American International University-Bangladesh

Bachelor Thesis 2016-17: Analyzing the Goal Contribution of English Club Midfielders

Thesis
Submitted By

Chowdhury, Koushik	**[13-23200-1]**
Chowdhury, Abrar Hamid	**[13-23872-1]**
Nayeem, Mohammod Mahmodul	**[13-23389-1]**
S.M Mahmudul, Hasan	**[12-23670-1]**

Supervisor

Mr Sharfuddin Mahmood
Assistant Professor
Department of Computer Science
Faculty of Science & Information Technology
American International University-Bangladesh

October-2016

Acknowledgements

We owe the gratitude to Mr. Sharfuddin Mahmood, Assistant Professor, Department of Computer Science, American International University-Bangladesh (AIUB) for his guidance and support from the beginning of the idea. His encouragement, feedback, and valuable suggestions were really appreciated.

We extend our gratitude to the entire academic and staff people of AIUB for their unconditional love, memories, and support throughout the academic life that gave us strength to complete the research.

We also would like to thank our families and friends, who have been on our side from the beginning of the journey.

Abstract

Football is the most popular sport, followed by all the countries. Professional clubs spend huge amounts of money to fetch good players for competition at the highest domestic levels. In its efforts to achieve performance-related goals, clubs have to ensure that they identify the right players. In this research, a data mining technique is used to predict the goal contribution of midfielders from English football clubs, which assists the club in making player selection.

With the help of five football seasons (English Clubs: 2011-12 to 2015-16) data, an attempt has been made to identify the potential of midfielders by using a classification model based on characteristics like weather, ground, competition level, timing of matches, opposition strength, substitution status, player form, and position of players. In data mining, various techniques were tested, such as BayesNet, Naïve Bayes, Naïve Bayes Multinomial, Logistic Regression (multinomial), Multilayer Perceptron, Random Forest, J48, and SMO. Out of these, tool SMO, which solves the quadratic programming problem related to Support Vector Machines (SVMs), gave the best result.

The performances were mainly assessed with a confusion matrix that offered information about the prediction of performance by midfielders and the events that included them. In the case of this experiment, SMO was seen to be better off than other algorithms and therefore most suitable as a classifier. Thus, this work focuses on presenting the benefits of data mining approaches, with consideration given to Sequential minimal optimization (SMO) in identifying player performance and advancing the area of sports analytics studies.

Contents

List of Figures

List of Tables

1 Introduction

Data mining, also known as knowledge discovery, is the method of analyzing big data from different points of view in order to find valuable knowledge that can be used in decision-making in various spheres, including business, healthcare, and sports. Data mining has a number of approaches that are still considered to be an emerging discipline that has the capability to develop as a new field of study and has a number of uses in areas of classification, clustering, and association rule mining [1]. More in particular, classification techniques have been employed to great extent for developing the models wherever predictions of future trends from historical patterns are possible.

In the context of sports, it is possible to find numerous applications of data mining that can improve decisions concerning players, a team strategy, and a result of the match. Due to the nature of this topic, football has been studied relatively well, given that it is the world's most popular sport. As football clubs pour their financial resources towards players more and more efficiently, strategies that base the chances of a club's success on data are likely to be adopted. This paper seeks to apply data mining on the goal contribution of midfielders in English football teams with a view to determining which of the players is most likely to contribute more in terms of goals given particular conditions of match play, position, competition level, etc.

The sample data for this research covers data from five football seasons (2011-12 to 2015-16) of 10 football clubs in England and several midfield players of these 10 clubs. There are various classification algorithms like BayesNet, NaïveBayes, NaïveBayesMultinomial, Logistic Regression (multinomial), Multilayer Perceptron, Random Forest, J48, SMO, etc., and all these were used on this data. The purpose was thus to establish which of the algorithms offered the highest predictive power of goal contribution amongst the midfielders. Out of these algorithms, SMO, which is a support vector machine implementation using sequential minimal optimization, proved to be the most efficient for this purpose.

The overall evaluation criterion employed here is the confusion matrix that measures the performance accuracy of the predictions. Classification carried out by the model built with SMO offered the highest values of accuracy

and reliability compared to the other algorithms. Thus, this paper enriches the field of sports analytics by illustrating how data mining methods can be used successfully in football with a focus on the prediction of a player's performance. With this kind of focus on midfielder's goal contribution, this similar work is useful information for clubs seeking to improve their selection methods and the development of the best performances for future target matches.

2 Background

2.1 Literature Review

Data mining is also known as knowledge extraction, which has greatly assisted business organizations in making analyses on large data sets and coming up with meaningful information. Data mining is used for different purposes in different fields, and there are lots of contributions from the side of the researchers in different fields like prediction, performance measurement, forecasting, etc. The major objective in most data mining research is to predict future occurrences given past records by employing different algorithms and frameworks. In this section, we aim to demonstrate that through the given analysis of various areas, it is possible to identify examples of data mining applications as well as to demonstrate how this study advances the existing knowledge on data mining techniques. Yeom and Bentley (1995) have noted that a lot of work has been done while identifying the predictors of employee performance and its relation to compensation. For instance, some of the works that have applied data mining have been in determining the probable salaries for the employees depending on their performance, the job posts they hold, and other categories [1]. This line of research seeks to understand the factors that affect an employee's job performance, which results in compensation. Such studies fall under the general category of data mining, where different data classes are classified in order to predict the outcome, which is an area of focus and keen interest in our work on the goal contribution of football players. The applicability of data mining has also moved to the area of social networks and opinion mining. Some of the researchers have used data mining techniques to develop algorithms for sentiment analysis and opinion mining on social networking sites, which is an area of study that is relatively new and growing rapidly. By evaluating large data sets from social media sites, researchers have been able to determine crowd moods or trends from patterns observed on user intercourse [2]. While this work is mainly related to the context of social networks and sentiment analysis, the main approaches considered here, such as classification and prediction in data mining, are directly related to sports informatics, where players' attribute data is employed for predicting their performance.

