**INTRODUCTION**

In Europe, football is the most popular sports followed by millions of people, where most countries have football competitive leagues consisting of 20 teams each season. The top five European leagues are the English premier league, La-Liga, Ligue 1,Bundesliga, and Serie A. Football match has three possible outcomes in each game; away team win, home team win, or the game ends in a draw. Amongst the top five European leagues, the Italian Serie A is ranked third behind the English premier league and La Liga. Teams such as Juventus, Milan, Inter Milan, Napoli, and Roma have earned Serie A its treasured status due to their frequent European football participation against other top ranked football teams across different leagues.

**MOTIVATION**

Machine learning methods can be applied to different fields, including sports. On the example of English Premier League it is shown that it is possible to find a classifier that predicts the outcome of soccer matches with the precision of more than 60 percent Yezus A, (2014).There are lots of claims from multiple studies about the prediction. This project focus in the predictions accuracy of the results of the Serie A league by using various machine learning methods. This will fill the gap in the betting world for most people who are participating in the betting world. The project depending on the prediction accuracy, will help the participants of the betting world and the fans for the serie A league to have the expectations before the game starts using the algorithm. Predicting the outcomes of a football match has always been something spectators enjoy doing, by coming up with anticipated starting lineups, team form, and other different statistics to support their predictions. Similarly, football match betting has been carried out with a great deal of interest by fans, of which some of them rely on betting winning’s as their primary source of income. If there is a way to evaluate different football match statistics to assist in making predictions about match outcomes, it will be highly advantageous to the betters to help in making informed betting’s.

**DATA COLLECTION**

The dataset was collected from the following source: https:www.football\_data.co.uk/data.php.

Out of five high ranked leagues in Europe, we have selected one league (Serie A) to work with to implement the prediction algorithm. We have combined ten (10) seasons from season 2011/2012 to season 2020/2021 containing information of Serie A 20 teams. The datasets from football data were obtained individually as a csv file, each season file contained hundred-rand (100) features. Although the data contained 100 features, we only utilized 19 features.

**The features used are presented in the following table below alongside with their meaning:**

|  |  |
| --- | --- |
| HomeTeam | Name of the team playing home |
| AwayTeam | Name of the team visiting |
| FTHG | Full Time Home Team Goals |
| FTAG | Full Time Away Team Goals |
| FTR | Full Time Result |
| HTHG | Half Time Home Team Goals |
| HTAG | Half Time Away Team Goals |
| HS | Home Team Shots |
| AS | Away Team Shots |
| HST | Home Team Shots on Target |
| AST | Away Team Shots on Target |
| HC | Home Team Corners |
| AC | Away Team Corners |
| HF | Home Team Fouls Committed |
| AF | Away Team Fouls Committed |
| HY | Home Team Yellow Cards |
| AY | Away Team Yellow Cards |
| HR | Home Team Red Cards |
| AR | Away Team Red Cards |

**An overview of tools used in the process of implementing algorithm.**

1. Pandas \*loading the data, data wrangling and data manipulation

2. Matplotlib & Seaborn\* data visualizations

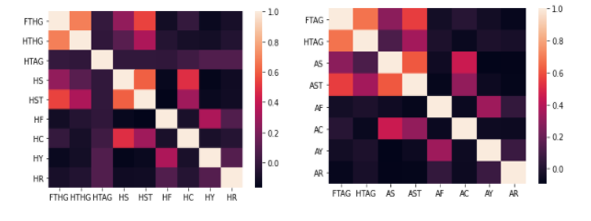
3. Xgboost\*libraries for classifier, metrics, and model evaluation

