**Predicting outcomes and scores for EPL**

* **Problem Definition:** Predicting the football match outcomes and scores for English Premier League. The objective of this model is to predict the outcomes and scores for English Premier League (EPL) by using Data Mining and Machine learning techniques. By using these approaches, the most optimal model will be able to provide relative accuracy.
* **Background/literature review:** The EPL is the most popular and best soccer league in the world. It had the most lucrative TV deals in all of soccer. Having a model with multiple data points will soccer enable clubs to game plan according for each opponent. Furthermore, due to its popularity a large gambling community has formed much of the interest. Thus, having a predictive model to determine the outcome of matches is a great idea. This will give those who wager on the matches an outlook as to the predictions of the upcoming games and which team to place their wager on.
* **Data mining technique:** Logistic regression, Decision Tree Classification and Support Vector Machine (SVM).

**Machining learning technique:** Artificial Neural Network (ANN)

* **Description of design:** For predicting the result, Logistic regression, Support Vector Machine (SVM), and Decision Tree Classification were used. Based on which model gives the best predictions, that model will be implemented. For predicting the scores, the most optimal model to use is Multiple Linear Regression.
* **Explanation of data (file format, independent/dependent variables, source)**

**File format:** The data is structured in a CSV file format, so it can be imported into the model.

**IDV:** Away Team Shots on Target (AST), Home Team Fouls (HF), Away Team Fouls (AF), Home Team Corners (HC), Away Team Corners (AC)

**DV**: Full Time Results (FTR), Full Time Home Team Goal (FTHG), Full Time Away Team Goal (FTAG),**Source:** <https://datahub.io/sports-data/english-premier-league#data>

* **Identification of Training and Testing data**

**Training set:** Data set is split in to 80% to create the training set i.e. (800 rows of 19 variables).

**Testing set:** Data set is split in to 20% to create the testing set i.e. (201 rows of 19 variables)

* **Testing and Validation Procedure:** After training the model with the training test. The model will predict the test set and the confusion matrix will be created to compare with the actual data. The model should predict scores of matches with relative certainty.
* **Results and incites:** The results from the test set had compared with the actual scores and the result from the data set**.** The ANN model was able to predict the outcome with 98% accuracy.
* **Citations**

##### Nivard van Wijk. (December 2012). Soccer Analytics Predicting the outcome of soccer matches.

##### Retrieved from <https://beta.vu.nl/nl/Images/werkstuk-wijk_tcm235-315158.pdf>

##### Aditya Srinivas Timmaraju. Game ON! Predicting English Premier League Match Outcomes

##### Retrievedfrom [http://cs229.stanford.edu/proj2013/TimmarajuPalnitkarKhannaGameON!PredictionOf EPLMatchOutcomes.pdf](http://cs229.stanford.edu/proj2013/TimmarajuPalnitkarKhannaGameON!PredictionOf%20EPLMatchOutcomes.pdf)