Predictive Analysis of IPL Match Winner using Machine Learning Techniques

Ch Sai Abhishek, Ketaki V Patil, P Yuktha, Meghana K S, MV Sudhamani

Abstract*: *Artificial intelligence (AI) can be implemented using Machine Learning which allows the computing to potentially robotically study and improve from its previous experiences without being manually typed. Data can be accessed and used by the computer programs developed using Machine learning. This paper mainly focused on implementation of machine learning in the arena of sports to predict the captivating team of an IPL match. Cricket is a popular uncertain sport, particularly the T-20 format, there's a possibility of the complete game play to change with the effect of any single over. Millions of spectators watch the Indian Premier League (IPL) every year, hence it becomes a real-time problem to compose a technique that will forecast the conclusion of matches. Many aspects and features determine the result of a cricket match each of which has a weighted impact on the result of a T20 cricket match. This paper describes all those features in detail. A multivariate regression-based approach is proposed to measure the team's points in the league. The past performance of every team determines its probability of winning a match against a particular opponent. Finally, a set of seven factors or attributes is identified that can be used for predicting the IPL match winner. Various machine learning models were trained and used to perform within the time lapse between the toss and initiation of the match, to predict the winner. The performance of the model developed are evaluated with various classification techniques where Random Forest and Decision Tree have given good results.

Keywords*: *Cricket prediction, Decision Trees, KNN, Logistic Regression, Multivariate Regression, Random forest, SVM, Sports Analysis.

I. INTRODUCTION

The main aim is to use Machine learning to develop the computer programs which will be capable of retrieving data and using it for self-learning. The procedure of learning commences with observations of data, such as instances, direct experience, or training, in order to identify some patterns in statistics and take improved decisions in the future built on the samples that are provided. Machine Learning primarily aims at eliminating the human intervention or assistance by allowing the computers learn automatically and adjust its actions accordingly. The advancement in computing in the recent years, has made it increasingly easy to acquire in-depth information.

Revised Manuscript Received on December 15, 2019.

Chakka Sai Abhishek, Pursued Bachelor, of Technology In Information Science Engineering In RNS Institute of Technology.

Ketaki Vinod Patil Pursued Bachelor of Technology In Information Science Engineering In RNS Institute of Technology.

Yuktha P,Pursued Bachelor of Technology In Information Science Engineering In RNS Institute of Technology,

Meghana K S, Pursued Bachelor of Technology In Information Science Engineering In RNS Institute of Technology.

Dr. M V Sudhamani, Professor And Hod, Dept. of ISE, RNSIT.

As a consequence, the fact of having both live and historic data has made Machine Learning quite popular in the fields of sports analytics [1–5]. Sports Analytics is a method of collecting and analyzing historical game information to derive essential knowledge from it, with the aim that it will promote successful decision-making.

Machine learning in sports arena, both off-the-field and on-the-field, can be used effectively on different occasions.

A team's performance and its outcome against an opponent can be efficiently predicted using the proposed model. This model primarily focuses on the healthy growth and productivity of team owners and other investors in the industry. Here, analysis is done by using certain machine learning classification techniques, like Decision Trees, Logistic Regression, Support Vector Machine, K-Nearest Neighbors and Random Forest.

One of the greatest successful football clubs in Portugal, Sport Lisboa e Benfica [6-9] implements machine learning in information processing techniques for making decisions , demonstrating the importance of machine learning in athletic analysis. The club not only tracks but also evaluates virtually each part of the game, together with their habit of resting, drinking, and practicing. After capturing raw player data, different models are programmed to analyze data to maximize game preparation and create custom training schedules. The data coming from the built models allow players to constantly improve their performance by incorporating machine learning and predictive analytics. Decisions including player substitution, holding a player in the lineup and leaving a player at bench can be made by the team coach depending on the analysis of the facts obtained.

The dataset used in this work is collection of different match plays, there are around 675 match details with complete information about the match winner, location toss winner, team names and other important attributes. The matches are from 2007 – 2018. This dataset has helped us achieve our main aim of our project.

II. LITERATURE REVIEW

In the last few years, Major League Baseball (MLB) has realized huge development in the sport technology domain [7,8]. Sometimes ball-by-ball information and simple insights about the game prove to be of great importance. These are typically not noticeable by human observation. Hence, Professional MLB teams apply various machine learning techniques to obtain and utilize the data.

Predicting the result of the game, classifying whether a team can deliberately allow a player walk in bat and categorizing non-fastball pitches by type of field, etc., are few of the sorting issues resolved with help of machine learning in baseball sport [9].

