DECISION TREE

INTERVIW QUESTION'S:

1. What are some common hyperparameters of decision tree models, and how do they affect the model's performance?

Some common hyperparameters of decision tree model's are:

- **max_depth**: The maximum depth of the tree. Limiting the depth can prevent overfitting by ensuring the tree does not become too complex. A shallow tree might underfit, while a very deep tree might overfit the training data.

- **min_samples_split**: The minimum number of samples required to split an internal node. Higher values can result in a more generalized model, while lower values can lead to overfitting.

min_samples_leaf: The minimum number of samples that a leaf node must have. Increasing this value can smooth the model and help avoid overfitting.

- **max_features**: The number of features to consider when looking for the best split. Limiting the features can lead to a more robust model by reducing variance.

- **criterion**: The function used to measure the quality of a split. Common options include "gini" for the Gini impurity and "entropy" for information gain. Different criteria can lead to different tree structures and performance.

- **splitter**: The strategy used to choose the split at each node. Options include "best" (the best split) and "random" (a random split). This can affect both performance and computation time.

2. What is the difference between the Label encoding and One-hot encoding?

Difference Between Label Encoding and One-Hot Encoding

Both label encoding and one-hot encoding are techniques used to convert categorical variables into numerical format, but they serve different purposes and have distinct characteristics:

- **Label Encoding**:
  - Each category is assigned a unique integer. For example, categories "Red", "Green", and "Blue" might be encoded as 0, 1, and 2, respectively.
  - **Use Case**: Suitable for ordinal data where the categories have a natural order (e.g., "Low", "Medium", "High").
  - **Limitation**: Can mislead models into thinking there is a relationship between the integer values (e.g., 1 is closer to 2 than to 0), which is inappropriate for nominal data.
- **One-Hot Encoding**:
  - Each category is transformed into a binary column. For the same example, "Red", "Green", and "Blue" would be represented as:
    - Red: [1, 0, 0]
    - Green: [0, 1, 0]
    - Blue: [0, 0, 1]
  - **Use Case**: Ideal for nominal data where no intrinsic order exists.
  - **Limitation**: Increases the dimensionality of the dataset, which can lead to the "curse of dimensionality" if there are many categories.

Summary

1. **Hyperparameters**: Key hyperparameters of decision trees include `max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`, `criterion`, and `splitter`, which affect model complexity and performance.
2. **Encoding**: Label encoding assigns unique integers to categories, suitable for ordinal data, while one-hot encoding creates binary columns for each category, ideal for nominal data without implied order.