Data mining techniques have especially found relevance in the healthcare field in areas of medical diagnosis and treatment prognosis. Medical data can be particularly useful, as it is possible to classify individuals based on their disease and offer a suggestion on how to overcome it. For example, some researchers have used the decision tree algorithm, including the use of J48, to develop computerized systems that can predict the existence of diabetes in patients [3]. This approach is analogous to that applied in sports science, where characteristics of performance are captured in order to predict values like goals scored. The capacity to sort data by certain characteristics (like medical signs, players's positions) is considered a crucial advantage and potential of data mining in various fields.

Another significant field of investigation is the comparison of one decision tree algorithm with another one. Decision trees that include the ID3 and C4.5 have been widely incorporated in various sectors because of their efficiency and ease in dealing with classification problems. These algorithms have been compared in various setups based on their performance and, more importantly, how well they handle big data sets [4]. In these studies, the goal is to draw a comparison between the different algorithms to find out which algorithm provides the most accurate prediction, which is an important question when using football analytics.

Our WEKA workbench contains the data and classifiers such as BayesNet, NaïveBayes, NaïveBayesMultinomial, Logistic Regression, Multilayer Perceptron, Random Forest, J48, and SMO. Out of these algorithms, SMO, which involves Support Vector Machines (SVM), offered the highest balance as far as classification is concerned. SVMs have been well researched and are popular for their capability to deal with a large number of features and non-linearity. Of these, SMO is optimized for training SVMs and thus appropriate to use in classifying player data for our study. Picking up from the analysis, the study conducted here provided evidence that SMO was instrumental in sports analytics and notably bringing out the data contributions that speak to player's performances.

The research highlighted above demonstrates how data mining techniques are useful in different fields, as shown above. Data mining is a powerful tool that can be used in everything, ranging from predicting the performance of an employee or a group of employees to coming up with diagnoses of a certain disease or even predicting the results of certain sports by analyzing the data that has been collected. Our research extends this line of work by employing state-of-the-art data mining methods for goal prediction in football. Thus, attention is paid to a certain group of players (midfielders) and several characteristics; thus, we expand the coverage of the existence of sports analytics as a science and true possibility of data mining during the

decision-making process within football teams.

2.2 Data Mining

Data mining is the process of computation where patterns are discovered with the help of large data sets. It extracts information from the data set and transforms it into a structure. There are mainly two types of techniques used in data mining. One is clustering, and another one is classification.

Classification is a branch of supervised learning technique in which class labels are predetermined. This technique is also known as working with supervised data or with labeled data. Labeled data can be categorized into two types: Categorical and Inferential On the basis of the established research objectives and hypotheses, the study used two types of data analysis techniques that include categorical and numerical.

While **clustering** is the process of categorizing related objects in one cluster according to the features that define the objects. This technique comes under unsupervised learning, which means that the class labels and the dataset are not specified. Unsupervised learning works with the instances that are pure unlabeled data that refers to instances of categorized data that are also non-numerical. As opposed to the k-nearest neighbor, or KM method, the results of the clustering are not skewed by the input variables in which we are interested, which is dependent in this scenario.

2.3 WEKA and Python

Weka is a machine-learning tool. It was implemented in Java and developed by the University of Waikato, New Zealand [5]. It is a free tool. It contains different types of classification and clustering algorithms. With Weka, one can easily analyze data and can make a decision with that analysis. With its visualization tools and algorithm, a prediction model can be made. Along with WEKA, we also used Python for basic graph generation. Though working in WEKA is relatively simpler than Python, there are some tasks one cannot do in WEKA, such as different graph generation for evaluating the outcome and project showcasing. At the end of the report, a simple inferance based on the best classifier has been displayed with the help of Python.

2.4 English Club Football

Football is the major sport in England. It was rumored that the first evident football match had been played in 1170 [6].

> "After dinner all the youths of the city goes out into the fields for the very popular game of ball [6]" **William FitzStephen (died c. 1191)**

First-team football has been played since 1581 [6], but football was developed in the 19th century. Modern football was first introduced in 1863 [6], but League was started in 1888, which was created by William McGregor, who was a director of Aston Villa [8]. The league was based on the principle of promotion and relegation format. There are different levels based on that format. Level one (top level) is known as the Premier League, containing 20 clubs. Our focus is on the 2011-12 to 2015-16 season and the ten elite football clubs in England.

2.5 Midfielder

The midfielder role is very important in a team. A midfielder can make his team win through his performance. A midfielder is a position in a football game between the forwards and defenders. There are various types of midfielders. Central midfielder, attacking midfielder, defensive midfielder, and wide midfielder. A box-to-box midfielder is known as a center midfielder. They should have good abilities of passing, tackling, breaking the defense, and shooting. Good dribbling capability is also required. They run to the opponent's area with the ball to score or make the ball to the forward. Left and right side midfielders are considered wide midfielders. They are positioned near the touchlines. Sometimes they need to cross the ball to the opponent's area. Very good running skills are required for that position. Attacking midfielder is positioned between forward and center midfielder. They are known as playmakers. They are acquainted for their dribbling skills, shooting from ranges, and accurate passing abilities. They can play in right, left, and center. The defensive midfielder role is to play in the midfield in a defensive mood and help the defenders of the team to stop the attack of the opponent. Very good tackling skills are required for that position. Defensive midfielders are not the focus of this research because they have very few possibilities to contribute to a goal. Scoring goals or assisting goals are not their responsibilities.

