A Survey Paper on Using Machine Learning Techniques for Soccer Result Prediction

Minh Vo
Computer Science
Earlham College
801 National Road West
Richmond, Indiana 47374
mbvo14@earlham.edu

***Abstract*— This paper covers the basics of the framework for a soccer (English Premier League) results prediction system/project using Machine Learning techniques. First, the paper introduces the idea of using technology in sports, especially in soccer. Then, it goes into details about each of the components of the framework. The end goal of the research within this paper is to produce an understanding of the procedure of taking in public data of the English Premier League results, processing them with Machine Learning techniques, and determining the results of the future matches. This can be predicting the match results as home wins, ties, or away wins. Another way to determine the outcome of future matches is to anticipate the exact scores of the match, which can include the number of goals scored by each team per half.**

I. Introduction

With the non-stop improvements in technology, more and more fields are trying to apply computer science to achieve their goals in a more efficient and less time consuming way. Sports are no outsiders to this group of fields. In sports, especially in soccer, technology has become an essential part. Soccer experts now make use of technology to evaluate a player's or a team's performances. Other than using their experience and their management abilities after many years being parts of the game, the soccer coaches also use statistics from data providers to improve their knowledge of their own players and teams so that they can come up with different strategies/tactics that bring them closer to the wins. Besides coaches, soccer analysts also make use of the data to predict results in the future as well as evaluate new talents emerging from the scene. This is where Machine Learning techniques can become useful. Machine Learning is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction [10]. Therefore, Machine Learning can be used as the learning strategy to provide a sport prediction framework for experts.

II. Framework and Methodology

The framework for predicting soccer match results in the English Premier League include the following components: knowledge about the sport and the English Premier League, data understanding and preparation, feature extraction, training and testing, and finally, performance evaluation. This framework slightly follows the one proposed by Bunker and Thabtah [10].

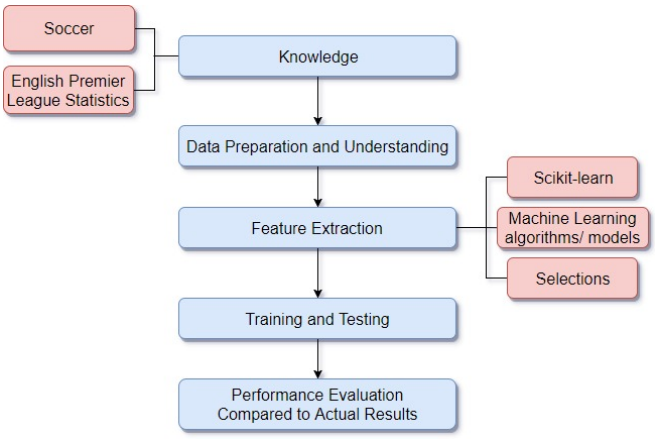

Figure 1: Framework of the English Premier League match results prediction. This includes Knowledge about soccer and the English Premier League, Data Preparation and Understanding, Feature Extraction, Training and Testing, Evaluation - Compared to Actual Results.

A. Knowledge about The Sport and the English Premier League

To do any type of learning in a particular field, which is soccer in this case, one needs to understand the basic ideas of the sport as well as the league from which the data are being retrieved and used. The main characteristics about soccer and the English Premier League, of which the results are the data to be used, are explained in this section. Here, we only introduce simple knowledge of the sport and the league that should be grasped before actually going into the more detailed statistics of the sport in the Data Understanding section.

1) What is Soccer?:

Soccer (or football) is a sport that involves two teams, each of which has eleven players on the pitch, including one goalkeeper. The basic idea of soccer is that to win a match, a team needs to score more than the other team. Every team has one manager and a few coaches (decided by the manager), who come up with formations, strategies, training drills, etc., to increase the chance of winning.

2) What is the English Premier League?:

The English Premier League is the English soccer league that is at the highest order in the English soccer league system. It has been around for more than twenty years. The Premier League is often regarded the most exciting soccer league on the planet, with fierce competitions from the La Liga, which is the top-level League in the Spanish soccer league system. The English Premier League has 20 teams competing for the Championship. Every team plays each of the other 19 teams in the league twice per season, one at home and one away. This means that each team has to play 38 games per season.

