Artificial Intelligence Masterclass

Classification & Regression Trees (CART)

H.M. Samadhi Chathuranga Rathnayake

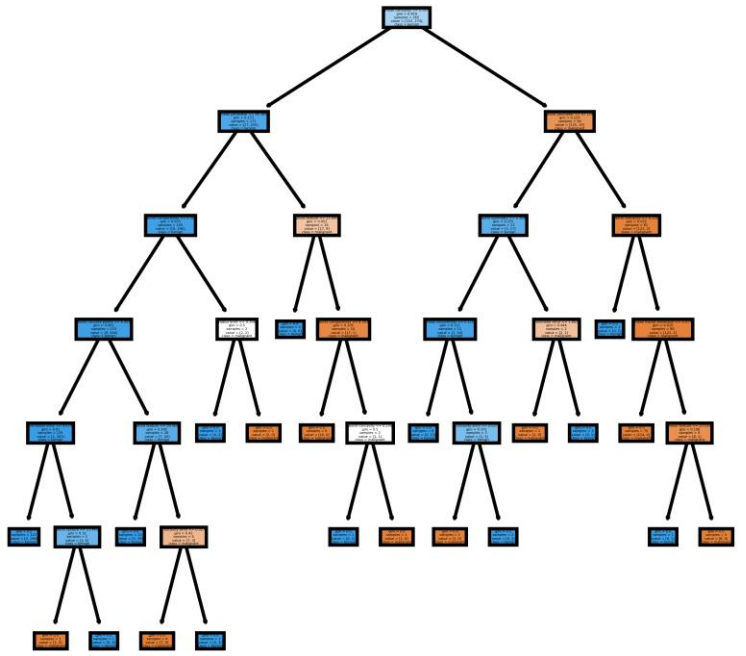
M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Classification & Regression Trees (CART)

A Decision Tree (CART) is a simple representation for classification as well as for regression. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.

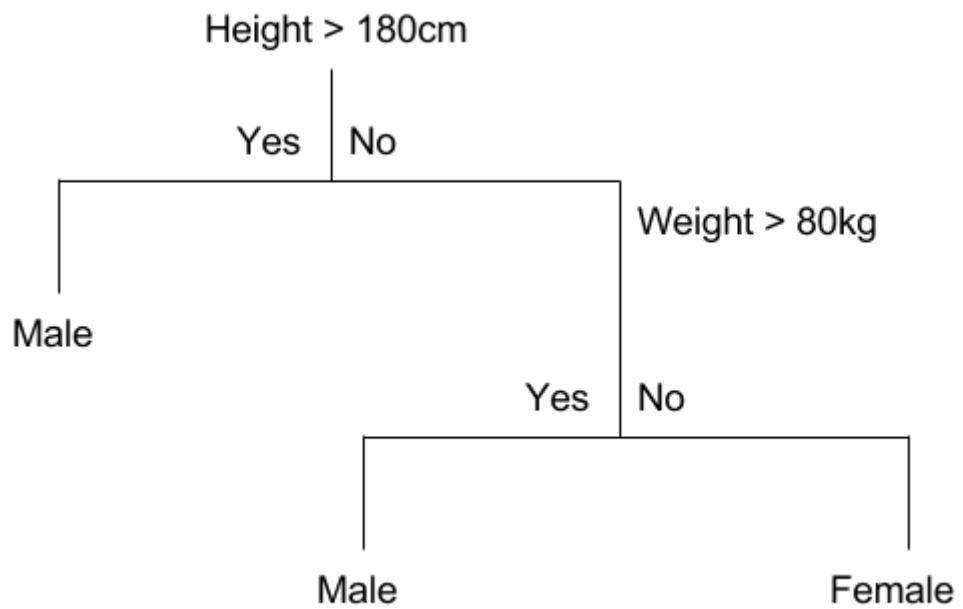
There are two types of Decision Trees we can identify.

