Random Forest. Homework #5

Name:_____SBU ID:_____

Please include (1) Rmd file; (2) Output from Rmd with answers to all the questions asked

Please upload your homework solutions to the Brightspace by 8:30am on Wednesday, March 8, 2023.

Random Forest with the Titanic Data – Classification Task

The Titanic.csv data we will use for our homework is taken from the Kaggle competition site (https://www.kaggle.com/c/titanic) where it was called the train.csv. **We will treat this dataset as our entire data** because we do not know the survival status in the Kaggle test.csv data. Our Titanic data has 891 passengers and 12 variables:

- *PassengerId*: Passenger ID: 1- 891
- *Survived*: A binary variable indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our response variable
- *Pclass*: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
- *Name*: A field rich in information as it contains title and family names
- *Sex*: male/female
- *Age*: Age, asignificant portion of values aremissing
- *SibSp*: Number of siblings/spouses aboard
- *Parch*: Number of parents/children aboard
- *Ticket*: Ticket number.
- *Fare*: Passenger fare (British Pound).
- *Cabin*: Cabin number
- *Embarked*: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)

First, one must clean the data and decide which variables to exclude from our analysis. My recommendation is that we exclude *PassengerId*, *Name*, *Ticket*, and *Cabin* in the ensuing analysis. Next, please note that *Age* has many missing values – my suggestion is to delete those with missing values. Now after the data cleaning step, your task is to split the data randomly into training (75%) and testing (25%), first build the best random forest to predict passenger survival using the training data, then use the out-of-bag (OOB) data to measure its performance, and then use that model to predict whether each passenger in the testing data survived or not. Please use *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. For the entire dataset, please perform the data cleaning as instructed before; namely, exclude the variables *Name*, *Ticket*, and *Cabin* and delete missing values in the variable *Age*. Please report

how many passengers 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 passenger survival 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 the survival of passengers 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
    a. *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.
    b. *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*.

6. In a classification task using the random forest, suppose we have 36 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?