The teams do not just compete for the Championship, they also compete to stay in this highest ordered league of the system. Three teams that end up in the eighteenth, nineteenth and twentieth positions will be relegated to the league that is right below the Premier League in the ranking order. These three teams will be replaced by three teams from the second-level league. The top two teams of the lower league will get the automatic promotions while teams in the next four positions on the table will have to compete in play-offs for the final promoted spot in the top league. Other than contesting for the league title and to stay in the league, teams also compete to get the top positions that will let them play in the European competitions (UEFA Champions League and UEFA Europa League) in the following season. The most successful team in the English Premier League to this day is Manchester United with twenty times winning the Championship title.

B. Data Preparation and Understanding

Before going into learning Machine Learning techniques as well as deciding which methods to use, researchers have to get adequate understandings of how to obtain the data and what each set of data describes. There are a few things to be considered when you retrieve the data, which include the validity of the source, the correctness of the data, making sure that the data do not get lost during retrieving process, etc. After the right source has been found and the data have been retrieved, it is essential that you have good knowledge of what the data represent, or in other words, what they tell about the subject, in this case, soccer matches.

1) Data Preparation:

Detailed statistics for specific players are very important in determining teams' strengths and values. Certain "key" players can turn the matches on their heads and bring their teams the more benefiting results. However, the data providers do not want to make this type of statistics public, and often "sell" the data for high prices. This is one struggle/obstacle that many researchers have faced. While it may be more interesting and even more efficient to look into players' statistics to predict match outcomes, there have been many researches/papers that look into result prediction from data of past results.

There are a few sites that actually provide data publicly although the data provided are only on the statistics of matches. Buursma in [4], Constantinou et al. in [5], Timmaraju et al. in [7], Ulmer et al. in [8], and Tax and Joustra in [9] all did their researches/works using the data from one source, the Football-Data site, http://www.football-data.co.uk [11]. The database contains all full-time and half-time score lines and the bookmaker odds for most of the matches [4]. The site provides data for matches in the English Premier League since season 1993 - 1994. However, from season 1993 - 1994 to season 2000 - 2001, the data only contain half-time results and

full-time results, which are inadequate for the usage of Machine Learning. From season 2000 - 2001 until the current season, the site provides data with much more diversity than before. The data contains the number of shots, shots on target, fouls, corner kicks, cards, etc., which will be listed and explained in the Data Understanding section below. Thus, this data source is the perfect fit for the purpose of applying Machine Learning techniques to find patterns and features that may affect the future match outcomes. The data are provided in the form of CSV files. Each season's data are included in a separate file instead of one big CSV file. This makes it easier and more efficient when retrieving and preprocessing the data.

2) Data Understanding:

Using Machine Learning techniques on public league data to predict results in soccer requires one to have more than just the basic knowledge above. Following are some of the most useful terms in soccer statistics/data:

- Home Team/Away Team: The home team gets to play on home field and gets more attendance at the game than the other team. On the other hand, the away team has to travel to the stadium of the other team to play. This can be described as an advantage for the home team as they get more mental encouragement from the home supporters. The home advantage, or away disadvantage, is important in soccer as the home team can outperform the away team even in situations where the away team is ranked/rated higher than the home team. Thus, this plays an important role in predicting match outcomes [9].
- Total shots: This number is the record of the number of attempts to score of each team. The number of shots can tell which team attacks more, which may mean that that team is the better side in the match.
- Shots on target: Any goal attempt that goes into the net, would have gone into the net but for being stopped by a goalkeeper's save or would have gone into the net but for being stopped by a defender who is the last man [6].
- Corner kicks: A corner kick is a set-piece/free-kick taken from a corner of the field. It is given to a team in situations where the ball goes over the end line and the final touch is by the other team's player. The number of corner kicks may tell if a team is attacking a lot or not as they are taken on the opponent's half of the field.
- Fouls: Any infringement that is penalized as foul play by a referee [6]. This statistic can be interesting as it can reflect whether a team plays with a defensive mentality or not. The team that fouls more in a match (by a significant amount) is usually the team that plays defensively in order to to avoid losing to the opponent and thus, get at least one point.
- Yellow cards: A yellow card is usually given after a foul as a warning. This is decided by the referee.
- Red cards: A red card, like a yellow card, is usually given after a foul. A straight red card can be given to a player if the foul is very bad. Two yellow cards also result in a red card. A player who gets a red card will be removed from the match and his/her team will have to play with one fewer player for the rest of the game. This is a huge disadvantage in soccer.
- Results: Soccer results at league levels can be divided into three types. The first one is the home team wins. The second one is the away team wins. The final one is a tie, in which both teams score the same amount of goals, which can also be zero.
- Points: The winning team gets three points, the losing team gets two points. If the match ends up in a tie, each of the two teams gets one point. The current season performance for a soccer classification task can be modeled in one feature: points collected in previous matches [9].
- Rankings: The rankings are determined by the points they get in the season/previous matches. Most of the time, the teams that are rated higher/stronger than other teams will get higher rankings on the table as they win more games/lose fewer points. Teams that are widely known as the strongest in the league or have good financial state to buy good players should have good rankings on the table. For example, in the current season (2017-2018) of the English Premier League, the "big" teams such as Manchester City, Manchester United, Tottenham Hotspur, Chelsea, Liverpool, and Arsenal all hold the top 6 positions in the table, with their points ranging from 19 to 31. Thus, the positions/rankings of teams on the table are essential as they reflect the strength and preparation of teams in the current season.

