Conceptual Questions

1. What is a decision tree?

- A decision tree is a supervised learning algorithm used for classification and regression tasks, which splits data into subsets based on feature values.

2. What are the main components of a decision tree?

- Root Node: Represents the entire dataset.

- Internal Nodes: Represent decisions based on feature values.

- Leaf Nodes: Represent the final output (class or value).

3. What is the difference between classification and regression trees?

- Classification trees predict a category or label, while regression trees predict a continuous value.

4. What is a splitting criterion in a decision tree?

- A metric used to decide the best split at each node, such as Gini Index, Information Gain, or Mean Squared Error (for regression).

5. What is meant by overfitting in a decision tree?

- Overfitting occurs when the tree becomes too complex, capturing noise instead of the underlying pattern.

Mathematical and Practical Questions

6. How is Information Gain calculated?

- Information Gain = Entropy(parent) - [Weighted average of Entropy(children)].

7. What is the Gini Index, and how is it used?

- The Gini Index measures impurity or diversity in a dataset. A lower Gini Index indicates a better split.

8. What is entropy in the context of a decision tree?

- Entropy is a measure of uncertainty or randomness in a dataset. Lower entropy indicates more homogeneity.

9. How do you prevent overfitting in decision trees?

- Techniques include pruning, limiting tree depth, setting a minimum number of samples per split, or using ensemble methods like Random Forest.

10. What is pruning in a decision tree?

- Pruning is the process of removing branches that have little importance to reduce overfitting and improve generalization.

Implementation Questions

11. How do you implement a decision tree in Python?

- Steps:
  - Import: `from sklearn.tree import DecisionTreeClassifier` (or `DecisionTreeRegressor`).
  - Fit the model: `model.fit(X_train, y_train)`.
  - Predict: `model.predict(X_test)`.

12. What is the role of max_depth in a decision tree?

- It controls the maximum depth of the tree, preventing overfitting by limiting the tree's complexity.

13. How do you evaluate the performance of a decision tree?

- Metrics include accuracy, precision, recall, F1-score for classification, and RMSE or MAE for regression.

14. Can a decision tree handle categorical data?

- Yes, but the data must often be encoded (e.g., one-hot encoding or label encoding).

15. What are the advantages of using decision trees?

- Easy to understand, handles both numerical and categorical data, and requires little data preprocessing.

Limitations and Scenario-Based Questions

16. What are the limitations of decision trees?

- Prone to overfitting, instability (small changes in data can lead to different trees), and biased splits for features with more categories.

17. How does a decision tree compare to logistic regression?

- Decision trees are non-linear models and can capture more complex relationships, while logistic regression assumes a linear relationship between features and the log odds.

18. What is the role of feature importance in a decision tree?

- Feature importance quantifies the contribution of each feature to the decision-making process.

19. How do decision trees handle missing data?

- Some implementations can handle missing data by splitting on surrogate splits or ignoring missing values during splitting.

20. When would you prefer not to use a decision tree?

- When interpretability is critical, or if the dataset is too small, as decision trees can overfit easily.