Universiteit Leiden

Opleiding Informatica

A Comparison of Methods for Predicting Football Matches

Name:	David B. Ekefre
Studentnr:	s1470655
Date:	14/02/2016
1st supervisor:	Jan N. van Rijn
2nd supervisor:	Dr. A.J. (Arno) Knobbe

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Data analysis (or Data mining) has become a major part of our businesses, governments and lives. The concept of analyzing the things that happen around us in our everyday life could be said to be human nature — making sense of our environment, the people around us, how we and/or the people around us live their lives and drawing conclusions from these observations. Data mining, in this sense, is observing a stream of data or a static dataset, using algorithms to make sense of the data or find patterns, have a target or conclusion in mind you want the algorithm to draw from the data (inductive) or you allow the algorithm to draw its own conclusions from the data (deductive); whichever way, new insights and conclusions are drawn.

Data mining is now being used in sports. In Association Football (football), the sports seen as the most popular sports in the world, data mining has been a big part of the sport as a cooperate business, a betting business, gaming industry business.

There are different methods that have been proposed for predicting a football match but in this thesis paper, we do not propose one method or any method for predicting a football match. We show a comparison of different methods using feature selection from mainly historical data; and show if using football team formations (or tactics) as dependent features add any significant value to the predicting a football match. However, the team formation data used in this thesis has limited information due to the difficulty in acquiring football data in the world today.

Contents

1 INTRODUCTION

Data analysis (or Data mining) has become a major part of our businesses, governments and lives. The concept of analyzing the things that happen around us in our everyday life could be said to be human nature — making sense of our environment, the people around us, how we and/or the people around us live their lives and drawing conclusions from these observations. Data mining, in this sense, is observing a stream of data or a static dataset, using algorithms to make sense of the data or find patterns, have a target or conclusion in mind you want the algorithm to draw from the data (inductive) or you allow the algorithm to draw its own conclusions from the data (deductive); whichever way, new insights and conclusions are drawn. Data mining is said to give answers to questions you never thought were there.

Data mining is now being used in sports. In Association Football (football[1]), the sports seen as the most popular sports in the world, data mining has been a big part of the sport as a cooperate business, a betting business, gaming industry business. Data mining is a big part of the football cooperate business by determining the worth of players bought into a club and sold out of a club, by determining how players' health are managed individually as well as collectively, how players are told to play and many more. The football betting industry business, however, has been an active business long before data mining was introduced to it but as every other industry, with the introduction of technology and more specifically, data mining, there becomes an exponential growth in the amount of business that can be done in a quicker but more efficient way. Having said this, data mining in the football betting industry has brought about more efficient ways of predicting the outcome of a single football game, to the extent that even the minor details that go on in a football game like the amount of corners, who gets the first goal kick or which player gets the first booking of the game is being predicted on a daily basis; league positions are also being predicted and bookmaker odds have become more profitable with the use of data mining techniques amongst many advantages data mining has brought to the football betting industry business. In the football gaming industry, data mining has been instrumental in determining the realism of in-game tactics — how tactics are executed in-game by video game teams are becoming more and more like how tactics are being executed in-game in real life.

[1] From henceforth in this paper it will be called football

1.1 Problem Statement

Companies and organizations like OPTA, STATS and ProZone [3, 27] gather and analyze sports data in high volumes, velocities and varieties. These companies gather football data such as players' performances, heat map activities, historical data amongst many other football data they gather. They gather these data to create new insight into player and team performances, tactical insights, team strengths and match predictions. Data mining techniques like classification and regression models have been used to make match predictions. Predicting the outcome of a football match prior to when that football match is played can be said to be a million dollar industry, hence, football match predictions is a very interesting subject in the area of sports data analysis.

Papers like have researched on the efficiency of football match predictions, more specifically by bookmakers of the betting industry. Different methods have been proposed over the years on how to predict a football match. Some methods are derived using just quantitative data and some using just qualitative data while others use a mixture of both quantitative and qualitative data in predicting football matches. A quantitative dataset would be a dataset that consists of historical data like head-to-head results, recent performances or shots on target while a qualitative dataset would consist of expert knowledge on team and player performance, player availability or media pressure. Although, these methods propose one way, respective to their various methods, making a comparison of various methods that can be used to predict the outcome of a football match is a not so common approach to football match prediction research papers.

1.2 Comparison of Methods & Feature Selection

In this paper, we do not propose any specific methods for predicting football matches but rather explore various methods using data mining techniques. We compare these various methods using different data mining classifiers (algorithms) and present the results in a chart. The methods we explore are a product of feature selection. We select a number of features, features that we presume could affect the outcome of a football match, use either any or a combination of these features for predictions.

The team formations or tactic, however, is slightly different in the sense that it is used contingent upon other features. This is due to the limitation of the team formation data we had obtained for this thesis. Gathering football data is a million dollar business, and for the fact that this is a master's study thesis, there was only so much that could be spent on gathering the data needed and used for this thesis. The team formation data used in this thesis can be said to be just a string of numbers, for example, (4-4-2), and is used contingent on any or a combination of other features. Features like recent performance (form) or shots. We were also able to determine, to some extent, in which match scenarios the team formations data would have more value. Match scenarios like when a top league team plays a bottom league team or a middle league team plays a bottom or top league team — these match scenarios are determined by the teams' total form values.

Despite the limited data we were able to gather for this thesis, we pose two research questions. These questions are:

1. Which features or combination of features would produce better predictions?
2. Do team formations add significant value to football match predictions?

In the next chapter, chapter 2, we will talk about previous literature work done regarding football match predictions. In chapter 3, we will show our experiment results and comparisons using different features and in chapter 4, the discussion chapter, we will give insights and pose questions based on the results shown in the chapter 3. Finally, in chapter 5, we will conclude this thesis paper and suggest future research in the area of comparing various methods and the use of team formations in football match predictions.

2 LITERATURE

The outcome of football matches have been predicted for a very long time, mostly by intuition or "gut" feeling. The first "scientific" model and research for predicting the results of football matches was the Poisson model which is modelled after the actual Poisson Distribution named after French mathematician Siméon Denis Poisson.

A Poisson Distribution according to Investopidia, is a "statistical distribution showing the frequency probability of specific events when the average probability of a single occurrence is known" [19]. The Poisson model used in predicting football match outcomes is mainly based on the how much ball possession each team has [22, 24, 9], in respect to scoring when with the ball and not conceding when not with the ball; not necessarily about keeping the ball for a certain amount of time. In [34], the author introduced the notion of maximum likelihood estimates, using the 1973 National Football League (NFL) outcome as dataset to conclude, among other things, that the present ranking system of the 1973 NFL season was not a fair ranking system, going ahead to show that the Poisson model — maximum likelihood is a better ranking system; whereas [24] calculates, using a bivariate Poisson model [9], the maximum likelihoods on four parameters — home attack (α), away defense (β), home defense (γ) and away attack (δ), he came to the conclusion that the relative strength of a team's attack is the same whether home or away, same about teams' defense.

The author of [25] uses goals as the parameter to predict game results by proving goals in football matches are distributed by negative binomial distribution, although at that time he had not named the distribution but rather called it the "modified poisson distribution"; not until [30] proved the same conclusion not only in football matches but in other ball games, came up with the name of the distribution as "negative binomial".

Unlike the Poisson models that predict the number of goals scored and conceded, which in turn are used indirectly to predict match outcomes, other models are restricted to predicting just the match results i.e win, draw or loss; of course these models consists of various explanatory variables (or features) and are typically ordered probit or logit regression models.

The impact of specific variables or factors on match outcomes have been taken into consideration in a number of studies. Home advantage has been a forefront factor in many of these studies; the authors in [2] found empirical evidence that the adoption of artificial playing surfaces, adopted by a few teams during the 1980s and 1990s, led to a sufficient amount of home advantaged wins. Similarly, home advantage has been linked to the importance of crowd size and crowd density by [26, 28].

While investigating the efficiency of bookmakers' odds, some other studies have created models for forecasting football match outcomes, using various factors in their models. For example, [21] takes into consideration certain differences in teams' performance, for example, the difference in teams' average point per game and cumulative points over the season; he also took the teams' cumulative and average goal differences, as well as bookmakers' odds as factors in creating an ordered probit regression model. In [13, 17], the authors use a combination of three factors: *teams' quality* — dependent on how far back and in what division played in prior to the match in question will data be taken into consideration; *recent performance* — this includes information regarding the home teams' most recent home results and away results as independent variables, same for the away teams; the third factor used is a combination of different other factors, factors like match significance, cup competition involvements, geographical distance and crowd attendance relative to league position.

Models that use team quality ratings or team rankings, as it is also called, have also been considered in some studies. Two of the most popular rating systems used in predicting match outcomes are: The ELO ratings system, which was initially developed for calculating the strength of chess players [10], which has subsequently been adopted by other sports including football; and the FIFA/Coca-Cola World rankings. The adopted ELO rating system assigns, and on a continual basis, updates points to a team dependent on a win, draw or loss, it also includes an interesting variant that allows the rating update coefficient $\boldsymbol{k}$ to depend on the goal difference, thus rewarding a 3-0 win more strongly than a 2-1 win. The authors in [18] used the ELO[2] ratings to predict the outcome of matches and concluded that the ELO ratings appear to be useful in encoding information of past results, but their forecasts were still under par compared to bookmakers' odds. In [33], the authors evaluate the efficacy of the FIFA/Coca-Cola World Ranking for predicting World Cup results and concluded that the FIFA/Coca-Cola World Ranking are an effective means of predicting World Cup results. In [23], the authors have also modeled a forecasting system based on ELO ratings along with the FIFA/Coca-Cola World rankings for the EURO 2008 tournament and concluded that bookmakers' forecasts outperformed their forecasts.

Most of the variables or factors mentioned so far are quantitative variables; variables derived solely from statistical and historical data, for the use of football match predictions. We have yet to study papers that offer the qualitative side of football predictions. According to [6], "a numerical approach to forecasting offers one perspective on fixed sports betting, qualitative research and judgment another", proposes that "the most effective approach to sports prediction is likely to make use of both quantitative and qualitative information". The subjective or qualitative information mentioned in this research paper may refer to information derived from expert opinion.

[2] The ELO mentioned here and hereafter means the adopted ELO rating system for football.

Machine learning techniques are at the forefront of making use of both quantitative and/or qualitative information for predicting football match outcomes. In [20], the authors concluded that an expert constructed bayesian network model (see Section E for model) built for predicting football match outcomes, compared to other machine learning models built using the same data, performed generally superior. Their bayesian network model was based almost exclusively on subjective judgment. Some of the subjective judgment or information used by the authors include: the identification of 'special' players, the inclusion or exclusion of these 'special' players on match-day to measure the quality of both teams and the venue where the game is being played. The authors in [4] carried out a performance prediction experiment during the European Football Championships in 1996 by relying on the knowledge of four predictors, using evidence theory, an integrating framework for predictions. These four predictors, as called by the authors, which are mainly subjective (qualitative) information are: missing key players, home advantage, pressure by media and public and performance of past matches. Although the authors in [4, 20] ended up with positive findings, we can already see the complexity that is found in deriving qualitative information, as compared to that of deriving quantitative information; this complexity can be a limitation to making predictions solely on qualitative information and while they mainly used subjective information for their models, some other studies have had a more balanced approach in using both the qualitative and quantitative side of information for predictions. For example, in [35], the authors use machine learning techniques such as fuzzy rules, neural networks and genetic programming techniques for prediction based on a vector of features. The features used by the authors are the differences of: player statuses — injured players, disqualified players, of both teams; score of both teams over the last five games; team rankings in the current championship; home factor of both teams — the fraction of team total points over team home games; and lastly, goal difference, which is goals scored minus goals conceded, of both teams within a 10 years period. They concluded that the genetic programming technique was superior for predicting correct results to the other two methods. The authors in [31] also concluded by claiming that acceptable match simulation results can be obtained by tuning fuzzy rules using tournament data. The fuzzy rules used were determined by parameters of fuzzy-term membership functions and rules weighted by a combination of genetic and neural optimization techniques. More importantly for this research, the match outcome influencing factors used by the authors were the results of: five previous matches for each team; and the last two head to head results of both teams in question.

In [7], the authors compare the model proposed with and without subjective information and concluded that the subjective information improved the model such that the posterior forecasts were on par with bookmakers' performance.
The model, which they call 'pi-football' (v1.32), develops predictions based on four generic factors for both the home and away teams; these four factors are: team strength, form, psychology and fatigue. The first component is mainly derived from objective information while the subsequent components, mainly from subjective information.

The authors, after satisfying all four components (see Section F of the Appendix), attempt to assess the quality of their forecast model with and without the subjective information against bookmakers' accuracy and profitability. For the accuracy measurement, they chose to use the Rank Probability Score (RPS), and for the profitability measurement, they perform a betting simulation that satisfies the following standard betting rule: *"for each match instance, place a 1-pound bet on the outcome with the highest discrepancy, of which the pi-football model predicts with higher probability, if and only if the discrepancy is greater or equal to 5%"*.

They divide their forecast into two separate forecasts: (1) the objective forecast generated at component 1, (2) the rather subjective (revised) forecast after including components 2, 3 and 4. They are both tested against a normalized bookmakers' forecast and they concluded that the accuracy of objective forecast was significantly inferior to the bookmakers' performance but with the improvement on the objective forecast (which forms the subjective forecast), the accuracy becomes on par with the bookmakers' performance. Suggesting that the bookmakers also make use of subjective information. This supports the claim of [6], that the most effective approach to sports betting is to make use of both qualitative and quantitative information.