Additionally, cricket also uses sports analytics to forecast a match's outcome while the action is underway or even before the match has begun [10-14]. Also issues such as forecasting a player's runs or wickets intended for a match, built on his / her previous performance, are fascinating issue that must be focus on. Few practical tools employed in cricket consist of WASP (Winning and Score Predictor) [15], a tool that forecasts a score and potential result of a restricted match over cricket, i.e. one-day or t-20. This tool was first introduced by Sky Sports New Zealand during an ongoing T-20 match in 2012. Software such as Hawk-Eye [22–23], which records a ball's trajectory and reveals the most statistically significant direction visibly, has also been formally used since 2009 as the Umpire Decision Review System. Similarly, this computer-assisted intelligent technology is also used by other sports such as tennis, badminton, snooker. In the further sections of this paper we have briefly discussed about the importance of machine learning in sports, proposed methodology.

III. MACHINE LEARNING IN SPORTS REVIEW

With AI and Machine learning application, Machine learning is changing the technology in every possible field, machine learning can help real time problem in the society. It has a profound impact in the field of sports, provided in various sports like cricket, baseball, football and others. organizations can use the data during the game play to improve every area of their players move. From player recruitment to player performance to ticket sales, predictive analysis can help make targeted decisions and strategic changes that impact every area of sports organization.

In the past few years, importance of sports has emerged as an important element in our society, even in Olympics sports helps towards community health and productivity. Using machine learning we can improve the way players involvement and necessary improvements can be made using powerful machine learning techniques, these techniques help to improve the results over the previous results which where achieved without using any machine learning techniques. Detailed data about the players performance will be analyzed and improvement areas can be predicted using machine learning. In the study “Predicting the match outcome in one day international cricket matches” [17] this paper proposed a basic solution towards forecasting the winner of a cricket match he also has given a discussion of how machine learning algorithms can change the future of the means players evaluate their play in the game. This method has many benefits over the existing method where manual effort to find the flaw in the game play is reduced which makes it easier for computers to efficiently find the improvement strategy. Even during the 1960’s Arthur Samuel used a intelligent approach to gaming by the methods of artificial intelligence in the game of checkers to provide the number of moves for the player can win the match against his opponent. In the research paper [12] discussed they have proved that Bayesian networks has performed over other machine learning techniques. This paper has provided a solution to manual effort to improve over the results obtained previously. This Bayesian network prediction has provided a result of 59.21% accurate over the other methods. Machine learning has always provided a prominent solution towards the way gaming can be monitored.

In the paper [8] analysis of baseball game is shown where the game play can be evaluated by using machine learning techniques like SVM, K nearest neighbors which helped over the existing methodologies, this method involved extracting 75% of the highlights of the match and then deep evaluation is carried out. This change in gameplay evaluation helped the players to find out new techniques to win over the opponents. They even proposed the way game play results can be interpreted pictorially and provided the way the analysis of a match is given out.

Supervised algorithms were used during the game prediction, like logistic regression, Random forest, SVM , KNN which are explained deeply within this paper.

IV. PROPOSED WORK

From the Literature survey, it can be seen that a machine learning model can be built which will be able to forecast the result of the match even beforehand the match begins. There are many formats of playing cricket, among them T20 format is the one that has many turnarounds such that it will be very tough to foresee the champion until the last ball. So, it is quite complex to predict the winner.

Most of the statistical work in sports is performed using regression and classification tasks, both of which are subject to supervised learning. Simply put, y = f(x) which is a predictive value learned from a dataset constructed by the learning data: D = ((X_1,Y_1), (X_2,Y_2), (X_3,Y_3), ... (X_n, Yn)). Supervised learning can be divided based on output to two categories as classification, regression. So, this problem is a classification problem. So, we apply several classification algorithms on the cricket dataset, evaluate the results and select the most appropriate model that gives greater accuracy.

Figure – 1 basic methodology of the paper

The above figure 1 represents the basic methodology for our research work. This represents multiple methods like data collection, data exploration, data cleaning, data preprocessing, model development and model evaluation. These methods play a major role in predicting the output for our work. These methods are explained in detail in this paper in further sections. There are various algorithms can be used to solve the real time problems in machine learning. These algorithms take a predefined input of game play and its previous experiences and provide an accurate output, these algorithms differ in task and the operations involved to give solution to the problem, these algorithms are discussed here:

Decision tree

Decision tree is involved in both regression and classification, this methodology is used to depict the various decisions taken to provide the necessary result. This results in a tree of choices selected. Decision tree has more impact over other machine learning algorithms and has provided more accurate results. is calculated using entropy by classifying into two major types as "yes" or "no" the mathematical equation is given below in the equation (1):

$$H(S) = \Sigma - p(c) \log_2 p(c) \quad ---(1)$$

$$C = \{ \text{yes, no} \}$$

Random Forest

Random forest is uniquely the most prominent machine learning algorithms which consists of multiple decision trees together. Here every single individual tree will explore more deeply about its class predictions and the class with most of the votes, becomes our model's actual prediction.

