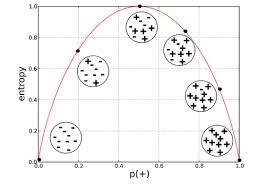
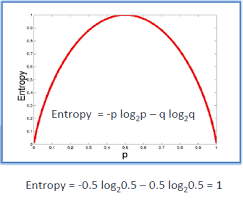
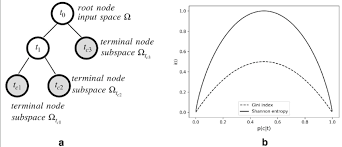
***Tree 1 -> F1 is the root, F2 and F3 are the leaves (F1- 9y/5n , F2 – 6y/2n, F3 – 3y/3n)***

1. **Entropy (Shannon Entropy)** ([Video](https://www.youtube.com/watch?v=1IQOtJ4NI_0))
   1. Helps in calculating the purity of the subset i.e. it helps in understanding whether the subset is pure or not
   2. Splitting in a D-Tree continues till a pure subset is achieved
   3. Value ranges from **0 to 1**, **Entropy = 1** means **completely impure** i.e. **50%** of both classes, pure subset at **Entropy = 0** i.e. just a single class exists in the subset



* 1. **Entropy Formula** : H(S) = -P(A)\*log2 P(A) - P(B)\*log2 P(B) {For a 2 o/p class subset}
     1. Here H(F2) = -6/8\*log26/8 – 2/8\*log22/8 = 0.81
     2. H(F1) = 0.94
     3. H(F3) = 1

1. **Information Gain** ([Video](https://www.youtube.com/watch?v=FuTRucXB9rA&t=626s))
   1. It is used for deciding what feature is to be used as a root node and what is to be used as leave nodes
   2. Higher the Information Gain , better the split
   3. Formula for calculating Information Gain :
      1. Gain(S,A) = H(S) - { [∑|Svalue|/|S|]\*H(Svalue) }
      2. 0.91 – (8/14)\*0.81 – (6/14)\*1
      3. It can be understood as a method of identifying the split which results in the maximum entropy reduction so that a system of purity can be reached
2. **Gini Impurity** ([Video](https://www.youtube.com/watch?v=5aIFgrrTqOw))
   1. Helps in calculating the purity of the subset i.e. it helps in understanding whether the subset is pure or not
   2. Splitting in a D-Tree continues till a pure subset is achieved
   3. GI = 1 - ∑(p^2) = 1 – [ P(A)^2 + P(B)^2]
   4. It is preferred over Entropy because it is **computationally efficient**
      1. **No Log**
      2. **Max Value till 0.5**

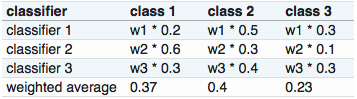


1. Splitting Numerical Data ([Video](https://www.youtube.com/watch?v=5O8HvA9pMew&t=18s))
   1. All the values in the column will be sorted in ASC
   2. A threshold value will be calculated i.e. let’s say a column has data ranging from

1 to 10, One tree would have threshold as 1 other 2 other 3..and so on

* + 1. Hence **numerical attributes make the decision tree very slow** because for every scenario a tree would be constructed and information gain would be calculated accordingly hence the one with max gain is selected

1. Measures used for splitting a tree
   1. ID3 (Iterative Dichotomiser 3)
      * Use Shannon Entropy function and Information gain as metrics only for Classification
      * C4.5 is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals.
      * Algorithm creates a multiway tree
   2. CART (Classification and Regression Trees)
      * Gini Index(Classification) as metric for Classification
      * Reduction in Variance : Used for continuous target variables (regression problems)
      * it constructs binary trees.
      * Obtained tree is pruned by cost-complexity Pruning.
2. Ensemble Techniques ([Video](https://www.youtube.com/watch?v=KIOeZ5cFZ50))
   1. It means combining multiple models
   2. Bagging (Bootstrap Aggregation)
      1. **Row Sampling with Replacement**: Initial Dataset D on a part of which **multiple different models** will be trained and tested [-> *Bootstrap*] and a score will be generated or an output class will be declared (in case of classification problems), the mean/median score of all the scores and the highly voted output will be selected in respective scenarios of regression and classification [->*Aggregation*]
      2. ***Note- “With Replacement”*** *means Dataset shared to 2 models is different, there might be some repeated records but not entirely the same*
      3. **Random Forest**
         * **Row Sampling with Replacement and Feature Sampling with Replacement**: Initial Dataset D on a part of which **multiple** **Decision Tree models** will be trained and tested [-> *Bootstrap*] and a score will be generated or an output class will be declared (in case of classification problems), the mean/median score of all the scores and the highly voted output will be selected in respective scenarios of regression and classification [->*Aggregation*]
         * Whenever a Decision Tree is developed to it’s full depth it is highly overfitted ( Low Bias, High Variance) ; **Since in Random Forest the output which is taken is of multiple DTrees (by voting) , High Variance is transformed to Low Variance**
         * Random Forest can also be used to select the most important features ([Article](https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f))
      4. Hard Voting vs Soft Voting
         * For Bagging, the final output is selected using voting :
           + **Hard Voting:** The predicted class label for a particular sample is the class label that **represents the majority (mode) of the class labels predicted by each individual classifier**.
           + **Soft Voting**: the output class is the prediction **based on the weighted average of probability given to that class**. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the **average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier**.
           + Another Example :



1. Pruning
   1. We need to prune the tree in such a way to reduce the prediction error rate by dealing with Overfitting.
   2. Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.
   3. The dual goal of pruning is the ***reduction complexity of the final classifier as well as better predictive accuracy by the reduction of over-fitting and removal of sections of a classifier that may be based on noisy or erroneous data.***
   4. Pruning is carried out from the leaves to the root.