Regarding the profitability measurement, the subjective forecast was tested against the maximum (best available for the bettor) bookmakers' odds, the mean bookmakers' odds and the most common[3] bookmakers' odds. Their conclusion was that pi-football was able to generate profit via longshot bets; longshot bets or longshot bias have been identified in [8, 14, 16] as a bias against bettors for bookmakers' profitability, implying that the pi-football

[3] The most common as at the time the article was written.

would have generated higher profits if there were less biased odds. They also tested the objective forecast against the three bookmakers' odds and it was proven, although not much detail was given in the paper about the experiment involving the objective forecast, to result in losses. This again suggests that the use of subjective information (qualitative) and objective information (quantitative) together gives better performance in sports prediction, either for measuring the accuracy or the profitability of any forecast model.

3 EXPERIMENTS

This chapter shows the various experiments carried out in this research and the respective results. These experiments are carried out using Weka [37][4] and Python [36][5]. The dataset is obtained from [11] and [12].

There are different ways in which the quality of a forecast model can be assessed. In particular, we consider just the accuracy for the purpose of this research. The machine learning techniques used for these experiments are: k-Nearest Neighbor (IBk) [1], Random Forest [5], J48 [29] and Boosting (AdaBoostM1) [15].

The attributes (features) used in these experiments are described under Section C of the Appendix.

3.1 Dataset

Our dataset consists of data derived from the past three football seasons (2012/2013, 2013,2014 and 2014/2015) from the five (5) big leagues in Europe, which are the English Premier League, German Bundesliga, French Ligue 1, Italian Serie A and the Spanish La Liga, in no order of relevance. Before going into the main experiments of this research, we would like to point out the diversity of the different leagues in our dataset based on one attribute, which is the total form. The total form is representative of the cumulative points gathered by each team in the course of a season; these points (cumulative) are then divided by the number of games played (also cumulative).

The figure below shows two column bars for each of the five leagues; the first bar represents the default accuracy value of that league while the second bar shows the accuracy value based on team league positions of the league in question.

[4] Version 3.7.13

[5] Version 2.7.10

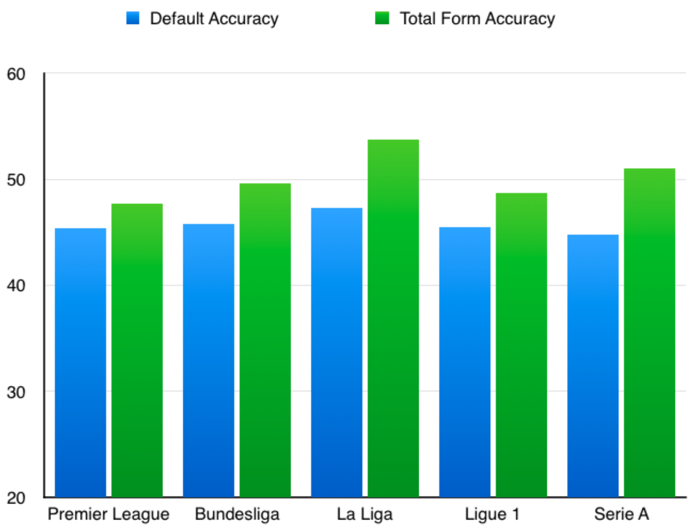


Figure 1: The Big Five (5) accuracy comparison

3.2 Team Form

This part of the experiments is carried out on the team forms. The team forms are representative of a team's recent performance. The team form values will be represented using four (4) different representations, which will be explained below.

The team form consists of the Home and Away team forms as shown below:

1. **Home Team Form:** The total amount of points the home team got from the last n (the value of n will be determined experimentally) games.
2. **Away Team Form:** The equivalent information for the away team.

The representations of the Team Forms are:

1. **Representation 1 ($r1$):** This represents the numeric values of the team forms, normalized to interval [0,3].
2. **Representation 2 ($r2$):** This represents the discretized value of the team forms. We had reason to believe that the classifiers do not distinguish between values well enough while using $r1$, so we discretized $r1$ using the set of rules shown in Table 1.

Numeric Values	**Discretized Values**
[0,1>	Bad Form
[1,2>	Good Form
[2,3]	Best Form

Table 1: Numeric & Discretized Values

3. **Representation 3 (*r*3):** This represents the subtracted value between the home team form and away team form. This subtracted value is normalized to the interval [-3,3]; a negative value means away team superiority and a positive value means home team superiority while zero means an equal advantage.
4. **Representation 4 (*r*4):** This represents the discretized values of $r3$. This representation will be discretized by equal frequency into three bins.

The results are shown in the figures below.

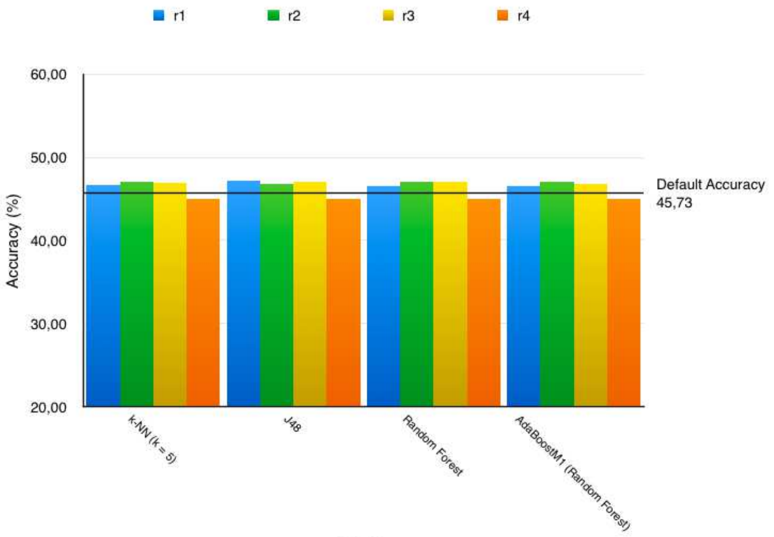


Figure 2: $n = 3$

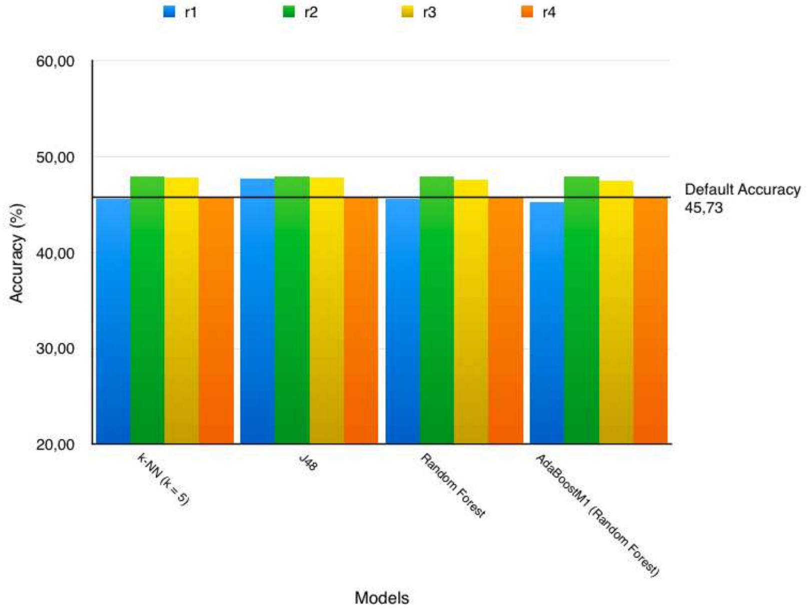


Figure 3: $n = 5$

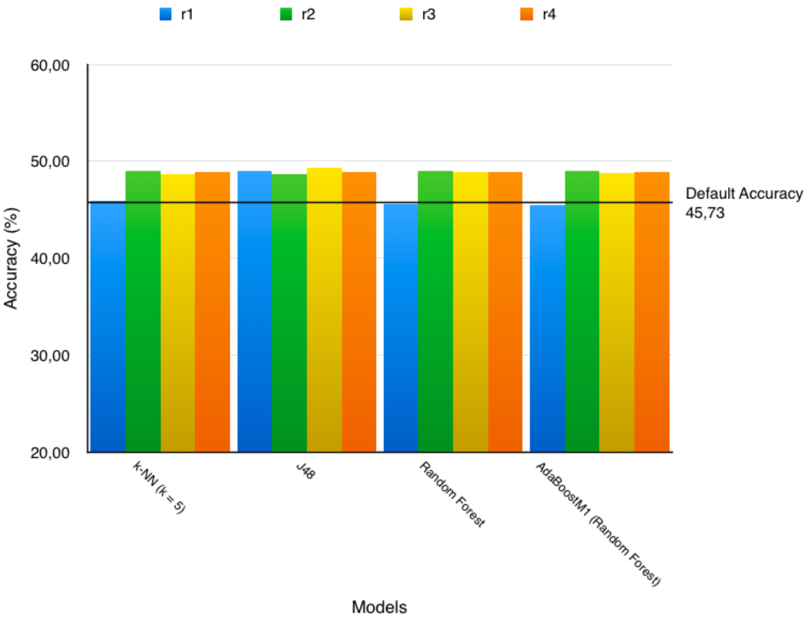


Figure 4: $n = 9$

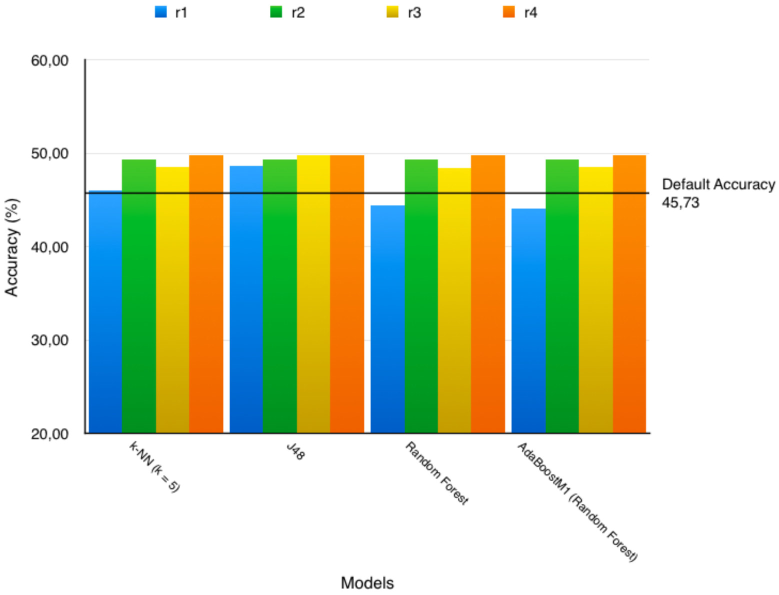


Figure 5: $n = 11$

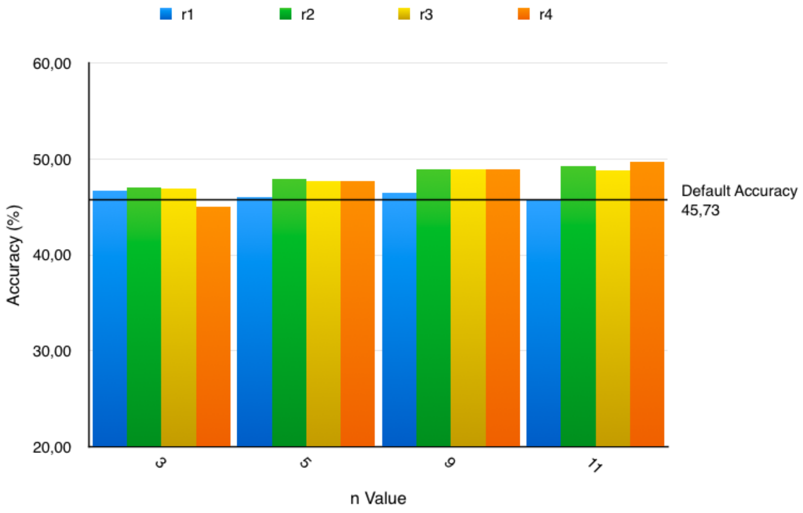


Figure 6: Average n value from all four (4) classifiers)

Figures 2 through 5 show the different accuracy values gotten from the four different classifiers from the various values of n used. Figure 6 shows the average accuracy value of the various values of n compared against the default accuracy.

3.3 Team Home & Away Form

In the previous experiment, we used the overall team forms. In this part of the experiments, we use the teams' home and away forms. The home and away forms are representative of the teams' recent home performance and recent away performance respectively. Their values will also be represented by the four (4) representations used in the Team Form experiments.

The Team Home and Away Forms consists of:

1. **Home Team Form:** The total amount of points the home team got from the last n games.
2. **Home Team Home Form:** The amount of points gotten by the home team from the last n home games.
3. **Home Team Away Form:** The amount of points gotten by the home team from the last n away games.

4. **Away Team Form:** The equivalent information for the away team form; **Away Team Home Form** and **Away Team Away Form**.

