levels, may not capture complex relationships well.
4. **How can you determine the optimal hyperparameters for a Decision Tree?**
   - **Answer:** Hyperparameters can be optimized using techniques like Grid Search or Random Search, which involve training the model with different combinations of hyperparameters and evaluating performance using cross-validation.
5. **What is the significance of pruning in Decision Trees?**
   - **Answer:** Pruning involves removing branches from the tree that provide little predictive power. It helps to reduce overfitting by simplifying the model and improving generalization to new data.
6. **Explain the concept of Gini impurity and how it affects decision-making in a tree.**
   - **Answer:** Gini impurity measures the purity of a node by calculating the probability of incorrectly classifying a randomly chosen element. It is used to evaluate the quality of a split, with lower Gini impurity indicating a better split.
7. **What are the main differences between Decision Trees and Random Forests?**
   - **Answer:** Decision Trees are single models that can be prone to overfitting. Random Forests are an ensemble method that combines multiple Decision Trees to improve accuracy and robustness by averaging their predictions.
8. **How does the `splitter` parameter affect Decision Tree training?**
   - **Answer:** The `splitter` parameter determines the strategy used to choose the split at each node. Options include "best" (chooses the best split) and "random" (chooses the best split among a random subset).
9. **What are some common applications of Decision Trees?**
   - **Answer:** Decision Trees are used in various applications such as credit scoring, medical diagnosis, customer segmentation, and recommendation systems.
10. **Can Decision Trees be used for regression tasks? If so, how?**
    - **Answer:** Yes, Decision Trees can be used for regression tasks. In this case, the tree predicts a continuous value rather than a class label. Splits are based on minimizing variance or mean squared error rather than class impurity.