Analysis on Sports Data Match Result Prediction Using Machine Learning Libraries

S Kevin Andrews1*, K Lakshmi Narayanan2, K Balasubadra3, M S Josephine4
1Department of Computer Applications, Dr.MGR Educational and Research Institute, Chennai, Tamil Nadu, India
2Department of Electronics and Communication Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
3Department of Information Technology, R.M.D Engineering College, Kavaraipettai, Tamil Nadu, India
4Department of Computer Applications, Dr.MGR Educational and Research Institute, Chennai, Tamil Nadu, India
Email: 1stevekevins89@gmail.com, 2kyelyen@gmail.com,
3kbalasubadra@gmail.com,
4josejbr.07@gmail.com

Abstract. Sports research and betting have been powered by quick exposure to the Internet and machine learning popularity. Football is considered the most popular game in 200 countries and contrasts to other sports. It is considered much more diverse and complicated, making soccer an enticing area to do research. A variety of methodologies and methods are used for the production of prediction systems. We expect the outcome of a match between the Premier League and a home side. The projections are based on numerous significant evidences from the Premier League’s previous seasons. These essential characteristics would possibly decide the result of a match. We use three different algorithms to predict the machine learning techniques and then choose from those three the best algorithm for predicting the label.

Keywords: Football, Premier league, Machine learning, prediction and countries

1. Introduction
The transcendent world of technology from Internet technology has introduced and generated the human brain’s intriguing modern philosophy to the very height. Also, there has been an intriguing trend concerning the global strategy, as the key areas in consumer goods are a prime example. In terms of the revolution of time, some of the most unmistakable transcendence recognized in the history of the Internet often ranked in the status where the work was carried out faithfully and acted as the basis for progress. Sports show its ability. The Internet can transition from football to Football to baseball to linking points regardless of sport. This infiltration among sports contributed to its very presence as the "Internet of sport." The Internet will transition to many different fields of play.

One of the main transformations of the "Sports Internet" has placed the numerous professional athletes’ financial performance in the codex. The business has progressed along the way towards grandiose growth. Activities and issues of concern such as the comparative calculation of athletes were scarcely carried out in the old age in an orderly and timely fashion. The regular review of these
dilemmas was easy to enforce and easy to observe with Internet power aid. We can eventually realize that the Internet in sports was very helpful for fantasy sports, considering the simplicity with which people might get into a living, taking a few moments in their life and all that seems unlikely or very difficult to conquer.

2. Literature Survey

Machine Learning (ML) is a smart technique [1] that has demonstrated promising results in classifying and forecasting areas. Sport estimation, because of the vast monetary sums involved in betting, is one of the expanding fields of successful prevision. Furthermore, club administrators and owners search for classification structures to grasp and devise plans for matches. These models are based on several variables that participate in sports, such as past match results, team success metrics and opponent information. This research contrasts the various classification techniques done using inertial and magnetic sensor units on humans' chest, arms, and legs. Every unit has a three-axial gyroscope, accelerometer and magnetometer. The feature collection [2] derived from Naive Bayesians, artificial neural networks (ANN) and dissimilarity classification, is considered three forms of decision-making: Gaussian mixture models and SVMs.

Classification is used from raw sensor data that utilize the key feature analysis. We launch a predicting method designed to take advantage of a machine learning sports [3] betting market. Three primary components are contributed. Second, previous efforts to learn model outcomes maximized the model's statistical accuracy as the only criterion. In comparison, we decrease the model's association with the bookmaker's forecasts available through reported odds. Therefore, our method [4] is a means of 'useful' the bookmaker, and we have demonstrated that an integrated model enables improved profit generation. This article provides an important scene recognition and structure analysis system for sports footage, with tennis and baseball examples. The predetermined temporal grammar repeated events with clear characteristics, and a fixed number of views will characterize Sports Footage. Our methodology incorporates domain-specific awareness, supervised machine-learning and automated multi-level function analysis [8,9].

The use of a compact domain achieves Real-time treatment efficiency. Tactics as a core component are the success of Elite soccer today. However, there have been just a few extensive scientific studies on squad tactics until recently. One explanation was the lack of accessible, usable knowledge. This condition [5] has improved lately with the advent of sophisticated monitoring technology. Instead, it is becoming impossible to handle the volume of data available. This article
discusses how these issues are solved by new advances in big data technology from industrial data processing domains. The Indian Premier League (IPL) is one of the world's most successful cricket competitions, with an increased financing level, spectatorship and the annual betting market for IPL. Since cricket is a very dynamic game, bettors and bookies are encouraged to bet on match outcomes, and it is a game where ball by ball varies[10,11]. This paper [6] explores cricket match forecasts dependent on historical machines' learning technologies.