The representations are:

1. **Representation 1 ($r1$):** This represents the numeric value of the team home and away forms, normalized to interval [0,3].
2. **Representation 2 ($r2$):** This represents the discretized values, using the set of rules shown in Table 1.
3. **Representation 3 ($r3$):** This represents the subtracted value between the home team form and away team form; the home team home form and away team away form; the home team away form and away team home form. This subtracted value is normalized to interval [-3,3]. A negative values means away team superiority and a positive value means home team superiority.
4. **Representation 4 ($r4$):** This represents the discretized values of $r3$. This representation will be discretized by equal frequency into three bins.

The results are shown in the figures below.

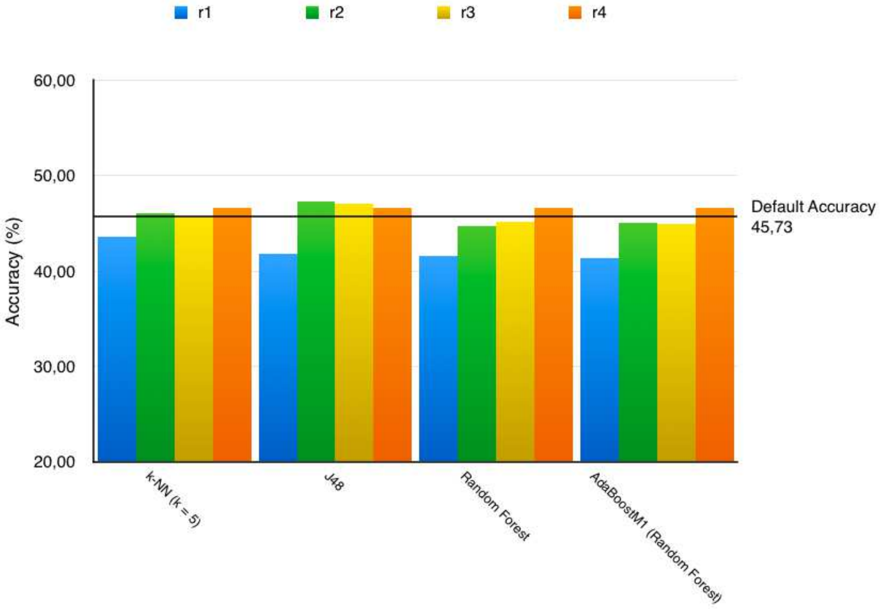


Figure 7: $n = 3$

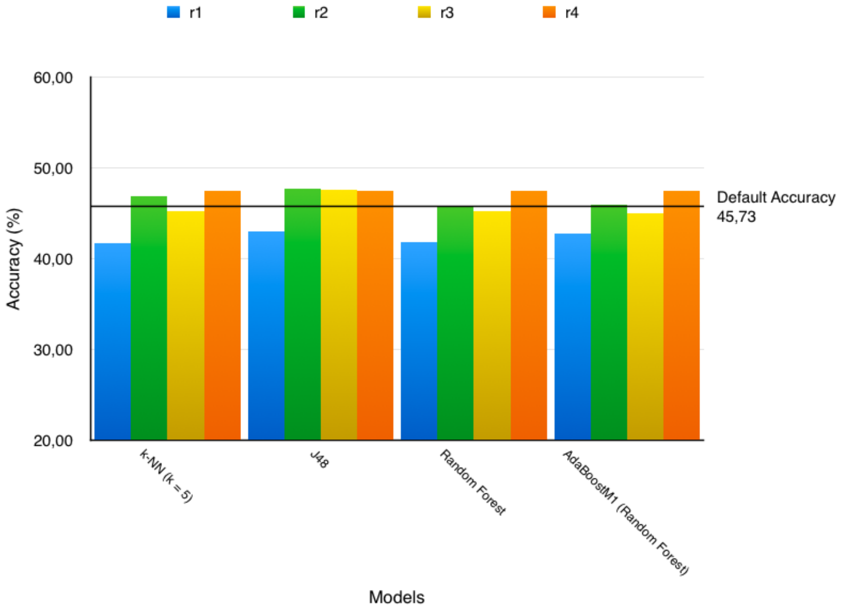


Figure 8: $n = 5$

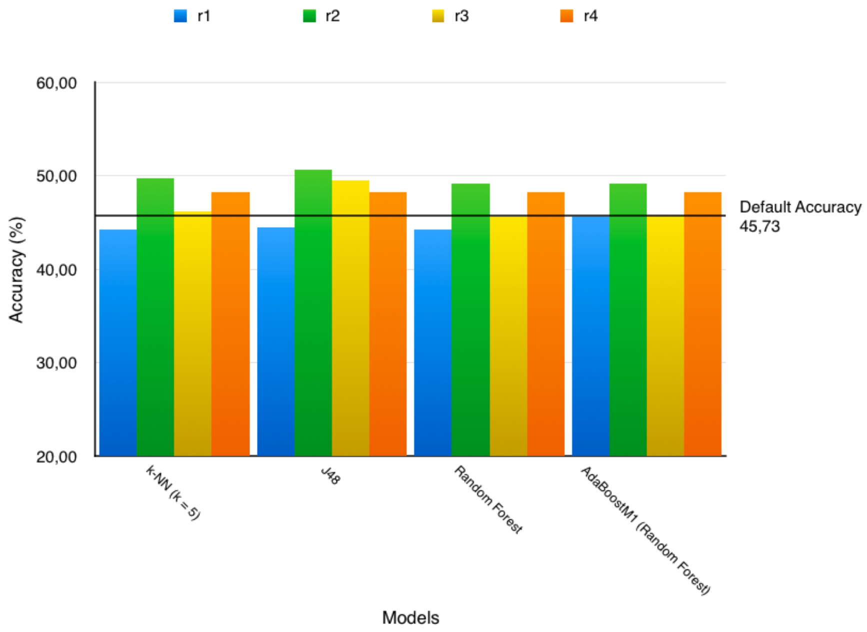


Figure 9: $n = 9$

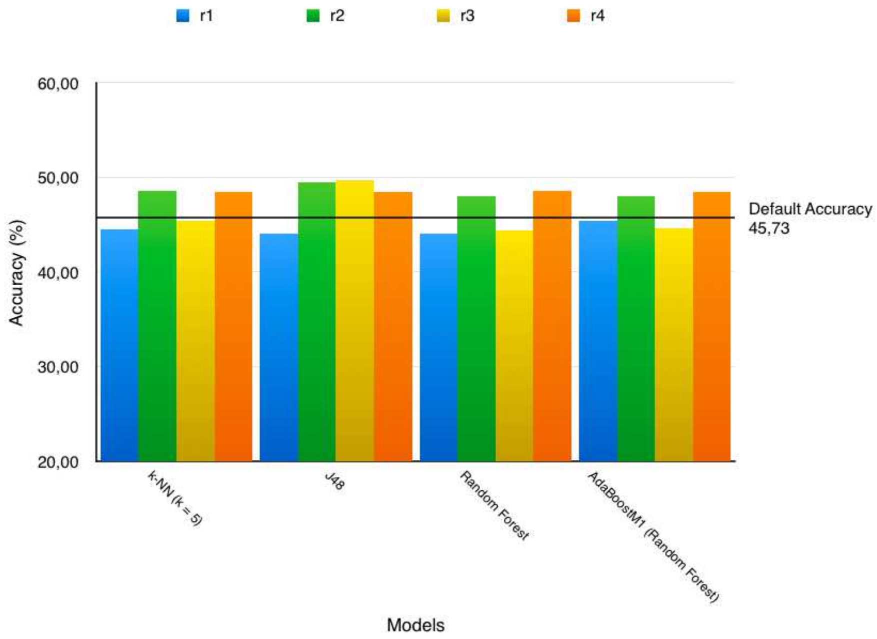


Figure 10: $n = 11$

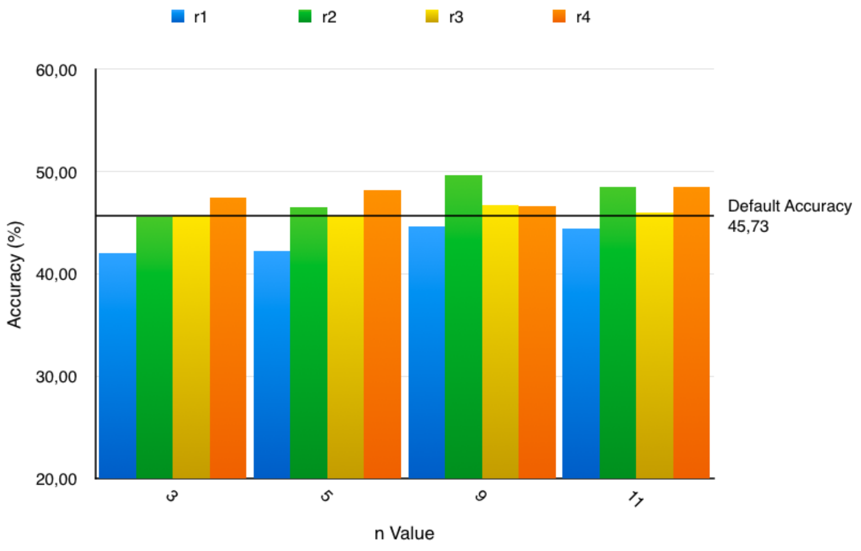


Figure 11: Average n value from all four (4) classifiers

Figures 7 through 10 show the different accuracy values from the four classifiers comparing the various values of n used. Figure 11 shows the average accuracy value of the various values of n compared against the default accuracy.

3.4 Match Statistics

In this part of the experiment, we will use a combination of features we call Match Statistics. Based on the superiority of the discretized values over the numeric values from the previous experiments, the Match Statistics values will be represented using only Representation 4. As mentioned in the earlier experiments, Representation 4 is discretized by equal frequency into three (3) bins. The value of n used for team forms is set as 9 and 11, also because of their superiority over the the two other values of n tested in the previous experiments.While for the shots and goals attributes only 9 is used as the n value.

The Match Statistics features consists of:

1. **Goals:** The subtracted difference in goals from home and away teams over 9 games.

2. **Shots:** The subtracted difference in shots from home and away teams over 9 games.
3. **Shots on target:** The subtracted difference in shots on target from home and away teams over 9 games.
4. **Goals Ratio:** The subtracted difference in goals ratio from home and away teams over 9 games.
5. **Shots Ratio:** The subtracted difference in shots ratio from home and away teams over 9 games.
6. **Form:** The subtracted difference in points from home and away teams over 9 games (represented as **Form-9**); 11 games (represented as **Form-11**) and all subsequent games (represented as **Total Form**).

The attribute selections (or combinations) used are:

1. **Attribute Selection 1 ($a1$):** Total Form, Form-9 and Form-11.
2. **Attribute Selection 2 ($a2$):** Total Form, Shots, Form-9 and Form-11.
3. **Attribute Selection 3 ($a3$):** Total Form, Goals, Form-9 and Form-11.
4. **Attribute Selection 4 ($a4$):** Total Form, Shots on Target, Form-9 and Form-11.
5. **Attribute Selection 5 ($a5$):** Total Form, Shots Ratio, Form-9 and Form-11.
6. **Attribute Selection 6 ($a6$):**Total Form, Goals Ratio, Form-9 and Form-11.

The results are shown in the figures below.

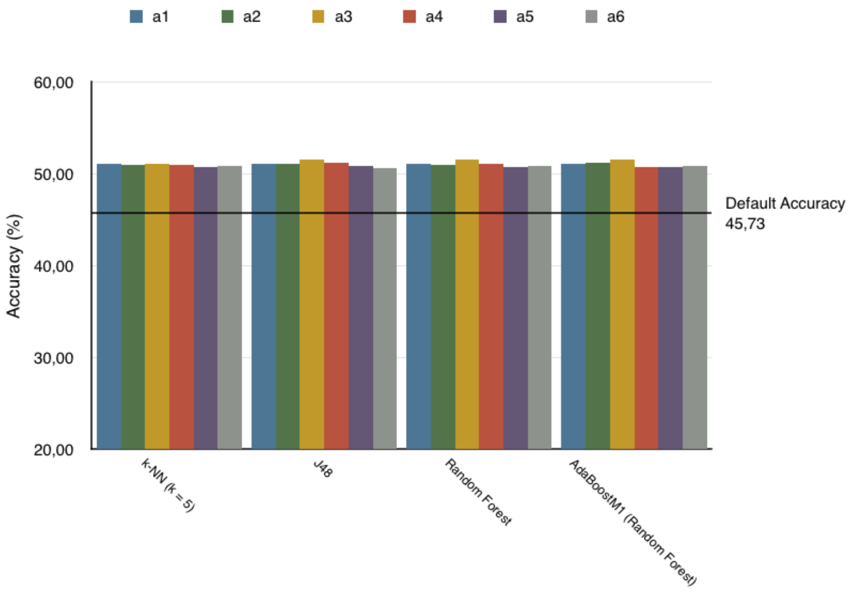


Figure 12: Accuracy of all attribute selections from all four (4) classifiers

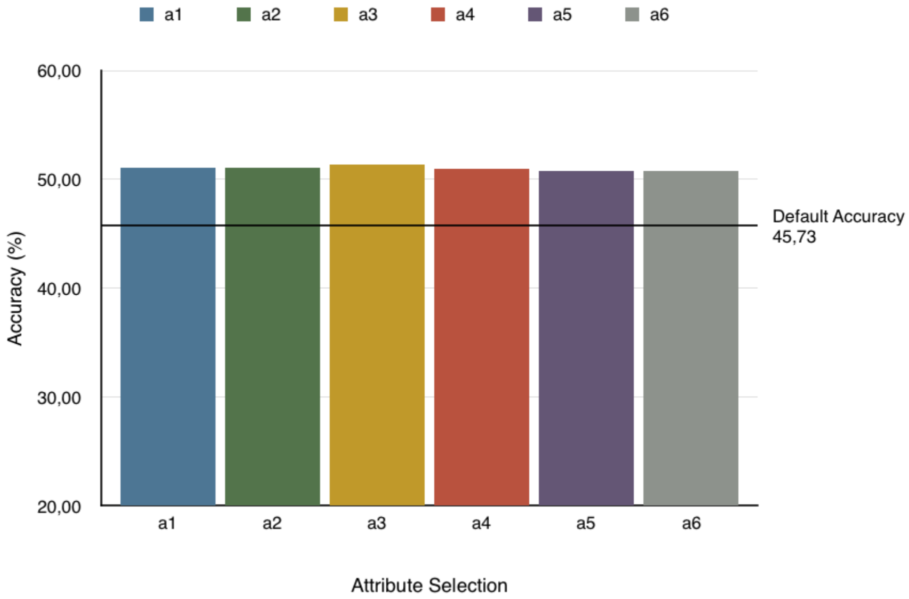


Figure 13: Average accuracy from all four (4) classifiers of the different attribute selections

Figure 12 shows the different accuracy values from the four classifiers comparing the different attribute selections against the default accuracy. Figure 13 shows the average accuracy figure of the various attribute selections compared against the default accuracy.

