Decision Tree

* Decision Tree is a Supervised learning technique that can be use for both classification and Regression problems, But mostly it is preferred for solving Classification problems. It is a tree structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision node and Leaf Node.** Decision nodes are used make any decision and do not contain any further branches.
* The decision or the test are performed on the basis of features of the given dataset.
* It is called a decision tree because, similar to as tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**

**Why use Decision Tree?**

* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
* The logic behind the decision tree can be easily understood because it shows a tree like structure.

Decision Tree Terminologies

 **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

 **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

 **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

 **Branch/Sub Tree:** A tree formed by splitting the tree.

 **Pruning:** Pruning is the process of removing the unwanted branches from the tree.

 **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

**How does the Decision Tree algorithm work?**

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and further. It continues the process until it reaches the leaf node of the tree.

* **Step-1:** Begin the tree with root node, says S, which contains the complete dataset.
* **step-2:** find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
* **Step-3:** divide the S into Subsets that contains possible values for the best attributes.
* **Step-**4:Generate the decision tree node, which contains the best attribute.
* **Step-5:** Recursively make new decision trees using the subsets of the dataset created in setp-3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**Attribute Selection Measures**

**While implementing a Decision tree, the main issue arises that how to select the best attribute for root node and sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM.** By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular technique for ASM, which are:

* **Information Gain**
* **Gini index**

1.Information Gain

* Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
* It calculates how much information a feature provides us about a class.
* According to the value of information gain, we split the node and build the decision tree.
* A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first.

**Entropy:** Entropy is a metric t measure the impurity in a given attribute. It specifies

Randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

2. Gini Index:

* Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
* An attribute with the low Gini index should be preferred as compared to the high

Gini index.