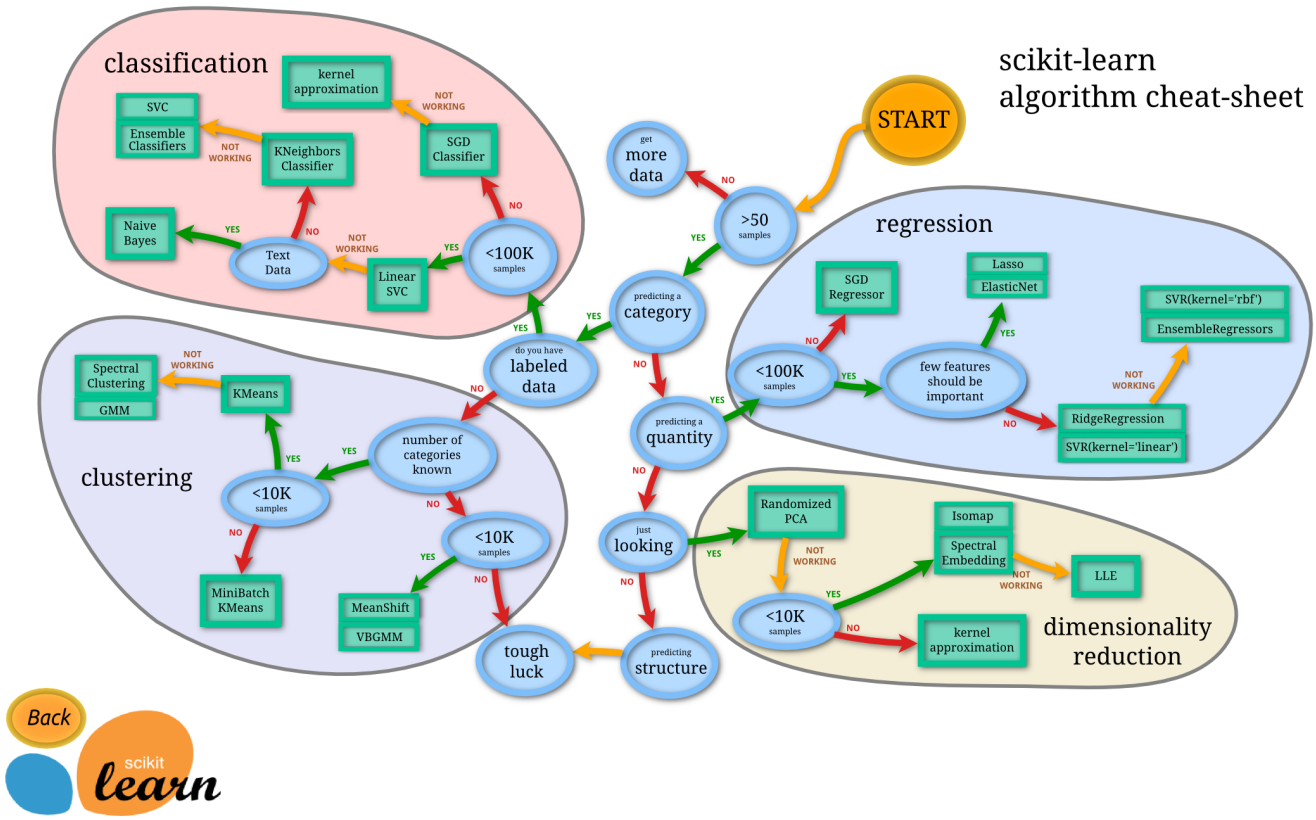

Figure 2: Often the hardest part of solving a machine learning problem can be finding the right estimator for the job. Different estimators are better suited for different types of data and different problems. The flowchart is designed to give users a bit of a rough guide on how to approach problems with regard to which estimators to try on your data [13].

C. Feature Extraction

1) Scikit-learn - The Tool:

Nowadays, there are a lot of tools that have been developed for the purpose of analyzing statistics and Machine Learning. A few outstanding softwares to be mentioned are The Shogun Machine Learning Toolbox and Waikato Environment for Knowledge Analysis (WEKA). The Shogun Machine Learning Toolbox does not provide as many common models and model-tuning, meta-methods as WEKA. On the other hand, WEKA provides users with several graphical user interfaces that make it easier for the users to use. However, this feature is also a disadvantage. This is because with the main usage being clicking and dragging instead of purely coding, it is hard to automate the whole process of data mining as the users would have to choose everything manually. A brilliant solution for this problem is Scikit-learn.

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems [14]. The most important characteristic of the Scikit-learn is that it provides the users with the algorithms for machine learning tasks including classification, regression, dimensionality reduction, and clustering. The Scikit-learn official site also help users get a clearer ideas of what to do/which algorithms to choose according to their dataset and purposes by equipping a detailed flowchart, which is called Scikit-learn algorithm cheat-sheet. This cheat-sheet can be seen in the figure above. It also provides modules for extracting features, processing data, and evaluating models [12].

Scikit-learn is popular for academic research because it has a well documented, easy-to-use, and

versatile API [12]. Furthermore, with its wide range of sets of tools and algorithms, Scikit-learn can be considered the ultimate tool for researchers who intend to use machine learning algorithms on their data in order to find specific patterns for future predictions. With Scikit-learn, creating scripts to run many lines of code that carry out the whole process of data mining and machine learning can become easier.

2) Machine Learning Algorithms/ Models:

Machine Learning is the concept of learning from the data provided/given. It gives the computers the capability to learn without having to be set up manually, which means that the computer will determine the next steps to take from the past events/data. This idea is the same as the idea of how human learn from experiences (or mistakes) to make decisions in the future. The fundamental goal of machine learning is to generalize, or to induce an unknown rule from examples of the rule's application [12].

There are two main categories of Machine Learning: supervised learning and unsupervised learning. There is one more category which is not as common as the two previous ones, it is reinforcement learning.

In supervised learning, a dataset is given as training data. The labeled training data is analyzed to create a function that can determine the next instances from examples of the right answers.

Unsupervised learning, on the other hand, serves the purpose of attaining patterns in the data provided. This means that in unsupervised learning, the computer attempts to discover relationships in the data that is unlabeled.

Reinforcement learning can be considered the middle-ground between supervised learning and unsupervised learning. Reinforcement learning, or semi-supervised learning, takes in and uses both supervised and unsupervised data. A reinforcement learning program receives feedback for its decisions, but the feedback may not be associated with a single decision [12].

