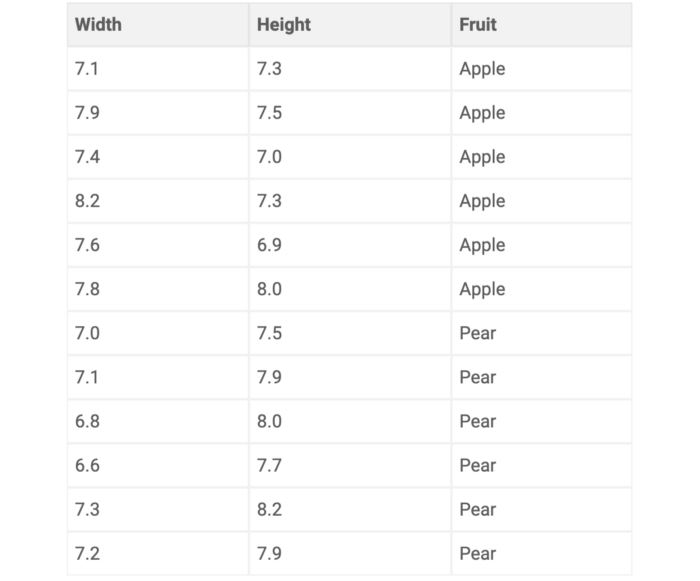
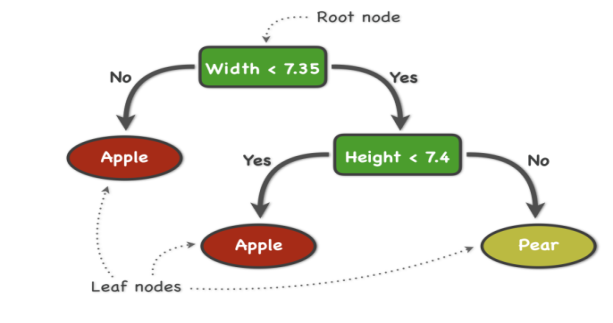
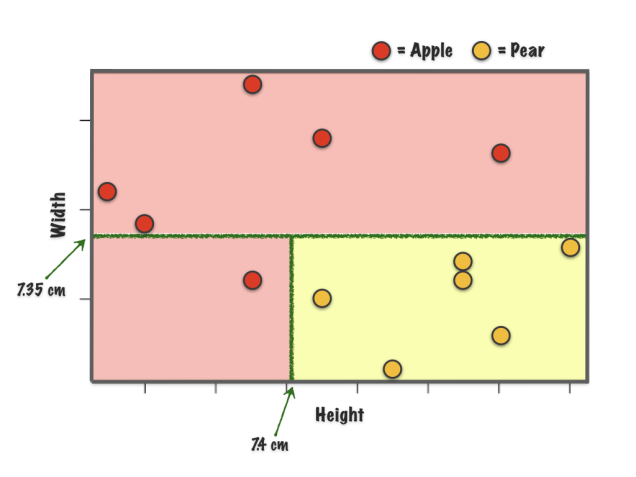
Decision tree for Numeric data:



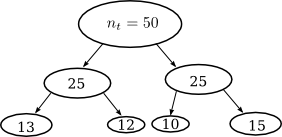




**Min node size :**

the **minimum node size**, in the example below the minimum node size is 10. This parameter implicitly sets the depth of your trees. i.e. min number of datapoints in each node. i.e Min no. of data points for split

So once we reach to this node size we stop splitting further



**Maximal depth :**

So maximum depth is a part of parameter tuning, but generally each feature may contains 2 to 3 splits, not more than that.

We are going to understand what is entropy, information gain and gini index,

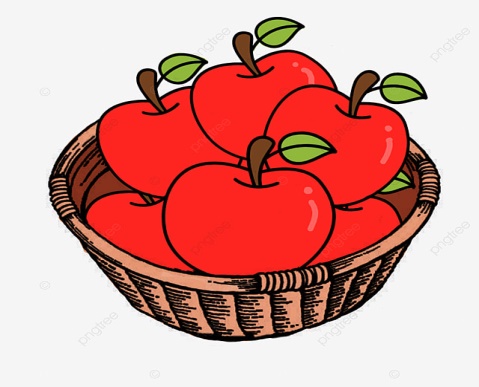
**Pruning :**

Pruning a decision tree means to remove a subtree that is redundant and not a useful split and replace it with a leaf node.

Sometimes, what happen when we build decision tree unnecessary we go on splitting the features but sill we are not able to get the required outcome, so in that case our tree becomes bulky and not useful also, so in that case we use to trim some part of our decision tree, i.e nothing but pruning.

**Entropy :**

Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data.



