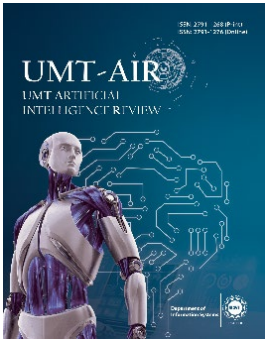


UMT Artificial Intelligence Review (UMT-AIR)
Volume 1 Issue 1, Spring 2021
ISSN(P): 2791-1276 ISSN(E): 2791-1268
Journal DOI: https://doi.org/10.32350/UMT-AIR
Issue DOI: https://doi.org/10.32350/UMT-AIR/0101
Homepage: https://journals.umt.edu.pk/index.php/UMT-AIR

Journal QR Code:

Article: **Predictive Analysis of PSL Match Winner Using Machine Learning Techniques**

Author(s): Muhammad Awais[1], Khushal Das[2]

Affiliation: [1]Software Engineer, Audience, Pakistan
[2]Game Developer, Finz Games, Pakistan

Article QR:

Muhammad Awais

Citation: A. Muhammad, and D. Khushal, "Predictive analysis of PSL match winner using machine learning techniques," *UMT Artificial Intelligence Review,* vol. 1, pp. 74–85, 2021. https://doi.org/10.32350/UMT-AIR/0101/05

A publication of the
Dr Hasan Murad School of Management
University of Management and Technology, Lahore, Pakistan

PREDICTIVE ANALYSIS OF PSL MATCH WINNERS USING MACHINE LEARNING TECHNIQUES

Khushal Das[1]*, Muhammad Awais[2]

ABSTRACT: Cricket is one of the most celebrated open-air sports generating a huge amount of measurable information. As Pakistan Super League (PSL) games grow in popularity, the potential predictors which influence the outcomes of the matches need to be investigated. PSL data spanning the years 2016-2019 and containing the specifics of the players, match venue details, squads, and ball-by-ball details was collected. Subsequently, it was analyzed to draw various conclusions that offer assistance to enhance a player's performance. Due to the expanding number of matches day by day, it is difficult to oversee or extricate valuable data from the accessible information of all the matches. This paper presents the preprocessing of the data, as well as its visualization and prediction. It centers on measuring the results of PSL matches by applying the existing information mining algorithms. It includes factors such as team 1, team 2, toss winner, and toss decision to predict the match winner with the help of the Random Forest algorithm.

KEYWORDS: K-Nearest Neighbours (KNN), machine learning techniques, Naïve Bayes, Random Forest, PSL prediction

I. INTRODUCTION

Recent advancements in technology have allowed the simultaneous collection and processing of large amounts of information, collectively known as big data. Currently, data science is being used in almost every walk of life to achieve better and consequential results. The goal is to use statistical and machine learning algorithms on the available data for creating computer programs that are able to retrieve and use data for understanding and self-learning. This process begins by data observation, for instance, events, direct knowledge or training, for the reorganization of certain statistical trends and to make informed decisions based on the samples collected in the future. The

[1] Department of Information Systems, Dr Hassan Murad School of Management, University of Management and Technology, Lahore, Pakistan
*Corresponding Author: f2020313013@umt.edu.pk
[2]Department of Information Systems, Dr Hassan Murad School of Management, University of Management and Technology, Lahore, Pakistan
*Corresponding Author: f2020313003@umt.edu.pk

main goal of machine learning is to eradicate the need for human interference or help by enabling computers to gain knowledge automatically and to change their task as per requirement. In recent years, developments in computing have made this all progressively easier to obtain as well as to process large amounts of data containing detailed information.

As a result, the availability of both live and historical data in the field of sports analytics has made machine learning very applicable and useful [1–5]. Sports analytics is based on the gathering and analysis of historical game information in order to extract essential knowledge from it, needed to facilitate good decision-making.

