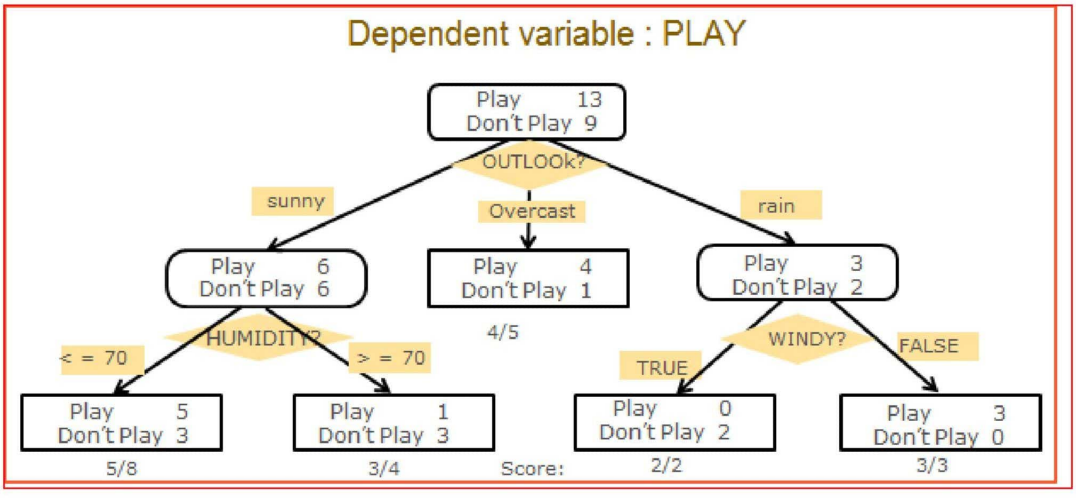
Decision trees and rule induction methods are capable of culling through a set of predictor variables and successively splitting a data set into subgroups in order to improve the prediction or classification of a target (dependent) variable

As such they are valuable to data miners faced with constructing predictive models when there may be a large number of predictor variables and not much theory or previous work to guide them.



How DT is different from Traditional statistical Modelling

Traditional statistical prediction methods (for example, regression, logistic regression or discriminant analysis) involve fitting a model to data, evaluating fit and estimating parameters that are later used in a prediction equation. Whereas Decision tree or rule induction models take a different approach. They successively partition a data set based on the relationships between predictor variables and a target (outcome) variable. When successful, the resulting tree or rules indicate which predictor variables are most strongly related to the target variable.

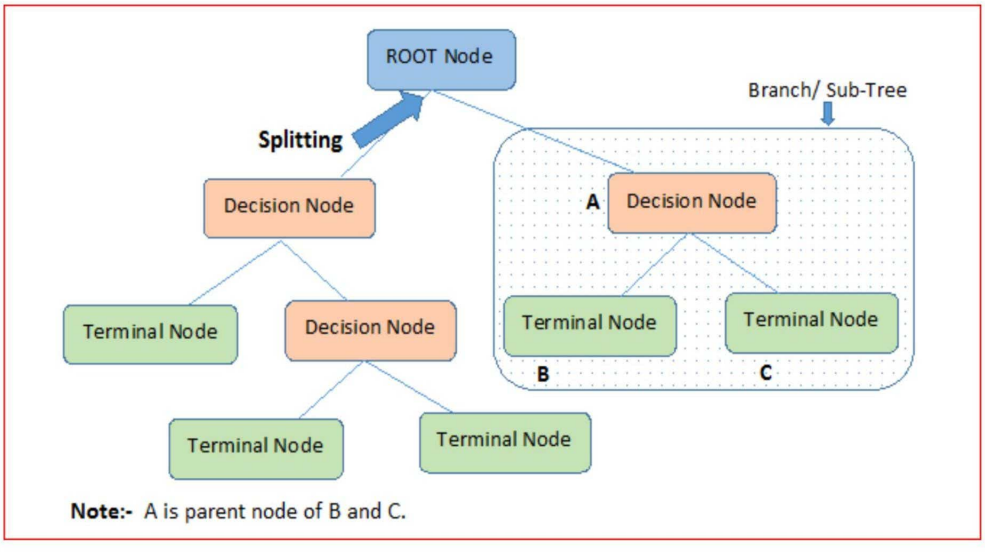
Tree building starts by finding the *variable/feature* for **best split**.

They also find subgroups that have concentrations of cases with desired characteristics. Decision trees represent a set of decisions

These decisions(Splits) generate rules for classification/Prediction of a dataset using the statistical criterions such as

- Entropy
- information gain
- Gini index
- Chi-square test,..etc

Basic terminology used in DT



- **Root Node:** Entire population or sample further gets divided into two or more homogeneous sets.
- **Parent and Child Node:** Node which is divided into sub-nodes is called parent node, whereas sub-nodes are the child of parent node.
- **Splitting:** Process of dividing a node into two or more sub-nodes.
- **Decision Node:** A sub-node that splits into further sub-nodes.
- **Leaf/ Terminal Node:** Nodes that do not split.

- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. (Opposite of Splitting)

Types of Decision Trees

Types of decision tree are based on the type of target variable we have. It can be of two types:

Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example:- In above scenario of student problem, where the target variable was "Student will play cricket or not" i.e. YES or NO.

Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Advantages

1. **Easy to Understand:** Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
2. **Useful in Data exploration:** Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. You can refer article (Trick to enhance power of regression model) for one such trick. It can also be used in data exploration stage. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
3. **Less data cleaning required:** It requires less data cleaning compared to some other modelling techniques. It is not influenced by outliers and missing values to a fair degree.
4. **Data type is not a constraint:** It can handle both numerical and categorical variables.
5. **Non Parametric Method:** Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Disadvantages

1. **Over fitting:** Over fitting is one of the most practical difficulty for decision tree models.
2. In case of imbalanced dataset, decision trees are biased. However, by using proper splitting criteria, this issue can be resolved.

Most Important Parameters

Minimum Samples Split: Minimum number of sample required to split a node. This parameter helps in reducing overfitting.

High value: Under fitting, Low value: Overfitting

Maximum Depth of a Tree: Most influential parameter. Gives limit on vertical depth decide upto which level pruning is required.

Higher value: Overfitting, Lower value: Underfitting

(c) **Maximum Features:** At each node, while splitting either we can chose best feature from pool of all the features or limited number of random features. This parameter adds a little randomness - good generalised model.