1. Data is pure
2. Probability of getting Apply is 1
3. Entropy is 0

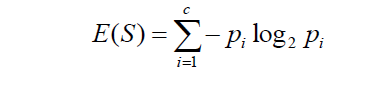


1. Data is not pure
2. Probability of getting fruit as an apple is less than 1
3. Entropy is greater than 0

We need to calculate Entropy, because initially we have a heterogeneous data, and eventually we need to bring it in a homogeneous one.

(i.e we need to remove impurity)

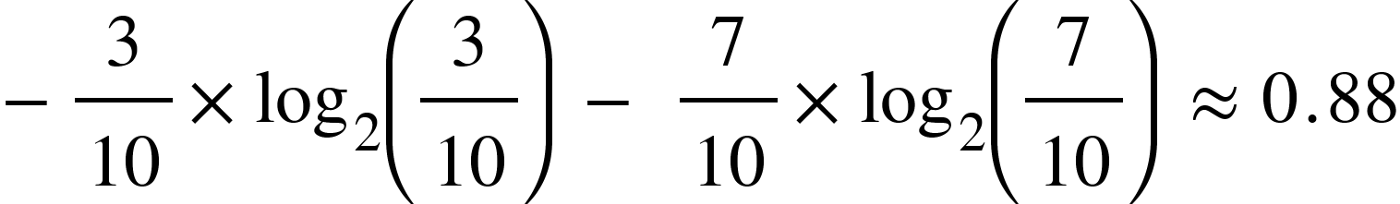
**The Mathematical formula for Entropy is as follows -**



Entropy : Sometimes also denoted using the letter ‘H’

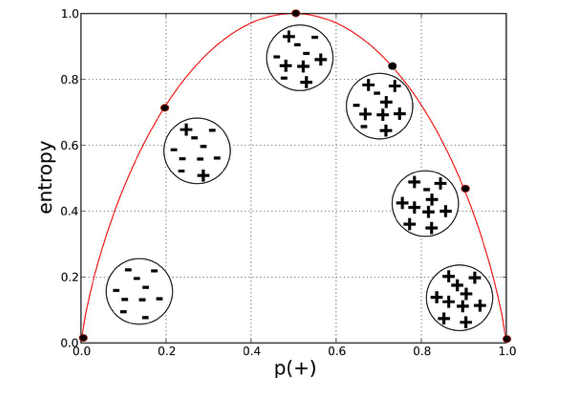
Where ‘Pi’ is simply the probability of an element/class(category) ‘i’ in our data. For simplicity’s sake let’s say we only have two classes, a positive class and a negative class. Therefore ‘i’ here could be either + or (-). So if we had a total of 100 data points in our dataset with 30 belonging to the positive class and 70 belonging to the negative class then ‘P+’ would be 3/10 and ‘P-’ would be 7/10

If I was to calculate the entropy of my classes in this example using the formula above. Here’s what I would get.



The entropy here is approximately 0.88. This is considered a high entropy, a high level of disorder ( meaning low level of purity). Entropy is measured between 0 and 1.(Depending on the number of classes in your dataset, entropy can be greater than 1 but it means the same thing , a very high level of disorder. For the sake of simplicity, the examples in this blog will have entropy between 0 and 1).

Take a look at this graph below.

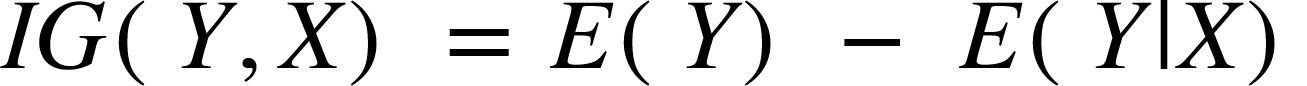


The x-axis measures the proportion of data points belonging to the positive class in each bubble and the y-axis axis measures their respective entropies. Right away, you can see the inverted ‘U’ shape of the graph. Entropy is lowest at the extremes, when the bubble either contains no positive instances or only positive instances. That is, when the bubble is pure the disorder is 0. Entropy is highest in the middle when the bubble is evenly split between positive and negative instances. Extreme disorder , because there is no majority.

When the p(+) is 0.5 entropy is very high which is equal to 1

Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty.

Now we know how to measure disorder. Next we need a metric to measure the reduction of this disorder in our target variable/class given additional information ( features/independent variables) about it. This is where Information Gain comes in. Mathematically it can be written as:



**Information Gain from X on Y**

We simply subtract the entropy of Y given X from the entropy of just Y to calculate the reduction of uncertainty about Y given an additional piece of information X about Y. This is called Information Gain. So, higher the value of information gain, lower the value of Entropy and variable with highest IG will be considered as a root node.