**Project (Due May 19th)**

In this project you will use the methods taught in this class to analyze a dataset. The dataset you analyze must have a binary qualitative $y$-variable.

**DATA**

Datasets can be obtained from the data.world repository found here:

https://data.world/datasets/classification

- Note that the dataset you choose must meet the following criteria:
  1) It has a column to use as a qualitative response (y) variable which has EXACTLY TWO values.
  2) It has at least two quantitative columns that could be used as predictor variables.
  3) It has at least four columns that could be used as predictor variables total.

- Note that in the link above, the available datasets have been pre-filtered to include "Classification"-appropriate datasets, however a vast number of datasets in the link I sent will not meet these criteria.

  **YOU MUST EMAIL ME the names of five possible Datasets you are interested in studying.**

**I WILL NOT ALLOW THE SAME DATASET TO BE USED BY MORE THAN ONE PERSON. The datasets will be rationed out on a first-email first-serve basis. I will let you know which dataset you have once I received your email.**

**ASSIGNMENT**

You are to come up with a well-selected model to predict the class of a binary $y$-variable. In this case, **well-selected** means that it is the best-performing model out of many that were tried. The write-up must consist of the following sections:

**Introduction/Description:** Discuss what the dataset is and what it consists of and what the goal of analyzing it is.

**Methodology:** Describe the different models/methods that were tried through the course of the analysis and how you model-performance was assessed. Do not state any results in this section.

**Analysis:** Describe what the outcome of each model/analysis was, including relevant diagrams and output. You are not responsible for using every topic from class, however most should be covered.

**Conclusion:** Describe the model that we should use for predicting the class of the $y$-variable and assess its performance one final time.

**Appendix:** Include all code used in the analysis along with some description of what the code is doing.

**Project (Due May 19th)**

**Required Analysis:** **You must use the following Model Types in your Analysis:**

Logistic Regression.

- Note that backward and forward variable selection must be carried out using AIC/BIC and cv-prediction error. For each subset of variables that is found to be optimal by these methods, you must perform 4-fold cross-validation misclassification rate as per the "Required Model Assessment" section below.
- Using all available predictor variable columns obtain principle components for the quantitative predictor variables. Use the optimal number of principle components as well as the remaining qualitative variable in a new model and obtain the 4-fold cross-validation misclassification rate for that model.
- In each case, you must clearly identify the optimal models by writing out the estimated model so that it is clear what predictor variables are used.

Decision Trees.

- You must obtain 3 decision trees based on varying the control parameter. One of the decision trees must be "optimal" as it pertains to XERROR we discussed. Compute the 4-fold cross-validation misclassification rate for this model as per the "Required Model Assessment" section below. Identify the optimal model by stating the control parameter used and including the graph.

Support Vector Machines

- You must try 4 svm models based on varying the 'cost' and 'kernel' parameters (two different values of each). Compute the 4-fold cross-validation misclassification rate for this model as per the "Required Model Assessment" section below. Identify the optimal one of these.

K-nearest Neighbors

- You must try 50 different values for the number of neighbors used, computing 4-fold cross-validation misclassification rate for each. State the optimal number of neighbors used.

Cluster Analysis

- Try 4 different hierarchical models based on varying the distance measure used between observations and distance measure used for the clusters. Compute the misclassification rate for each model on the entire dataset (no need to do cross-validation here).

Random Forests

- Obtain 1 Random Forest Model by creating 1000 trees where we randomly select 50% of the variables in each iteration. Compute the misclassification rate according to this Random Forest Model on the entire dataset (no need to do cross-validation here).

**Project (Due May 19th)**

***Required Model Assessment:***

Your models must be assessed **using 4-fold cross-validation.** *If the number of observations in your dataset is not divisible by 4, you can remove up to 3 observations from the end in order to get it divisible by 4.*

**Advice:**

Start early.

Do not write a lot. This is a technical report and simply requires you to (succinctly) state the analysis you did, the output you observe, and the conclusions you draw. You should not be turning lines of code anywhere in the body of the assignment (aside from the appendix). Refer to each model using notation, name, and the values of the parameters.