**Homework**

**Performance measurement, Imbalance correction and Imputing missing values**

1. Load the Wisconsin breast cancer dataset from blackboard
2. How many missing data points are there and in which columns are they missing?
3. Impute for the first column with missing values using the mean and round to an integer
4. Impute for the second column with missing values using KNN with three nearest neighbors and round to integer
5. Impute for the third column with missing values using a regression to predict the third column based on uniformity\_of\_cell\_shape, marginal\_adhesion and normal\_nucleoli. Round to integers.
6. Build a decision tree model to predict class using 80% training data and five fold cross validation with three repeats. Hint: Use method =” rpart” in caret
7. Show the accuracy results for each resample (fold) in the training data
8. Do the following performance measures for the test dataset
   1. Compute all the accuracy measures (Accuracy, sensitivity, specificity etc)
   2. Plot the ROC curve
   3. Plot the lift curve
   4. Compute the AUC
9. Is the cancer data imbalanced by the class feature? What is the percentage of the majority class and the percentage of the minority class?
10. Now re-build the decision tree model above after correcting for imbalance using SMOTE.
11. Did any of the accuracy measures improve? If so, which ones?

**Data Wrangling**

1. Install and load the library nycflights13
2. Load the datasets/tables flights and airlines
3. Add full airline name from the airlines table to the flights table that keeps all the records in the flights table by using the appropriate join
4. Now add the destination latitude and longitude to the flights table from the airports table by using the appropriate join