2.6 Attribute Selection

Nine attributes were selected (8 input variables and 1 output variable.). Time, ground, position, competition, weather, opposition strength, substitution status, player form, and class. Analysis has been done to predict the goal contribution based on these attributes.

Attribute	**Value**
Time	0-15, 15-30, 30-45, 45-60, 60-75, 75+
Ground	Home, Away
Position	AML, AMC, AMR, ML, MC, MR
Competition	Premier League, Europe, Cup
Weather	Clear/Sunny, Rain/Cloudy, Windy/Snow
Opposition strength	Weak, Medium, Strong
Substitution status	Starter, Substitute
Player form	Poor, Average, Good
Class	Low, Below Average, Average, High

Table 2.1: Attributes and their values.

1. **Time**: Specific time where a player makes a goal contribution.
    - 0-15: 0 to 15 minutes.
    - 15-30: 15 to 30 minutes.
    - 30-45: 30 to 45 minutes.
    - 45-60: 45 to 60 minutes.
    - 60-75: 60 to 75 minutes.
    - 75+: 75 plus minutes.
2. **Ground**: where has been played.
    - Home or Away.
3. **Position**: Specific position of a midfielder
    - AML: Attacking Midfielder Left.
    - AMC: Attacking Midfielder Center.
    - AMR: Attacking Midfielder Right.
    - ML: Midfielder Left.
    - MC: Midfielder Center.
    - MR: Midfielder Right.
4. **Competition**: Competition usually an England club participates in.
    - Premier League: Major competition.
    - Europe: UEFA Champions League + UEFA Europa League.
    - League: League Cup, FA Cup.

5. **Weather**:
    - Clear/Sunny: Can be clear or sunny.
    - Rain/Cloudy: Can be rainy or cloudy.
    - Windy/Snow: Can be windy or snow.
6. **Opposition strength**: How you classified the opposite team.
    - Weak, Medium, Strong.
7. **Substitution status**:
    - Starter (start from the beginning).
    - Substitute (start from the bench).
8. **Player form**: In the last 5 games.
    - Poor, Average, Good.
9. **Class**: Contribution rate (%). This contribution rate was calculated based on the overall goal contribution, such as goals or assists.
    - Low: 0-10.
    - Below Average: 10-20.
    - Average: 20-30.
    - High: 30+.

3 Methodology

3.1 Data Classification Process

Two-step has followed in data classification [1]. The first step is known as the learning step. It explains a predetermined set of classes that is designed based on analyzing a set of data instances. Each instance belongs to a predefined class. In the final step, the data is tested using different data mining techniques that are used to calculate the classification accuracy. After that, a model is designed that predicts the future based on the past data. There are several data mining techniques for classification. In our research, eight data mining classifiers have been used as follows: BayesNet, NaïveBayes, NaïveBayesMultinomial, Logistic Regression (multinomial), Multilayer Perceptron, Random Forest, J48, and SMO. These classifiers are most frequently used in previous data mining research [1]. [2] [3] [4].

3.2 Data Collection

One of the major parts of this research was data collection. Data was collected from whoscored.com [6]. From 2011-12 to 2015-16 season data was collected. Several midfielders and ten English clubs were observed. For example, each of the teams participates in a different competition in a season, but not all the clubs participate in the same competition. Common competitions are the Premier League, FA Cup, and League Cup, where some of the clubs compete in UCL (championship league) or UEL (European league) based on their previous season ranking in the Premier League. 4 out of 10 clubs that we selected for the data collection never played in UCL or UEL from 2011-12 to 2015-16. The following table shows the details of each club's involvement in competition in these 5 seasons.

Club Name	Competition (in 5 seasons [2011-12 to 2015-16])
Liverpool	Premier League, UCL, UEL, League Cup, FA Cup
Chelsea	Premier League, UCL, UEL, League Cup, FA Cup
Arsenal	Premier League, UCL, UEL, League Cup, FA Cup
Manchester United	Premier League, UCL, UEL, League Cup, FA Cup
Manchester City	Premier League, UCL, UEL, League Cup, FA Cup
Tottenham Hotspur	Premier League, UEL, League Cup, FA Cup
Leicester City	Premier League, League Cup, FA Cup
West Ham United	Premier League, League Cup, FA Cup
Stoke City	Premier League, League Cup, FA Cup
Everton	Premier League, League Cup, FA Cup

Table 3.1: Club Names and Competitions: Participated in 5 Seasons

3.3 Process of Data Collection

Whoscored [7] has been followed throughout the data collection. Data was collected match by match with the existence of selected midfielders. In the match center option, midfielder position, weather, ground, competition, and contribution time (if contributed) have been viewed. With respect to games, data were collected player by player. A simple illustration has been given.

