# Quiz # 5. AMS 580

**Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dear all, you have the entire lecture time to do this quiz. It is open book however you must do the work entirely on your own. Please turn on your video and mute your audio.**

**Please submit your Quiz in two documents: (1). The Rmd file, and (2). The output file (word or pdf) – to the SBU Brightspace. If you have difficulty submitting to Brightspace, you may then email the solutions (with the subject of “Quiz 5, AMS 580”, and in one email with the two attachments), to your TA Ian at:** [**weihao.wang@stonybrook.edu**](mailto:weihao.wang@stonybrook.edu)

#### Part I. Random Forest with the GreatUnknown Data – Classification Task

The **GreatUnknown.csv** data contain 12 predictors and one binary response variable **y** (= 0 or 1), which is the true class label**.**

For this dataset, **sensitivity** is defined as a case labeled 1 being classified to label 1, while **specificity** is defined as a case labeled 0 being classified to label 0.

1. For the entire dataset, please perform the data cleaning as instructed before; namely, delete observations with missing value(s). Please report how many cases (namely, data points) are left after this step. Then please use the random seed 123 to divide the cleaned data into 75% training and 25% testing.
2. Please first build the best random forest to predict ‘**y**’ 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.
3. Next please use this random forest to predict ‘**y**’ in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.
4. 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.
5. Please show the importance of each variable in percentage based on *MeanDecreaseAccuracy*.

#### Part II. Random Forest with the QuestionMark Data – Regression Task

The **QuestionMark.csv** data contain 14 predictors and one continuous response variable **y**.

Our goal is to establish a regression model predicting the response variable using the predictors – and in particular, using various modern and classical techniques to perform variable selection.

1. Please perform data cleaning by checking whether there are any missing values and if so, please delete observations with missing values. Please report how many observations with missing values we have in our dataset. Please use the random seed 123 to divide the data into 95% training and 5% testing.
2. Please first build the best random forest to predict ‘**y**’ using the training data. Please use 10-fold cross-validation to obtain the best tuning parameter **mtry**, which is the best number of random variables to select (without replacement) for each tree node.
3. Next please use this random forest to predict ‘**y**’ in the testing data. Please compute the **RMSE** for the testing data.
4. Please use the **randomForest** function to refit the training data using the optimal **mtry** parameter and with 500 trees.
   1. Please print out the model summary, and write down the **RMSE ()** of the **out of bag (OOB)** samples.
   2. Show the variable importance values.
   3. Plot the variables importance measures using **MSE** and **Node Impurity**.

A picture containing text, outdoor, water

Description automatically generated