CAP 5610 Assignment 03
Student: Lam Nguyen
NID: la815794

Problem 1 (Written Question 5 points)
What are the advantages of using a Random Forest over a single decision tree?

A single decision tree has a tendency to overfit training data. A random forest is made up of multiple decision trees; each tree fitting a randomized subsample of data. Each of these decision trees can then be averaged together to create a prediction with much more accuracy since the overfitting of each tree can be balanced out by averaging. Multiple decision trees reduce variance and overfitting compared to a single decision tree.

Problem 2 (Written Question 5 points)
What is the difference between Random Forest and Gradient Boosting?

Both Random Forest and Gradient Boosting both use decision trees. However a Random Forest Classifier uses decision trees that are NOT correlated in order to make a decision. The results of one decision tree does not change the results of another decision tree within the random forest. In contrast, Gradient Boosting uses correlated Decision Tree. The results of the predictions of the initial decision tree will be attempted to be improved upon by the next Decision tree in order to attempt to reduce error.

Problem 2 (Coding Question 30 points)
For the Titanic challenge we need to guess whether the individuals from the test dataset had survived or not. Please:
1) Preprocess your Titanic training data; (3 points)
2) Select a set of important features. Please show your selected features and explain how you perform feature selection. (3 points)
3) Learn a decision tree model with the Titanic training data using Gini index, plot your decision tree; (4 points)
4) Apply the five-fold cross validation of the decision tree learning algorithm to the Titanic training data to extract average classification accuracy (using max_depth=10); (5 points)
5) Apply the five-fold cross validation of the random forest learning algorithm to the Titanic training data to extract average classification accuracy (using n_estimators=200); (5 points)
6) Which algorithm is better, Decision Tree or Random Forest? (5 points)
7) What are your observations and conclusions from the algorithm comparison and analysis? (5 points)