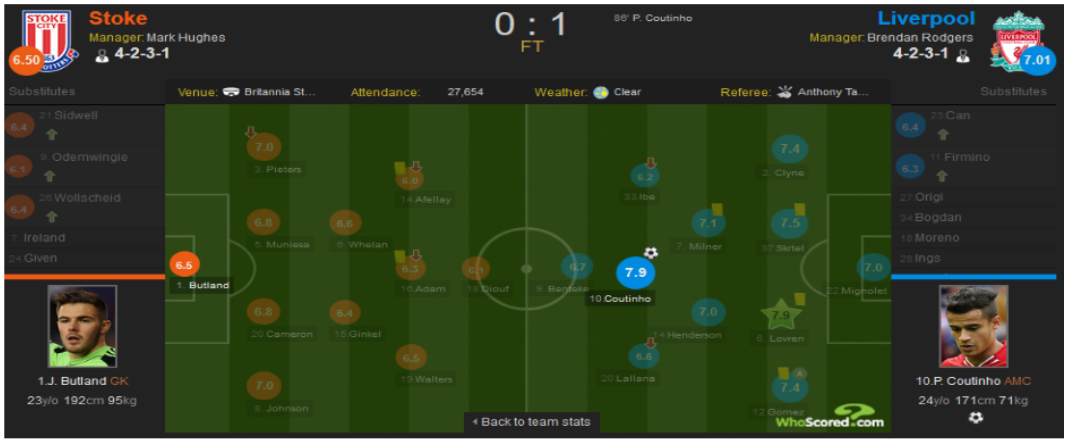

Figure 3.1: A simple illustration of data collection

If we take account of Fig. 3.1, it was a Premier League game, the weather was clear, and the match was being played in Stoke City's home ground, which is away for Liverpool. We categorized Stoke City as a weak opposition. We collected the information of Coutinho here who played the match as an AMC who started the match from the beginning (starter), and according to previous five match records, we identified his form as average, and he made a goal contribution by scoring a goal after 75 minutes.

Time	Ground	Position	Competition
75+	A (Away)	AMC	PL (Premier League)
Weather	**Opposition Strength**	**Substitution Status**	**Player Form**
Clear	Weak	Starter	Average
Total	1 (Goal)		

Table 3.2: Player Performance Overview

3.4 Data Preparation

After collecting the data, all player data was merged based on their similarities. Similarities included the value of the attribute. Then the merged data has been transferred to a csv sheet. The data has been partitioned into two datasets. The first one includes the data of percentages that has some zero percentage data, and the second one is designed by removing the zero percentage data. Classes have been set for every instance. That means classes were predefined. After that, CSV sheets are designed based that.

These CSV sheets were prepared, and they were converted into '.arff' format because Weka accepts only '.arff' format. Each of the datasets was used in an individual experiment. The first dataset has 2460 instances, and the second dataset has 755 instances. The class is the conversion rate, for example, how many contributions a player makes in a game within the time interval mentioned in the 'TIME' input variable. Class 1 is the lowest, and class 4 is the highest in terms of contribution rate.

Data Set 1: with 0% instances
Data Set 2: by removing 0% instances

3.5 Analysis with different Data Mining techniques

3.5.1 10 Fold Cross-Validation

Cross validation is a data mining technique. It evaluates predictive models by partitioning the data set into training sets to train the model. It gives the presumption to predict the model that is more accurate. One of the very common techniques is K-fold cross-validation. For classification problems, K-fold cross validation is used. For 10-fold validation, K = 10. It disintegrates the data set into 10 sets of size n/10. It trains on nine datasets and tests on one. It repeats 10 times to calculate the estimation. To do work with 10-fold

cross-validation, the Weka tool has been used. The idea was collected from references [1] and [2]. Results are given below.

Note: [8]
Accuracy = (TP + TN) / (P + N)
Precision = TP/ (TP+FP)
Error Rate = 100-Accuracy
Where T stands for true, P stands for positive, and N stands for negative.

Table 3.3: 10-fold cross-validation (**Dataset 1)**

Algorithms	Accuracy	Precision	Error Rate
BayesNet	0.972	0.95	0.028
NaïveBayes	0.964	0.939	0.036
NBMultinomial	0.946	0.898	0.054
LRMultinomial	0.919	0.892	0.081
Multilayer Perceptron	0.901	0.872	0.099
Random Forest	0.893	0.874	0.107
J48	0.987	0.974	0.013
SMO	0.988	0.981	0.012

Table 3.4: 10-fold cross-validation (**Dataset 2**)

Algorithms	Accuracy	Precision	Error Rate
BayesNet	0.912	0.841	0.098
NaïveBayes	0.896	0.817	0.104
NBMultinomial	0.877	0.787	0.123
LRMultinomial	0.849	0.771	0.151
Multilayer Perceptron	0.862	0.826	0.138
Random Forest	0.872	0.864	0.128
J48	0.936	0.903	0.067
SMO	0.943	0.919	0.057

3.5.2 Split 66.0% train, remainder test

In WEKA, a typical way of model assessment is the partitioning of data into training sets and testing sets that is used to measure the performance of a given machine learning model. 66% of training entails that 66% of data is used to train the model so that the model can learn the relationships and

patterns about the data. Hence, the rest 34% is set apart for testing, which also enables the evaluation of the model on unseen data to measure its ability to generalize. It also gives an opportunity to train the model on a large part of the data while at the same time testing its ability on a different set of data. By applying the 66/34 split, one is able to determine various performance indexes, such as accuracy, precision, and others, and avoid the problem of overfitting and guarantee that the results of the assessment conform to the model's performance on real data. Results are given below for both dataset 1 and dataset 2.

Table 3.5: Split 66.0% train, remainder test (**Dataset 1**)

Algorithms	**Accuracy**	**Precision**	**Error Rate**
BayesNet	0.954	0.909	0.046
NaïveBayes	0.947	0.898	0.053
NBMultinomial	0.927	0.858	0.073
LRMultinomial	0.918	0.865	0.082
Multilayer Perceptron	0.902	0.894	0.098
Random Forest	0.893	0.844	0.107
J48	0.991	0.982	0.009
SMO	0.993	0.984	0.007

Table 3.6: Split 66.0% train, remainder test (**Dataset 2**)

Algorithms	**Accuracy**	**Precision**	**Error Rate**
BayesNet	0.849	0.745	0.151
NaïveBayes	0.811	0.721	0.189
NBMultinomial	0.769	0.650	0.231
LRMultinomial	0.787	0.748	0.213
Multilayer Perceptron	0.803	0.794	0.197
Random Forest	0.812	0.784	0.188
J48	0.934	0.901	0.066
SMO	0.943	0.913	0.057