Machine learning can be used successfully on multiple occasions in sports, both off-the-field and on-the-field. The success of a team and its result against another team can be forecasted effectively through the projected model. The proposed model focuses primarily on the healthy development of players and on increasing their competitiveness for the facilitation of team owners and other financiers or investors. The analysis was conducted using various classification methods designed for machine learning, for example, Gaussian Naive Bayes, K-Nearest Neighbours (KNN) and Random Forest.

Portugal's football club Lisboa e Benfica [6-9] is one of their most successful football clubs. It incorporates machine learning for decision-making based on available information using data processing and prediction techniques, showing the significance of machine learning in sports research. This sports club not merely records but also estimates almost everything about players on- and off-the-field, including their habits of relaxing, drinking, and training, practically each part of the game. Diverse models are designed based on raw data to optimize game planning and to create personalized training timetables. By integrating machine learning and predictive analytics, data analyzed by the developed models enables players to continuously enhance their performance. Depending on the interpretation of the information gathered, team manager is assisted in decisions such as player replacement, keeping a player in the squad, parting a player on the bench and the time at which a player should be introduced onto the field.

In the current project, the data set used emerged from the various matches played, with about 146 match specifics providing complete information about the winner of the match, the winner of the position toss, names of the teams and other significant characteristics. The

dataset used in our project contained information about the matches of PSL held from the year 2016 through 2019. Using this data set, we achieved the above mentioned primary objective of our project.

II. Literature Review

One of the prime sports leagues, that is, Major Baseball League, has progressed tremendously in the area of sports technology in recent years by incorporating ball-by-ball statistics and basic awareness into the game. These are usually not observable otherwise and this is where machines come into play. In order to obtain and use the relevant data, experienced teams of technical experts implement different machine learning skills to better train their players with improved results.

Some of the categorizing or sorting problems resolved with the aid of machine learning in the game of baseball [9] include forecasting the games' outcomes, categorizing if a team will purposely permit a player to walk in to bat, and classifying non-fastball pitches by field type. Moreover, sports analytics are also utilized in cricket to predict the result of a match while it is still in progress or even when the match is yet to begin [10-14]. Concerns such as predicting the number of wickets taken or the runs scored during the match are also interesting problems which must be focused. WASP (Winning and Score Predictor) [15] is a realistic method used in cricket. It forecasts the score and the possible outcome of a limited over match of cricket, including a T20 or a one-day match. Sky Sports New Zealand, for the first time, developed such a tool in 2012.

Hawk-Eye, which is a software, can track the ball's trajectory and visibly shows the best statistically important track. It has also been formally used in the Umpire Decision Review Method since the year 2009. Likewise, this computer-aided intelligent technology is also used in other sports such as badminton, tennis and snooker. In this paper, we explore briefly the relevance of machine learning for sports and the methods suggested for its enhancement.

III. Machine Learning

Modern autoencoders are networks that attempt to encode and With the application of machine learning and Artif icial Intelligence (AI), we can address the real-time problems of the society. In sports, they have a thoughtful influence. During the game, teams utilize data to enhance on-the-field performance of their players. Such predictive analysis helps them to make directed decisions and structural adjustments which affect various aspects of their sports association, from recruiting players to

match results, from indulging more fans to high sales of match ticket.

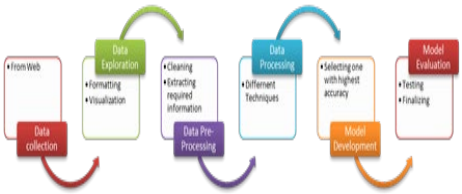


Fig.1. Research Work Flow

Sports have been an important factor affecting our culture over the past few years. Participation in the Olympics contributes to community wellbeing and productivity. Player performance and participation in the match can be enhanced using powerful machine learning techniques which help to enhance their previous performance, while necessary changes can be made. Without the use of machine learning, a player's strengths and weaknesses cannot be addressed in detail and areas for improvement cannot be identified.