- Classification Trees
- Regression Trees



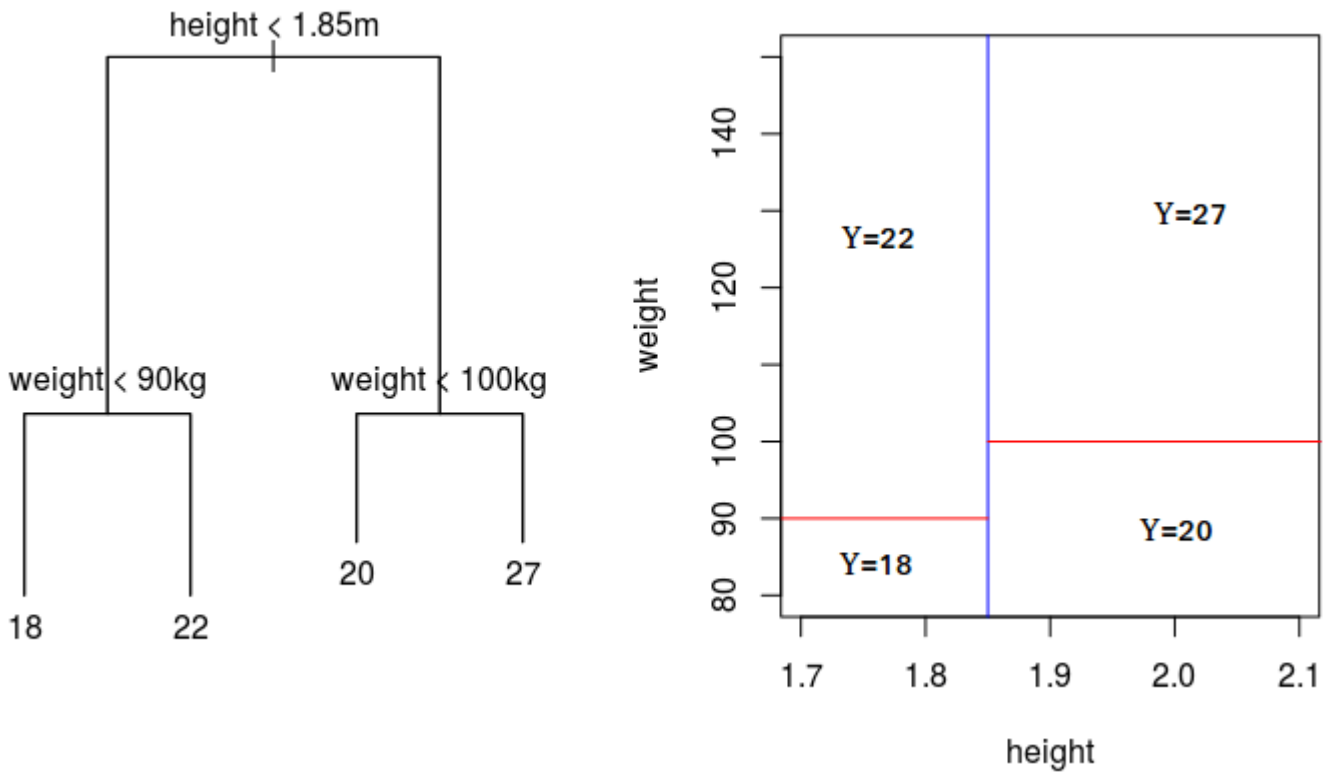
Classification Trees

Here the response variable is a categorical variable. So the main objective is to classify observations. Consider the following example.



