Stat 6910 Final Project

Jakob Johnson and Jennifer Vance

**Data/Problem:** For the final project, we have chosen a dataset that contains statistics about European soccer matches from 2008 to 2016. It also contains information, from this same timeframe, about teams, individual players, leagues, and countries. Our goal is to predict the outcome of soccer matches given information about the game. This is useful because it could be used to predict the outcome of a game currently in progress or the attributes of a game that tend to lead to a win. Also, incorporating individual player data, we might be able to predict which players will perform better in fantasy leagues. The dataset is downloaded as an SQLite database, and contains an enormous amount of information. Most of which will be very difficult to parse and may not be useful for predictions.

<https://www.kaggle.com/hugomathien/soccer/>

**Methods:** This is a classification problem with three outcomes (home win, away win, draw). We will use Decision Trees, Random Forests, and Neural Nets. We will also try a “Super Learner” on top of the other models to see if we can further increase accuracy.

**Division of Work:** Not sure exactly who will be doing what, but Jakob will be doing Neural Nets and Jennifer will be doing Random Forests. I think that most of the work, as well as most of the predictive accuracy increases will come from data cleaning and clever variable uses, such as foul types or something.