3.5 Team Formation

The team formations will be introduced in this set of experiments. It will be experimented on as a dependent feature on the total form feature to show what impact adding team formations would have in predicting the outcome of football matches.

Before any experiments on prediction were done, we realized that in our dataset we had 20 unique team formations and out of these 20 unique formations, 13 were used by less than 10% of the dataset while out of these 13, 4 were used by less than 1%. For this reason we decided to partition the formations into 10 different partitions.

We split the dataset into three bins based on Representation 3 of the total form values, first using binning by equal width and then binning by equal frequency.

3.5.1 Binning by equal width

The total form values are between the interval -3 and 3; for the binning by equal width, the values are then normalized between the values 0 and 3. The bins are described briefly below:

1. **Bin 1 (*b*1):** This bin represents total form values of 0 to -1 and 0 to 1 and is represented in the interval [0,1>.
2. **Bin 2 (*b*2):** This bin is represented in the interval [1,2> and represents total form values of -1 to -2 and 1 to 2.
3. **Bin 3 (*b*3):** This bin is represented in the interval [2,3] and represents total form values of -2 to -3 and 2 to 3.

These three bins mentioned above were then experimented on, with and without the inclusion of the team formations. The statistics and results are shown in the table and figures below respectively:

Bins	Number of Instances
[0,1>	4614
[1,2>	761
[2,3]	30

Table 2: Binning by equal-width Statistics Table

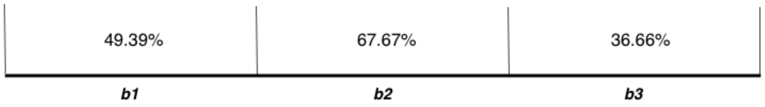


Figure 14: Bins 1 to 3 without formation

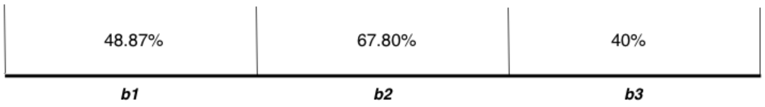


Figure 15: Bins 1 to 3 with formation

Figure 14 shows the accuracy of the three (3) bins by equal-width without the inclusion of formations while Figure 15 shows the accuracy with the inclusion of formations. While Figures These results were obtained using the J48 classifier.

3.5.2 Binning by equal frequency

The original total form values are used. The statistics of the bins are given in the table below.

Bins	Number of Instances
[-3,-0.278>	1804
[-0.278,0.268>	1799
[0.268,3]	1802

Table 3: Binning by equal-frequency Statistics Table

Figure 16 shows the accuracy of each bin without team formations; and Figure 17 shows the accuracy with the inclusion of team formations. This experiment was done using the Random Forest classifier.

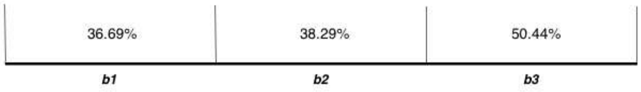


Figure 16: Bins 1 to 3 without formation

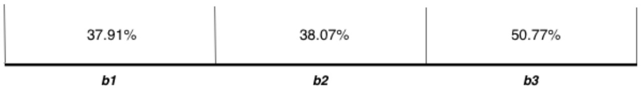


Figure 17: Bins 1 to 3 with formation

3.6 Formation Clustering

As mentioned in the previous experiment, the team formation experiments, the number of unique team formations found in our dataset is 20. Out of these 20 unique formations, 13 of them are used by less than 10% of the dataset while out of these 13, 4 of the team formations are used by less than 1%. For this reason, as mentioned earlier, we decided to partition, in the case of this experiment — cluster the formations.

This experiment will be divided into two parts:

1. The Formation Clustering: this part will show the results derived from clustering the team formations.
2. Comparison of Accuracy: this part will show the difference in accuracy levels between the clustered formations generated earlier, the partitioned formations generated in the 'Team Formation' experiments section and the original dataset team formations.

3.6.1 Formation Clustering

The team formations will be clustered using the hierarchical clustering method. The formation clustering will be based on two sets of attributes:

- shots and;
- shots on target.

Both clusters will be made using the average linkage method and the euclidean distance function.

The dataset had to be aggregated to the team formations for this experiment. For every unique team formation used by the home teams and the away teams respectively, the average total home shots, the average total away shots, the average total home shots on target and the average total away shots on target were calculated. These aggregated datasets, for both the home and away team formations, are then combined into one new aggregated dataset that includes the combined team formations' average total home and away shots and the average total home and away shots on target values, making this new dataset look like this:

Formations	Average of HS	Average of AS	Average of HST	Average of AST
3-3-3-1	28.77	22.6	9.63	8.57
3-4-1-2	28.92	25.89	13.29	8.69
3-4-2-1	27.63	24.57	12.47	9.20
3-4-3	27.86	23.64	9.66	8.85
3-5-1-1	28.45	23.44	9.83	8.1190

The table above shows an example of what the dataset now looks like. The formations column consists of the team formation used by both the home and away teams, for example, row 1 of Table 3.6.1 tells us that the formation '3-3-3-1' used by both home and away teams together has the average total home shots value of '28.77'. The clustered results are shown in the dendrograms below:

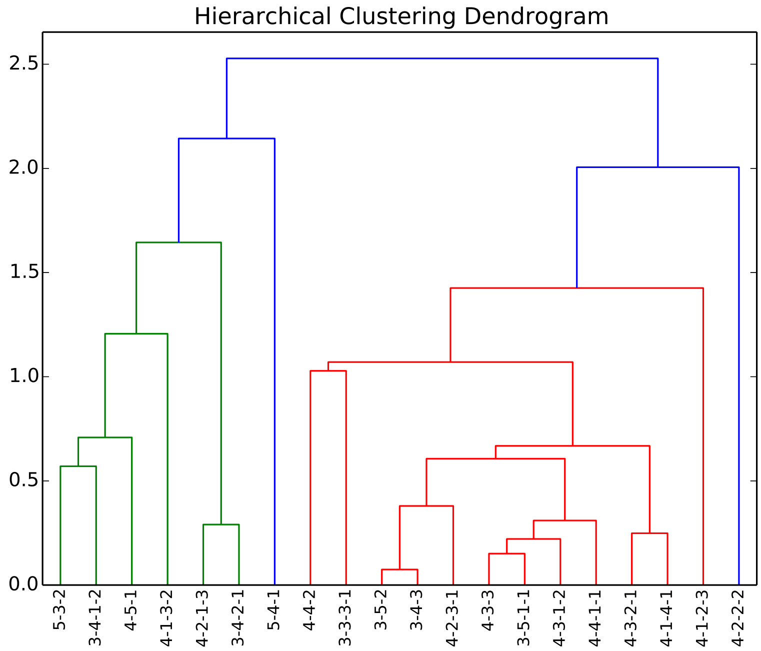


Figure 18: Shots-based Dendrogram

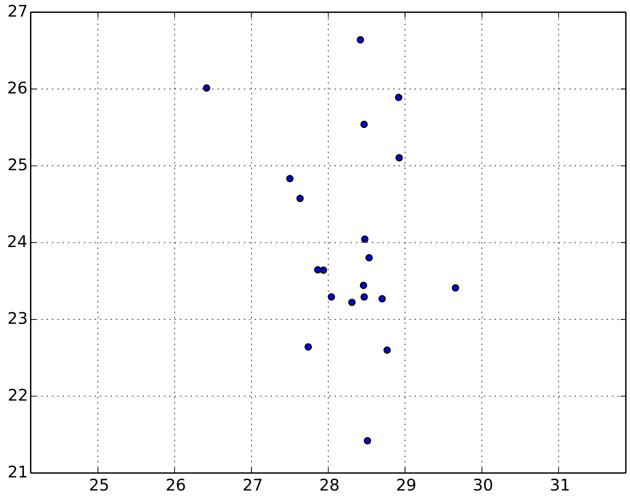


Figure 19: Shots-based 2-d Scatter Plot

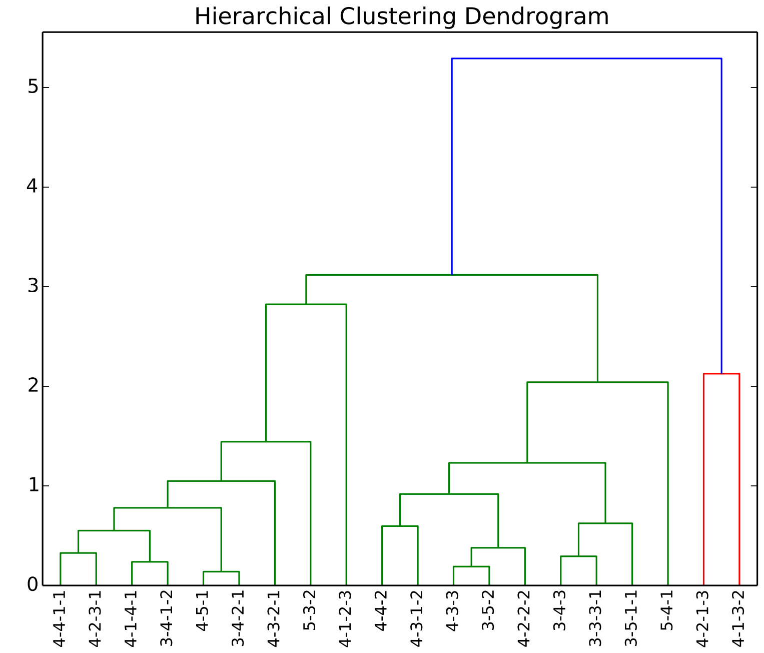


Figure 20: Shots-on-Target-based Dendrogram

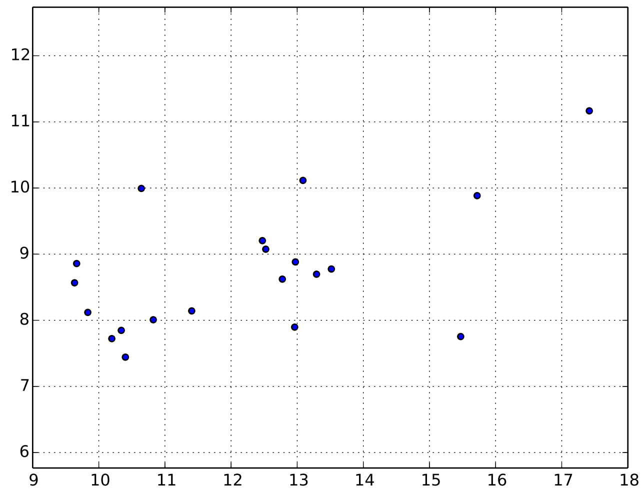


Figure 21: Shots-on-Target-based 2-d Scatter Plot

Figure 18 shows the team formations clustered based on home/away shots, Figure 19 shows the points on a scatter plot of home shots against the away shots; while Figure 20 shows the team formations clustered based on home/away shots on target and Figure 21 shows the points on a scatter plot of home shots on target against the away shots on target. Both dendrograms, however, show the whole hierarchy of clusters — all clusters from the least clusters with the least number of merges to the clusters with the most number of merges. Whereas, the figures below show the dendrograms at 10, 5 and 3 clusters respectively, for each of the shots-based and shots-on-target-based dendrograms.