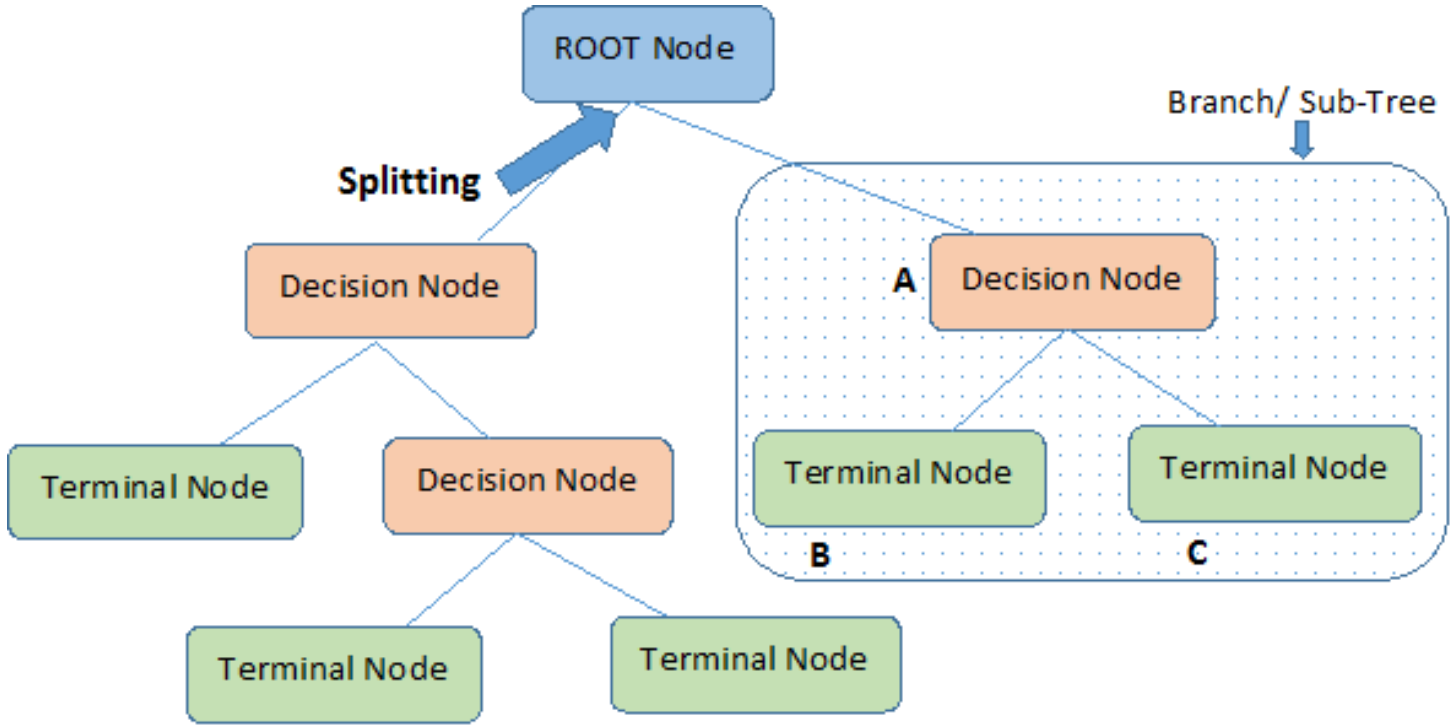
Regression Trees

Here the response variable is a numerical variable. So the main objective is to predict observations. Consider the following example. Here the average value is returned such that the criteria is satisfied.



Important Terminologies Related to Decision Trees

Consider the following diagram. All the terminologies are given.



Note:- A is parent node of B and C.

Splitting in Decision Trees

Decision trees use multiple algorithms 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.

There are several techniques we can use to do this.

- Gini Impurity
- Information Gain
- Reduction in Variance

Gini Impurity

This says that if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

- It works with categorical target variable “Success” or “Failure”.
- It performs only Binary splits
- Higher the value of Gini higher the homogeneity.
- CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to Calculate Gini for a split

