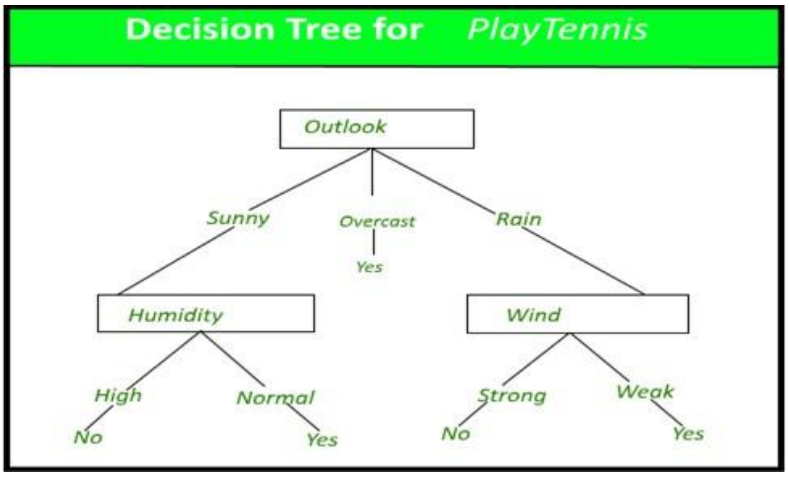
Decision Tree

- Decision Tree is the most powerful and popular tool for classification and prediction.
- A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



A decision tree for the concept PlayTennis.

Construction of Decision Tree:

- A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning.
- The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.
- The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.
- Decision trees can handle high-dimensional data.
- In general decision tree, classifier has good accuracy.
- Decision tree induction is a typical inductive approach to learn knowledge on classification.

Short note on Decision Tree:-

- A decision tree which is also known as prediction tree refers a tree structure to mention the sequences of decisions as well as consequences.
- Considering the input X = (X1, X2,… Xn), the aim is to predict a response or output variable Y.
- Each element in the set (X1, X2,…Xn) is known as input variable.It is possible to achieve the prediction by the process of building a decision tree which has test points as well as branches.
- At each test point, it is decided to select a particular branch and traverse down the tree.
- Ultimately, a final point is reached, and it will be easy to make prediction.
- In a decision tree, all the test points exhibit testing specific input variables (or attributes), and the developed decision tree is represented by the branches.
- Because of flexibility as well as simple visualization, decision trees are mostly probably deployed in data mining applications for the purpose of classification.
- In the decision tree, the input values are considered as categorical or continuous.
- A structure of test points (known as nodes) and branches is established by the decision tree by which the decision being made will be represented.
- Leaf node is the one which do not have further branches. The returning value of leaf nodes is class labels while in some cases they return the probability scores.
- It is possible to convert decision tree into a set of decision rules.
- There are two types of Decision trees: classification trees and regression trees
- Classification trees are generally applied to output variables which are categorical and mostly binary in nature, for example yes or no, sale or not, and so on.
- Whereas regression trees are applied to output variables which are numeric or continuous, for example predicted price of a consumer good.
- In variety of situations, it is possible to apply decision tree. It is easy to represent them in a visual way, and the analogous straightforward.
- Also as the result is a sequence of logical if-then statements, there is no any presence of underlying assumption regarding a linear or nonlinear relationship between the input variables and the response variable.

Decision Tree Representation:

- Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.
- An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute as shown in the above figure.
- This process is then repeated for the subtree rooted at the new node. The decision tree in above figure classifies a particular morning according to whether it is suitable for playing tennis and returns the classification associated with the particular leaf.(in this case Yes or No).

For example, the instance

(Outlook = Sunny, Temperature = Hot, Humidity = High, Wind = Strong)

- would be sorted down the leftmost branch of this decision tree and would therefore be classified as a negative instance.
- In other words, we can say that the decision tree represents a disjunction of conjunctions of constraints on the attribute values of instances.

(Outlook = Sunny ^ Humidity = Normal) v (Outlook = Overcast) v (Outlook = Rain ^ Wind = Weak)

Gini Index:

- Gini Index is a score that evaluates how accurate a split is among the classified groups.
- Gini index evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes.
- In this case, we want to have a Gini index score as low as possible.
- Gini Index is the evaluation metrics we shall use to evaluate our Decision Tree Model.