Lesson 6.01
Decision Trees (CART)

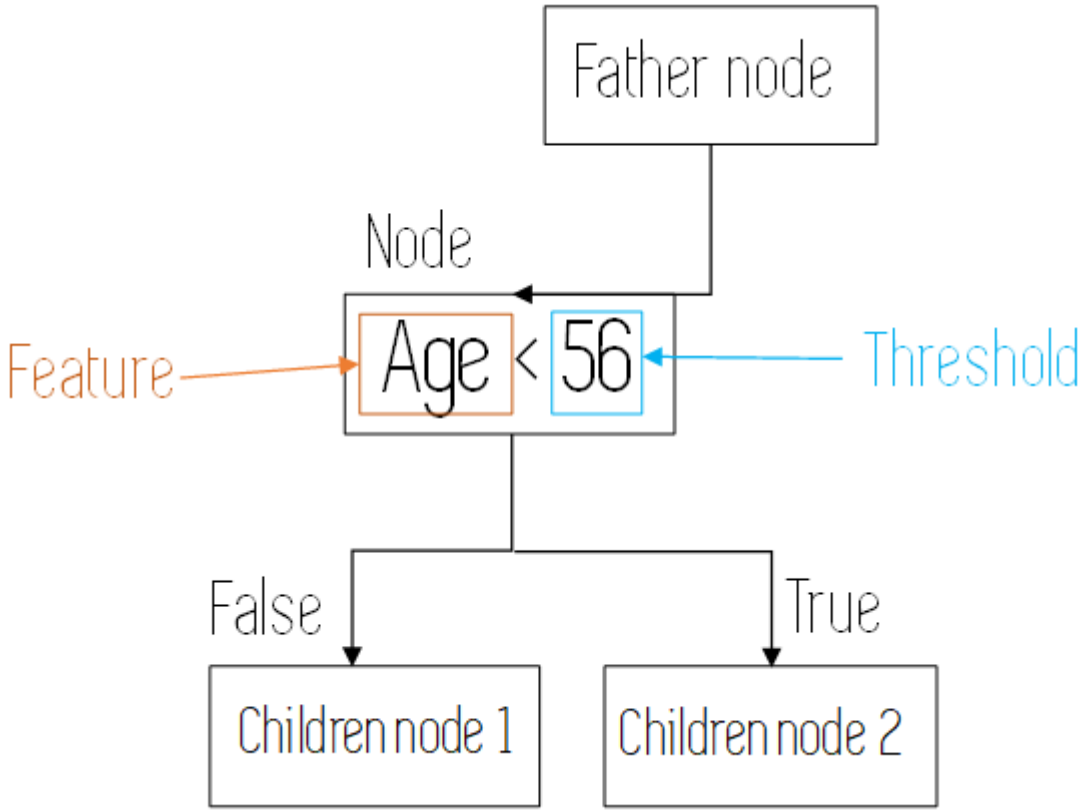
Decision Trees (DTs) Overview

- Decision Trees are supervised ML algos that can be used for classification and regression

- Decision trees are also called "Classification and Regression Trees," sometimes abbreviated "CART."

