Indian Premier League Match Analysis and Prediction (Group No. 11)

Akhileshkumar Mutuguppe Subbarao
School of Engineering and Computing
Dublin City University
Dublin, Ireland
akhileshkumar.mutuguppesubbarao2@mail.dcu.ie

Amruth Hebbasuru Suryanarayana
School of Engineering and Computing
Dublin City University
Dublin, Ireland
amruth.hebbasuru2@mail.dcu.ie

Kadesh Basavaraj Huddar
School of Engineering and Computing
Dublin City University
Dublin, Ireland
kadesh.basavarajhuddar3@mail.dcu.ie

Naveen Garaga Krishnamurthy
School of Engineering and Computing
Dublin City University
Dublin, Ireland
naveen.garagakrishnamurthy2@mail.dcu.ie

***Abstract*— Sports is one of the prime topics of discussions in today's world, attracting millions of followers. Cricket is considered to be most exciting and unpredictable game. It's becoming a billion dollar market for many as they gamble financially, expecting to be able to earn income. The gambling market should be on hike every year as there is a great deal of worry about spot fixing. Because cricket is known to be gentleman game, it becomes an important aspect of how great it would be if we could predict or understand the game's outcome, making it more interesting to see if their most loved and followed team could possibly win the match and invariably the tournament. So, this work focuses on predicting Indian Premier League (IPL) match-winning result, considering different influencing factors using machine learning algorithms such as Logistic Regression, Random Forest and K-Nearest Neighbors. Paper also provides the analysis of why IPL has achieved such huge success in short period and for team owners to look for best players in auction. We are therefore putting forward a model for viewers to know the possible probability of winning any T-20 style game.**

*Keywords—**Indian Premier League, cricket, T20, Random Forest, Logistic Regression***

I. Introduction

Cricket is most famous and much loved game after soccer. It is widely followed sports in South Asia, Australia, The Caribbean and GB, with a devotee base of about 2.5 billion . Cricket is played in 3 formats at international level - Test, ODI and T20. Cricket's popularity grew when ICC (International Cricket Council) began the idea of fast cricketing in the form of matches (T-20). In 2007, the first twenty-20 world cup was held in South Africa which was won by India, this increased game's popularity in India. BCCI (Board of Control for Cricket in India) cashed the opportunity and formed a league known as Indian Premier League (IPL) in 2008 and received ICC approval.

IPL is one of the finest twenty-20 cricket competition at present involving hundreds of foreign stars, and it’s thrilling structure has altered cricket money aspects. The gross IPL revenue was estimated at 3.2 billion US dollars in 2014. There are 8 teams competing with each other in the first stage in each IPL season, 4 teams go to the eliminator round after the first stage and 2 teams go to the final match after the eliminator round and finally there will be one winner [1].

The result of matches is very significant for all stakeholders because of the presence of income, team spirit, city loyalty and a large fan following. This, in effect, depends on the intricate rules governing the game, team luck (Toss), players 'skill and success on a given day. Various other natural parameters, such as the player-related historical data, play an integral role in predicting a cricket match outcome[3]. A way to predict match result between different teams can help in the team. The research provided in this paper can be used to determine players performance. This analysis includes visualization of the results of the players. Using IPL T-20 variables relating to batsmen and bowlers’ statistics, a variety of apt variables have been defined with elucidating power over auction values. In addition , several predictive models are also built to predict the outcome of a match, based on past performance of each player as well as some match-related data[2]. The built models can help decision-makers determine the power of a team against another during the IPL matches.

Seeing this from a business perspective, the key purpose is to enhance predictive accuracy using various advanced technology that allow a better system to be put in place to help people make future bets with prior knowledge. Betting industry has also given its marketplace stand attracting gamblers with better odds, and this model will provide insights on match result providing an idea of the likelihood before making any possible bets. In addition, this study will also help IPL franchises and coaching staff look for better players in the auctions year after year based on the tournament strength for their respective teams.

