1. Choose your dataset

My project idea is to create an English Premier League (soccer) prediction bot that will bet using its prediction of the outcome of each game. I chose to use the dataset on the Premier League 18/19 from the following websites: http://www.football-data.co.uk/englandm.php and https://www.soccerstats.com/trends.asp?league=england. In fact, it provides detailed and insightful statistics about every game such as the number of goals each team scores in every game, the number of shots on target, the number of goals scored on free kicks, the number of goals scored on corners, and many more! All of those features provide important information about the chances of a team against another. In addition, the website updates its dataset weekly. Therefore, it is possible to have access to the latest statistic for a more precise and up to date prediction. The EPL is regarded as one of the best and most exciting leagues in the world due to its highly competitive and unpredictable nature. In parallel, it contains the top clubs and best players in the world which makes it very entertaining.

2. Methodology

The goal of this program is to predict one of the three outcomes of an EPL championship game. We want to determine the full time result which can either be a win, a draw or a loss. Therefore, it is a multinomial classification problem. A number of models can be used for this project such as logistic regression, SVM and XGBoost. I will train my dataset with these 3 models and choose the best one. After that, I will optimize the hyperparameters of the classifier that I will choose.

From the dataset, the most important information is the full time result of the past games. In addition, the data tells us the number of occasions each team has per game (i.e. # of corner kicks, # of free kicks, # of shots on target, etc.). Whether a team is home or away would always have an impact on the match since the team that is playing home gets more moral support and is less fatigue, because they don't need to travel. On the second website, information on the number of clean sheets per team as well as the likelihood of a team to score per game at "home" or "away" are given and updated weekly. These data will help me profile each team and its chance to win against other teams in the league.

Ultimately, I want to create a bot that analyzes each feature and bet automatically based on his prediction. However, since I have to consider the time factor, a webapp might be the alternative. It would take the user's entries (which will strictly be names of EPL teams) and output a prediction.