In sports in general, and in soccer in particular, data is usually carefully labeled. This means that data in soccer has some kind of meaningful tag, or label, such as number of shots, fouls, points, etc. The choice of machine learning category can also be looked at in more details from the cheat sheet that is provided by the Scikit-learn site (figure 2). Therefore, the most common machine learning techniques in sports/soccer belong to the category of supervised learning. In the following part of the paper, some of the common machine learning algorithms that are the basics of many others used for soccer prediction will be explained.

- Decision Trees:
  The way a decision tree works can be easily inferred from its name. A decision tree implements a tree-like graph of decisions and the consequences that come after those decisions.
  The initial node of the tree (the highest node) can be called the root. The ending node(s) of the tree (the bottom one(s)) can be called the leaf, or leaves if there are more than one final node, which is also the most likely situation. The nodes that bridge the root node to the leaf node(s) can be called branch nodes, which are not possible results but the determining factors of the result.
  Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute [1]. Starting from the root, the attributes of the root will be examined. After that, each branch of the initial node will be another "root" for further attributes and branches and the previous process will be repeated. The decision tree algorithm will stop after all decisions have been made and the final result has been determined.

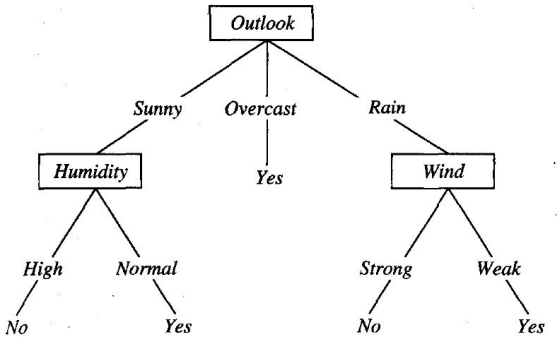

Figure 3: A decision tree for the concept PlayTennis. An example is classified by sorting it through the tree to the appropriate leaf node, then returning the classification associated with this leaf (in this case, Yes or No). This tree classifies Saturday mornings according to whether or not they are suitable for playing tennis [1].

The collection of a large number of decision trees

can make up a random forest. This is an ensemble learning method for classification that operates by constructing a set of decision tree while training and then outputs the class which is the mode of the classes output by individual trees [6]. The random forest method's purpose is to avoid or overcome the habit of overfitting to the training set of the decision trees. In [8], Ulmer et al. used random forest because they believed that many results did not reflect correctly the matches they were trying to predict due to the upsets in the data.

- Bayesian Networks:
  A Bayesian network is a graphical probabilistic belief network that represents the conditional dependencies among uncertain variables, which can be both objective and subjective [5]. Bayesian networks provide a means for representing, displaying, and making available in a usable form the knowledge of experts in a given field [3].
  Looking into Bayesian learning methods, it is important to consider every training instance as each of those can increase or decrease the correctness of a hypothesis. Moreover, combining the newly examined data with the past observation is the way to establish a final and concrete evaluation of a hypothesis's correctness. From the hypotheses that have been examined/observed and evaluated, Bayesian networks will attempt to make predictions in a percentage format. New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities [1].

• *Product rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B $P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$
• *Sum rule:* probability of a disjunction of two events A and B $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
• *Bayes theorem*: the posterior probability of h given D $P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$
• *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then $P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$

Figure 4: Table - Summary of basic probability formulas. [1].

The core and underlying factor of Bayesian networks is the Bayesian theorem, which supplies a calculation method of the posterior probability P(h|D). To make the calculation, three pieces of information need to be provided. The first one is P(h), the prior probability of hypothesis h. This probability provides information about any chance that h may be a correct hypothesis. The second one is P(D), the prior chance that training data D will be examined. Next, we will write P(D|H) to denote the probability of observing data D given some world in which hypothesis h holds [1]. With the three probabilities above, one final formula that tells us about the posterior probability is produced. It is the third formula in the figure/table 4.
In [5], Constantinou et al. presented a new Bayesian network model for forecasting the outcomes of soccer matches in the distribution form of {p(H), p(D), p(A)}; corresponding to home win, draw and away win. They made use of previous information, current information and subjective information (expert's) to produce a non-symmetric Bayesian parameter learning procedure.
One of the most common learner in the Bayesian learning methods is the naive Bayes learner. This is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions [6]. This means that the naive Bayes classifier makes the assumption that all instances or characteristics are dependent on each other.