The rest of the paper is organized as follows: Section - II. deals with previous work related to analysis and prediction of matches. Section - III presents the Data Mining Methodology and steps taken to arrive at the results, Section- IV briefs about the results, findings and subsequent discussions. Section V draws conclusions and future works about the research done.

II. RELATED WORK

Parker et al have printed a blueprint for IPL auction valuation of players. Their thought-about model features like player experience, player's previous bid value, strike rate etc.[1]. Authors looked at restricted overs cricket for inspecting and selecting batsmen. They printed a fresh live P(out), i.e. probability of getting out and used a 2-D graphic illustration of strike rate on one axis and strike rate on another. They then define the batsmen's range criteria assisted by P(out), strike rate, and batting average. They used a replacement graphical diagram with strike rate on one axis and thus the chance to get out on the opposite, similar to the risk– come paradigm used in portfolio analysis, to gain useful, direct and comparative insights into batting results, significantly in the one-day game sense[4].

Authors classified all-rounders into four varieties of victimization Naïve Thomas Bayes classification: all-rounder batting, performer, under-performer and all-rounder bowling. The study helps identify factors that are to blame for bowler performance in Twenty20 cricket. The selectors also use the findings to settle out of a bunch of potential players for a given team on the bowlers. This data and target players could be used by the franchisee of IPL groups when bidding. Analysis of Bowlers performance victimization Combined Bowling Rate applied mathematics measurements assisted performance assessment and related studies are a fertile area for future study[5]. Islam et al. [6] analyzed sport information, especially for statisticians, cricket is a remarkable field. The success of a team depends, potentially, on whether they play reception or away. A superior team is deemed once performs competitively while loving the reception or away. On the opposite, the impact of the environmental advantage is often attributed to a team winning on a structured base reception and not playing fairly as they play abroad.

Authors suggested a model of machine learning, predicting the English-twenty over the county cricket cup. Main motive behind his research was to use multi-step approach to support the gambling industry. For the 2009 to 2014 season, authors took the archive dataset available on cricinfo.com and estimated additional features such as strike rate for predicting outcomes. They implemented four classifiers, namely Naive Bayes, Logistic Regression, Gradient Enhanced Decision Trees and Random Forests, and assessed the advantages of home and away teams[7]. Here authors studied data from Indian Cricket team's One Day International Cricket matches and mined various association rules using attribute-based market basketball tools. They looked whether it’s a home team or away team and the outcome of the game to enforce rules that they concentrate on different issues like toss outcome, toss winner, decision to bat first or not, which two teams are going to play at [8]. Authors are looking at the growing number of matches every day, it is difficult to handle or collect helpful data from all matches obtainable information. They present an information mental image associate method in nursing prediction within which an ASCII text file, distributed and non-relational information, is used to stay the information associated with cricket matches and players in the IPL (Indian Premier League). This information is then used to visualize the success of past players. Furthermore, the information is used to predict a match's end result through various approaches to machine learning. The proposed method will prove helpful in choosing the best team for team management within the player auctions. Authors discuss the issue of forecasting the results of the associate IPL[9].

Simulator considers the probability of basic features such as fielding, batting, over, inning in the current scenario. Result was determined by considering both win game and lose game perspective by taking two samples using Binomial test and Bayesian model with different variations[10]. Lemmer developed two models and used them to forecast match results during a Twenty20 cricket game. The goal was to look for a technique that could complete incoherent outcomes. The live modified accuracy of the predictive performance is illustrated and shown to convey a successful predictive results assessment[11]. Sankaranarayanan et al. [12] for modeling and predicting ODI matches, authors used data processing techniques. Defined a predictive model by subsection of match factors with clustering and regression algorithms that analyze historical Cricket game knowledge and on-site match conditions for forecasting game progression and the final outcome of One Day International game. To track the progression, the Ridge method and arbitrary attribute selection are used to predict the runs scored within the innings and based on Milestone Reaching Ability (MRA), which is the aggregation of all the qualities of batsmen which consists of opening batsmen, middle-order batsmen, all-rounders, wicket-keeper, and tail-enders. Researchers used details from previous matches such as average runs scored by the team in an innings, average variety of wickets lost by the team, etc. to model the match state.