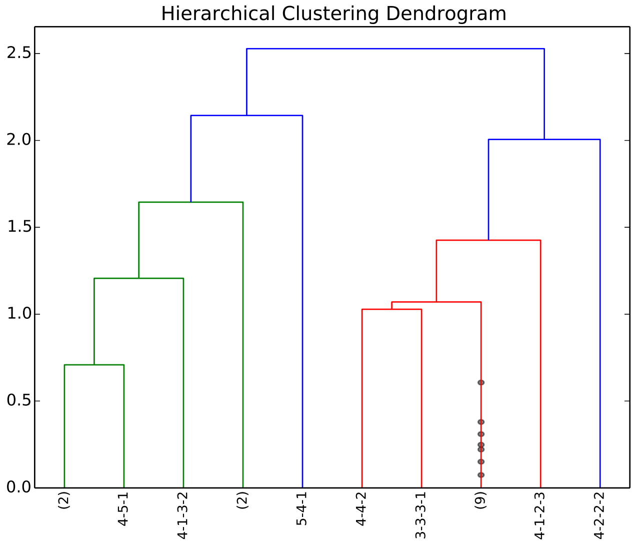


Figure 22: Shots-based (10 clusters) Dendrogram

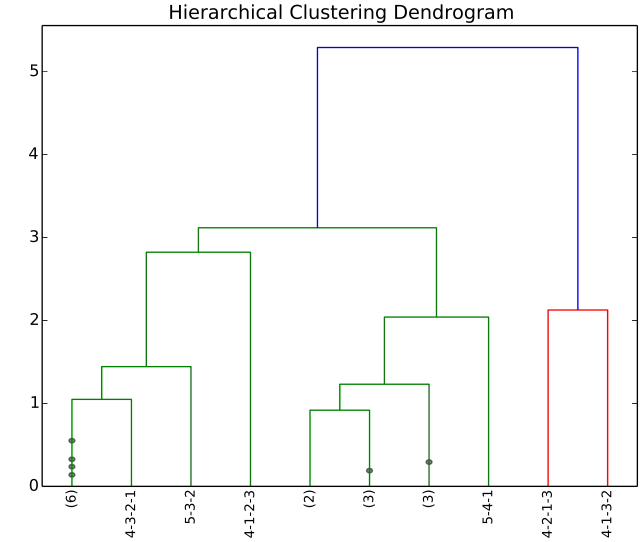


Figure 23: Shots-on-Target-based (10 clusters) Dendrogram

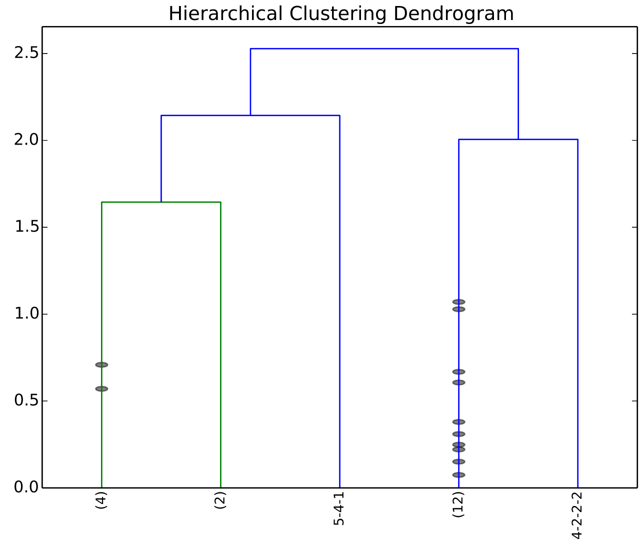


Figure 24: Shots-based (5 clusters) Dendrogram

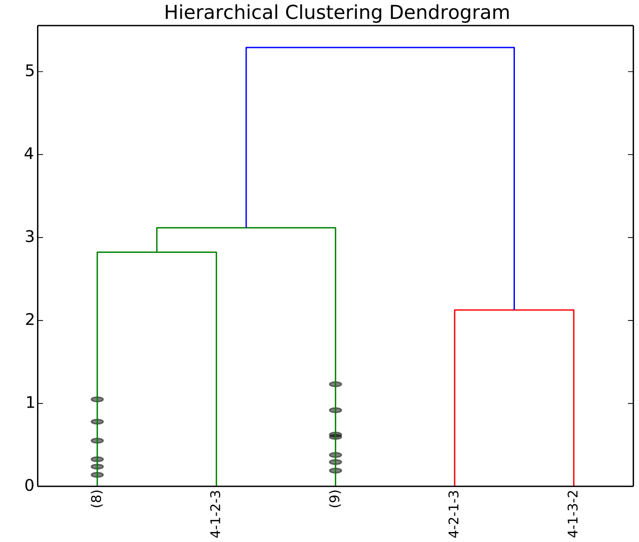


Figure 25: Shots-on-Target-based (5 clusters) Dendrogram

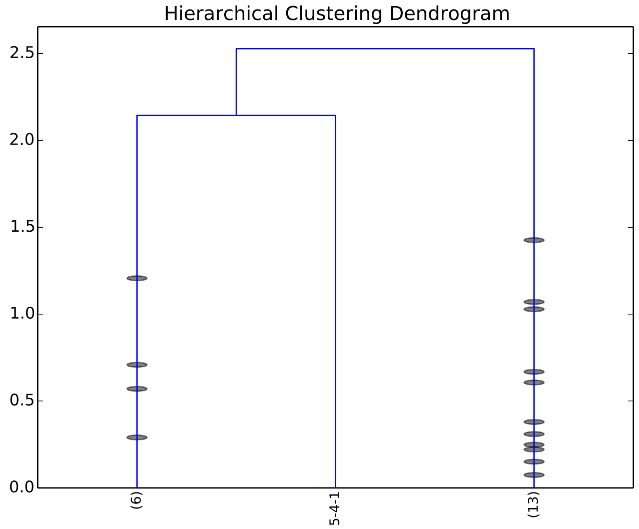


Figure 26: Shots-based (3 clusters) Dendrogram

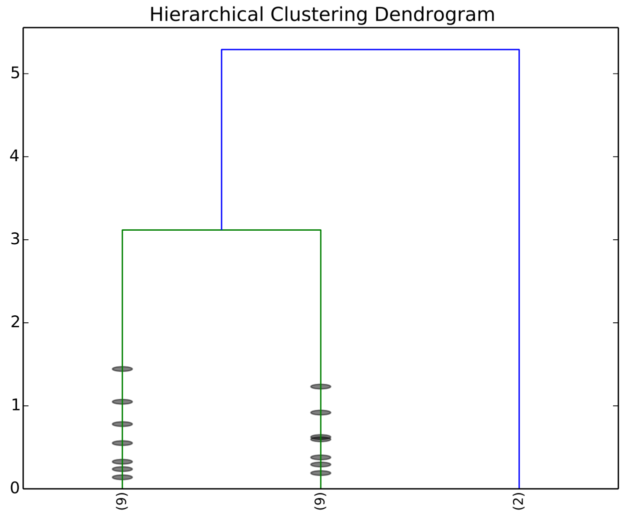


Figure 27: Shots-on-Target-based (3 clusters) Dendrogram

3.6.2 Comparison of Accuracy

Before making the classification problem and predicting the results for any of the clustered datasets used for the comparison of accuracy below, the clustered formations as shown in the figures above are then used to replace their respective team formations in the original dataset as the new home and away formation attributes. For example, the shot-based (3 clusters) dataset would have only the formations '5-4-1', '4-2-2-2' which represents every clustered formation in the first cluster in the dendrogram (from the right), and '4-5-1' which will represent all the clustered formations in the last cluster.

These datasets are then used to make comparison between the accuracy levels using the different cluster levels given earlier, the partitioned formations given in the 'Team Formation' experiments (also shown in Table 6 under the Appendix) and the original dataset team formations. The comparison results are given below, using the J48 classifier.

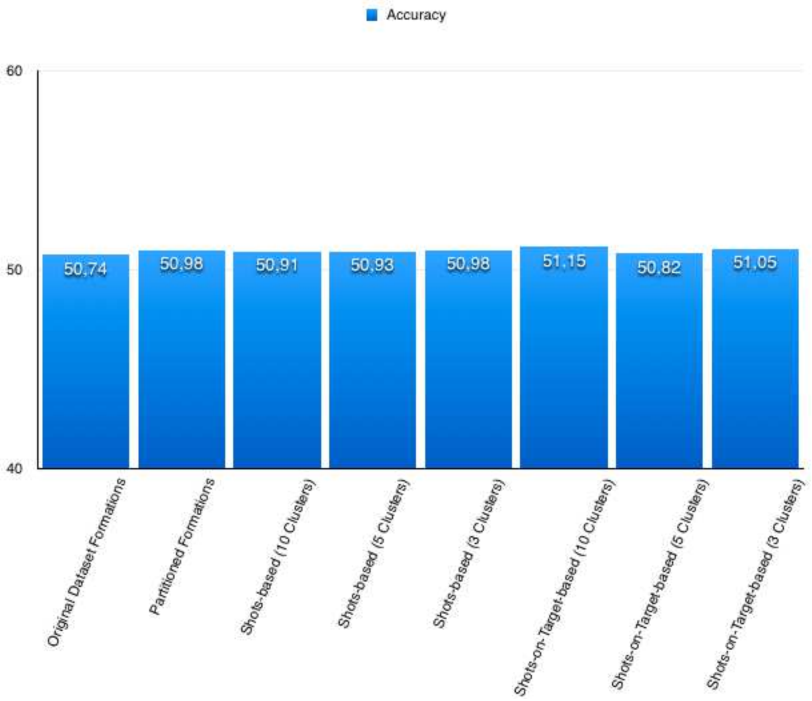


Figure 28: Accuracy Comparison of Formation-driven Predictions

Figure 28 shows a column chart that represents the different accuracy levels of various formation-driven match predictions. All predictions are made with the total form attribute. The 'Original Dataset Formation' consists of the original 20 unique team formations; the 'Partitioned Formations' consists of the 10 partitioned formations shown in Table 6; the clustered formations, 10, 5 and 3 consists of the 10, 5 and 3 clustered formations respectively as show earlier in this section.

3.7 Shot-Driven Results

The shots driven results experiment, as the title of the experiment suggests, is an experiment that predicts the outcome of a football match based on the final amount of shots in a game. Previous experiments done in this research paper have been experiments that predict the outcome of a football match based on the final score of the game. That is, a home win suggest the home team scored more goals than the away team, an away win suggests the opposite, while a draw suggests both home and away teams score an equal amount of goals. Whereas, in this experiment, a home win suggests the home had more shots in the game, an away win, the opposite, while a draw suggests both teams had an equal amount of shots.

The graph below shows two default accuracy bars and two total form accuracy bars — the first of either of the default accuracy and total form bars show the accuracy based on the final score of a game while the second of each bar shows the accuracy based on the number of shots had by each team. This is to show a comparison of using the shot driven result as opposed to using the goal driven result.

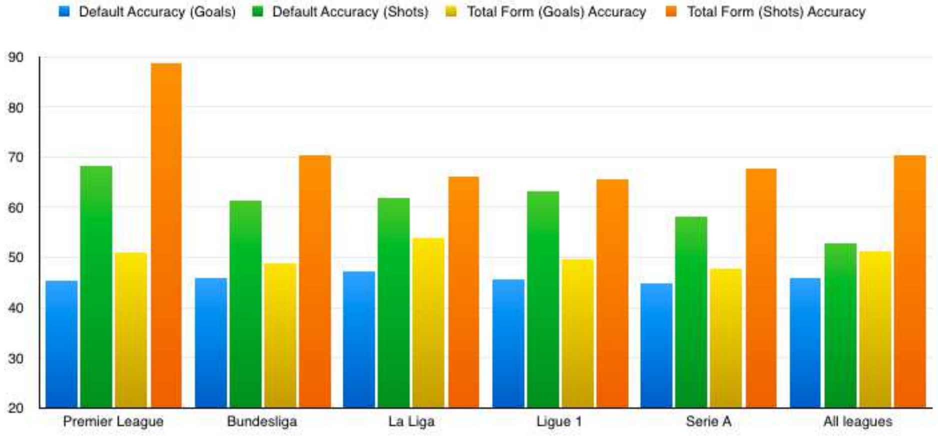


Figure 29: Comparison of Shot-driven Results and Goal-driven Results

3.8 Learning Curve

The learning curve given below is to experiment on what amount of data is best to use in order to get optimal results. The J48 classifier was used on the attribute selection $a3$ from the Match Statistics experiments to create this learning curve.

The learning curve shows the accuracy at different number of instances. It starts with just ten percent (10%) of the dataset and works its way up through ninety percent (90%) of the dataset.

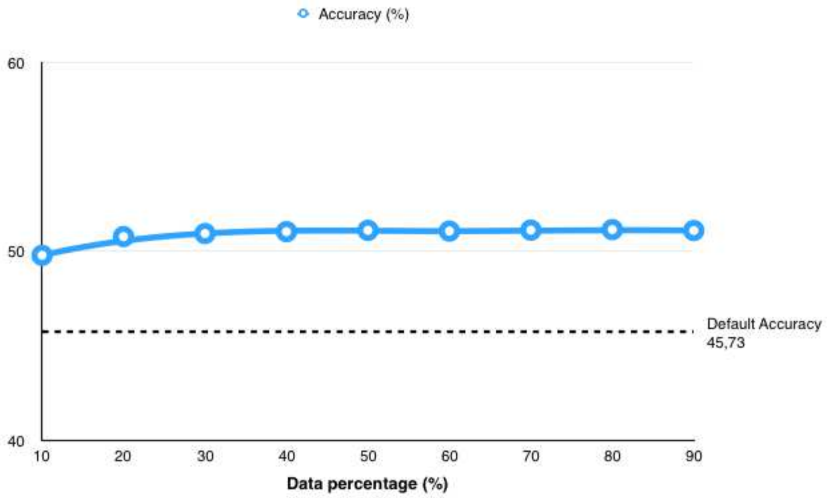


Figure 30: Learning Curve of instance,accuracy

4 DISCUSSION

In this section of this thesis paper, we will give our insights on the experiments and results gotten and shown in the experiments section.