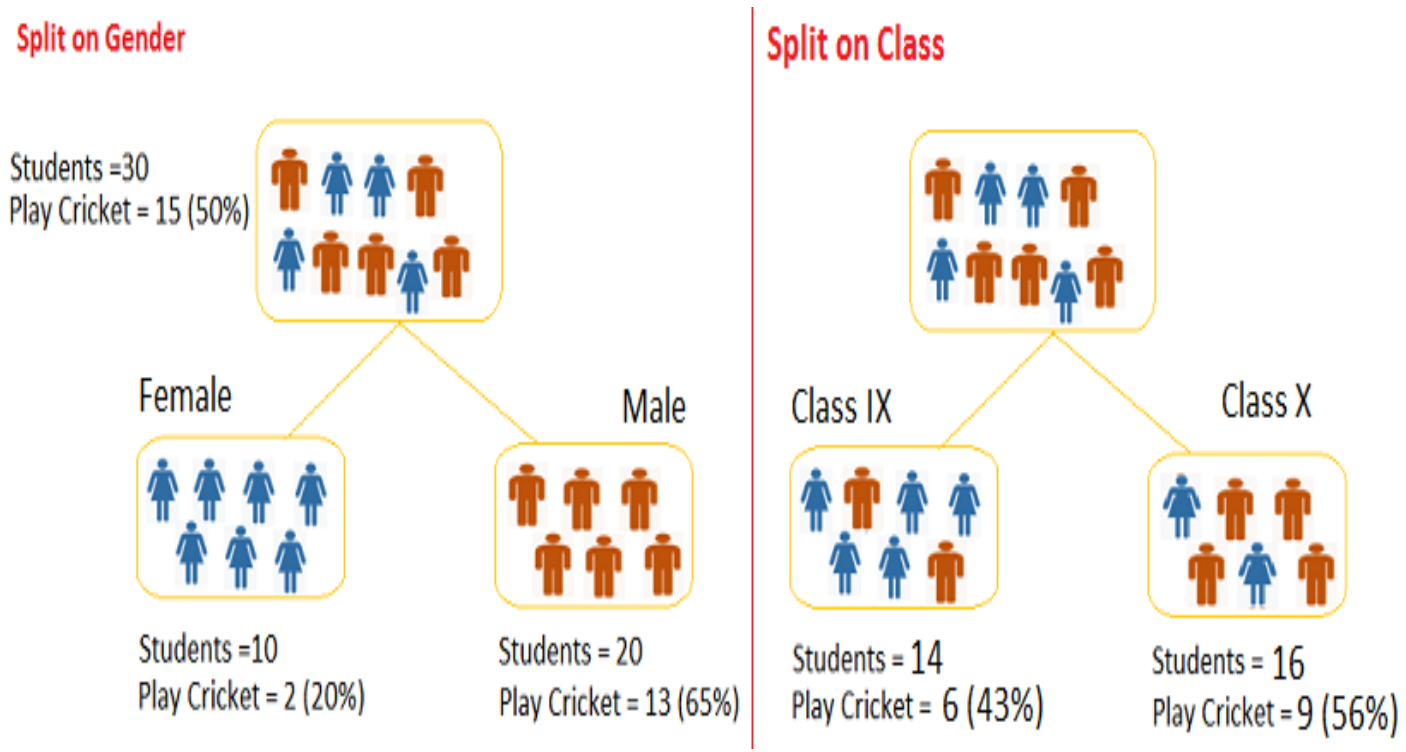
- Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure

$$p^2 + (1 - p)^2$$

- Calculate Gini for split using weighted Gini score of each node of that split.

Gini Impurity

Consider the following example. Here 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. The objective is to identify most homogeneous subgroups.



Gini Impurity

Consider the following example. Here 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. The objective is to identify most homogeneous subgroups.

Split on Gender:
Calculate, Gini for sub-node Female = $(0.2)^2+(0.8)^2= 0.68$
Gini for sub-node Male = $(0.65)^2+(0.35)^2= 0.55$
Calculate weighted Gini for Split Gender = $(10/30) \times 0.68 + (20/30) \times 0.55 = \mathbf{0.59}$

Similar for Split on Class:
Gini for sub-node Class IX = $(0.43)^2+(0.57)^2= 0.51$
Gini for sub-node Class X = $(0.56)^2+(0.44)^2= 0.51$
Calculate weighted Gini for Split Class = $(14/30) \times 0.51 + (16/30) \times 0.51 = \mathbf{0.51}$