Coincidence IPL info. Filter-based approaches such as correlation-based function collection, knowledge gains (IG), relief, and wrapper have established the data's characteristics. First of all, In order[12,13] to develop predictive models of distinctive features derived from the filtering methods, machine learning techniques like Naive Bayes, Random Forest, K-Nearest Neighbour (KNN) and Model Trees were introduced. The idea of data mining [7] is the processing of new data from large datasets. Knowledgeable Discover Database (KDD) is also known. In recent years, KDD has grown rapidly in the industry. It is most helpful [14,15] in all sorts of industrial ads since it is sorting. Before the advent of data mining; however, business marketing is sluggish, and business marketing relies more on TV advertising, advertisers, executives etc.

3. Methods And Materials

We will analyze English Premier League clubs, work on their data from the past decade and create sports betting applications to win and losing teams. We would use a step-by-step approach to evaluate
the past outcomes of both teams. We will forecast the results between 0% and 100% (normalized to 0-1) for the teams’ outcomes in the last 10-15 years, and we are therefore going to calculate all the results of the upcoming matches in the future. These findings will refer to all future and continuous installations generated by the Premier League Board in England.

4. Existing System
Our research investigates the emanating effect of Internet sports betting worldwide and how various consumers examine and affect the variables (age limit, inappropriate bookmakers, and independent transactions among people). Poor customer service bookmakers influence the entire system structure, and money engagement is originally a bad solution. Both issues have been taken together. Three usually threatens society in the betting process and are against those rules and other laws. With a secure approach to transaction processes, we can incorporate a stable, secure and maintainable online betting method, with various extensions, without losing sight of performance-precision parameters by any brute force review of both positive and negative aspects.

5. Implementation
Our article’s basis is the stats that win or lose percentage, lose or win the team and achieve outcomes. The users will review these findings to put their bets on the respective teams. Data such as points criteria, victory and loss of the past team, cumulative yellow and red cards received by a team, average referee actions for a team and referee assigned for individual matches, etc., are also aspects of data sets obtained the last 15 seasons. This broad data collection also contributes to improving forecasts by generating more stable and exact outcomes, so that more benefit than loss can be produced. Moreover, because of the need for the instant dissemination of data and knowledge that the main part of the technology satisfies, data processing can be transcendent by utilizing machine learning tools and precise data sets.

The focus of this project is on the outcome of the soccer matches. A formal approach is also important because it has more analytical bases than experimental bases. The basic system used is the CRISP-DM (Figure 1) platform for data mining across the industry. CRISP-DM is powerful and hence gives a stronger means of forecasting the results.

![Figure 1: Architecture of CRISP-DM framework](image-url)
6. Proposed Methodology

The CRISP-DM structure contains six main stages. The system CRISP-DM is modular, so phases cannot be used in sequence. The steps, in a sense, are mutually related. Data collections in data processing have continuous observation. The diagram’s outer circle reflects them, which is primarily used to develop the machine configuration and maximize the performance. As already illustrated, the above-mentioned structure is flexible, and the actions must not be taken in a strict order but must be taken in a particular manner. The phases of the framework appear in the following diagram Figure 2.

![Figure 2: Flow chart for the proposed methodology](image)

7. Dataset Exploration

In the data extraction process, the function subsets are generated. The subsets can have characteristics such as goals or team ranks. The attributes are split into subsets, such as the odds ratio or other scientists' professional views. However, in this project, only internal features and not professional perspectives are being carried out, so this project focuses not on feeling research. Match-related attributes concentrate only on the arithmetic part, i.e., the discrepancy between the aims, the objectives scored, and the research part's external focus. These two types of features function independently, but when combined, the full result is obtained. Figure 3 represents the structure of dataset exploration process.

![Figure 3: Structure of dataset exploration process](image)
8. Algorithms Module

8.1. Logistic regression:
The regression of logistics is focused on logistical activities. It is a blueprint for statistics. This logistical role model is a binary variable. Even if we look at regression analysis, logistic regression can be called binomial regression, mainly used to approximate the logistics model parameters. In terms of mathematics, the dependent model variable will obtain two values 1 or 0, and the predictor variable represents these values. The variables with value 1; if the logarithm of odds is combined, too.

8.2. SVM:
Support Vector Machines are supervised learning models (SVMs or the Support Vector Network). Vector machines support similar learning algorithms that can be used for two things, i.e. classification and regression analysis. The basic functionality of SVM is that as it is applied to current groups, it categorizes input data and all other additional data. The map's data doors are traced and separated using hyperplane to make the distance between the two groups clear and sufficient. According to its attributes, new data is then mapped in its class without impacting the other class in any way.