Representations 1 ($r1$) and 2 ($r2$) used in the Team Form and the Team Home and Away Form experiments show the values of the home and away teams as two separate attributes, one for each while representations 3 ($r3$) and 4 ($r4$) show the values of both the home and away teams in one attribute as the subtracted difference between both teams, a negative value means the home team superiority and positive value means away team superiority as already mentioned in the Experiment sections. The reason for using representations 3 and 4 was to show a clear superiority of either the home team or the away team represented in one single value so this superiority could be easily observed by the algorithms.

4.1 Team Form

In the Team Form Experiments, we opted to find the best possible value for n — n is how many games back we choose to go prior to match-day to calculate the form of a team, not regarding any degree of uncertainties. The results shown in Figure 6 indicate that the value of n, going as far back as 9 games and 11 games gave better average accuracies than going back as far as 3 and 5 games. This could also indicate that increasing the value of n seems to give better results; this may be due to the fact that this is quantitative data and the farther back we go in retrieving data, the better and richer the data gets.

In the four attempts on finding the best value for n, the accuracy values derived showed no clear advantage from any of the four classifiers used. The results, however, show that Representation 1 is clearly worse than the other three representations and the discretized representations ($r2$ & $r4$) perform better than the numeric representations ($r1$ & $r3$). Whereas, on a more general overview, the results from all four representations are better than the default accuracy, but only slightly better. This means that, based on our experiments and the results derived from these experiments, using team forms alone is not good enough to predict the outcome of a football match.

4.2 Team Home & Away Form

In this experiment, we introduced the home and away forms of the team. We presumed that including these extra attributes, the algorithms would have a richer dataset and understanding of the team forms. By distinguishing between the home and away forms, we presumed it could predict, depending on if the teams in question are playing at home or away and based on its home or away form, the outcome of the game much better.

The results, regardless, show in Figure 11 that the value of n at 9 and 11 games still perform better than at 3 and 5 games, supporting our earlier discovery that increasing the value of n seems to give better results. It also shows and supports the discovery that the discretized representations perform better than the numeric representations. This may be as a result of one of these two reasons:

- we only used tree classifiers and its supposed that tree classifiers work better with discretized variables than with numeric variables or
- discretized variables have a better expression as to distinguishing between its value than numeric variables.

On average, representations 2 and 4 perform clearly better than the baseline as compared to the other two representations.

However, despite our earlier presumption that adding both the home and away forms would give a clear advantage and yield better predictions, oddly this was not the case. This disadvantageous performance might be as result that using just team forms, with or without including the home and away forms, is not just enough to make good predictions on the outcome of a football match.

4.3 Match Statistics

In the Match Statistics experiments we added a number of other attributes to the Team Form attributes. One notable attribute included in these experiments was the 'Total Form' attribute. The total form attribute calculates the cumulative form of each team; this cumulative calculation could be used to show teams that are high or low in the the overall league positions,based on league points gathered, thus indirectly indicating the "top", "middle" and "bottom" league teams after about half the number of matches played in the course of a single season. Other attributes used were shots, shots on target, goals, goal ratio and shot ratio.

These five attributes are five of the more simple ways to determine the winner of a football game. It is quite self-evident and self-explanatory that roughly 80% of the time one team has more shots than the other team, the team with more shots is more likely to win the game; more shots on target would even give that team a higher chance percentage of winning the game. However, shots ratio, which is the amount of shots on target to shots, provides better probabilities or likelihoods than both shots and shots on target. Shots ratio provides more richness to the data because the team with a higher shots ratio is more likely to win the game. This is not the case for the team with higher shots or shots on target.

The results gotten from the Match Statistics experiments were a lot better than the previous two experiments clearly because of the new attributes added. Although the new attributes used do not vary that much in relevance, they were each used with the team forms and total forms in order to give the attribute selections a wider range of relevance. The results, however, produced similar results.

4.4 Team Formation

The team formation experiments were a bit different compared to the other experiments, this is because team formations are a very dynamic attribute in a football match. Football is a very dynamic game, dynamic in the sense that: one minute the odds could be with team **a** and in the next minute team **b**. Likewise, team formations, in a single game, a team can change its starting formation (the formation the team starts the game with) more than once. These team formation in-game changes could swing the odds back to the team's favor, hence changing the whole dynamic of the game from that change of formation. However, the team formation experimented with in this thesis reflects only the starting team formations, and does not reflect this dynamic nature of team formations just discussed.

As mentioned in the experiment section of the team formation, we identified 20 unique team formations in our dataset. Out of these 20 unique team formations, 13 of them were used by less than 10% of the dataset while 4 of them were used by less than 1% and for this reason alone we felt the need to partition the team formations. We opted to partition them mathematically using integer sequences found at [32] where we discovered that there are 51,724,158,235,372 ways to be exact to partition 20 unique team formations mathematically. So instead of trying out over 51 trillion ways to partition the formations, we decided to use our expert knowledge worth over 10 years of watching football matches in the comfort of our local pubs or living rooms in partitioning the team formations. This was a more simpler method and it resulted in making ten (10) different partitions of what we assume are the main formations and the lesser ones based on the amount of time each one had been used in our dataset[6].

4.4.1 Binning by equal width

We used the total form attribute from the match statistics because as we mentioned earlier, the total form attributes can be used to establish the "top", "middle" and "bottom" teams after a good number of matches and based solely on points for the purpose of this research. Also mentioned earlier, the team positions can only be established and seem reliable after about half of the games in a season has been played. Although the length of this period or games taken to have a reliable league position table might slightly differ per league depending of the competitiveness of that league. We showed in Figure 1 how diverse the different leagues found in our dataset can be in regards to predictions. Regardless of the diversity, we divided the dataset

[6] The partitions can be seen in Table 6 under Section B of the Appendix

into three bins — bins that should each represent the value of games played by and against teams from different league positions. These three bins were:

- Bin 1: Bin 1 represented the total form value of games found between the interval 0,-1 and 0,1. Bin 1 had the most number of instances (4614);
- Bin 2: Bin 2 represented values found between the interval 1,2 and -1,-2. Bin 2 had 761 instances;
- Bin 3: Bin 3 represented values found between the interval 2,3 and -2,-3. Bin 3 had the least number of instances (30).

We decided to make these three different bins so we could know how much of an impact team formations[7] would have on each of the three situations found in the three bins because team formations, without subjective information or meta-data regarding its dynamism, could be just a string of numbers that might add zero to no value in the prediction of a football match. The results, however, gave some interesting insight.

Just before we give our insights on the results, we would like to explain what scenarios or situations we expected to be found in each of the three bins and what were the actual situations found in each of the bins.

In bin 1, we expected that games which resulted between the total form value of 0 to 1 and 0 to -1 were games played by teams of somewhat equal league positions regarding to points accumulated week in week out through the course of a single season. For example, we would expect games like top team versus top team, middle team versus middle team and bottom team vs bottom team. Whereas, bin 2 would have games played by teams of one category higher in league position against one category lower, for example, top teams versus middle teams or middle teams versus bottom teams while bin 3 would have games played by teams of the two extreme league positions like the top teams versus the bottom teams, thus the low amount of instances.

However, because these league positions are determined solely on the number of cumulative points gathered by a team, we expected a few exceptions in all of the three bins, exceptions that should be found in the first five months of each season (which usually begins in the middle of August up till the middle of January) when the number of games are not enough to establish reliable league positions. For example, in bin 1 where we expected

[7] From this point henceforth, team formation refers to the starting team formation that remain the same through the entire match, entailing one way of playing the formation through the match also

games played by teams of equal league positions, we found exceptions like the game played between Burnley and Manchester United on the 30th of August, 2014, a bottom league team against a top league team game; or the game played against Inter Milan and Sassuolo on the 14th of September, 2014, a top league team against a middle league team; of course there were many other instances found like this which most of them were played in the first five months of the season with not enough games to establish reliable league positions. Also in bin 2, we found exceptions like the game played against Bayern Munich and Dortmund on the 1st of November, 2014, which is a game played between teams of equal league positions; or the game played between Reading and Chelsea on the 30th of January, 2013, a bottom league team against a top league team. Finally, in bin 3, because of the rare case of teams getting the values between 2 to 3 and -2 to -3, we expected two things, that:

- the number of instances will be the least amongst the three bins and;
- as mentioned earlier, it would be a representative of games played between the top league teams and the bottom league teams.

Eventually, yes the number of instances was the least compared to the other bins but the latter was not so. The instances found in bin 3 were not representative of games played between the top and the bottom league teams alone but rather games played only in the first five months of the seasons. This eventual discovery became the deciding factor on how good and easy the prediction on bin 3 should have been and what it actually was.

The results shown in Figures 14 and 15 show the difference in accuracy levels with and without the inclusion of formations in each of the three bins. In bins 1 and 2, the accuracy remained more or less the same after the inclusion of team formations. In bin 1, the accuracy levels, both with and without the team formations are above the default accuracy of 45.57% but still not a good enough accuracy which entailed that instances in bin 1 are difficult to predict with or without the inclusion of team formations. While in bin 2 the accuracy levels, both with and without the inclusion of the team formations are about 20% more than the default accuracy, which was expected because of the type of games that are represented in bin 2. However, the inclusion of the team formations in both bins 1 and 2 had little or no effect, this could mean that matches played in these two bins are more likely to change formations during the course of a game or change how the formation is played during the course of the game through substitutions or coaching instructions.

Whereas, in bin 3, although both the accuracy levels were below the default accuracy meaning the prediction in this bin was very poor, there was about a four percent (4%) increase in accuracy level with the inclusion of team formations. We would expect the poorness of the prediction to be as a result of the low amount of instances used but as we showed in the learning curve in Figure 30, the accuracy level remains more or less the same regardless of the amount of instances in regards to this research. As we crossed out the number of instances being the reason for the poor predictions, we looked into what this bin was representative of and attributed the poor results to the fact that, unlike in bins 1 and 2 where the instances consisted of games that were played through the course of the entire seasons, all of the instances in bin 3 consisted of games played only in the first five months of the seasons, making the dataset very shallow in terms of the algorithm seeing a good enough picture to predict with. However, regardless of the poor accuracy results, the inclusion of team formations resulted in a four percent (4%) increase which could mean that teams do not change their formations or how the formations are played early on in the season as this bin consists of games played in the first five months of the seasons. Whereas, as the season progresses, teams change their formations or how their formations are played.

Below is a table showing the instances found in bin 3; the table also includes the final league position of both the home (**HTFLP**) and away teams (**ATFLP**) and the match-week[8] (MW) of each game played, as well as the result (**R**)of the game, the competition name, date of the match, home (**HT**) and away (**AT**) team names and total form (**TF**) value of each match. This table shows two things mentioned earlier regarding bin 3, that:

- bin three is representative of games played within the first and fifth months of the season and;
- bin three can not be said to be representative of games played between the top league teams and the bottom league teams

[8] The Bundesliga has 34 match-weeks while the other four leagues have 38 match-weeks

MW	Date	Competition	HT	AT	TF	HTFLP	ATFLP	Result
2	14/09/14	Serie A	Napoli	Chievo	3	5th	14th	A
3	21/09/14	Serie A	Roma	Cagliari	2.5	2nd	18th	H
2	23/08/14	Premier League	Aston Villa	Newcastle	3	17th	15th	D
2	23/08/14	Premier League	Chelsea	Leicester	2	1st	14th	H
2	23/08/14	Premier League	Swansea	Burnley	3	8th	19th	H
2	24/08/14	Premier League	Hull	Stoke	3	18th	9th	D
2	24/08/14	Premier League	Tottenham	QPR	3	5th	20th	H
3	30/08/14	Premier League	Man City	Stoke	2.5	2nd	9th	A
14	28/11/12	Premier League	Swansea	West Brom	2	9th	8th	H
2	30/08/14	Bundesliga	Leverkusen	Hertha	2	4th	15th	H
3	12/09/14	Bundesliga	Leverkusen	Werder Bremen	2	4th	10th	D
3	14/09/14	Bundesliga	Ein Frankfurt	Augsburg	2	9th	5th	A
2	15/08/14	Ligue Un	Caen	Lille	2	13th	8th	A
2	17/08/14	Ligue Un	Bordeaux	Monaco	3	6th	3rd	H
2	17/08/14	Ligue Un	St Etienne	Reims	2	5th	15th	H
14	24/11/13	Ligue Un	Nantes	Monaco	2	13th	2nd	A
5	25/09/14	La Liga	Valencia	Cordoba	2	4th	20th	H
20	11/01/13	Ligue Un	Paris SG	Ajaccio	2.153.846	1st	20th	D
2	13/09/14	Serie A	Empoli	Roma	-3	15th	2nd	A
2	14/09/14	Serie A	Lazio	Cesena	-3	3rd	19th	H
2	14/09/14	Serie A	Parma	Milan	-3	20th	10th	A
7	18/10/14	Serie A	Sassuolo	Juventus	-2.5	12th	1st	D
2	30/08/14	Bundesliga	Schalke 04	Bayern Munich	-3	6th	1st	D
2	30/08/14	Bundesliga	Wolfsburg	Ein Frankfurt	-3	2nd	9th	D
2	16/08/14	Ligue Un	Toulouse	Lyon	-3	17th	2nd	H
2	30/08/14	La Liga	Cordoba	Celta	-3	20th	8th	D
2	31/08/14	La Liga	Elche	Granada	-3	13th	17th	D
2	31/08/14	La Liga	Sociedad	Real Madrid	-3	12th	2nd	H
4	21/09/14	La Liga	Levante	Barcelona	-2.666.667	14th	1st	A
2	25/08/13	Premier League	Cardiff	Man City	-2.102.564	20th	1st	H

Table 4: Bin 3 instances

4.4.2 Binning by equal frequency

While the binning by equal width experiments were aimed at creating different match scenarios through each of the bins and comparing the accuracies from predicting these three different match scenarios with and without the inclusion of team formations; the binning by equal frequency was a more scientific experiment. In the binning by equal frequency experiments, the total form values, although used as the binning values, are ignored, in regards to the aforementioned statement that this experiment was completely scientific. Hence, bins 1 to 3 represent no match scenarios, neither presumed nor factual.