4 Evaluation

4.1 Comparison of experiment mode by accuracy

4.1.1 For Dataset 1

The performance of several algorithms was compared based on the 10-fold cross-validation accuracy and 66% split accuracy, to bring out some points. J48 and SMO displayed the best accuracy in both evaluation metrics, recording 10-fold cross-validation accuracies of 98.7% and 98.8% and 66/100 split accuracies of 99.1% and 99.3%, respectively. This means that J48 and SMO are the most effective at generalizing over different subsets of data as well as on a held-out split for prediction purposes. BayesNet and NaïveBayes also show good results, with 10-fold cross validation accuracies of 97.2% and 96.4%, respectively, and respectedly 66% split accuracies being equal to 95.4% for the first one, while the second one returned an accuracy that reached only 94.7%. However, their performance is not as high compared to J48 and SMO in terms of predictability. The next two algorithms are

Table 4.1: Comparison of Accuracy: 10-fold Cross-Validation vs Split 66% (**Data set 1**)

Algorithms	**10-fold CV Accuracy**	**66% Split Accuracy**
BayesNet	0.972	0.954
NaïveBayes	0.964	0.947
NBMultinomial	0.946	0.927
LRMultinomial	0.919	0.918
Multilayer Perceptron	0.901	0.902
Random Forest	0.893	0.893
J48	0.987	0.991
SMO	0.988	0.993

NBMultinomial with 10-fold CV accuracy at 94.6%, whereas LRMultinomial has a 10-fold CV accuracy of 91.9%. In general, these results are lower than those obtained from the best-performing algorithms, although they indicate reasonable accuracy still exists within them all together (which may cause

confusion). Therefore, multilayer perceptron did slightly worse, having 10-fold CV equal to 90.1% along with corresponding splits reaching just over 90%. They did not capture data complexity satisfactorily because two other models presented moderate performance rates, such as Random Forests, having it around 89% for both experiments.

Overall, J48 and SMO have the highest accuracy among all the algorithms presented in the work and clearly enforce their ability to predict results with high efficiency, while the rest of the algorithms also state essential levels of efficiency and high results in the classification tasks.

4.1.2 For Dataset 2

Similarly for Dataset 2, SMO and J48 performed better for both experiments, though SMO is slightly better than J48. It is also noticable that except BayesNet, J48, and SMO, other algorithms provide accuracy less than 90% for 10-fold CV. NBMultinomial and LRMultinomial provide accuracy less than 80% for Split 66% experiment. If we compare the accuracy from dataset

Table 4.2: Comparison of Accuracy: 10-fold Cross-Validation vs Split 66% (**Data set 2**)

Algorithms	**10-fold CV Accuracy**	**66% Split Accuracy**
BayesNet	0.912	0.849
NaïveBayes	0.896	0.811
NBMultinomial	0.877	0.769
LRMultinomial	0.849	0.787
Multilayer Perceptron	0.862	0.803
Random Forest	0.872	0.812
J48	0.936	0.901
SMO	0.943	0.943

2 with dataset 1, we can observe that for the 10 fold CV and Split 66% experiment, accuracy is decreased highly. For example, if we take our best two classifiers, such as SMO and J48, we can observe that accuracy dropped 4.5% for SMO and 5.1% for 10-fold CV. Similarly, the accuracy dropping in Split 66% is also highly noticeable.

4.2 ROC Graph

ROC stands for receiver operating characteristic and is a very important measure that is used to compare the performance of binary classification

models where every record needs to be classified as either positive or negative [8]. It defines the so-called true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis, revealing the strengths and trade-offs between sensitivity and specificity.

It is important to understand here that ROC (Receiver Operating Characteristic) curves are inherently proposed for binary classification problems, but these can be nicely used for multiclass classification problems by the technique called 'one-vs-rest' or 'one-vs-all' [9]. In this approach, the multiclass problem is transformed into several two-class problems. In particular, if the problem is of the classification kind, K curves of receiver operating characteristic (ROC) are constructed, each of which the curves relates to one of the K classes. In particular, each binary problem involves the separation of the class of interest from all the other classes. This is done by using the label binarizing function, which compares each of the true labels to all other classes, resulting in a binary form of the multiclass target variable for each class. For each of these binary issues, the ROC curve is calculated by plotting the true positive rate, known as sensitivity, and the false positive rate, which is equal to 1-specificity, at various thresholds.

This leads to the generation of an ROC curve for each of the classes, where ROC really stands for Receiver Operating Characteristics, meaning the ability of a classifier to distinguish between a specific class and the entire other classes. The average area of the ROC curve, which is known as the AUC of each class, is then computed to ascertain the capability of the classifier to correctly categorize instances of the class in the test set as well as the class itself when compared with other classes. By so plotting these curves, one gets a view of the classifier performance across all the classes in case of a multiclass classification problem so that one can be in a position to determine whether the classifier performs well as expected in this type of classification problem. This is helpful in understanding how the classifier performs as compared to each class, despite the fact that for binary classification, ROC curves are common.

4.2.1 For Dataset 1

Since from the accuracy we can see that both J48 and SMO provide the highest accuracy, and also the accuracy of BayesNet is close to both J48 and SMO, we drew the ROC curve on these 3 classifiers. Also, we only consider the result from 10 fold CV, since in this experiment mode, the better results were observed than Split 66% experiment.

