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

Ans:

Decision Tree models have several hyperparameters that can significantly affect their performance, complexity, and generalizability. Tuning these hyperparameters is crucial to strike a balance between underfitting and overfitting. Below are the most common hyperparameters, along with their effects on the model.

1. Max Depth **(**max\_depth**):**

* Specifies the maximum depth of the tree.
* Lower max depth can prevent overfitting by limiting the complexity of the tree, but may lead to underfitting.
* Higher max depth allows the tree to capture more intricate patterns in the data, but increases the risk of overfitting

1. Min Samples Split **(**min\_samples\_split**):**

* The minimum number of samples required to split an internal node.
* Larger values can prevent overfitting by requiring more samples to split, resulting in fewer splits.
* Smaller values allow the tree to create more splits, capturing more detail but potentially overfitting.

1. Min Samples Leaf **(**min\_samples\_leaf**):**

* The minimum number of samples required to be at a leaf node.
* Larger values result in fewer, larger leaf nodes, which can prevent overfitting but may lead to underfitting.
* Smaller values allow for more granular splits, which can increase the risk of overfitting.

1. **Max Features (**max\_features**):**

* The number of features to consider when looking for the best split.
* Smaller values reduce the model's complexity and variance but may miss important patterns.
* Larger values increase the model's ability to capture patterns but also increase the risk of overfitting.

1. **Max Leaf Nodes (max\_leaf\_nodes):**

* Limits the number of leaf nodes in the tree.
* Creates simpler models with fewer leaf nodes, which can prevent overfitting but may lead to underfitting.
* Allows more leaf nodes, capturing more detail but increasing the risk of overfitting.

1. **Criterion (criterion):**

* The function to measure the quality of a split. Common criteria include "gini" for Gini impurity and "entropy" for information gain.
* Typically results in faster computation and can create purer nodes.
* May produce more balanced trees but can be computationally intensive.

1. **Splitter (splitter):**

* The strategy used to choose the split at each node. Options include "best" and "random".
* Always selects the best split, which can result in better performance but slower training.
* Selects a random split, which can speed up training but may reduce accuracy.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Label Encoding | One-Hot Encoding |  |
| Representation | Assigns a single integer to each category | Creates multiple binary columns |  |
| Order | Can introduce an artificial order | Preserves the categorical nature of the data |  |
| Dimensionality | Increases the number of features by 1 | Increases the number of features significantly |  |
| Suitability | Suitable for ordinal data | Suitable for nominal data |  |