A fundamental approach to predict a cricket match winner was suggested in the paper "Predicting the match outcome in one day international cricket matches" [17]. It also addressed how machine learning algorithms would change the future of how players' judge their gameplay. This technique has many benefits over the current technique, where manual work is reduced to find faults in gameplay, making it easier for the management to effectively determine the enhancement policy. Arthur Samuel introduced the smart method to gaming using artificial intelligence techniques in the game of checkers during the 1960s and vouched for the inclusion of a number of moves for any player to win against his rival.

The authors demonstrated that Bayesian networks perform better than other machine learning techniques in the research paper "Using mathematics and statistics to comprehend data from baseball, football, basketball, and other sports." This paper offered a remedy for the manual attempts aimed to improve the previously obtained findings. Such prediction of the Bayesian network delivered a 59.21 percent accurate outcome as compared to the other approaches.

The way gaming can be tracked has always been an influential approach in machine learning. In the baseball game paper [8], it was shown that gameplay can be evaluated using machine learning techniques, such as SVM and KNN, supported with the current procedures. These techniques involve the retrieval of 75 percent of a match's highlights and the subsequent application of a deep assessment method.

Improvements in the assessment of gameplay help the players to find fresh tactics to secure victory over their opponents. They also suggest how the outcomes of the game can be

represented pictorially and provide the manner in which a match's analysis is given out. During game prediction, we used supervised algorithms such as Random Forest, Naïve Bayes and KNN classifier, which are comprehensively explained later in this paper.

IV. Proposed Work

Literature review reveals that it is possible to build a machine learning model that can forecast the results of a match even before its commencement. In cricket, there are different formats, including the T20 format, which have a number of turnarounds. This scenario makes it challenging to predict the winner till the last ball of the match. Therefore, predicting the winner is very complicated. Regression and classification tasks are used to do a great deal of mathematical work in sports, all of which are subject to supervised learning.

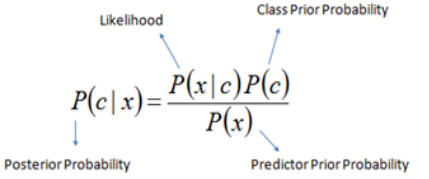
Supervised machine learning can be split into two groups, that is, classification and regression, based on performance. This problem is of classification. Therefore, on the PSL dataset, we applied multiple classification algorithms, analyzed the outcomes and picked the best suitable model with the highest accuracy.

For the current research work, the above figure reflects the basic methodology. It reflects numerous stages including the collection of data, exploration of data, data cleaning, data processing, and creation of model evaluation. These stages, also explained later in this paper, are fundamental in our work.

There are a number of algorithms available to resolve the real-time problems of machine learning. These algorithms consider the predefined gameplay input and their former experiences to yield a precise result. These algorithms vary in performance based on the nature of problem.

A. Naïve Byes Classifier

It comprises various classification algorithms centered on Bayes' theorem.



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

Fig.2. Naïve Bayes Classifier

This pool of algorithms shares a common principle, which is the strong independence of each pair of features being classified. Bayes' theorem calculates posterior probability P(c|x) from P(c), P(x), and P(x|c). Naive Bayes classifier

presumes that the impact of a predictor's (x) value on a given class (c) is independent of the values of other predictors. This is referred to as the conditional independence of that class.

B. Randon Forest

Random Fortvfest, primarily based on Decision Tree, is the most eminent machine learning algorithm and includes multiple decision trees. It obtains prediction from each decision tree and estimates the best one by voting. Moreover, by averaging the result, it reduces overfitting.

The output of classification problems are usually discrete values which are completely different from each other. This algorithm includes constructing a decision tree for every sample and generating prediction. Finally, prediction results are voted and one with the most votes is selected.

Test error, also known as expected prediction error, can be computed any time using equation (2) mentioned below, where E is the error and L represents the values of the data.

$$\mathrm{Err}(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\{L(Y, \varphi_{\mathcal{L}}(X))\},$$

C. K-Nearest Neighbors

K-nearest neighbors (KNN) algorithm is a simple, yet powerful machine learning algorithm which can be implemented to solve both the problem of regression and the problem of classification.

