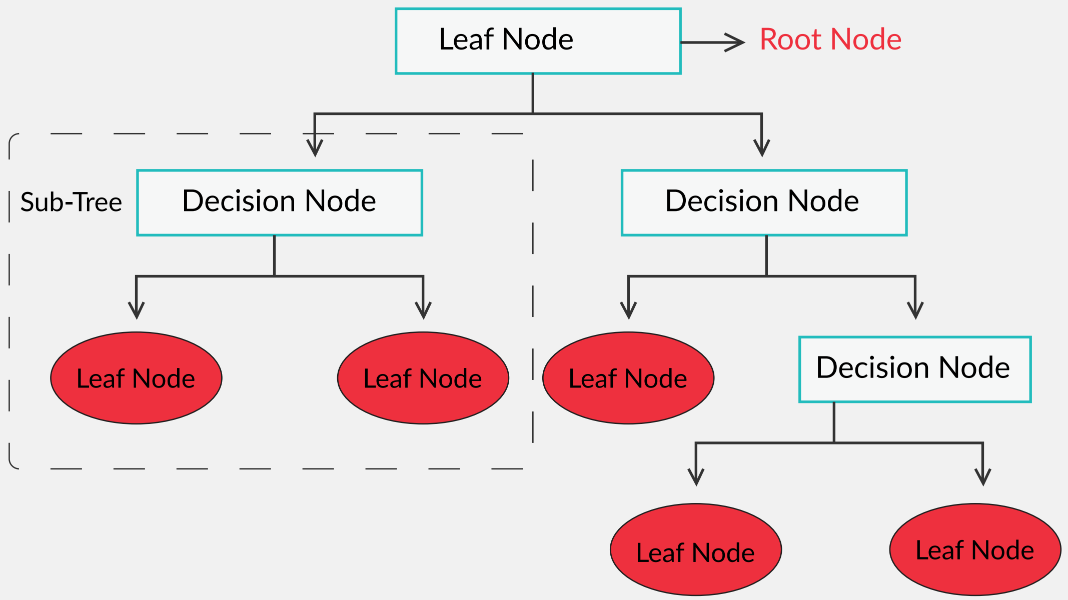
Classification Using Decision Trees

- Decision trees are supervised classification algorithms. They work for both categorical and continuous data.
- A decision tree is a binary tree that recursively splits a data set until data with only one type of class (leaf nodes) remains.
- A leaf node contains only one type of class and, hence, cannot be split further.
- Decision trees can be thought of as a set of nested if and else statements. The conditions for these if and else statements are modelled as a tree. The model learns which features are to be considered and decides the threshold for each node; hence, this is machine learning.
- Decision trees focus on one split at a time. They do not backtrack and change previous splits. This is a type of greedy algorithm.



Common Interview Questions:

1. How does a decision tree work on numerical and categorical data?
2. Why do decision trees have a low bias and a high variance?
3. What is the effect of outliers on the decision tree?
4. Do missing values affect the decision tree?
5. What is pruning, and when is it used on decision trees?
6. What is information gain, and when should it not be used?
7. Which algorithms are used in decision trees?
8. How does feature selection occur in decision trees?
9. In which scenarios will decision trees not work well?
10. Which machine learning technique uses decision trees?

Terminology

Root Node : It is the starting node of the decision tree and represents the entire data.

Splitting : This is the process of a parent node dividing into multiple child nodes. This division occurs based on features and the threshold obtained using machine learning algorithms and metrics, such as information gain and Gini Impurity.

Decision Node : This is the node where splitting occurs (i.e., the decision is taken). Leaf/Terminal Node This is the last node that cannot be split further.

