# Home Work # 6. AMS 597

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# The homework is due on Monday, April 27, 2020, at 12noon. Solutions will be emailed to you by 1pm. The homework will not be graded and all complete submission will warrant full points. Quiz 7 will be given on Tuesday, April 28, 2020, in class, covering the same materials – for the entire class duration; and will be graded. So please study hard from now on. Please email your completed homework to your TA at: [song.jiecheng@stonybrook.edu](mailto:song.jiecheng@stonybrook.edu)

# Please include (1) R code

# (2) Output from R

# (3) Answers to all the questions asked

#### Random Forest with the Spam Data – Classification Task

How to separate Spam e-mails from non-Spam e-mails? Thanks to the advancement in text mining, we can now readily generate a set of relevant numerical attributes to help in such classification tasks. The Spambase.data we will use for our homework is taken from the UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/spambase>). **Y**our task is to split the data randomly into training (75%) and testing (25%), first build the best random forest to predict Spam e-mails using the training data, then use the out-of-bag (OOB) data to measure its performance, and then use this random forest model to predict whether each e-mail in the testing data is Spam or not. Please use the *randomforest* function in R to build the random forest classifier.

Please review the following website for related methods and concepts:

<http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/140-bagging-and-random-forest-essentials/>

1. Please use the random seed 123 to divide the cleaned data into 75% training and 25% testing.

1. Please first build the random forest to predict Spam e-mail using the training data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy using the out of bag (OOB) samples.
2. Next please use this random forest to predict whether each email in the testing data is Spam or not. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.
3. Please plot the variables importance measures using
   1. *MeanDecreaseAccuracy*, which is the average decrease of model accuracy in predicting the outcome of the out-of-bag samples when a specific variable is excluded from the model.
   2. *MeanDecreaseGini*, which is the average decrease in node impurity that results from splits over that variable. The Gini impurity index is only used for classification problem.
4. Please show the importance of each variable in percentage based on *MeanDecreaseAccuracy*.
5. In a regression task using the random forest, suppose we have 26 variables (as predictors) in the original data set – then at each node split, what is the number of variables we should (as commonly recommended) to select, at random, to be considered for that node split?

