←Go Back to Practical Data Science

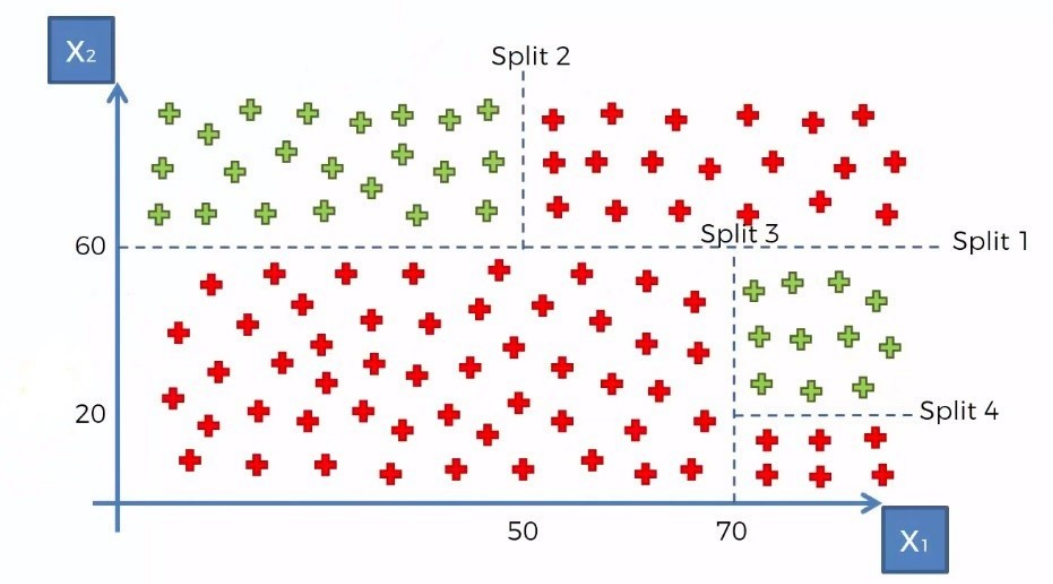
Course Content

Decision Trees

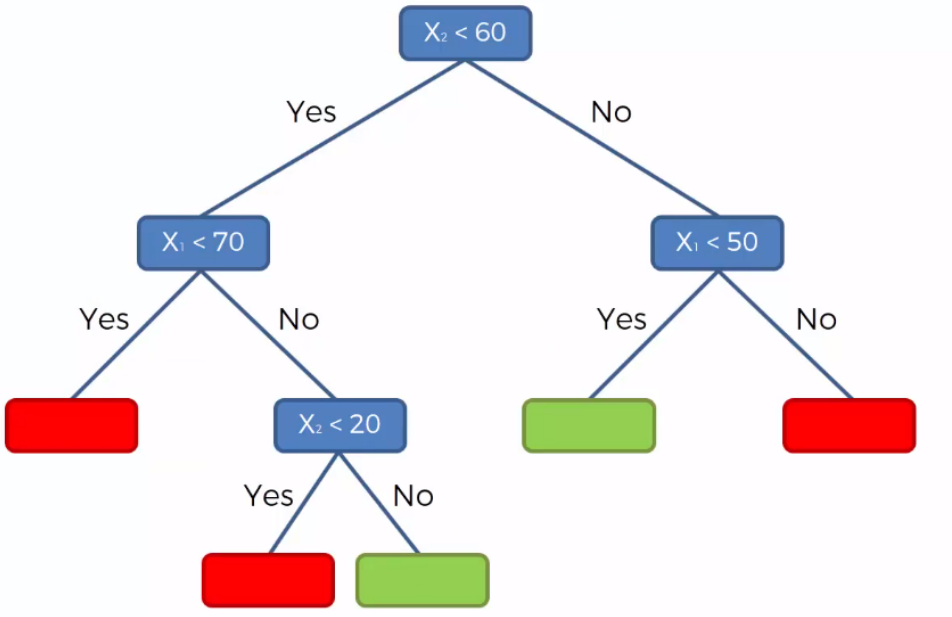
Introduction To Decision Trees

Decision trees are tree-based models that help in making a decision in both regression and classification problems. To make a decision, they use a hierarchical structure and split the dataset into smaller subsets.

For example: Suppose there is a dataset with two independent variables, X_1 and X_2, and one dependent variable with two classes, red and green, as shown in the below picture.



Let's look at what the decision tree looks like for this example.



The decision tree will make the first split where X_2<60, the second split is where X_1<70, and similarly the third and fourth split are where X_2<20 and X_1<50, respectively. After making splits 1 and 2, there are only red data points in the left region where X_2<60 and X_1<70, this

region is pure as it has only one type of class. If a region is not pure, it is further split into sub-regions.

Formally, the decision tree splits the data based on different splitting methods. One of the most commonly used methods is Entropy and Information Gain. They are defined as follows:

Entropy: Entropy is the measure of randomness or impurity contained in a dataset. Mathematically, it can be written as:

$$H(Y) = -\sum p_i \log_2(p_i)$$

Where, Y is the target variable and p_i is the probability of a class i in the data.

Information gain: It is the measure of the information gained by adding a feature/independent variable or, in other words, reduction in the impurity after adding a feature. We simply subtract the entropy of Y given X from the entropy of Y to calculate the reduction of impurity about Y given an additional piece of information X.

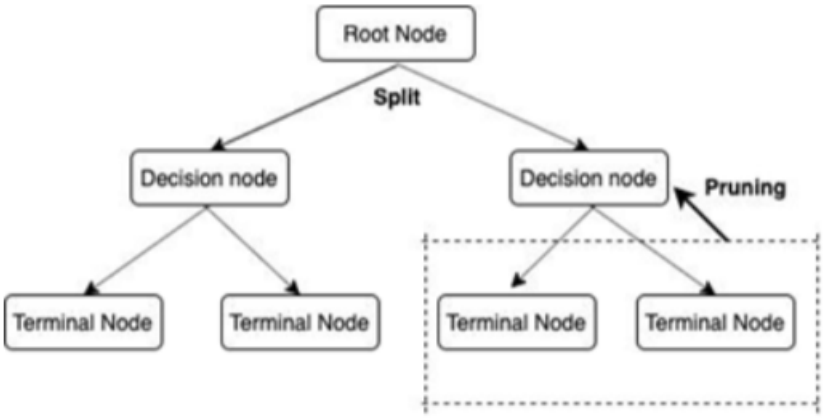
Mathematically, it is given as:

$$Information\ Gain = H(Y) - H(Y \mid X)$$

Where, Y is the target variable, X is the independent variable.

Important terminology:

- **Root Node:** The root node is from where the decision tree starts. It represents the entire population or samples which get divided into two or more branches.
- **Branch or Sub-Tree:** A part of the entire decision tree is called a branch or sub-tree.
- **Splitting:** Dividing a node into two or more sub-nodes based on if-else conditions.
- **Decision Node:** A sub-node that splits into further sub-nodes. In simple terms, every node is a decision node, except for leaf nodes.
- **Leaf or Terminal Node:** This is the end of the decision tree where it cannot be split into further sub-nodes.
- **Depth of the tree:** The depth of a decision tree is the number of nodes from the root node down to the furthest leaf node. The below tree has a depth equal to 2.



Next >