Classification Using Decision Trees

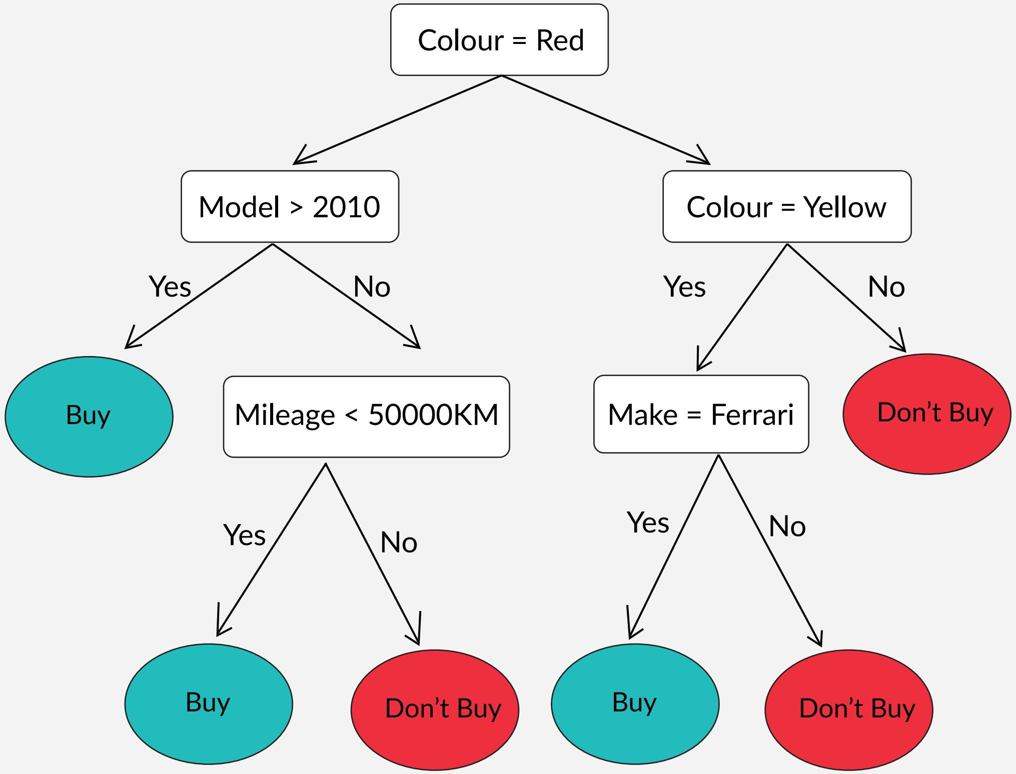
Creation of decision trees

Decision trees are supervised classification algorithms. They work for both categorical and continuous data.

Each split divides the tree into several distinct, non-overlapping subspaces. The model tests all the features and threshold values to find the optimal split that minimises the cost function.

The model compares all possible splits at a node and will choose the split that maximises information gain.

This process continues until leaf nodes that cannot be divided further.



Entropy

Entropy is the measure of information contained in a state. Mathematically,

$$E = -\sum_{i=1}^{N} P_i \log_2 P_i$$

If the entropy is high, we are uncertain about the randomly picked data point class. If the entropy is low, we can be confident about selection of the random data point class.