Passi and Pandey[13] have showed that selection of players is one of the most critical tasks in any sport and cricket is no exception. The players 'success depends on various factors such as the opposition squad, the venue, his current form etc. The team management, the coach and the captain pick 11 players from a squad of 15 to 20 players for each match. To pick the best playing 11 for each match, they evaluate different characteristics and player statistics. Each batsman contributes by scoring as many runs as possible and each bowler contributes by taking as many wickets as possible and conceding minimum runs. This paper attempts to forecast players 'success as how many runs each batsman will score, and how many wickets each bowler will take for both teams. Both the problems are treated as problems of classification where the number of runs and the number of wickets is categorized into specific categories. Authors used naïve bays, random forests, multiclass SVM classifiers and decision tree classifiers to create prediction models for both issues. For both the issues, the random Forest classifier was found to be the most accurate. Random forests are a set of decision trees where each tree is dependent on a random vector sampled independently and with the same distribution of all the trees in the forest.

III. DATA MINING METHODOLOGY

Following section deals with the architecture of proposed data mining methodology.

Block Diagram is show in Figure – 1. Steps starts with Data Collection, followed by Pre – processing of collected data, next is data transformation of raw datasets to get insights, later different attributes are analyzed to come up with patterns and data visualization and others are used for partition of dataset in training and testing data. Last is training and classification to get the results.

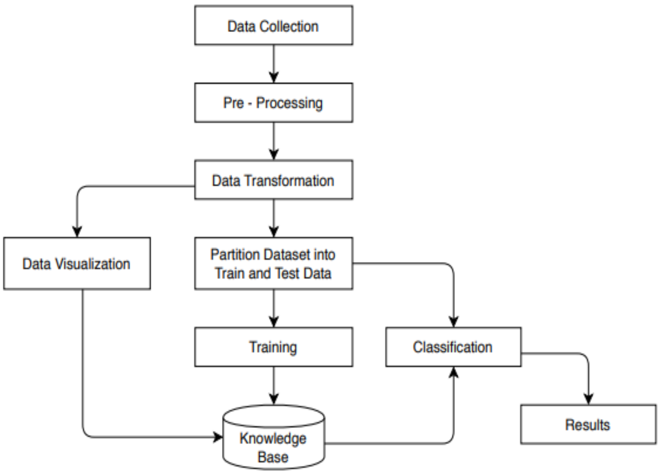
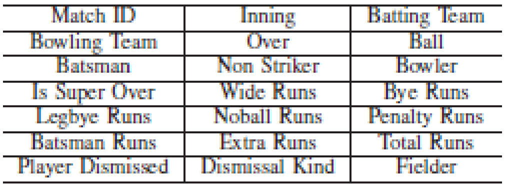


Figure - 1

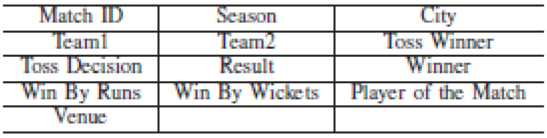
A. Data Collection

To evaluate and forecast the outcome of the IPL T-20 match, data sets are obtained from Kaggle.com The dataset contains descriptions of the past matches from 2008-2019. Each row in the dataset represents ball by ball information played with 21 attributes (Deliveries Data) in both inputs as given in Table-I. As shown in Table-II, Dataset also provides results of each IPL match with additional 14 attributes (Matches Data). Deliveries Data-dataset contains all matches ball by ball record, and also includes the corresponding data dictionary. Matches Data-dataset includes extra meta-data for each match played. This data file includes extra data for all matches in the Deliveries system, such as match venues, who won the match, win margin, toss decisions, etc.