Classification type of problem always have a discrete value as the output which are completely different to each other. The main strategy behind random forest is that it divides the whole strategy into multiple trees resulting in various solutions resulting in the most prominent tree path as the final accuracy. This helps in many classification algorithm, to classify various object depending their behavior. Here the expected prediction error is calculated for every time, this error is also known as test error, this equation (2) is given below.

$$\mathrm{Err}(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\{L(Y, \varphi_{\mathcal{L}}(X))\}, \quad ----(2)$$

Here E respresents the error, L represents the data values.

K Nearest Neighbors

This algorithm is possible for classification as well as regression type of problems, this algorithm is one of the prominent in machine learning since it is a non – parametric way where there won't be any expectations about the distribution of data. In supervised learning KNN is used in powerful application like pattern identification, data mining and intrusion detection. KNN is completely robust, it calculates the distance between the test data and the input and gives the prediction accordingly. One of the equations used for finding the distance between the input and test data is shown in the equation (3) below

$$D_H = \sum_{i=1}^{k} |x_i - y_i| \quad ----(3)$$

Support Vector Machine

Support vector machine is used prominently for supervised algorithms, they work clearly well when there's a strong boundary of separation between classes, they're additionally effective in high dimensional spaces. SVM is more memory efficient. They are suitable for more large data sets. They work well in unstructured and semi structured data like text, images and trees. They have generalization in practice. They are good for text classification.

V. METHODOLOGY

A. *Data Collection*

The Indian Premier League's official website[21] is the principal basis of data for this project. The data was webscrapped from the website and kept in the appropriate format using a python library called beautifulsoup. The dataset has the columns regarding match-number, IPL season year, the place where match has been held and the stadium name, the match winner details, participating teams, the margin of winning and the umpire details, player of the match. Indian Premier League was only 11 years old, which is why, after the pre-processing, only 634 matches were available. Here, some of the columns may contain null values and some of the attributes may not be required for match winner prediction which is discussed in data preprocessing.

B. *Data Preprocessing*

Here, in this step we have tried to explore more in the dataset to find any anomalies present, every dataset might have certain defects which have to be regulated to make it a standard form for performing calculations. Defects can be like having null values in certain attribute values or like having empty values in the certain required attributes. This step provides us a detailed format or understanding the dataset and presenting in a structured format which easy to process.

(i) Data cleaning

There are some null values in the dataset in the columns such as winner, city, venue etc. Due to the presence of these null values, the classification cannot be done accurately. So, we tried to replace the null values in different columns with dummy values.

(ii) ***Choosing*** *Required Attributes*

This step is the main part where we can eliminate some columns of the dataset that are not useful for the estimation of match winning team. This is estimated using feature importance. The considered attributes has the following feature importance.

C. *Model Development and Evaluation*

Here, we have developed a generic model and applied all classification methods. The detailed procedure is as follows:

Team2	0.254593
Team1	0.223901
Venue	0.168813
Toss Winner	0.164288
City	0.154104
Toss Decision	0.034301

Algorithm: Model Evaluation Algorithm:

```
for for each sample iteration do
    Split data into train and test set.
    Train the model on training set using all features.
    Predict on the test set.
    Calculate features ranking.
    for for each subset of features S_i, where i = 1 to S do
        Keep the S_i most important features.
        Train the model on training set using S_i features.
        Predict on the test set.
    end
end
Calculate performance of the model over S_i using held-back samples.
Identify appropriate number of features.
Identify final list of features to keep in the model.
Train the model using the optimal set of features on original training set.
```

The above algorithm is mainly describes the procedure for the work in a pseudo code format, initially for every iteration, the data is split into training data and test data, we train the model using certain features and use it to predict the testing data , then we calculate the performance of the system.

The above procedure evaluates the classification model and calculates the accuracy.

The various classification models used are: Logistic Regression, Gaussian Naïve Bayes Classifier, KNN (K Nearest Neighbor) algorithm, Support Vector Machines, Gradient Boost Algorithm, Decision Trees and Random Forest Classifier. Among these methods the random Forest and Decision tree has given good results.

