Course Name: Data Science: Process and Tools (DATASCI400)

1. **Project Problem**: Create a Data Flow Diagram to model the task of helping a group of investors open a new restaurant in the greater Seattle Area.
   1. Model type: Date Flow Diagram
   2. Data Source: Yelp (as a model, not actually done as part of data flow diagram)
   3. No data source, just an exercise in creating a data flow diagram
   4. Final Conclusion/Result – this was a filtering problem rather than a predictive problem
      1. Method(s) used: NA
      2. Evaluation metric (accuracy, precision, etc): NA
2. **Project Problem**: Prepare a data set for Machine Learning
   1. Model Type: none – just cleaning up the data
   2. Data Source: UCI Heart Disease
   3. URL : data (url: <http://archive.ics.uci.edu/ml/datasets/heart+Disease>)
   4. Final Conclusion/Result: outliers and aberrant data were removed. Values were normalized to 1. Categorical data was binned. New categorical variables were constructed. The old categorical data was removed
      1. Methods: use mean +/- 2 standard deviations to remove outliers
      2. Methods: remove NAs
      3. Normalize data (value – mean)/standard deviation
      4. Binned data turned to categorical variables
      5. Evaluation: goals achieved
3. **Project Problem**: Predict heart attacks using heart disease data
   1. **Model Type**: KNN and SVM
   2. Data Source: UCI Heart Disease
   3. URL : data (url: <http://archive.ics.uci.edu/ml/datasets/heart+Disease>)
   4. Final Conclusion/Result: ﻿SVM achieved better results than Random Forest

For the SVM - TP, TN, FP, FN: 57 , 140 , 15 , 23

Accuracy rate: 0.8382978723404255

Error rate: 0.16170212765957448

Precision: 0.79

Recall: 0.71

F1 score: 0.75

AUC 0.9

﻿

For Random Forest

Accuracy rate: 0.7787234042553192

Error rate: 0.16170212765957448

Precision: 0.67

Recall: 0.7

F1 score: 0.68

TP rates: [0. 0.2 0.44 0.7 0.86 0.94 0.98 1. ]

FP rates: [0. 0.04 0.12 0.18 0.35 0.55 0.78 1. ]

Probability thresholds: [2. 1. 0.83 0.67 0.5 0.33 0.17 0. ]