Figure 4.1 represents a Receiver Operating Characteristic (ROC) curve for Dataset 1, displaying the performance of three different classification algorithms: To solve the problem, several classifiers, such as BayesNet, SMO, and J48, have been applied with four different classes, namely, low (1), below-average (2), average (3), and high (4). Different curves on the graph show the performance of the classifier for the different classes; TPR is plotted on the Y axis and FPR is plotted on the X axis. The ROC curve is different from the normal curve and is widely used in performance measurement of the classification algorithms with a special focus on binary and multi-class classification, as described in this graph. Closeness to the upper left corner of the plot indicates better performance of the classifier. In figure 4.1, the Area

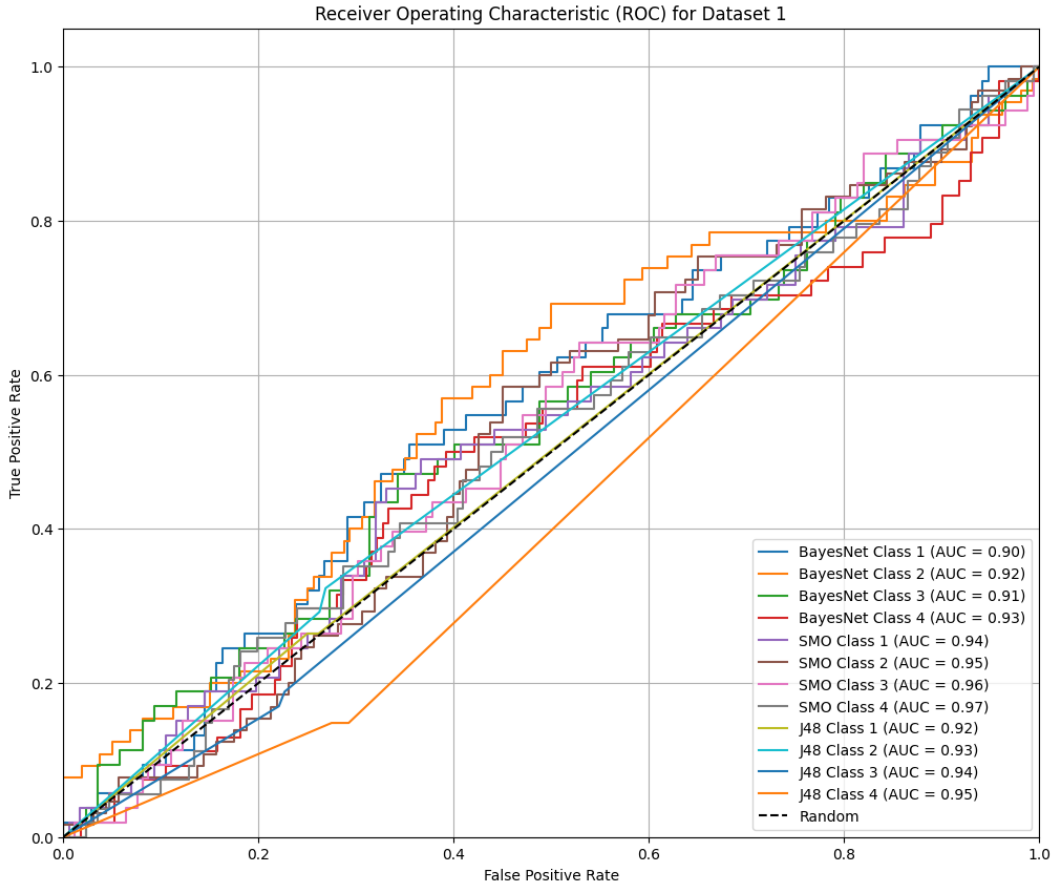

Figure 4.1: ROC curves for SMO, J48, and BayesNet (10 fold cross validation, dataset 1)

Under the Curve (AUC) of every classifier and class is shown that directs towards the efficiency value. The value of AUC for the classifier should be high; the greater the value of AUC, the better the result obtained by the classifier. As for the AUC values indicated below, they vary between 0.90 and 0.97, which affirms the fact that the majority of the models performed well. It is also important to make a note that SMO Class 4 has the highest

AUC of 0.97, while the lowest value of SMO was obtained from Class 1 with a value of 0.94. Thus, This model show a very good working of sensitivity and specificity for these classes. The dotted black line here denotes the random classifier or the baseline. The lines of all classifiers are above this line, which gives a deeper insight into the fact that all the sent trained models are performing well. It is thus also an effective means of ascertaining the extent to which these models are able to differentiate between true positive observations and false positives at various classes.