Gini Impurity

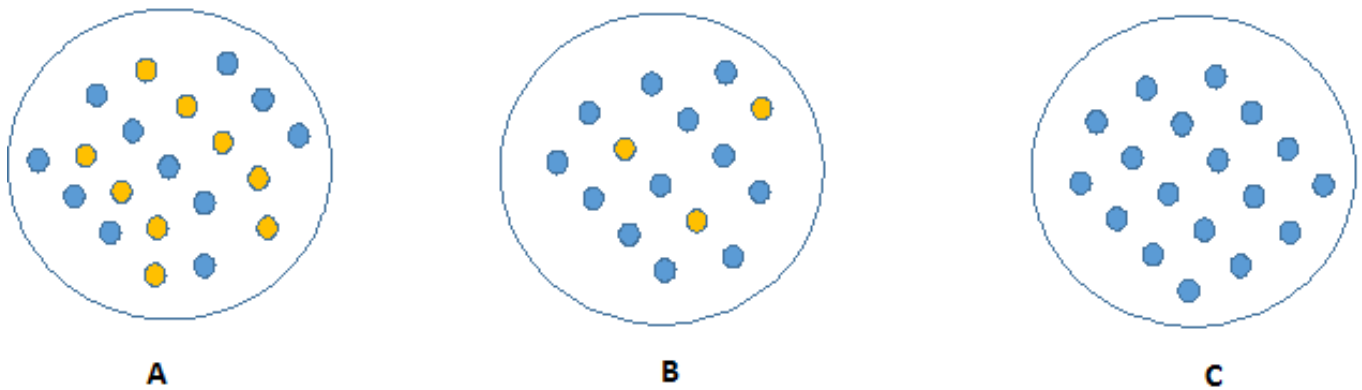
We can see that Gini score for Split on Gender is higher than Split on Class, hence, the node split will take place on Gender. We can work with the Gini Impurity value as well. It is nothing but,

$$Gini\ Impurity\ =\ 1 - Gini$$

So if we go with the Gini Impurity, we select the lowest Gini Impurity. In that case also, here we have to select Gender.

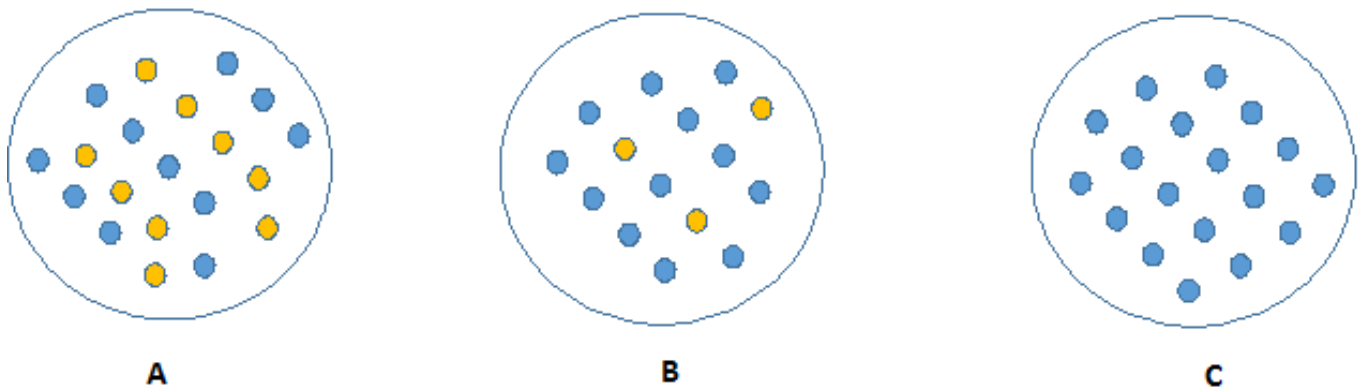
Information Gain

Think which can be described easily among following groups.



Information Gain

Think which can be described easily among following groups.



Answer is C because it requires less information as all values are similar. On the other hand, B requires more information to describe it and A requires the maximum information. In other words, we can say that C is a Pure node, B is less impure, and A is more impure.

Information Gain

we can build a conclusion that less impure node requires less information to describe it. And, more impure node requires more information. Information theory is a measure to define this degree of disorganization in a system known as Entropy. If the sample is completely homogeneous, then the entropy is zero and if the sample is an equally divided (50% – 50%), it has entropy of one.