- The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from training data

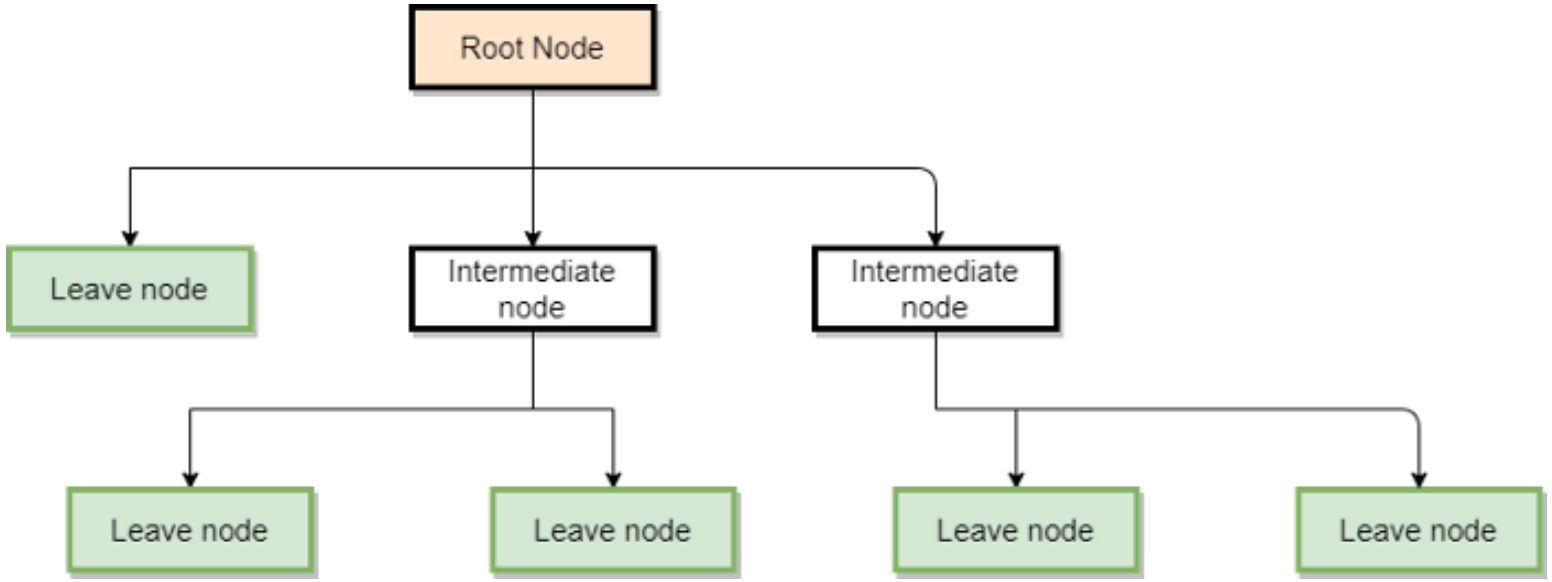
Decision Trees Overview



Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

Decision Trees Overview



Root Node: Is the node that starts the graph. In a normal decision tree it evaluates the variable that best splits the data.

Intermediate nodes: These are nodes where variables are evaluated but which are not the final nodes where predictions are made.

Leaf nodes: These are the final nodes of the tree, where the predictions of a category or a numerical value are made.

How does Decision Trees decide where to split?

- The decision of making strategic splits heavily affects a tree's accuracy
- Decision trees use Gini among other algos to decide to split a node in two or more sub-nodes
- The creation of sub-nodes increases the homogeneity of resultant sub-nodes.
- In other words, we can say that purity of the node increases with respect to the target variable.
- Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

Gini

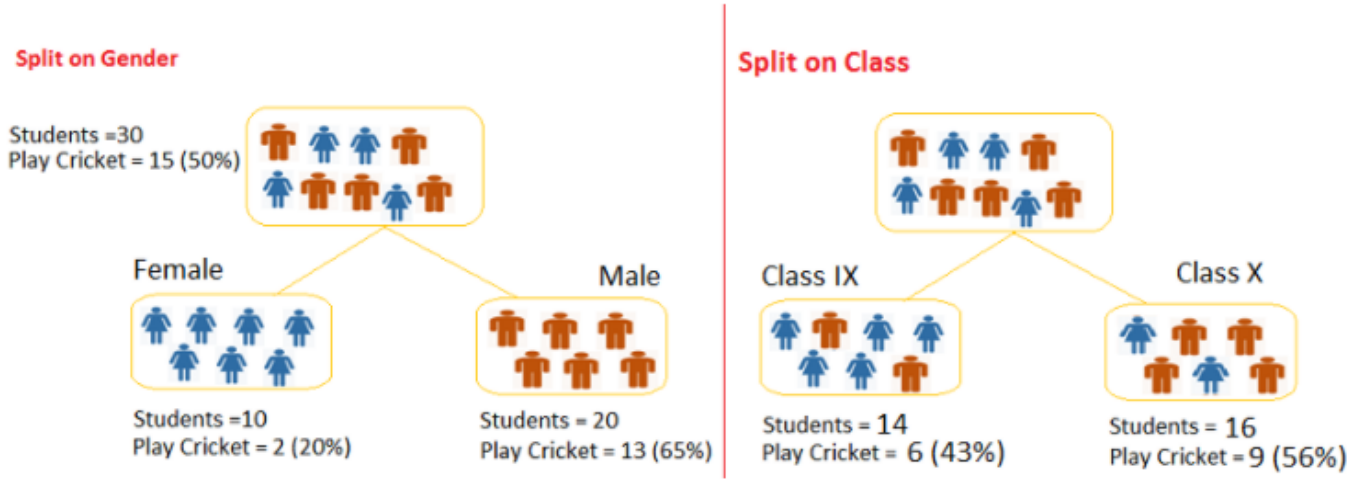
- Gini says if we select two items from a homogeneous population at random, then they must be of same class and probability of this is 1
- It works with categorical target variables e.g. True/ False
- It performs binary splits
- Higher the value of Gini, higher the inequality of that class

Gini

- **Steps to Calculate Gini for a split**
  - Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
  - Calculate Gini for split using weighted Gini score of each node of that split

- **Example (Next Slide)**
  - where we want to segregate the students based on target variable (playing cricket or not).
  - we split the population using two input variables Gender and Class.
  - We want to identify which split is producing more homogeneous sub-nodes using Gini

Example: – Referring to example used above, where we want to segregate the students based on target variable (playing cricket or not). In the snapshot below, we split the population using two input variables Gender and Class. Now, I want to identify which split is producing more homogeneous sub-nodes using Gini .



Split on Gender:

1. Calculate, Gini for sub-node Female = (0.2)*(0.2)+(0.8)*(0.8)=0.68
2. Gini for sub-node Male = (0.65)*(0.65)+(0.35)*(0.35)=0.55
3. Calculate weighted Gini for Split Gender = (10/30)*0.68+(20/30)*0.55 = **0.59**

Similar for Split on Class:

1. Gini for sub-node Class IX = (0.43)*(0.43)+(0.57)*(0.57)=0.51
2. Gini for sub-node Class X = (0.56)*(0.56)+(0.44)*(0.44)=0.51
3. Calculate weighted Gini for Split Class = (14/30)*0.51+(16/30)*0.51 = **0.51**

Above, you can see that Gini score for *Split on Gender* is higher than *Split on Class,* hence, the node split will take place on Gender.

Decision Tree Model Assumptions

- In the beginning, the whole training set is considered as the **root.**
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are **distributed recursively** on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.