Match ID	Inning	Batting Team
Bowling Team	Over	Ball
Batsman	Non Striker	Bowler
Is Super Over	Wide Runs	Bye Runs
Legbye Runs	Noball Runs	Penalty Runs
Batsman Runs	Extra Runs	Total Runs
Player Dismissed	Dismissal Kind	Fielder

Table – 1 : Deliveries Data Features



Match ID	Season	City
Team1	Team2	Toss Winner
Toss Decision	Result	Winner
Win By Runs	Win By Wickets	Player of the Match
Venue		

Table – 2 : Matches Data Features

B. Pre – Processing

Collected dataset at its initial stage has a raw data table which needs to be pre-processed to remove irrelevant information. Pre-processing step cleans the dataset by deleting data which is not useful for performance. Data that have not been declared or identified results are removed during the pre-processing stage. As given features in one dataset are not sufficient to predict the model, in addition, the combination of all features given in the dataset is not sufficient to analyze and predict results so required features are added which can play a major role in predicting the model. Of time the result may be in favor of different team.

C. Data Transformation

The data given in the dataset is of a categorical nature because of which classification will be complex. It can also influence the process of classification leading to erroneous prediction. All categorical data in the dataset is transformed to numeric in this step and is standardized on scale basis.

D. Data Visualization

In Data visualization step detailed analysis is made on research questions like What makes Indian Premier League so popular? Which team is winning high scoring matches? Bowlers who excel better in powerplay and death overs. Which bowlers will be best bid during future auctions?
The Graphical and Statistical investigation is finished with the assistance of seaborn and Matplotlib. This product is utilized for business insight applications and in addition for measurable examination. It has the arrangement to interface with any database and furthermore to transfer documents for performing factual investigation on them.

E. Training and classification

For training and classification, we have used three algorithms i.e. Logistic Regression, Random Forest and KNN.

1) Logistic Regression : Logistic regression uses maximum likelihood estimation for transforming the dependent variable into a logistic variable. Logistic regression uses the linear regression function to estimate the value of dependent variable by estimating the parameters for the linear equation. As shown in equation below α, b1, b2...bn are the parameters to be calculated using the training data and the equation will then be used to predict P(X) which represents the dependent variable value if values of features x1,x2,…xn are given.

$$P(X) = \alpha + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n,$$

Logistic regression is used when the output is binary i.e. of the form yes or no, 0 or 1, etc. As the output obtained by the above equation is a real valued number it needs to be converted in a form appropriate for making the prediction. Hence for this purpose logic or the sigmoid function is used to convert the output of linear regression into a probability value and its equation is as follows:

$$Q(X) = \frac{1}{1 + e^{-(a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}}.$$

$$0 \leq Q(X) \leq 1$$
$$-\infty < P(X) < +\infty$$

As the value of Q(X) is a number between 0 and 1 this value can be considered as the probability of the particular outcome. For example, if the output is 0.8 it means there are 80% chances of getting the output as 1 and it can therefore be safely predicted that for the given set of input attributes the output would be 1 [15].

2) Random Forest : Noteworthy enhancements in order exactness have come about because of growing a group of trees and giving them a chance to vote in favor of the most well-known class. So as to develop these outfits, frequently arbitrary vectors are created that oversee the development of each tree in the gathering[13]. An early precedent is arbitrary part choice where at every hub the split is chosen indiscriminately from among the K best parts produces new preparing sets by randomizing the yields in the first preparing set. Another methodology is to choose the preparation set from an irregular arrangement of weights on the precedents in the preparation set.

The normal component in these methodology is that for the kth tree, an irregular vector k is produced, free of the past arbitrary vectors 1,...,k−1 however with a similar dispersion; and a tree is developed utilizing the preparation set and k , bringing about a classifier h(x, k) where x is an info vector[13].