Entropy can be calculated using formula:-

$$Entropy = -plog_2(p) - (1-p)log_2(1-p)$$

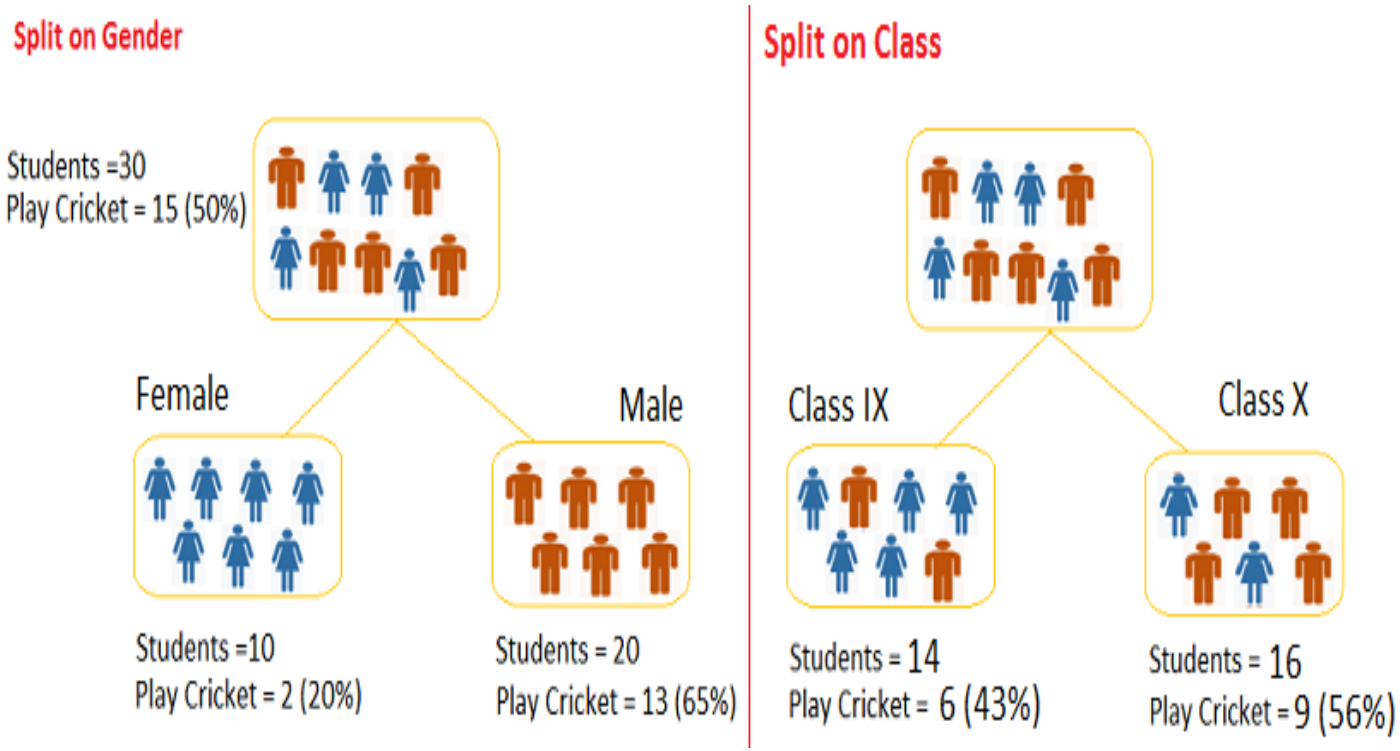
Here the p is the probability of success. The lesser the entropy, the better it is.

Steps to calculate entropy for a split:

- Calculate entropy of parent node
- Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

Information Gain

Consider the above example



Information Gain

Consider the above example

Entropy for parent node = -(15/30) log2 (15/30) – (15/30) log2 (15/30) = **1**. Here 1 shows that it is an impure node.

Entropy for Female node = -(2/10) log2 (2/10) – (8/10) log2 (8/10) = 0.72 and for male node, -(13/20) log2 (13/20) – (7/20) log2 (7/20) = 0.93
Entropy for split Gender = Weighted entropy of sub-nodes = (10/30)*0.72 + (20/30)*0.93 = **0.86**

Entropy for Class IX node, -(6/14) log2 (6/14) – (8/14) log2 (8/14) = 0.99 and for Class X node, -(9/16) log2 (9/16) – (7/16) log2 (7/16) = 0.99.
Entropy for split Class = (14/30)*0.99 + (16/30)*0.99 = **0.99**

Information Gain

Information gain is considered here as,

$$Information\ Gain = Entropy(Parent) - Entropy(Split)$$

Maximum information gain is given in the lowest Entropy. Entropy for the split of the Gender is the lowest. So select the split with Gender.

