5)

The importance of a feature is derived from the probability of reaching the node that makes a decision about the feature summed over all trees in the ensemble. The probability for a node is derived by counting how many samples reach that node. (https://stackoverflow.com/questions/15810339/how-are-feature-importances-in-randomforestclassifier-determined)

6)

To test the random forest classifier, we used the python sci-kit learn implementation on the wine quality dataset. This is a quiet famous dataset from the UCI machine learning repository. The dataset is available for red and white wines, but we only focused at the white wines. The goal for the dataset is to predict the quality of a wine on a discrete scale from 0 to 10. Each integer can be seen as its own class. The dataset has 11 variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) that can be used to predict/classify the quality of the wine. The accuracy of our classifier is 0.623, using cross-validation (test size = .4). The most important predictors are: free sulfur oxide(0.10), density(0.11) and alcohol(0.11). We compared our results with a SVM. Not only does the SVM have a lower accuracy(0.509) it also takes significantly more time than the random forest.