Decision Tree

**Interview Questions:**

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

-->

1. max\_depth

Description: Limits the maximum depth of the tree (i.e., the longest path from the root to a leaf node).

Effect: Setting a low max\_depth can prevent over-fitting by restricting the model complexity, which may improve generalization but can lead to under fitting if set too low. A high or unlimited depth allows more complex patterns but increases the risk of over-fitting, especially with noisy data.

2. min\_samples\_split

Description: The minimum number of samples required to split an internal node.

Effect: Higher values prevent the model from creating nodes that capture very specific, potentially noisy patterns, reducing over-fitting. Smaller values allow the tree to grow more branches, capturing fine details of the data, but at the risk of over-fitting.

3. min\_samples\_leaf

Description: The minimum number of samples required to be at a leaf node.

Effect: Higher values prevent the model from creating leaf nodes that represent very specific cases, which can improve generalization. Lower values allow leaf nodes with fewer samples, capturing more specific patterns in the data but increasing the risk of over-fitting.

4. max\_features

Description: The maximum number of features considered when splitting a node.

Effect: Limits the number of features to evaluate at each split. This can introduce randomness into the tree, especially in ensemble methods like Random Forests, improving generalization and reducing overfitting. Setting it too high can increase the risk of overfitting, while very low values may lead to underfitting.

5. max\_leaf\_nodes

Description: Limits the total number of leaf nodes in the tree.

Effect: A lower limit on leaf nodes reduces complexity, making the model more generalizable, but may cause underfitting. Higher values allow more complex trees, which might overfit if not controlled properly.

6. criterion

Description: Defines the function to measure the quality of a split (e.g., "gini" or "entropy" for classification).

Effect: The choice of criterion can influence the splits made by the tree. Gini impurity (default in classification) tends to create purer splits faster, while entropy may lead to a more balanced tree, potentially at the cost of computation time.

7. min\_impurity\_decrease

Description: Sets the minimum decrease in impurity required to make a split.

Effect: Higher values make the tree more conservative, splitting only when a significant reduction in impurity occurs, which can reduce over-fitting.

8. splitter

Description: Specifies the strategy used to choose the split at each node ("best" for optimal split, "random" for a random split).

Effect: "Best" leads to more accurate, potentially deeper trees, while "random" introduces more randomness, which can improve generalization in ensemble methods by reducing correlation between trees.

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

-->

| **Feature** | **Label Encoding** | **One-Hot Encoding** |
| --- | --- | --- |
| **Output Format** | Single column with integer labels | Multiple binary columns |
| **Best for** | Ordinal data with a natural order | Nominal data without an order |
| **Encoding Example** | Red = 0, Green = 1, Blue = 2 | Red = [1, 0, 0], Green = [0, 1, 0], Blue = [0, 0, 1] |
| **Interpretation Risk** | May imply unintended ordinal relationships | No ordinal relationships implied |
| **Dimensionality** | Doesn’t increase dimensionality | Increases dataset dimensionality |
| **Effect on Model** | Suitable for models that handle ordinal features | Suitable for most models but may cause sparsity with high cardinality |
| **Memory Usage** | Low, due to single-column representation | Higher memory usage due to multiple columns |