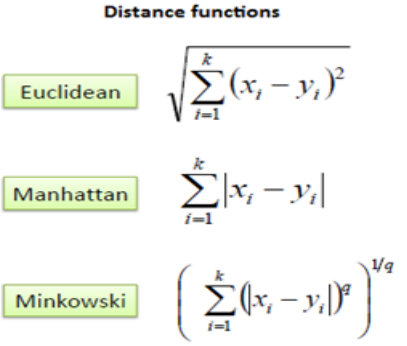


Fig.3. Distance Functions

The supervised KNN algorithm assumes that in close proximity, similar things occur. Hence, it classifies the data based on the distance between two points. Its various applications include intrusion detection, data mining, and pattern identification. KNN is non-parametric and totally tough. Various distance functions used in this algorithm are given below in equation (3).

V. METHODOLOGY

A. Data Collection

Data was obtained from cricsheet.org [20]. On this website, data is available in YAML format, which is a human-readable format.

There is a different file for each match containing information about PSL season year, details of match winner, player of the match, match number, location, name of the stadium, umpires' particulars, participating teams, and margin of winning. Since PSL is only five years old, the files of only 146 matches played during the five seasons are available. The data contains null values. For prediction analysis, certain attributes may not be needed.

B. Data Pre-Processing

Since the available data was in YAML format and contained a different file for each match, it needed strong preprocessing. First, we read it into python and normalized it using python library known as json [999]. Then, we converted it to dataframe [999] using another library called pandas [999]. Afterwards, we concatenated all of these files to construct a single dataframe [999] containing the data of all matches. Subsequently, we did exploratory data analysis to identify any existing anomalies. Usually, datasets contain flaws which need to be addressed before applying the algorithms. There could be several null values that affect the result. Data preprocessing yields a format that is easier to handle while using different algorithms.

1. Data Cleaning

Data may contain noise and missing values for different columns due to error(s) in recording or parsing. This results in improper classification. Hence, we replaced these values with dummy values such as mean and mode, depending on column types. Further, we removed columns that were completely null.

2. Choosing Required Attributes

In this step, we dropped certain columns of the data that did not help in prediction or only resulted in decreased accuracy. Feature importance was used to point out such columns.

3. Model Evaluation and Development

We applied different classification algorithms including KNN, Random Forest and Naïve Bayes classifier. The basic procedure followed in applying all of these algorithms is given below.

1. for for each sample iteration do
2. Split data into train and test set.

3. Train the model on training set using all features.
4. Predict on the test set.
5. Calculate features' ranking.
6. **for** for each subset of features Si, where i = 1 to S **do**
7. Keep the Si most important features.
8. Train the model on training act using Si features.
9. Predict on the teat set,
10. **end**
11. **end**
12. Calculate the performance of the model over Si using held-beck samples.
13. Identify an appropriate number of features.
14. Identify the final list of features to keep in the model.
15. Train the model using the optimal set of features on the original training set.

VI. RESULTS AND DISCUSSION

PSL dataset was used for applying different classification algorithms and Random Forest gave the highest accuracy. The top two highest accuracies were 50% and 60%, yielded by KNN and Random Forest, respectively. Classification report comprises F1score, recall, precision and accuracy. The representation of confusion matrix is as follows:

Table.1. Confusion Matrix

Confusion Matrix		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
Actual	Negative	False Positive	True Negative

a) True Positive is the number of correct predictions that an instance is negative.
b) False Negative is the number of incorrect predictions that an instance is positive.
c) False Positive is the number of incorrect predictions that an instance is negative.
d) True Negative is the number of correct predictions that an instance is positive.

Precision: It is the ratio of correctly predicted positive instances to the total number of positively predicted observations.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: It is the ratio of correctly predicted positive observations to total observations.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1-Score: It is the weighted average of recall and precision.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Accuracy: It is the average of recall and precision.

$$Accuracy = \frac{Precision + Recall}{2}$$

The model has an accuracy of 60.11% and classification report is given in Table 2. The accuracy for the prediction of match winner is shown for all the teams.