4.2.2 for Dataset 2

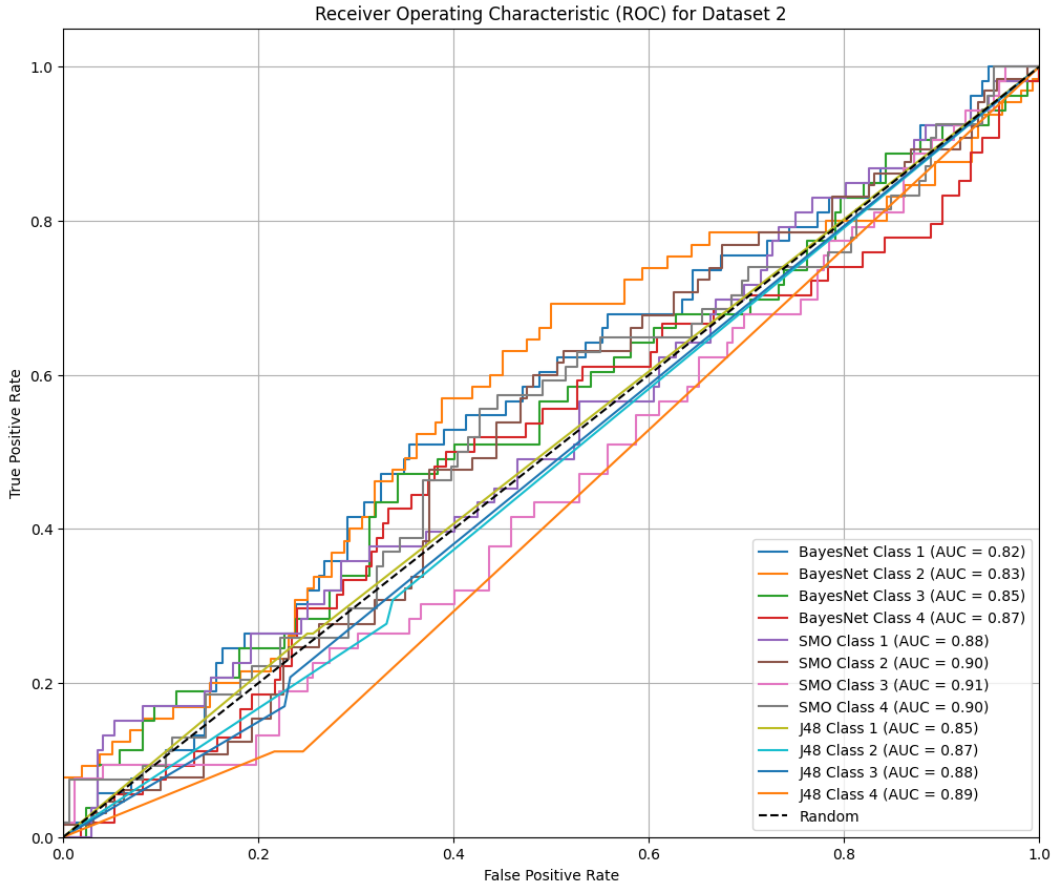

Figure 4.2: ROC curves for SMO, J48, and BayesNet (10 fold cross validation, dataset 2)

Similary for dataset 2, only 10-fold CV was considered, and also the ROC curve was generated based on BayesNet, SMO, and J48 results. AUC values are lower here than the previous ROC curve (Figure 4.1). It is mainly because dataset 2 has fewer instances than dataset 1. Therefore, the classifier got

fewer instances to train. Still, BayesNet, SMO, and J48 provide good AUC values in the range of 0.82-0.91.

Similarly, like Figure 4.1, SMO provides the best result among others. SMO class 3 provides the highest AUC value among the others, but in the previous curve (Figure 4.1), we notice SMO class 4 provides the highest AUC.

4.3 Test with Test data

For the purpose of the comparison, three classifiers, namely SMO, J48, and BayesNet, were tested on a subset of the test data. The purpose of this was to determine the ability of these models to classify unseen data after they have been trained on a different dataset. Selection of these classifiers was feasible owing to their excellent performances as compared to other classification models during preliminary testing. More so, it was evident that the SMO and J48 classifiers would successfully predict on other instances some of the instances that have high accuracy, making the models remarkable.

This work employed the testing phase to assess the accuracy of the classifiers in predicting the instances as follows: While in the testing phase, we randomly chose eight instances from the test set to test for the effectiveness of these classifiers. Both the classifiers predicted the class of the instances with one hundred percent accuracy and correctly classified all the eight instances at hand without any misclassification at all. This is evident in their ability to deliver accurate classification results, as will be demonstrated later in this paper. Nevertheless, BayesNet, which generally performs well for the dataset, misclassified the class for one instance, thereby indicating that even the classifiers with high levels of accuracy may occasionally end up misclassifying some instances.

Actual Result	**SMO Result**	**J48 Result**	**BayesNet Result**
High	High	High	High
High	High	High	High
Low	Low	Low	Low
Below-average	Below-average	Below-average	Below-average
Low	Low	Low	Below-average
High	High	High	High
Average	Average	Average	Average
Below-average	Below-average	Below-average	Below-average

Table 4.3: Comparison of Actual Results with **predicted** SMO, J48, and BayesNet Results

The test data set is a subset of the original dataset and was never used in any training stage or process. Its intention was for it to act as a test set designed to gauge the performance of the classifiers as they relate to unseen datasets. This type of testing is very essential when it comes to evaluating the ability of classifiers when they are used out of the training area.

This approach has aimed at evaluating the classifiers' ability to assign a correct class to the test instances compared to the true labeling of the test data set. In this experiment, the SMO and J48 models have generated the correct classes, while the BayesNet has just 1 incorrect classification, which is a note of interest because the efficiency of the models is quite sensitive to the classifier and the nature of the data.

4.4 Finding the best classifier

Therefore, based upon all the accuracy metrics and classifier statistics shown above, the most effective classifier for the specified datasets is SMO. Comparing the results obtained through two datasets, it can be concluded that the proposed SMO classifier is offering superior accuracy and better class discriminating ability as compared to other classifiers.

Specifically, for the first dataset, there is a really high accuracy of 0.988 achieved by SMO in 10-fold cross validation and 0.993 in Split 66% training. It puts SMO in the first place ahead of other classifiers, such as J48, which has slightly low accuracies of 0.987 and 0.991 respectively for both experiment mode. In the same respect, SMO also excels in Dataset 2, yielding an accuracy of 0. 943 that had been achieved in both 10-Fold CV and 66% Split, whereas J48 has lower averages of 0.936 and 0. 901, respectively.

It is pointed out specifically as to exactly what the AUC values are in the respective Figures (4.1 and 4.2). There are some observations that can be made: SMO once again yields the highest AUC values for Dataset 1 and competitive AUC values for Dataset 2. An AUC value of a classifier means its efficiency to classify the classes, which again reflect the robustness of the SMO classifier.

Therefore, by comparing the results of performance measurement, it has been identified that SMO is the most accurate classifier for the datasets. It outperforms other methods in terms of class discrimination, and it proves it has high performance in different datasets to classify your data sets comprehensively, making it the best option for your classification problems.