VI. RESULTS AND DISCUSSIONS

As discussed above, the IPL dataset was trained in different machine learning algorithms for the database that included all the match details from the launch of the Indian Premier League till 2018 and the highest accuracy is given by Random Forest Classifier and Decision Tree.

The Random Forest classifier and Decision Tree correctly predicted the outcome with the accuracy of 89.151% given the train data 70% and test data 30% of the entire dataset. Classification report consists of values for accuracy, precision, recall and f1-score, the explanation for which is given below. The confusion matrix can be graphically represented as:

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive	False Negative
	Negative	False Positive	True Negative

Precision:

Precision is the proportion of correctly predicted positive observations versus the total number of positively predicted observations. Precision talks about how precise a model is out of predicted positive, how many of them are actual positive.

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

Recall:

Recall is the proportion of number of correctly predicted positive observations versus total number of all actual observations.

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

F1-Score:

F1 Score is the weighted average of precision and recall.

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

Accuracy:

Accuracy can be calculated as the average of Precision and Recall.

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

The classification report of our proposed model is shown in the table 2 below. The accuracy for predicting the winner of the match is 90.1% and the accuracy for predicting the loser is 88.2%.The average accuracy for predicting the outcome of the match is 89.151%.

		Precision	Recall	F1-Score	Accuracy
Matches Won	**1**	0.8831	0.9189	0.9	0.901
Matches Lost	**0**	0.913	0.851	0.8809	0.882
Average		0.898	0.8969	0.8904	0.89151

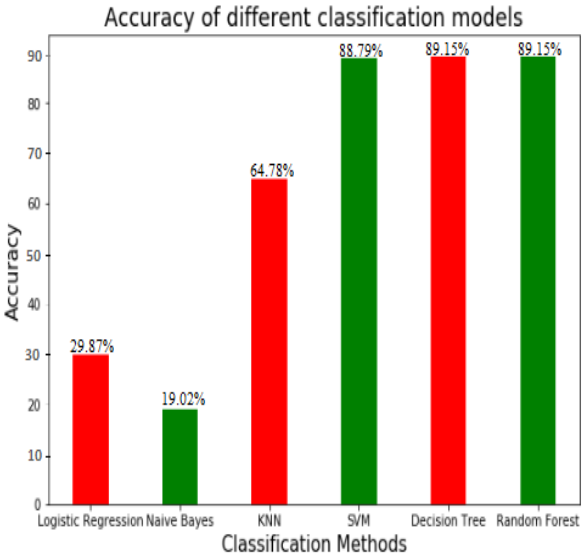
Table 2: precision recall and accuracy evaluation

Comparison of the performance in terms of accuracy given in different models is discussed in Table 3 below.

Table 3 Accuracy of various methods

Type	Logistic Regression	KNN	SVM Technique	Decision Tree	Random forest
Obtained accuracy score	29.84%	64.78%	88.79%	89.151%	89.151%

As seen from the table 3, based on the accuracy of the classification, Random forest and Decision Tree were the highest, followed by other methods such as Support Vector machine and K-nearest neighbor classification algorithms. The performance of Naive Bayes and Logistic Regression classifiers was poor in predicting the IPL match outcome. The bar graph plotted below shows the accuracy of different classification models.



VII. CONCLUSION

Predicting the winner in sports, cricket in particular is a challenge and very complex. But by incorporating machine learning, this can be made much simpler and easier. In this study, the various factors that influence the outcome of an Indian Premier League matches were identified. The factors which significantly influence the result of an IPL match included the playing teams, match venue, city, the toss winner and the toss decision.

A generic function for classifier model was designed to measure the points earned by each team based on their past performances, including team1, team2, venue of the match, toss winner, city and toss decision. Different classification-based machine learning algorithms were trained on the IPL dataset developed for this work. The methodologies used in our work to find the final evaluation are Logistic regression, Decision trees, Random forest and K-nearest neighbors. Among these techniques, the Random forest classifier and Decision Tree provided the highest accuracy of 89.151%.

For future work, we plan to expand our work using more attributes like the previous match score of the selected team and opponent team, the number of skilled batsmen in the opponent team, and more. The machine learning methods used in our research can also be used to predict the outcome in other outdoor sports such as football, baseball and more.

