Problem 1 (Written Question 5 points)

What are the advantages of using a Random Forest over a single decision tree?

Problem 2 (Written Question 5 points)

What is the difference between Random Forest and Gradient Boosting?

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)