Interview questions.

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

① - Max depth :-

* Controls the maximum depth trees
* A deeper tree capture more patterns but can lead to overfitting.

② . Min Sample Split

* The minimum number of samples requires to split an internal node.
* Higher values prevent overfitting by ensuring split.

③ . Min Sample leaf :-

* The minimum number of samples a leaf node must contain.
* A smaller number reduces overfitting and increases generalization but may also lower accuracy leaves.

④ Max features :-

* The number of features to consider when looking for the best split.

⑤ . Criterion :-

* gini : Measures impurity

*entropy:- measure the information gain.

*MSE (mean squred Error):- For regression tree.

⑥. Max leaf Nodes:-

* limits the number of leaf nodes.
* Helps prevent overfitting by controlling tree complexity.

⑦. Min impurity Decrease:-

* Prevents unnecessary split and controls overfitting.

Effect on Model Performance:-

① Overfitting vs underfitting:-

* Small, Large, large leads to underfitting
* large(max_depth), small (Min_sample_split leads of overfitting.

② Bias-variance Trade off:-

* High depth & small constraints reduce bias but increase variance.
* Regularization through constraints

③ Computational Efficiency:-

* Large trees are more computationally expensive.
* limiting & max_leaf_nodes improve efficiency

2) What is different between Label Encoding & one-hot encoding?

1) Label-Encoding.

* Assigns a unique interger category.

Ex:-

categories: [`Baran', `Hari', `Deepak']

Label Encoded: {`Baran': 0, `Hari': 1, `Deepak': 2}

* Pros:-

* Simple & memory-efficient

* Cons:

* Imposes an ordinal relationship between categories, which may not be meaningful.

* use case:-

* Suitable for ordinal categorical variable, where order matter (low to high)

2) One-Hot-Encoding:-

* convert each category into a separate binary column (0 or 1)

Ex:-

categories: ['Red', 'Green', 'Blue']

One-Hot-Encoded:

Red	Green	Blue
1	0	0
0	1	0
0	0	1

* Pros:-
  * No ordinary relationship assumption.
* Cons:-
  * Increases dimensionality, leading to the curse of dimensionality for large datasets.
* use case:-
  * Suitable for nominal categorical variable.