Math 5364 Final Project

1. You may work in groups of up to three people, and each group should turn in a single, professional quality report detailing your findings. At the end of your report, provide a breakdown of what each group member did. Each member is expected to contribute significantly to the project.

2. Each group will also give a brief presentation (around 10 to 20 minutes) on their project in November or December.

3. The goal of this project is to build a statistical model for predicting some variable Y from one or more variables $X_1, \ldots, X_p$ based on real data.

4. A good place to find data sets to work on is kaggle.com or the UCI machine learning repository. There are also a large number of data sets included in R.

5. The problem you work on can be a classification problem (categorical dependent variable) or a regression problem (quantitative dependent variable).

6. If you choose to work on a classification problem, build multiple classifiers using all of the techniques we have discussed (decision trees, k-nearest neighbors, naive Bayes, neural networks, support vector machines, random forests, bagging, and boosting).

7. If your project is a regression problem, you should also build multiple models using a variety of techniques. Many of the algorithms we have learned about can be adapted to regression problems, including decision trees, neural networks, support vector regression, and random forests. Naive Bayes and k-nearest neighbors can be applied to regression problems by discretizing the dependent variable.

8. All models built should be tuned and thoroughly assessed using appropriate metrics. Classifiers can be assessed with classification accuracy, sensitivity, specificity, precision, recall, and area under the ROC curve. Regression models can be assessed with root mean square error, RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$ and mean absolute error, MAE $= \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$