Reduction in Variance

Reduction in variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population:

$$Variance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

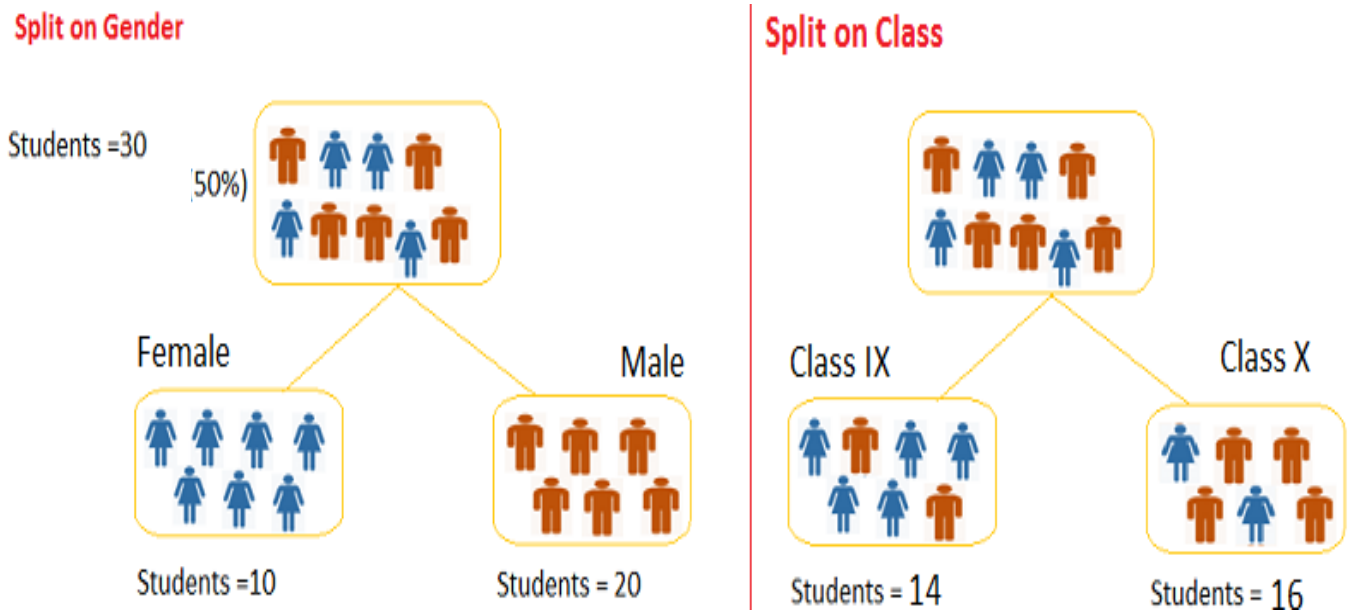
$\bar{x}$ is the mean of the values.

Steps to calculate Variance:

- Calculate variance for each node.
- Calculate variance for each split as weighted average of each node variance.

Reduction in Variance

Consider we have to create a regression tree for predicting Mathematics marks for above example.



Class Mathematics Marks Variance		Class Mathematics Marks Variance	
25		25	
Female Mathematics Marks Variance	Male Mathematics Marks Variance	Class IX Mathematics Marks Variance	Class X Mathematics Marks Variance
16	21	24	25

Reduction in Variance

Consider the above example.

- Variance for Split Gender = Weighted Variance of Sub-nodes = $\left(\frac{10}{30}\right) \times 16 \ + \ \left(\frac{20}{30}\right) \times 23 \ = \ \mathbf{20.67}$
- Variance for Split Class = $\left(\frac{14}{30}\right) \times 24 \ + \ \left(\frac{16}{30}\right) \times 25 \ = \ \mathbf{24.5}$

We can see that Gender split has lower variance compare to parent node, so the split would take place on Gender variable.

Advantages & Disadvantages

Advantages

- Easy to Understand
- Useful in Data exploration
- Less data cleaning required
- Data type is not a constraint
- Non-Parametric Method

Disadvantages

- Over fitting

Pruning

Overfitting is one of the key challenges faced while using tree based algorithms. If there is no limit set of a decision tree, it will give you 100% accuracy on training set because in the worse case it will end up making 1 leaf for each observation. Thus, preventing overfitting is pivotal while modeling a decision tree. This can be done through Pruning trees.

Techniques for reducing the complexity of trees.

- Maximum depth of tree
- Maximum number of terminal nodes
- Maximum features to consider for split