Given an ensemble of classifiers h1(x), h2(x), . . . ,hk (x), and with the training set drawn at random from the distribution of the random vector Y, X, define the margin function as:

mg(X, Y) = avk I(hk (X) = Y) − max j=Y avk I(hk (X) = j).

where I( ) is the indicator function. The margin measures the extent to which the average number of votes at X, Y for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. The generalization error is given by
PE∗ = PX,Y (mg(X, Y) < 0) where the subscripts X, Y indicate that the probability is over the X, Y space.
In random forests, hk (X) = h(X, k).

3) K – Nearest Neighbors : Due to its flexibility, non-parametric design and ease of execution, the KNN algorithm is commonly used in text categorization tasks. The algorithm seeks K's nearest neighbors of 'd 'from all training instances to classify a new document d. And it also considers the group which has the highest number of K nearest neighbors. The similarity score is determined in the first step between the test document d and each training document ti. In the second step, scores are determined for each category Cj by calculating weighted total of neighboring documents that belong to that category. By sorting the category scores, the final decision is given to the evaluation document which has the highest category score. The decision rule for KNN is written as

$$f(d) = \sum_{i=KNN} sim(d, t_1) y(t_t, C_f)$$

where f(d) is the label of category with respect to test document d, sim(d,ti) - Similarity between test document d and training document ti, Cj - Candidate category with respect to d, y(Li,Cj) the category score of training document di with respect to category Cj [14].

IV. EVALUATION AND RESULTS

Within this section we highlight the results obtained and then discuss the results of the proposed model. A detailed analysis of the IPL matches is also carried out and discussed further in the section. Every team plays one match in its home ground and its labeled as home match and another in opponent’s home ground which is labeled as away match. The dataset model is broken down into two parts. First set is referred to as the training dataset while second set is referred to as the Test Dataset. Training data set is equipped with logistic regression, KNN and random forest and information gained is used to predict test dataset results.

Results obtained are measured on the basis of accuracy, precision, recall and F1 score. The outcome is determined in terms of accuracy which is the percentage of team wins correctly categorized by classifiers versus total number of responses.

In modeling, we have worked on how accurate our model predicts the match winner based on each research question provided below.

1. Toss win match win : What is the match winning percentage of the team who win the toss?

2. Bat first : What are the chances of team batting first winning the match?

3. Home advantage : How does team playing in its home ground has advantage over away match?

4. Home advantage + toss winner : What is the percentage of team winning a match, if it plays in home ground and wins the toss?

S. No.	Research Question	Logistic Regression	KNN	Random Forest
1	Toss win match win	54.00	54.00	99.00
2	Bat first	17.00	34.00	89.00
3	Home advantage	23.00	32.00	89.00
4	Home advantage + toss winner	19.00	30.00	86.00

Table – 3 : Comparison of the 3 algorithms and accuracy

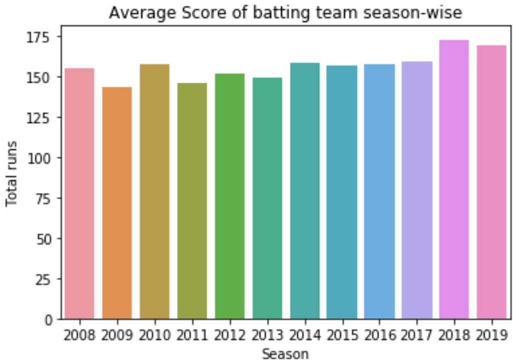
From above table it is evident that Random Forest gives more accuracy over other two algorithms.

Few match analysis is carried out with help of visualization and is shown below.