REFERENCES

1. P. Halvorsen, S. Sægrov, A. Mortensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, et al., "Bagadus: an integrated system for arena sports analytics: a soccer case study," Proceedings of the 4th ACM Multimedia System Conference, pp. 48–59, ACM, 2013
2. A. S. Forouhar, M. M. Kellogg, K. Ohiomoba, and E . Akhmetgaliyev, "Methods, systems and software programs for enhanced sports analytics and applications," May 14 2015.
3. K. Goldsberry, "Courtvision: New visual and spatial analytics for the nba," in 2012 MIT Sloan sports analytics conference, vol. 9, pp. 12–15, 2012.
4. M. Gowda, A. Dhekne, S. Shen, R. R. Choudhury, L. Yang, S. Golwalkar, and A. Essanian, "Bringing iot to sports analytics," in 14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI},2016
5. R. M. Rodenberg and E. D. Feustel, "Forensic sports analytics: Detecting and predicting match-fixing in tennis.," Journal of prediction markets, vol. 8, no. 1, 2014.
6. Wired, "The unlikely secret behind benfica's fourth consecutive primeira liga title," May 2017.
7. T. A. Severini, Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports. Chapman and Hall/CRC, 2014.
8. 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, vol. 11, no. 3, pp. 603–610, 2010
9. R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," SpringerPlus, vol. 5, no. 1, p. 1410, 2016
10. T. H. Davenport, "What businesses can learn from sports analytics," MIT Sloan Management Review, vol. 55, no. 4, p. 10, 2014.
11. G. Fried and C. Mumcu, Sport analytics: A data-driven approach to sport business and management. Taylor & Francis, 2016.
12. T. A. Severini, Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports. Chapman and Hall/CRC, 2014.
13. K. Koseler and M. Stephan, "Machine learning applications in baseball: A systematic literature review," Applied Artificial Intelligence, vol. 31, no. 9-10, pp. 745–763, 2017.
14. A. Bandulasiri, "Predicting the winner in one day international cricket," Journal of Mathematical Sciences & Mathematics Education, vol. 3, no. 1, pp. 6–17, 2008.
15. K. Koseler and M. Stephan, "Machine learning applications in baseball: A systematic literature review," Applied Artificial Intelligence, vol. 31, no. 9-10, pp. 745–763, 2017.
16. A. Bandulasiri, "Predicting the winner in one day international cricket," Journal of Mathematical Sciences & Mathematics Education, vol. 3, no. 1, pp. 6–17, 2008.
17. 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, vol. 5, no. 4, p. 480, 2006.
18. V. V. Sankaranarayanan, J. Sattar, and L. V Lakshmanan, "Auto-play: A data mining approach to odi cricket simulation and prediction," in Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 1064–1072, SIAM, 2014.
19. A. Kaluarachchi and S. V. Aparna, "Cricai: A classification based tool to predict the outcome in odi cricket," in 2010 Fifth International Conference on Information and Automation for Sustainability, pp. 250–255, IEEE, 2010.
20. E. Crampton and S. Hogan, "Cricket and the wasp: Shameless self promotion (wonkish)..",2016
21. http://stats.espncricinfo.com/ci/engine/records/index
22.] N. Owens, C. Harris, and C. Stennett, "Hawk-eye tennis system," in 2003 International Conference on Visual Information Engineering VIE 2016, pp. 182–185, IET, 2016.

AUTHORS PROFILE

Chakka Sai Abhishek has pursued bachelor of technology in Information Science Engineering in RNS Institute of Technology. He has won 2nd prize in IOT innovation Events and also presented in various technical research conferences.

Ketaki Vinod Patil pursued bachelor of technology in Information Science Engineering in RNS Institute of Technology. She has successfully presented a paper in IETE Sponsored Second National Conference on Emerging Trends in Engineering, Science and Technology and also won 2nd prize in IOT innovation event.

Yuktha P pursued Bachelor of Technology in Information Science Engineering in RNS institute of Technology, she has presented in various technical events and has also presented a paper in IETE Sponsored Second National Conference on Engineering Trends in Engineering, Science and Technology.

Meghana K S, pursued Bachelor of Technology in Information Science Engineering in RNS Institute of Technology.

Dr. M V Sudhamani, currently working as Dean-R&D, Professor and HoD, Dept. of ISE, RNSIT. She is having Teaching, Research and Industrial experience of 25 years. She has specialization in Image Processing, Content-based Image Retrieval, Advanced Algorithms and Databases. Guided and guiding candidates for Ph. D degree. She has carried out two research projects from VTU and AICTE. She has served as member of Board of Examiners (BOE) and Board of Study (BOS) member in VTU and other autonomous institutions across India. She has organized two international conferences ICDECS 2011 and 2015, and one more in December 2019.