**CS 395**

**Project Part 1**

**Eric Lipe - Kevin Ritter**

We chose the Players Underground dataset from Kaggle.com. There are 65,000 games played in the dataset with 100 players per game. Within each game statistics are collected for each player covering things such as damage dealt, head shots, kill place, ranking, team kills, etc. Each player is assigned an ID number and win place percentage to be used for evaluating the predictions.

The data within the dataset is represented in a few ways. The brunt of the dataset’s features are ratio data types. Out of the 29 features, 24 of them are ratio data types. The other types present in this dataset are interval and nominal. Four of the remaining five features are of type nominal and the one remaining feature is of type interval. The ratio features will technically have a range [0, ∞). These features will never be able to get to such a large value in practice, but theoretically they could. Three of the four nominal data types will have as many values as there are inputs. This is because these are identifiers and each one is unique to each input. The remaining nominal data type will be able to take one of six specific categories. The interval data type will take values [-1, ∞). This is because if data was missing for an input in this feature a ‘-1’ was substituted in.

* Due to the number of features present in the dataset we are going to link to Kaggle where features and data can be viewed within the browser.

<https://www.kaggle.com/c/pubg-finish-placement-prediction/data>

Transforming the data.

Preprocessing the data will be a big task in evaluating any model we build around the dataset. For all the ratio features we will be normalizing them. This will make the model perform better, but it will also make describing the results more interpretable. Based on the information available on the data, we may end up removing the feature that is of type interval. However, if we choose to keep it, we will either replace all the ‘-1’ values with the mean of the feature or simply replace them with a zero. This will convert the columns data type to ratio which we will then normalize to make interpretability better. For the nominal data types, we will completely remove all the features that are identifiers. These features will give no valuable insight into our target predictions. The remaining nominal data type will need to be one hot encoded. We need to do this so that this feature can be one of six values, where each of these values is categorical in nature.

The goal of the project is first to train a neural network to predict the winning percentages for each player given the features provided. The next phase of the project will be to reduce the features used in the model to determine if any, and how many of the features are unnecessary for the prediction. That is, how can we refine and simplify the model and still make an accurate prediction for each player. We then would like to use the model to predict winners in real time, given data streaming from a game.