5 Conclusion

5.1 Mini App Visualization

With the help of Python, the following mini-app was made. The interface was made with the help of an SMO classifier, which proved to be the best classifier for our dataset. Figure 5.1 shows two different cases. In the first case, it predicted class 4, which is high, which means a player with this variable setup has the opportunity to have a high percentage goal

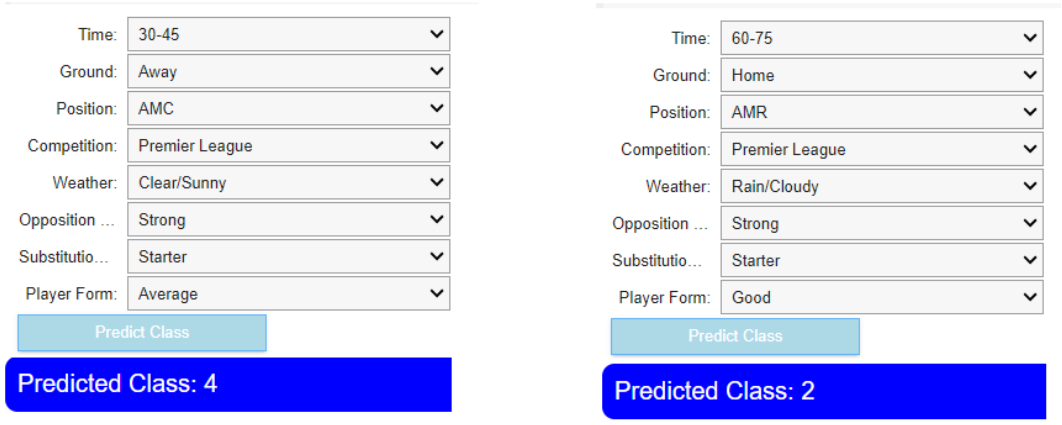

Figure 5.1: Visualization of prediction in two different cases

contribution. On the other hand, 2nd case predicted class 2, which is below-average. If we look carefully at Figure 5.1, it is noticeable that the variable setups are different. For example, time, ground, position, weather, and player form are not similar, though the opposition strength, substitute status, and competition are the same.

5.2 Future Work

Several additions and extensions are further suggested for the future research of the topic in question. First, data collection will be expanded beyond the first five seasons but rather up to more than 25 seasons. This will be an

increased dataset that will be used for the purpose of training, and this will make the training better and more accurate in its predictions.

Besides, other approaches that were not covered in the current research include data augmentation approaches. By applying several augmentation methods, it will be possible to achieve better results as more training samples are created and the model's ability to generalize will be strengthened.

Additionally, with the initial analysis carried out utilizing WEKA because of its simplicity, future experiments will therefore be done in Python. This shift will allow for other higher-level techniques of data analysis and visualizations using Python, which boasts numerous libraries for machine learning analysis and data manipulation.

Finally, if the new set of data, which is considered for being added to the training material, is large enough, the usage of a neural network will be discussed. It is possible that with the help of neural networks, the majority of patterns and relationships lying within the dataset can be identified and therefore provide an even more accurate estimation of the goal contributions of the midfielders. This advancement would be a step up towards progress in the study; better modeling would be used to further the assessment of football.

5.3 Final Thoughts

Thus, this study focused on proposing a model for forecasting the goal contribution of midfielders in English football clubs based on analyzing different classification models. Real data was collected for this research. Whoscored.com has been followed for data collection. Now a days, Club Football has become more professional, and it has achieved fantastic business status. The football clubs purchase a player based on their previous performance that is measured depending on some dimensions. This is one of the important reasons for taking the data set from English Club. This will help a lot for purchasing the right midfielder for the right team as well as helping the manager build his team.

After collecting the data, it has been partitioned into two datasets. Several classification methods have been tried for the data set. Datasets have been analyzed by WEkA with different data mining techniques, but the main challenge for this research was to choose the best classifier or technique. Different classifiers accuracy has been checked to find the best classifier, and the ROC graphs have been analyzed.

A thorough comparative study of various classifiers led to the following conclusion: the classifier that was most effective for this task was known as SMO, which is a support vector machine implementation using sequential minimal optimization. While there was a slightly close result from the J48 classifier, SMO was found to be better in the area of accuracy and predictability.

The paper underscores the reliability of SMO in categorizing and estimating goal contributions and therefore proves to be a worthy asset for football clubs targeting players based on previous performance. Although there is not much difference in accuracy between J48 and SMO, where SMO has slightly higher accuracy and slightly more consistency, it clearly reveals that SMO is more beneficial in this application.

Bibliography

[1] Al-Radaideh, Qasem A., and Eman Al Nagi. "Using data mining techniques to build a classification model for predicting employees performance." International Journal of Advanced Computer Science and Applications 3.2 (2012).

[2] Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." International conference on advances in ICT for emerging regions (ICTer2012). IEEE, 2012.

[3] Kaur, Harleen, and Siri Krishan Wasan. "Empirical study on applications of data mining techniques in healthcare." Journal of Computer science 2.2 (2006): 194-200.

[4] Hssina, Badr, et al. "A comparative study of decision tree ID3 and C4. 5." International Journal of Advanced Computer Science and Applications 4.2 (2014): 13-19.

[5] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[6] Football in England." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, https://en.wikipedia.org/wiki/Football_in_England.

[7] Whoscored. http://whoscored.com/

[8] Bramer, Max, and Max Bramer. "Introduction to data mining." Principles of Data Mining (2013): 1-8.

[9] Kumar, S. Mohan, and G. Balakrishnan. "Multi resolution analysis for mass classification in digital mammogram using stochastic neighbor embedding." 2013 International Conference on Communication and Signal Processing. IEEE, 2013.