Regardless, the results given in figures 17 and 16 show that the inclusion of team formations give little to none impact in any of the bins. This could be as a result of the limited information and shallowness of our team formation data. It could also mean that randomly selecting a match is not the most ideal when making predictions with team formations, making match scenarios, as is done in the binning by equal width experiments, gives more insight and added advantage on what to expect when making predictions with team formations.

4.5 Formation Clustering

In the attempts to reduce the number of unique team formations we had in our dataset from 20, we clustered the team formations using the hierarchical clustering method. We clustered the formations for the same reason we partitioned the formations in the team formation experiments, because out of the 20 team formations, 13 of them were used by less than 10% of the dataset, while of that 13, 4 were used by less than 1% of the dataset.

Although, the team formations had already been partitioned as shown in Table 6, the partitioned formations shown in the table did not achieve the main objective which was to reduce the number of total team formations used by less than 1% of the dataset. Moreover, partitioning the team formations, as shown in the table mentioned above, was based on little expert knowledge. Factors like:

- how easy would it be to change to any of the formations partitioned together?
- how comparable in playing styles are the formations partitioned together?

were taken into consideration while partitioning similar team formations together.

We decided to partition the team formations in a more scientific approach, through clustering — hierarchical clustering. These formations (clustered) were based on 1) shots, and 2) shots on target. Both shots and shots on target based clusters produced interesting results as shown in their respective dendrograms. The primary objectives for clustering the team formations were to 1) scientifically, partition the team formations based on historic data and feature selection and 2) to the aim of having a bigger impact than the team formations partitioned based on expert knowledge when included in making predictions.

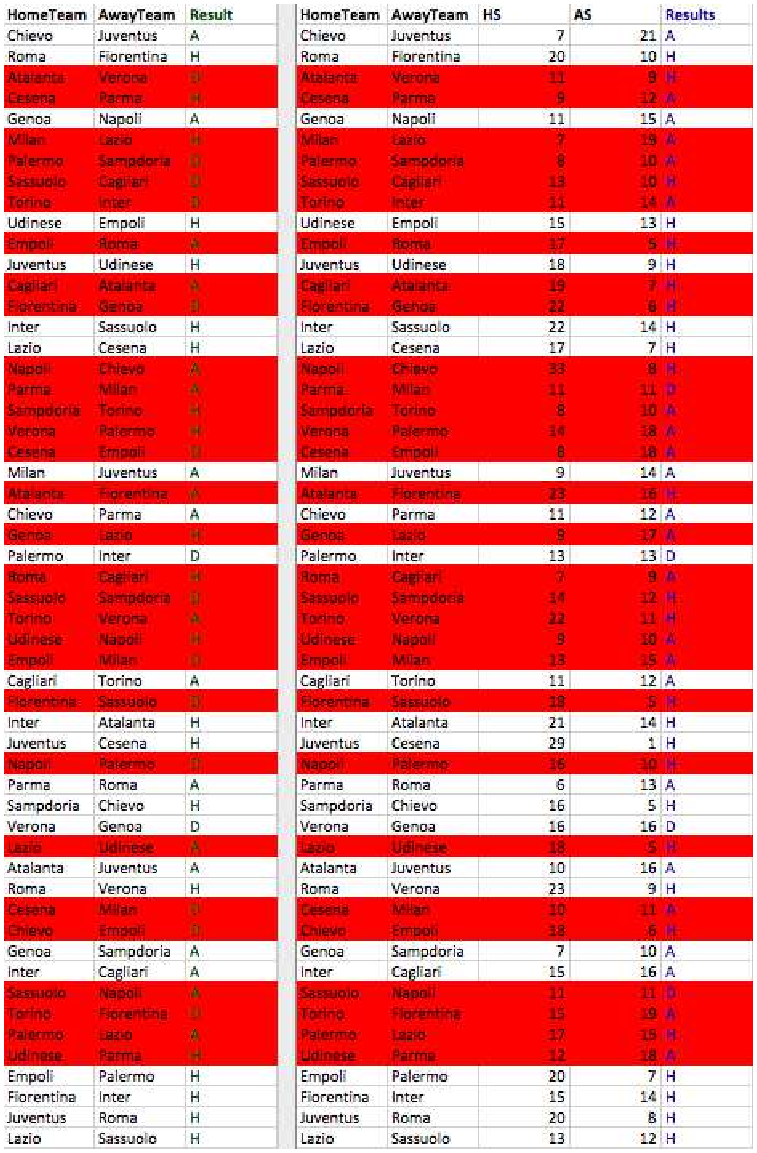
One of the two objectives mentioned earlier was more successful than the other. The formations were clustered based on feature selection, in our case, shots and shots on target. The comparison, however, was not as successful as we had presumed. We had presumed that the low impact both the original team formations and the partitioned (based on expert knowledge) had on the accuracy levels would be improved through a thorough scientific manipulation of the team formations.

As shown in Figure 28, the accuracy levels of all the various partitioned or clustered team formations were proven using the paired t-test[9] not to be statistically better or worse than the original dataset team formations. This statistically proven average accuracy levels, regardless the team formations being clustered or partitioned scientifically, based on expert knowledge or neither clustered nor partitioned, could be as a result of the limited information that can be derived from our team formation data (which has already been explained in the previous discussion section).

4.6 Shot-Driven Results

The shot-driven experiment was to make a comparison between the goal-driven results (which are the norm in this case) and the shot-driven results. As we explained in the experiment section of the shot-driven results, the aim of creating a shot-driven result dataset was to predict which team will have more shots at the end of the game. The team with more shots does not always win the game or when both teams have an equal amount of shots, the game does not always end as a draw — Figure 31 to support this argument. Figure 31 shows just over 1% of the whole dataset, the results on the left hand side (color-coded in green) show the goal-driven results while the results on the right hand side (color-coded in blue) show the shot-driven results and the rows filled with the red show the instances when a team has more shots but does not end up winning the game or when both teams have an equal amount of shots but the game does not end as a draw. Out of the 1%, just over 54% supports the claim made earlier that the team with more shots does not always win the game — which could mean that in approximately 54% of the time a team has more shots, that team does not end up winning the game. Hence, the comparison between the shot-driven results and the goal driven results should not be based on predicting the final outcome of the game.

[9] Paired t-test results shown under Appendix D



HomeTeam	AwayTeam	Result
Chievo	Juventus	A
Roma	Fiorentina	H
Atalanta	Verona	D
Cesena	Parma	H
Genoa	Napoli	A
Milan	Lazio	H
Palermo	Sampdoria	D
Sassuolo	Cagliari	D
Torino	Inter	D
Udinese	Empoli	H
Empoli	Roma	A
Juventus	Udinese	H
Cagliari	Atalanta	A
Fiorentina	Genoa	D
Inter	Sassuolo	H
Lazio	Cesena	H
Napoli	Chievo	A
Parma	Milan	A
Sampdoria	Torino	H
Verona	Palermo	H
Cesena	Empoli	D
Milan	Juventus	A
Atalanta	Fiorentina	A
Chievo	Parma	A
Genoa	Lazio	H
Palermo	Inter	D
Roma	Cagliari	H
Sassuolo	Sampdoria	D
Torino	Verona	A
Udinese	Napoli	H
Empoli	Milan	D
Cagliari	Torino	A
Fiorentina	Sassuolo	D
Inter	Atalanta	H
Juventus	Cesena	H
Napoli	Palermo	D
Parma	Roma	A
Sampdoria	Chievo	H
Verona	Genoa	D
Lazio	Udinese	A
Atalanta	Juventus	A
Roma	Verona	H
Cesena	Milan	D
Chievo	Empoli	D
Genoa	Sampdoria	A
Inter	Cagliari	A
Sassuolo	Napoli	A
Torino	Fiorentina	D
Palermo	Lazio	A
Udinese	Parma	H
Empoli	Palermo	H
Fiorentina	Inter	H
Juventus	Roma	H
Lazio	Sassuolo	H

HomeTeam	AwayTeam	HS	AS	Results
Chievo	Juventus	7	21	A
Roma	Fiorentina	20	10	H
Atalanta	Verona	11	9	H
Cesena	Parma	9	12	A
Genoa	Napoli	11	15	A
Milan	Lazio	7	19	A
Palermo	Sampdoria	8	10	A
Sassuolo	Cagliari	13	10	H
Torino	Inter	11	14	A
Udinese	Empoli	15	13	H
Empoli	Roma	17	5	H
Juventus	Udinese	18	9	H
Cagliari	Atalanta	19	7	H
Fiorentina	Genoa	22	6	H
Inter	Sassuolo	22	14	H
Lazio	Cesena	17	7	H
Napoli	Chievo	33	8	H
Parma	Milan	11	11	D
Sampdoria	Torino	8	10	A
Verona	Palermo	14	18	A
Cesena	Empoli	8	18	A
Milan	Juventus	9	14	A
Atalanta	Fiorentina	23	16	H
Chievo	Parma	11	12	A
Genoa	Lazio	9	17	A
Palermo	Inter	13	13	D
Roma	Cagliari	7	9	A
Sassuolo	Sampdoria	14	12	H
Torino	Verona	22	11	H
Udinese	Napoli	9	10	A
Empoli	Milan	13	15	A
Cagliari	Torino	11	12	A
Fiorentina	Sassuolo	18	5	H
Inter	Atalanta	21	14	H
Juventus	Cesena	29	1	H
Napoli	Palermo	16	10	H
Parma	Roma	6	13	A
Sampdoria	Chievo	16	5	H
Verona	Genoa	16	16	D
Lazio	Udinese	18	5	H
Atalanta	Juventus	10	16	A
Roma	Verona	23	9	H
Cesena	Milan	10	11	A
Chievo	Empoli	18	6	H
Genoa	Sampdoria	7	10	A
Inter	Cagliari	15	16	A
Sassuolo	Napoli	11	11	D
Torino	Fiorentina	15	19	A
Palermo	Lazio	17	15	H
Udinese	Parma	12	18	A
Empoli	Palermo	20	7	H
Fiorentina	Inter	15	14	H
Juventus	Roma	20	8	H
Lazio	Sassuolo	13	12	H

Figure 31: Shot-driven and Goal-driven Results

Figure 29 from the experiment section shows the "big five" leagues comparing the accuracy levels of the shot-driven results and the goal-driven results. The chart shows that the shot-driven results are much better than the goal-driven results. As we had mentioned earlier, predicting the final outcome of a football game is not a 'neutral ground' so to say for comparison between the shot-driven results and goal-driven results because goal-driven results are a 100% accurate regarding if a team has more goals, will always end up winning the game, whereas, it is not the same for the shot-driven results.

5 CONCLUSION & FUTURE WORK

This thesis paper was set out to show a comparison between different methods a football match can be predicted using feature selection, whether or not team formations add value in predicting a football match, what features would perform better than the others, what combination of features are best used together and their overall performances in predicting football matches. The general theoretical literature regarding football match predictions propose one method for predicting football games using quantitative data, qualitative data or a mixture of both, whichever type of data is used, a comparison of different methods and the use of team formations as a dependent feature to predict a football match is a gap usually left unanswered. This thesis paper sought to answer these two questions:

1. Which features or combination of features would produce better predictions?
2. Do team formations add significant value to football match predictions?

5.1 Conclusion

The results shown in this thesis paper show that:

- Using the total team forms produce better results compared to using a defined *n* form value;
- The total form, which could be equivalent to a team's league position or team strength is a good baseline feature for making predictions;
- Quantitative data like shots, shots on target, goals, goals ratio and shots ratio used in this paper produce similar results;
- Team formations — as just a string of numbers did not show to be of added value in making predictions;
- Clustering team formations based on historic data like shots, shots on target or goals do not produce realistic team formation clusters which could suggest the complexities that could be found in team formation.

The importance of football data in the world today is ever increasing, for this reason, conducting research in the area of sports data analysis is only as good as the data you can get your hands on. This, for us was a limitation encountered and need s to be considered in this research. Despite this limitation, we have been able to provide, using the data we had, insight regarding the research questions we posed at the beginning of this thesis paper.

5.2 Future Work

Team formations have shown to be as complex as the game of football itself, in the sense that team formations on its own has a number of other factors that influence how they are executed. For example, substitutions could affect how a team formation is played, injuries, player availability, team more, fatigue and so many more. If team formations are to become an independent feature in football match prediction, they need to be defined as a combination of other mini-factors like the factors mentioned earlier.