Table.2. Match Winner Prediction

	precision	recall	f1-score
1	0.75	1	0.86
3	0.67	1	0.8
4	1	1	1
5	1	0.71	0.83
6	0.5	0.5	0.5
accuracy	0.8	15	
macro avg	0.78	0.84	0.8
weighted avg	0.84	0.8	0.8

The accuracy of the three algorithms is compared in Table 3 below.

Table.3.Accuracy Comparison

Type	**KNN**	**Gaussian Naive Bayes**	**Random Forest**
Obtained Accuracy	49.99%	19.86%	60.11%

From Table 3, it is eminent that Random Forest has the highest accuracy, followed by KNN classifier. Gaussian Naive Bayes and other classifiers performed poorly in predicting the outcomes of PSL matches. Graphical representation of model performances is depicted below.

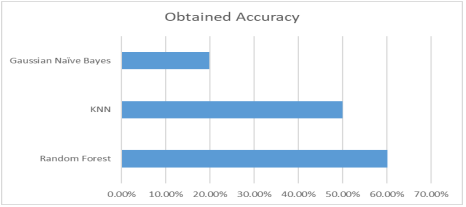


Figure.4.Obtained Accuracy

VII. CONCLUSION

Prediction of the winner in sports, specifically cricket, is a really complex and challenging task. However, by applying machine learning algorithms, it could be made less complicated. In this research, different factors influencing the outcomes of PSL matches were identified. The factors that considerably affect the results of a PSL match include toss winner, toss decision, participating teams, stadium, and location.

We designed a generic model based on team 1 and team 2, toss winner, toss decision, and venue and also predicted the match winner. Three classification models were trained on the PSL dataset. These included Gaussian Naive Bayes, KNN and Random Forest. Among these algorithms, Random Forest classifier gave the best result with an accuracy of 60.11%.

For future work, the plan is to add other attributes, such as the recent

history of the pitch, match day conditions, and the current form and experience of the players of both teams. The model designed in our project can be modified to be used for other sports, such as football and baseball.

References

1. R. Lamsal and A. Choudhary, Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning, Melbourne: University of Melbourne, 2018.
2. M. Bailey and S. R. Clarke, "Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress," Journal of Sports Science & Medicine, 2006.
3. Predicting the Winner in One Day International Cricket , "Predicting the Winner in One Day International Cricket," Journal of Mathematical Sciences & Mathematics Education.
4. C. M. Gil Fried, A data-driven approach to sport business and management, London: Routledge, 2016.
5. T. H. Davenport, "What Businesses Can Learn From Sports Analytics," MIT Sloan Review, 2014.
6. D. M. Robert Rein, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," Springerplus, 2016.
7. H. Ghasemzadeh and R. Jafari, "Coordination Analysis of Human Movements With Body Sensor Networks: A Signal Processing Model to Evaluate Baseball Swings," IEEE Sensors Journal , pp. 603 - 610, 2011.
8. T. A. Severini, Analytic Methods in Sports: Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports, New York: taylor & francis Group, 2014.
9. R. R. a. E. Feustel, "FORENSIC SPORTS ANALYTICS: DETECTING AND PREDICTING MATCH-FIXING IN TENNIS," Journal of Prediction Markets, pp. 77-95, 2014.
10. A. D. Mahanth K. Gowda, "Bringing IoT to Sports Analytics," NSDI, 2017.
11. K. Goldsberry, "CourtVision: New Visual and Spatial Analytics for the NBA," Harvard University, Cambridge.
12. R. Lamsal and A. Choudhary, "Predicting Outcome of Indian

Premier League (IPL) Matches Using Machine Learning," 2018.

13. P. Halvorsen, S. Sægrov, D. K. C. Kristensen and A. Mortensen, "Bagadus: An integrated system for arena sports analytics - A soccer case study," in Proceedings of the 4th ACM Multimedia Systems Conference, 2013.
14. e. a. Rabindra Lamsal, "Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning," DeepAI, 2018.