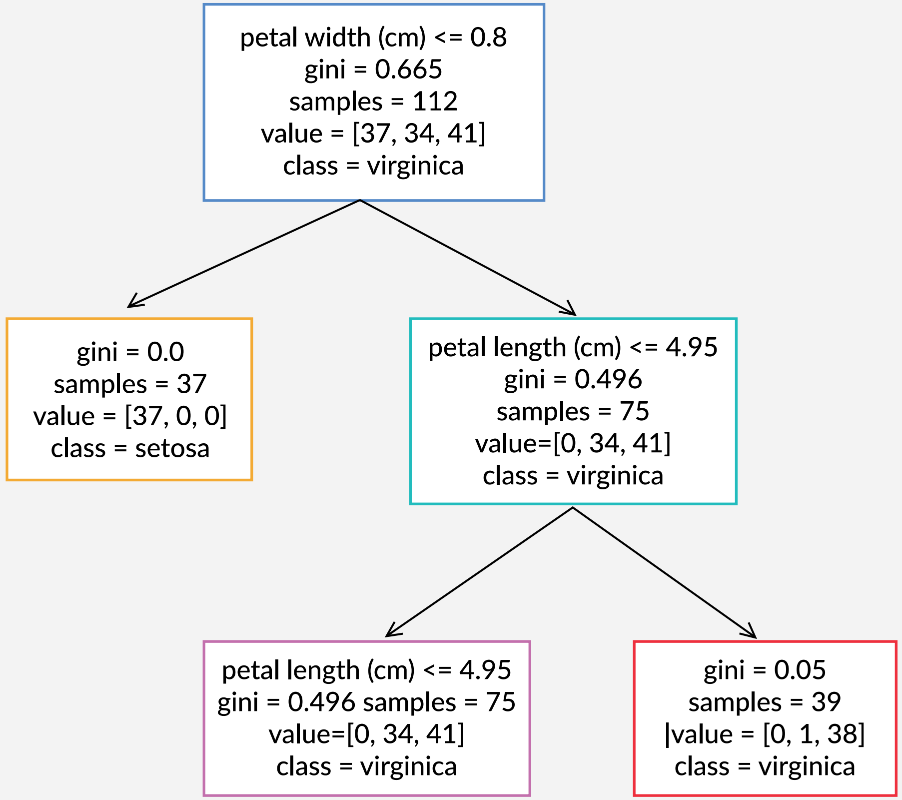
Classification Using Decision Trees

Gini Impurity

Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree

- It is a measurement of how often a randomly chosen element would be labeled incorrectly.
- The Gini impurity can also be used to evaluate possible split candidates and is calculated by the following equation:

$$Gini = 1 - \sum_{i=1}^{N} P_i^2$$



Information Gain

Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree

Mathematically,

$$Gain\ (S\ ,\ A)\ Entropy(S)\ -\ \sum_{v} \frac{|S_v|}{|S|}\ Entropy\ (Sv)$$

- The information gain can be calculated for a split by subtracting the weighted entropies of the children from the parent's entropy.
- Information gain for all possible splits is calculated at a node & then split which provided most information gain is selected.

Gini Impurity

- Gini Impurity is a measure used to build decision trees for determining how a data set's features should split nodes to form a tree.
- It is a measure of the frequency of a randomly chosen element being labelled incorrectly
- Gini impurity can also be used to evaluate possible split candidates and is calculated using the following equation:

$$Gini = 1 - \sum_{i=1}^{N} P_i^2$$

Classification Using Decision Trees

Choosing between Information Gain or Gini Impurity

Both can be used to create decision trees. The differences between the two approaches are listed below:

- Gini Index prefers large partitions and is relatively easy to implement, whereas information gain prefers small partitions
- Gini Index is used by CART(Classification And Regression Tree) algorithms.
- Information Gain is used in ID3(Iterative Dichotomiser 3) and C4.5 algorithms

CART vs ID3 Algorithms

CART and ID3 algorithms are commonly used to create decision trees.

Here are the common differences between the two:

- CART is used for classification and regression problems, whereas ID3 is used only for classification problems.
- CART uses the Gini Index, whereas ID3 uses entropy and information gain to perform splits.

Extension of decision trees

Decision trees are part of many algorithms and methods. The use property of decision trees and improve the model's accuracy as well.

Bagging

- It is also called “bootstrap aggregating.” It is used to reduce variance within a noisy data set
- We implement it by taking random samples with replacements from a data set and building decision trees.
- We repeat the process multiple times to generate many decision trees that are later aggregated.

Random Forests

- It improves the bagging process by decorrelating the trees (using the introduction of splitting on a random subset of features).
- For each split, only a certain number of features (say n) are randomly chosen.
- It is used for both classification and regression problems.

Ensemble Methods

- These techniques improve accuracy of a model by combining multiple models instead of using a single one.
- Bagging, boosting and stacking are its three main classes.