Based on our results, it is also recommendable to say that a mixture of qualitative data with quantitative data, which, according to general literature, has shown to always give better results, would produce a better comparison of methods for predicting a football match.

A
A description of Dataset

Meta feature	**Value**
Number of Instances	5478
Number of Classes	3
Per Class	
Home	2505
Away	1599
Draw	1374
Default Accuracy	45.73%

Table 5: Full Prediction Statistics

B
Team Formation Partitions

The team formation column consists of the main team formations that are being used; the partition column, the team formations that are partitioned into the main team formations; and the overall percentage of the new included main team formations. In any case where there are no team formations in the partitioned column, the main team formation remains as it were in the original dataset.

Team Formation	Partitioned Formations	Overall Percentage (%)
3-5-2	3-4-1-2	5.86
4-5-1	4-2-3-1; 4-1-4-1; 4-4-1-1; 4-3-2-1	59.13
5-3-2		0.61
5-4-1		0.26
3-4-3	3-4-2-1	2.79
4-4-2	4-1-3-2; 4-3-1-2	17.63
4-3-3	4-2-1-3; 4-1-2-3	11.3
3-3-3-1		0.19
3-5-1-1		1.17
4-2-2-2		0.56

Table 6: Team Formations Partitioned

C
A description of Attributes

The attributes are described in Table 7

Attribute Name	Description
HTF	Home Team Form
HTHF	Home Team Home Form
HTAF	Home Team Away Form
ATHF	Away Team Home Form
ATF	Away Team Form
ATAF	Away Team Away Form
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
FTR	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
TF	Total team form
SR	Shot ratio — Shots on Target divided by Shots
GR	Goal ratio — Goals divided by Shots on Target
FHT	Formation Home Team
FAT	Formation Away Team

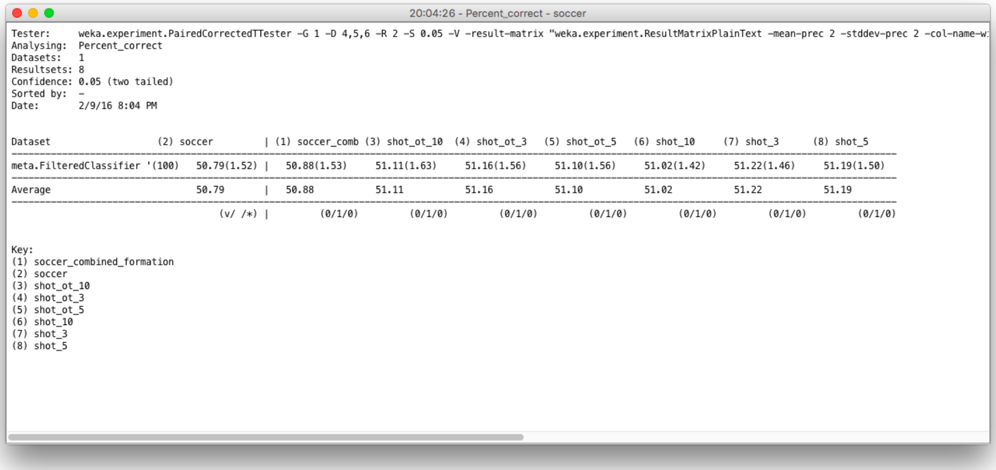
Table 7: Attributes of our dataset

D

Paired T-test Results

The figure below shows the paired t-test results for the Comparison of Accuracy results shown in Figure 28. The dataset labeled ‘soccer’ in the figure below represents the ‘Original Dataset Formation’ in the comparison figure, this dataset is used as the baseline for the comparison. The figure proves that none of the other datasets are statistically proven to be better or worse than the ‘Original Dataset Formation’ dataset.

This paired t-test experiment was carried out using Weka-3-7-13



```
Tester:     weka.experiment.PairedCorrectedTTester -G 1 -D 4,5,6 -R 2 -S 0.05 -V -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-w
Analysing:  Percent_correct
Datasets:   1
Resultsets: 8
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       2/9/16 8:04 PM


Dataset                   (2) soccer         | (1) soccer_comb (3) shot_ot_10  (4) shot_ot_3   (5) shot_ot_5   (6) shot_10     (7) shot_3      (8) shot_5
--------------------------------------------------------------------------------------------------------------------------------------------------------------
meta.FilteredClassifier '(100)   50.79(1.52) |   50.88(1.53)     51.11(1.63)     51.16(1.56)     51.10(1.56)     51.02(1.42)     51.22(1.46)     51.19(1.50)
--------------------------------------------------------------------------------------------------------------------------------------------------------------
Average                          50.79       |   50.88           51.11           51.16           51.10           51.02           51.22           51.19
--------------------------------------------------------------------------------------------------------------------------------------------------------------
                                     (v/ /*) |         (0/1/0)         (0/1/0)         (0/1/0)         (0/1/0)         (0/1/0)         (0/1/0)         (0/1/0)


Key:
(1) soccer_combined_formation
(2) soccer
(3) shot_ot_10
(4) shot_ot_3
(5) shot_ot_5
(6) shot_10
(7) shot_3
(8) shot_5
```

Figure 32: Paired T-test Results for Comparison of Team Formation Accuracy

E

Expert constructed Bayesian Network

Expert constructed bayesian network from [20] shown in figure below:

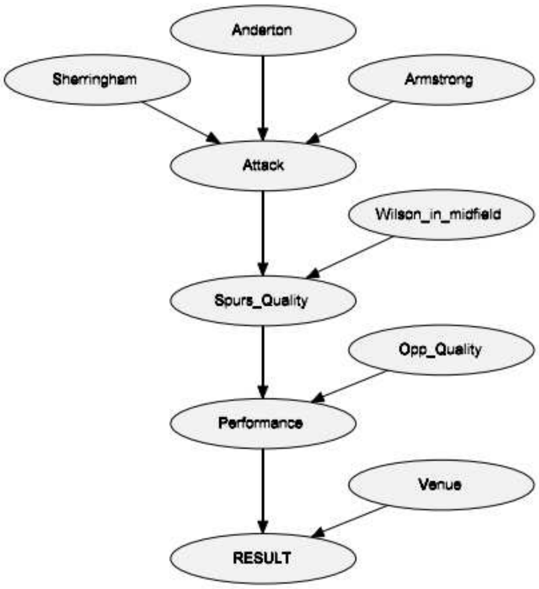


Figure 33: Expert constructed bayesian network for Tottenham Hotspur's performance

F A more detailed explanation on pi-football Components

Component 1: team strength, is derived mainly from objective information. Objective information like the number of total points accumulated in each of the five previous seasons called *previous information*, the amount of points obtained so far in the current season and points expected from residual matches, this is called the *current information*, and a third optional source, which is the *subjective information* that is derived when there is a dramatic change about the strength of a team just before the season starts. The example given of when such an instance occured was when Manchester City at the start of seasons 2009/10, 2010/11 and 2011/12, improving their strength by spending £160m, £77m and £75m respectively.

All three sources of the objective information just discussed are modeled to include a degree of uncertainty in them. For the previous information, the information becomes less important for each previous season, while the current information becomes more important after each successive match-week, and finally, the optional subjective information, this importance of this information is dependent on the degree of the expert's confidence regarding his/her indications. A high confidence breeds lower uncertainties, as a low confidence, higher uncertainties.

The subsequent components are mainly dependent on expert subjective information.

Component 2: team form, generates the team's recent performance which is the difference between the expected performance — represented by what the model had initially forecasted and its real performance during the five most recent match-weeks; this difference is then represented on a scale from 0 to 1, a value closer to 0.5 indicates the team's performing as expected and above 0.5 indicates performing higher than expected. Home advantage is also taken into consideration by assigning a heavier weight on home 'forms'. When this form has been generated, it is then revised according to subjective indications about the availability of certain players. These certain player (s) are categorized into (*i*) Primary key-player availability, (*ii*) secondary key-player availability, (*iii*) tertiary key-player availability(*iv*) remaining first team players availability and (*v*) first team returning players. The availability of these certain players, again are measured with the confidence levels of the expertise about the impact they can have on the outcome of the game.

Component 3: psychological impact, is indicated in two levels: level one is derived from the teams' head to head bias and the managerial impact; level two is generated from the teams' confidence, spirit and motivation level. The information accessed in level one is updated by that of level two, for both teams and regarded as the teams' psychological impact. These indications gotten from the experts are also limited to a degree of uncertainty.

Component 4: team fatigue, is determined by the toughness or difficulty of the previous match, the number of days gap since that match, the number of rested first team players (if any) and the number of first team players that have participated in international matches (if any). This component is also divided into two stages, these are: Stage 1 and Stage 2. Stage 1 is a combination of a 'restness' micro stage which consists of the number of days since last match and the number of first team players rested during the last match; and the toughness or difficulty of the previous match. This first stage is then updated in the second stage, as is done in the psychological impact component. The information in the second stage is gotten from national team participation (if any).

References

[1] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185

[2] V Barnett and S Hilditch. The effect of an artificial pitch surface on home team performance in football (soccer). *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 39–50

[3] Carl Bialik. The people tracking every touch, pass and tackle in the world cup. http://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/.

[4] AG Biichner, Werner Dubitzky, Alfons Schuster, Philippe Lopes, PG O'Doneghue, John G Hughes, David A Bell, Kenneth Adamson, John A White, and John MCC Anderson. *Corporate evidential decision making in performance prediction domains.* Morgan Kaufmann Publishers Inc., 1997.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32

[6] Joseph Buchdahl. *Fixed odds sports betting.* High Stakes, 2003.

[7] Anthony C Constantinou, Norman E Fenton, and Martin Neil. pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339

[8] Alexis Direr. Are betting markets efficient? evidence from european football championships. *Applied Economics*, 45(3):343–356

[9] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280

[10] Arpad E *The rating of chessplayers, past and present.* Arco Pub., 1978.

[11] Football-Data. Football results, statistics and soccer betting odds data. http://football-data.co.uk/.

[12] Football-Lineups. Football lineups, tactics, transfers, injuries and tournament data. http://www.football-lineups.com.

[13] David Forrest, John Goddard, and Robert Simmons. Odds-setters as forecasters: The case of english football. *International Journal of Forecasting*, 21(3):551–564 0169–2070, 2005.

[14] David Forrest and Robert Simmons. Globalisation and efficiency in the fixed-odds soccer betting market. *University of Salford, Centre for the Study of Gambling and Commercial Gaming*, 2001.

[15] Yoav Freund and Robert E Schapire. *Experiments with a new boosting algorithm*, volume 96. 1996.

[16] John Goddard and Ioannis Asimakopoulos. Forecasting football results and the efficiency of fixedodds betting. *Journal of Forecasting*, 23(1):51–66

[17] John Goddard and Ioannis Asimakopoulos. Forecasting football results and the efficiency of fixedodds betting. *Journal of Forecasting*, 23(1):51–66

[18] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470 0169–2070, 2010.

[19] Investopedia. Poisson distribution. http://www.investopedia.com/terms/p/poisson-distribution.asp.

[20] A Joseph, Norman E Fenton, and Martin Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553

[21] Tim Kuypers. Information and efficiency: an empirical study of a fixed odds betting market. *Applied Economics*, 32(11):1353–1363

[22] Alan J Lee. Modeling scores in the premier league: is manchester united really the best? *Chance*, 10(1):15–19

[23] Christoph Leitner, Achim Zeileis, and Kurt Hornik. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008. *International Journal of Forecasting*, 26(3):471–481 0169–2070, 2010.

[24] Mike J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118

[25] M Moroney. *Facts from figures.* Penguin, 1951.

[26] Alan M Nevill, Sue M Newell, and Sally Gale. Factors associated with home advantage in english and scottish soccer matches. *Journal of Sports Sciences*, 14(2):181–186

[27] Paul Nicholson. Stats swoop to add prozone to its growing global power play. http://www.insideworldfootball.com/world-football/42-news/16956-stats-swoops-to-add-prozone-to-its-growing-global-power-play.

[28] Richard Pollard. Evidence of a reduced home advantage when a team moves to a new stadium. *Journal of Sports Sciences*, 20(12):969–973 2002.

[29] J Ross Quinlan. C4. 5: Programming for machine learning. *Morgan Kauffmann*, 1993.

[30] C Reep, R Pollard, and B Benjamin. Skill and chance in ball games. *Journal of the Royal Statistical Society. Series A (General)*, pages 623–629

[31] Alexander P Rotshtein, Morton Posner, and AB Rakityanskaya. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4):619–630 2005.

[32] N. J. A. Sloane. Bell or exponential numbers: number of ways to partition a set of n labeled elements. https://oeis.org/A000110.

[33] Koya Suzuki and Kazunobu Ohmori. Effectiveness of fifa/coca-cola world ranking in predicting the results of fifa world cuptm finals. 2008.

[34] Mark Thompson. On any given sunday: Fair competitor orderings with maximum likelihood methods. *Journal of the American Statistical Association*, 70(351a):536–541

[35] A Tsakonas, G Dounias, S Shtovba, and V Vivdyuk. *Soft computing-based result prediction of football games.* Citeseer, 2002.

[36] Guido Van Rossum. An introduction to python for unix/c programmers. *Proc. of the NLUUG najaarsconferentie. Dutch UNIX users group*, 1993.

[37] Ian H Witten and Eibe *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.