Random Forests Quiz

Q1. Say a parent node has 3 samples and makes a split with a Gini Impurity of .444. How many samples does each child node have?

Answer: 2 Samples and 1 Sample

Q2. Why are the decision boundaries created by Random Forests often smoother than those created by Decision Trees?

Answer: Since the decisions made by the various trees are averaged, the linear boundaries created by each single tree will be smoothed in aggreggation

Q3. Why can Random Forests be used on both Classification and Regression problems?

Answer: Although classification requires categorical answers and regression numerical, since random forests relies on splitting up datapoints to make decisions, as long as the categorical data can be made numerical (say through assigning categories ids), then Random Forest can function with issue.

Q4. What are the two random process that Random Forests utilize?

a) Bagging, Feature Randomization
b) Random Depths, Bagging
c) Feature Randomization, Random Depths
d) Random Test-Train Splits, Feature Randomization

Q5. Calculate the Gini Impurity and Entropy of a split of 72 samples into daughter nodes of 51 and 21 samples

Answer: Gini Impurity: ~.413, Entropy: ~.87

Q6. If you indefinitely continue to increase the number of Random Trees, would the accuracy of the model continue to rise.

Answer: No, there is only a certain number of splits you can make given a number of a finite dataset. Once the number of random trees is greater than the number of datapoints, the accuracy cannot rise.

Q7. Say you are using Scikit-Learn's DecisionTreeClassifier and Train_Test_Split. You desire to ensure that the decision tree you create on this machine can be perfectly replicated on other machines, which parameter would most help you acheive this? Why?

*Answer: This would be achieved with setting the random state parameter. This is becuse DTC relies on random processes for Bagging and Feature Selection. Test_Train_Split relies on random processes to select features.