- Artificial Neural Networks:
  Artificial neural networks method is considered one of the most popular machine learning algorithms that are used for sport outcomes prediction. It is a connection organization that was derived from the complicated biological neural network of many interconnected neurons, which develop animal brains.
  Neural network learning methods provide a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions [1]. An artificial neural network is consisted of many interconnected components that may have the same functionality as a function. Their main purpose is to take in a set of inputs/arguments and work their ways to a final outputs/results.
  As described above, inputs and outputs are separated. The inputs make up the first layer of the network while the outputs belong to the last layer. However, those two layers do not establish the

core of the algorithm. In order to attain the result, an artificial neural network makes use of a hidden layer. The power of Artificial Neural Networks comes from the non-linearity of the hidden neurons in adjusting weights that contribute to the final decision [10].

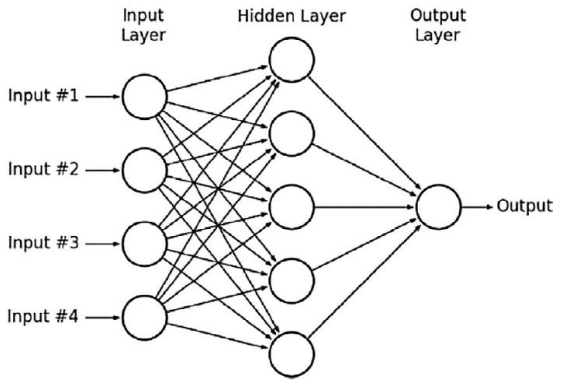

Figure 5: Example structure of an ANN with 4 input nodes in the input layer, 5 hidden nodes in the hidden layer and out output node in the output layer [2].

There can be many different layers that are organizations of neurons which stand between the input layer and the output layer. After receiving and processing the information from the training dataset, the artificial neural network model will be constructed. The information from the input carries important components (such as weights) which make up the basic of the artificial neural network classification construction. In other words, weights associated with interconnected components are continuously changing to accomplish high levels of predictive accuracy [10].
The inputs given will go through many layers and each layer can have a different transformation or alteration that affects the inputs. The adjustments done by the layers of the artificial neural networks algorithm is to satisfy the goal of achieving the accuracy that has been set by the user. This will be done again and again before the final results are produced. This also means that inputs may have to traverse these hidden layers numerous times, not just once. This may lead in some cases to the problem of overfitting, as well as wasting computing resources such as training time and memory [10].
The application of artificial neural networks algorithm is suggested by Kumar in [10] as many researchers have used this method for their problems regarding sports prediction and especially, soccer prediction.

- Linear Regression:
  Linear regression is the go-to algorithm when the problem that needs solving contain numeric values. When the outcome, or class, is numeric, and all the attributes are numeric, linear regression is a natural technique to consider [2]. Linear regression can be used to model a linear relationship between one scalar variable and one or more independent variables. Linear regression is a widely known concept in the statistics field as it performs the task of fitting a straight line through a set of points.
  The method that handles situations in which there is only one explanatory (independent) variable is called simple linear regression. Using training data to learn the values of the parameters for simple linear regression that produce the best fitting model is called ordinary least squares or linear least squares [12]. Ordinary least squares regression is one of the most popular methods that perform linear regression. It is done by first drawing a line, and then measuring the vertical distances between the line and the data points. After that, the total of the measured distances will be calculated. The final step is to display the result. The place at which the smallest possible sum of the gaps is found is represented by a fitted line.
  The use of linear regression can also be seen in classification problems, which utilize linear regression to predict the right class. The trick is to perform a regression for each class, setting the output equal to one for training instances that belong to the class and zero for those that do not [2]. This method is called classification via regression, which was used by Buursma in [4] and returned the best results, better than those of Bayesian networks and naive Bayes method.

