**Decision Trees and Random Forests**

External Resource #1 - <https://www.youtube.com/watch?v=DCZ3tsQIoGU> (decision trees)

External Resource #2 – <https://www.youtube.com/watch?v=D_2LkhMJcfY> (random forests)

Decision Trees

Decision trees are a relatively simple algorithm that splits data by “asking” a series of questions. In classification (which is what I will be needing for my project), it splits the data by about half at each question, so that when a test sample is passed through, it can easily classify it. In machine learning and the built-in algorithms, it takes the features and splits them along the axis so that the two resulting groups will be around the same size. The most notable pros and cons are listed below:

* Pros
  + Does not take a long time, since it is just repeatedly splitting the data into two groups
  + Easy to understand intuitively
* Cons
  + Can easily overfit the data by asking too many questions. Thus it is key to find a “good” stopping point, which would be a hyperparameter, but even then that hyperparameter only works on the training and validation datasets and not the test data
  + Outliers wreak havoc on the dataset, especially when there are few samples in the training dataset and the outlier of one class is located “inside” another class. The algorithm will get bogged down trying to classify those few ones, when in fact the analyst is likely looking for the “bigger picture”
  + It is not very good at regression, since the questions only break down the dataset into groups rather than finding some sort of continuous function

Random Forests

Random forests are essentially aggregated groups of decision trees. The algorithm will run a bunch of randomized decision trees on slightly different sections of the training data (they will and should overlap, but should not be exactly the same) that will split the data slightly differently, and then makes a composite out of them. For classification, it will determine the class of a sample based on the “majority vote” of the trees, and for regression, it will average the outputs of each of the decision trees.

* Pros
  + Much better at not overfitting the data
  + Not as affected by outliers as decision trees individually
  + Much better at regression and will come closer to a “smooth” line than a bunch of groups
* Cons
  + Takes much longer
  + Still not as good at regression as classification
  + The analyst cannot control the model almost at all