A decision tree leads to a prediction by asking a series of questions on whether you belong to certain groups. Each question must only have 2 possible responses, such as “yes” versus “no”. You start at the top question, called the root node, then move through the tree branches according to which groups you belong to, until you reach a leaf node. Decision trees are versatile, as they can handle questions about categorical groupings (e.g. male vs. female) or about continuous values (e.g. income). If the question is about a continuous value, it can be split into groups – for instance, comparing values which are “above average” versus “below average”.

The most important step in a decision tree is choose the root node, the primary attribute to split the data on. This is decided using a techniques called information gain/entropy, which answers the question of which node to split on- the node that best splits the dataset.

In this example we use the student pass data set from UCI repository. We pre-process the data to produce a final grad pass/fail (1/0). This is done using the rule (G1+G2+G3)>35. I choose 35 here because if you look at the dataset, that would more or less split the dataset into equal number of +/- examples.

We use the scikit learn library here.

The dataset is split into training and testing sets using an 80-20 divide. We use a cross validation of set 5 to make sure we just obtain the result by chance.

The tunable parameter of max tree depth can be found out using a for loop and then plotting the results which one has the least error.

Random forest is basically a ensemble classifier- a forest consisting of multiple trees. For he final result we take the majority vote from multiple tree outputs.

For this experiment we used the Caltech vision Birds data set. The goal is to predict the bird species given the features.

The dataset is pivoted so that each image is a row and the columns are the attributes and then joined with the class labels for the output.

Same as for the decision tree we split the dataset randomly into training and test sets. The model is trained using the random forest classifier, we use a loop to decide the best parameters-how many features each tree has and how many trees this forest will have.

The result on the test is not that promising and there are lots of confusion. On plotting and zooming in on the confusion matrix one can see that the most confusions occur for birds of the same species for example Brewer\_Sparrow and Chipping\_Sparrow etc. However things like Albatross or Bobolink etc are not confused.

We compare the same set using a decision tree and the results are much worse as compared to the random forest results.

Decision trees are prone to overfitting, especially when a tree is particularly deep, however they are easy to interpret and make for straightforward visualizations. In a Random Forest, the features are randomly selected in each decision split. The correlation between trees is reduces by randomly selecting the features which improves the prediction power and results in higher efficiency, this random selecting of features helps to overcome the problem of overfitting.