Assignment1

**1. Find a data set to work on.**

Find a unique dataset. Many tutorials use standard datasets like the iris, wine, etc. Try to be unique with your selection and choose something different (there are many possible datasets - therefore, if you try to find a unique dataset, the likelihood of two or more students choosing the dataset is very slim). If the dataset has been used in some online tutorial or Kaggle competition, other students may likely select this. Your mark will partially depend on the uniqueness of the dataset you choose. If you wish to 'claim' a dataset, email the professor and TA details about your chosen dataset. To be valid, your email must contain a header that indicates the course and the section you're in, and the contents of the email should include details about the data and where it will be obtained (with a link to where it resides).

**2. Create a notebook that pre-processes this data for model fitting**

In your notebook, analyze and process your chosen data. Identify your target variable and your input variables. Be sure to include details about what you observed, what changes you are making, how you are making these changes, and why you are making these changes. For instance, you might have a section in your notebook where you analyze the data to identify issues with missing data. Present your results, then argue for what (if anything) you need to do and why, and then show clearly the steps you undertook to accomplish this. Save the results into csv files (these files should therefore be pre-processed and ready for model fitting. Later model fitting notebooks should not need data manipulation/processing.

**3. Create a notebook within which you identify a high-quality predictive model using the pre-processed data.**

Identify what performance metrics you will use to measure the performance of your model(s). Explain why you've chosen this metric. You will fit a logistic regression, an SVM, and a decision tree model for modeling.  You will use random and grid searches to test a range of parameter values for each model. Your analysis must identify and save (for use in other programs later) the best-performing model you have found.

**4. Submit your analysis**

This must include your original data, data preprocessing notebook, and model fitting notebook. When evaluating your work, we will run the preprocessing notebook, assess the results, and then run the model fitting notebook (which must read the files output from the preprocessing python notebook).