8.3. XG boost:
It is an algorithm for applied machine learning that falls under the group-learning range. It employs a certain procedure, which involves constructing several variants of the datasets and combining different methods for forecasting results.

9. Results and Discussion
Football itself becomes unpredictable, and thus, certain characteristics must be shown to forecast the effect of an image accurately. There are a couple of reasons for a football game for ex, but there are still greater opportunities for Home Side to prevail, whether the remote team leads the table of the roster or is in very strong shape - 5 or 10 matches have not been beaten. All these considerations must also be taken into account to prepare the models for better predictions. Vast volumes of data must be generated with learning models, and the data from the past 12 Premier League season have been gathered to enhance the projections. Figure 4 displays the 2005-06 Premier League season data, while Figure 5 reveals the 2017-18 Premier League season data.
Figure 4: Data of 2005-2006 Premier League Season

Figure 5: Data of 2017-2018 Premier League Season
The training and test set confusion matrices for the three models used are as follows Figure 6, 7 and 8:

1. **Logistic regression:**

   ![Logistic Regression Confusion Matrix](image1)

   a. **Figure 6:** Training and test evaluation

2. **Support Vector Machine (SVM):**

   ![Support Vector Machine Confusion Matrix](image2)

   a. **Figure 7:** Training and test evaluation

3. **XGBoost:**

   ![XGBoost Confusion Matrix](image3)

   a. **Figure 8:** Training and test evaluation

It will be basic to use the name of two teams as an input, i.e. one team will be the home team, and the other team will serve as a remote team, and the projected winner with the respective probability will be the result that has been shown in Figure 9.
10. Conclusion
In design, this project is extremely scalable. Predictions may be made with the models that have been conditioned until now. Logistic regression that turned out to be the best prediction model can also give this prediction a probability. The model which provided the best results and the predictions at the lowest time was logistic regression under the given set of attributes. Logistic regression is the perfect algorithm for classification purposes. Therefore, logistic regression was chosen to forecast the outcome of future matches. Logistic regression with 4561 already played equipment was learned with a testing dataset. The preparation took just 0.0156 seconds. In just 0.0000 seconds, the predictions were made, and the f1 score associated with this model was 0.6119 and the precision 0.6563. The prediction was achieved only in 0.0000 seconds while the models were checked on the test data collection, and the f1 score was 0.5652, with the precision being 0.6000. The measured chance will then be transformed into odds and then used in the making of bets.

Reference
[1]. Bunker, R. P., & Thabet, F. (2019). A machine learning framework for sports result prediction. *Applied computing and informatics*, 15(1), 27-33.
[2]. Barshan, B., & Yüksek, M. C. (2014). Recognizing daily and sports activities in two open-source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11), 1649-1667.
[3]. Huňák, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2), 783-796.
[4]. Zhong, D., & Chang, S. F. (2001, August). Structure analysis of sports video using domain models. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.* (pp. 182-182). IEEE Computer Society.
[5]. Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5(1), 1-13.
[6]. Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2020). Sports analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics*.

[7]. Pattnaik, M. (2019). Infrastructure of Data Mining Technique with Big Data Analytics. *International Journal of MC Square Scientific Research*, 11(1), pp.23-30.

[8]. Amarasena, P. T., Kumara, B. T. G. S., & Jointion, S. (2019, March). Data Mining Approach for Identifying Suitable Sport for Beginners. In *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (pp. 57-62). IEEE.

[9]. Kaur, G., & Jagdev, G. (2020, February). Analyzing and Exploring the Impact of Big Data Analytics in Sports Science. In *2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)* (pp. 218-224). IEEE.

[10]. Siswanto, B., Ricky, M. Y., & Kerta, J. M. (2019, August). Item Rating and Football Score Priori Prediction Algorithm. In *2019 International Conference on Information Management and Technology (ICIMTech)* (Vol. 1, pp. 486-491). IEEE.

[11]. Tiwari, E., Sardar, P., & Jain, S. (2020, June). Football Match Result Prediction Using Neural Networks and Deep Learning. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 229-231). IEEE.

[12]. Pantzalis, V. C., & Tjortjis, C. (2020, July). Sports Analytics for Football League Table and Player Performance Prediction. In *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-8). IEEE.

[13]. Klivansky, D., & El-Hajj, M. (2018, November). Causes of Success in the La Liga and How to Predict them. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 1-8). IEEE.

[14]. Hervert-Escobar, L., Matis, T. I., & Hernandez-Gress, N. (2018, October). Prediction learning model for soccer matches outcomes. In *2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI)* (pp. 63-69). IEEE.

[15]. Barot, H., Kothari, A., Bide, P., Ahir, B., & Kankaria, R. (2020, June). Analysis and Prediction for the Indian Premier League. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1-7). IEEE.