- Logistic Regression:
  Logistic regression is a method that is used for classification tasks. The goal in classification tasks is to find a function that maps an observation to its associated class or label [12]. The model is useful for problems in which the dependent variable is

categorical.
Logistic regression is usually mistaken with linear regression. However, there are some characteristics of logistic regression that are distinct from linear regression's. In linear regression, the outcome is continuous, which means that the result can have any one of an infinite number of possible values. On the other hand, in logistic regression, the dependent variable does not have that infinite range. It only has a limited number of possible values.
Another difference between these two types of regression can be seen in the requirements of the outputs. In binary-variable problems, linear regression will attempt to predict the outputs as 0 or 1, which is the main form of results in yes or no, pass or fail questions. Instead of approximating the 0 and 1 values directly, thereby risking illegitimate probability values when the target is overshot, logistic regression builds a linear model based on a transformed target variable [2]. Moreover, in problems that require the dependent variables to belong to more than two possible outcomes, logistic regression is also used. This method is called multinomial logistic regression.
The outstanding difference between logistic regression and linear regression, namely multinomial logistic regression, provides another dimension for researchers in prediction problems. Timmaraju et al. used multinomial logistic regression in [7] as there were more than two possible outcomes. Other researchers such as Buursma in [4] and Tax and Joustra in [9] also used LogitBoost, a boosting algorithm that is based on logistic regression.

- Support Vector Machine:
  Support vector machine is a powerful model for classification and regression. From a given training dataset that contains examples that are components of two different categories, support vector machines select one of the two categories to assign new instances. Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible [2]. Then, the new instances will be mapped and to predict/decide which category they fit in, the only thing that needs to be considered will be the side of the gap to which they have been mapped. These systems transcend the limitations of linear boundaries by making it practical to include extra nonlinear terms in the function, making it possible to form quadratic, cubic, and higher-order decision boundaries [2].
  Although not many researchers have utilized the strength of support vector machine in sports prediction, there has been evidence that support vector machine performs better than some other commonly used machine learning algorithms. Ulmer et. al. in [8] reported that support vector machine was one of the best performing models in their three-class classification problem, with lowest error rates than the random forest as well as naive Bayes models.

3) Selections:

After gaining basic understandings of the data as well as the machine learning algorithms/models, we have to carefully consider the selections in the dataset that will best serve the purposes. An accurate feature set makes it a lot easier to predict the outcomes of matches [4].

One way to select the features or statistics that can have positive influence on the accuracy of the predictions in this problem is to follow the basic understandings of the data in soccer. For example, as explained the Data Understanding section, home advantage (or away disadvantage), the performances in recent matches (can be considered by the number of wins, ties, defeats or points), and the goals scored or goals conceded are deciding factors in the results of soccer matches. Buursma in [4] came up with a very basic set of features as following:

- (1) Goals scored by home team in its latest x matches
- (2) Goals scored by away team in its latest x matches
- (3) Goals conceded by home team in its latest x matches
- (4) Goals conceded by away team in its latest x matches
- (5) Average number of points gained by home team in its latest x matches
- (6) Average number of points gained by away team in its latest x matches
- (7) Number of home matches won by home team in its latest x matches
- (8) Number of away matches won by away team

in its latest x matches
- (9) Number of goals scored in home matches by home team in its latest x matches
- (10) Number of goals scored in away matches by away team in its latest x matches

However, the selection process must not stop here. The set of features that Buursma came up with only contains basic statistics of the game. Therefore, this set needs to be checked every time it is observed and based on the results, the list of features to be selected must be changed. This process needs to be done again and again so that the selections can be expanded.

Another way to approach the process of selections is to create feature subsets. Features in sport result data can be divided into several different subsets [10]. This can be done by separating match statistics and performance-related data. Match statistics include features such as the number of shots, goals, corners, cards, etc. Performance-related features may contain the points gained by each team in the last x matches, their standings on the table, etc.