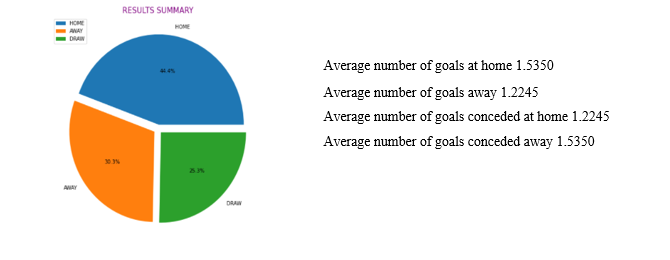
**DATA WRANGLING**

The seria A football database was imported and essential statistics on matches was read utilizing pandas . The data contains statistics about matches that have already played over several seasons, the data contains missing values of which the mode of the column containing missing values was used to fill the missing values. The results column is attached to the data frame -1 represent away team win, 0 represents draw, and 1 represent home team win based on the information provided by the full time results(FTR) from the data.

**EXPLANATORY ANALYSIS**

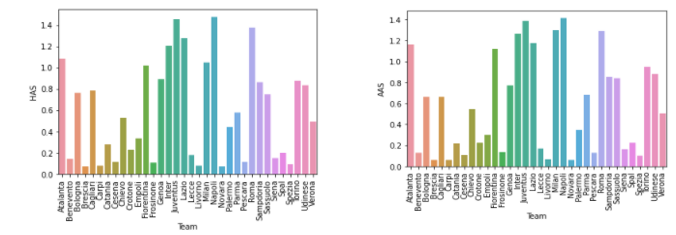
To begin, all features in the data were categorized according to home team status and away team status, then the home team features, and away team features were analyzed using the correlation metrics to investigation the strength of the relationship amongst the features. The pie chart was used to summaries the results of all 3787 football matches. The average number of goal scored by the away team and home team were computed to evaluate chances of scoring a goal as away team and as a home team given a particular match in Seria A.



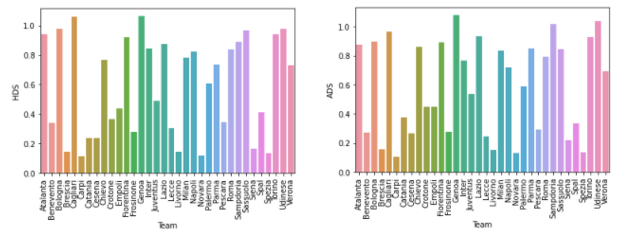


**HOME ADVANTAGE**

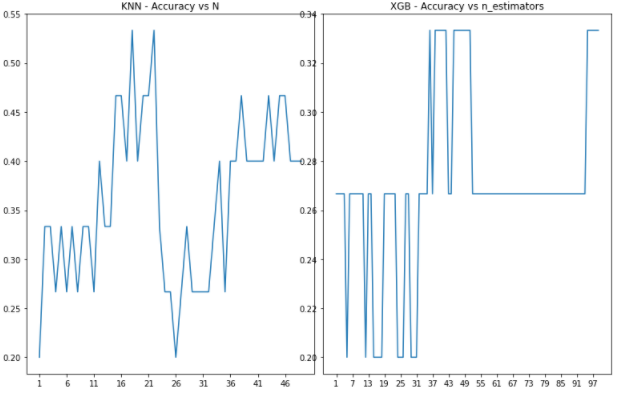
Looking at 34 different teams attacking strength, the attacking strength at home (HAS) is better than attacking strength away( AAS). The minimum attacking strength at home is 0.068811 for Novara and the maximum is 1.476002 for Napoli. The minimum attacking strength away is 0.0646974 for Novara and the maximum is 1.414706 for Napoli.



Checking the defensive strength at home (HDS) is better than defensive strength away( ADS). The minimum defensive strength at home is 0.112141 for Carpi and the maximum is 1.065344 for Genoa. The minimum defensive strength away is 0.106657 for Carpi and the maximum is 1.080337 for Genoa.