D. Training and Testing

After having attained a sufficient and efficient set of selections, we go into training and testing. The process during which we let the machine learning algorithm calculate the probabilities for the matches is called training [4]. It is important to preserve the order of the training data for the sport prediction problem, so that upcoming matches are predicted based on past matches only [10]. Normally, a large percentage of the data will be used for training while the remaining data will be used for testing.

Figure 6: Diagrammatic Representation of Ten-fold Cross Validation [10].

A common way to separate training and testing data is to use the ten-fold cross validation technique. It basically divides the data given into ten different sections and these sets of data will have the same size. Nine of the ten sets will be used for training and the one remaining section will be used for testing. This method also repeats the process of division ten times. A different set out of the ten sections of the data will be chosen as the testing data every time the process is repeated. The outcomes of the matches will be predicted/determined during the test phase. After that, they will be compared to the actual results to see the accuracy of the algorithms.

E. Performance Evaluation - Compared to Actual Results

After all the results have been calculated and predicted, we would need to evaluate the performances of the algorithms, which make up the core of the framework. Here, we can use actual results for the evaluation purpose. This is done by determining the outcomes of the matches based on the predictions of the algorithms and comparing those to the results in real life. This process can be considered the same as the training and testing process. We can just use the data of n-1 years for analysis and predict the results of the n-th year. The algorithms can be evaluated as successful if their accuracies are more than 70%, and as totally unsuccessful if they cannot get more than 50% of the results right.

III. Conclusions and Future Work

In conclusion, this is a detailed framework for using machine learning algorithms to predict the match results in the English Premier League. In soccer prediction so far, many features have been ignored. In many researches, the main sets of data are mainly those listed in the Selections section above, which regard the goals, the points, and the standings/rankings. A direction towards which should be headed in the future works is to include many features into the process, such as number of shots, shots on target, corner kicks, and cards (especially red cards). These are all instances that can may improve the accuracy of the predictions. Another improvement that can be made in the future is to predict the exact scores of the matches instead of categorizing the results as only wins, ties, and losses. This can be expanded and altered to be the framework for many other sports.

REFERENCES

[1] Mitchell, Tom M. *Machine Learning*. First edition, McGraw-Hill Education, 1997.

[2] Witten, Ian H., and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, 2005.

[3] Joseph, Anito, et al. "Predicting Football Results Using Bayesian Nets and Other Machine Learning Techniques." Knowledge-Based Systems, vol. 19, no. 7, 2006, pp. 544-553.

[4] Buursma, D. "Predicting Sports Events from Past Results Towards Effective Betting on Football Matches." Conference Paper, Presented at 14th Twente Student Conference on IT, Twente, Holland, vol. 21, 2011.

[5] Constantinou, Anthony C., et al. "Pi-Football: A Bayesian Network Model for Forecasting Association Football Match Outcomes." Knowledge-Based Systems, vol. 36, 2012, pp. 322-339.

[6] Kumar, Gunjan. *Machine Learning for Soccer Analytics.* Cambridge University Press, MSc thesis, KU Leuven, 2013.

[7] Timmaraju, Aditya Srinivas, et al. *Game ON! Predicting English Premier League Match Outcomes.* 2013.

[8] Ulmer, Ben, et al. *Predicting Soccer Match Results in the English Premier League.* Ph. D. dissertation, 2013.

[9] Tax, Niek, and Yme Joustra. "Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach." Transactions on Knowledge and Data Engineering, vol. 10, no. 10, 2015, pp. 1-13.

[10] Bunker, Rory P., and Fadi Thabtah. "A Machine Learning Framework for Sport Result Prediction." Applied Computing and Informatics, 2017.

[11] Football Betting — Football Results — Free Bets — Betting Odds. http://www.football-data.co.uk/. Accessed 9 Nov. 2017.

[12] Hackeling, Gavin. *Mastering Machine Learning With Scikit-Learn*. Packt Publishing, 2014.

[13] *Scikit-learn: Machine Learning in Python*. http://scikit-learn.org/stable/. Accessed 16 Nov. 2017.

[14] Pedregosa, Fabian, et al. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, Oct. 2011, p. 28252830.