**RESULTS CLASSIFICATION**



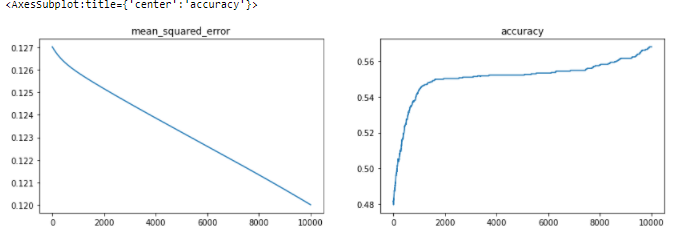
We added ten(10) new games with empty statistics including pending results, we aimed to predict the results of those games based on previous statistics using two classification algorithms namely, the Xgboost and K-nearest neighbors. Xgboost is the algorithm that has been dominating in applied machine learning for structured or tabular data whereas K-nearest neighbor is one of the simplest machine learning algorithm based on supervised learning technique. Xgboost classifier achieved 46.67% accuracy in classifying the results of the given new games, while the k-nearest neighbor classifier outperformed the Xgboost by achieving a fair accuracy of 53.33% .



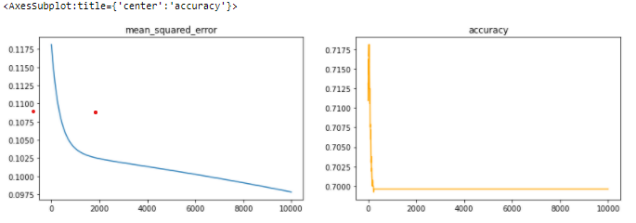
K-nearest neighbor classifier suggest that nine out of 10 home teams will win their new games whereas the Xgboost suggest that 7 out of 10 home teams will be victorious in their new games.



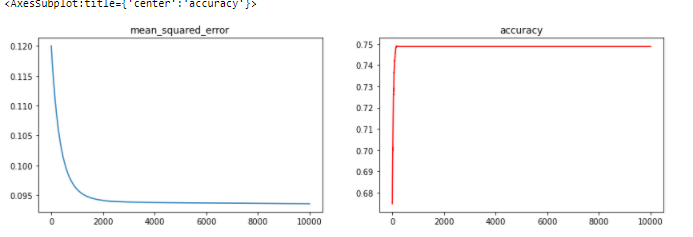
The neural network is used to train the data, then after we use the data to calculate the accuracy and the mean square error. From the data set we split the full time results column(FTR) into three columns namely, full time results for home win (FTR\_H), full time results draw (FTR\_D), and full time results for away win (FTR\_A). FTR\_H, FTR\_D, and FTR\_A were used as the response variables namely y1, y3, and y2, respectively. The remaining columns were used as explanatory variables .



**Results:** For FTR\_H we have used the learning rate of 0.01 with 10000 alterations to minimize the cost function as far as possible. The local minimum of a differentiated function is 0.120007 with the prediction accuracy of 56.79%.



**Results:** For FTR\_A we have used the learning rate of 0.01 with 10000 alterations to minimize the cost function as far as possible. The local minimum of a differentiated function is 0.097835 with the prediction accuracy of 69.96%.



**Result**: For FTR\_D we have used the learning rate of 0.01 with 10000 alterations to minimize the cost function as far as possible. The local minimum of a differentiated function is 0.093559 with the prediction accuracy of 74.87%.

CONCLUSION

Using previous data of ten seasons to predict future matches, we made use of classifiers (KNN and XGB) and the prediction using KNN showed 53.33% accuracy and XGB showed 46.67% accuracy. KNN is fairly good in predicting the games. With more information about each game from the past 10 years, we built a moderate accurate model. intuitively, teams with good attacking strength and defending strength are more likely to finish in a good position in a league. According to the results the chances of winning as a home team are greater than the chances of winning as away team since the average goal scored for home teams is greater than the average goal scored by away team. Additionally, we presume that it might be due to home advantages such as familiarity of pitch, home team weather conditions, spectators, etc. finally , given enough data set with relevant features it is possible to predict future matches.