1. What makes IPL so popular?

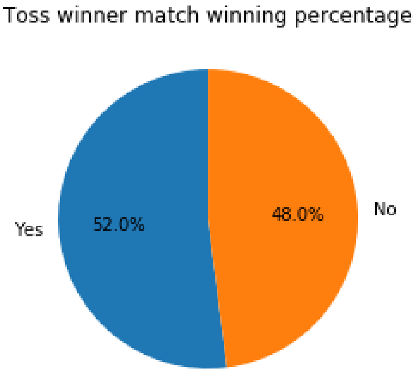
From the 3 plots below, we can infer three factors behind the popularity of Indian Premier League (IPL)

- Average score is high in every season



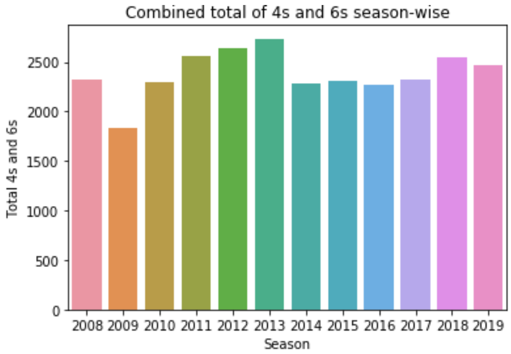
The run rate would usually not be as high as this format in a 50 over or a test match. Average innings score in 120 balls played here will be more than 140.

- A match can be unpredictable



Toss is a big factor in a cricket game as the team that wins the toss has high chances of winning. But we can infer from the plot that the match can go either way where winners of tosses have very slight advantage.

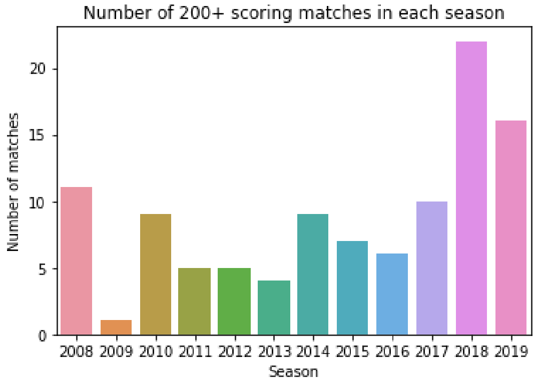
- Huge number of boundaries(4s and 6s) are hit



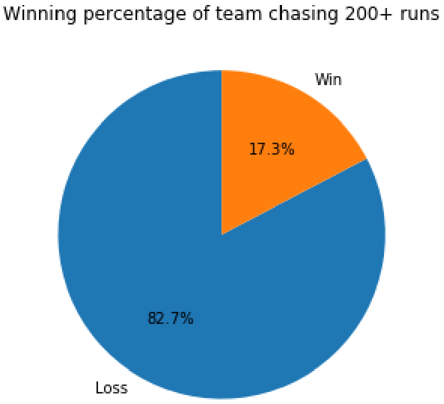
Over 2000 boundaries (4s and 6s) are scored in each season which is another explanation for high-score games and making the match fun to watch. South Africa hosted IPL in 2009 and it wasn't a big success because the pitches in South Africa weren't ideal for scoring many runs / boundaries. The position of the league played at can be attributed to a smaller number of boundaries in the year 2009.

IPL has huge cricket fan base because of many high voltage matches, and one cannot predict the match winner until very end ball. These things make this league super famous and attracts many fans.

2. Which team is winning the high scoring matches?



The first season had more than 10 200 + scores which contributes to the immediate success of the league. The second season has the lowest because it was held in South Africa. The 2018 season had the largest number of matches where more than 200 runs are scored. High match scores lead to the league's performance.



There is just 17.3% chance for the team chasing 200+ runs. This clearly shows that high scoring team with batting first have high chance of winning the match.

3. Bowlers who excel better in powerplay and death overs. Which bowlers will be best bid during future auctions?

In IPL powerplay refers to 1 - 6 overs and death over refers to 16 – 20 overs.

S. No.	Top 3 economic bowlers in powerplay	Top 3 economic bowlers in death overs
1	Bhuvaneshwar Kumar (B Kumar)	Jasprit Bumrah (JJ Bumrah)
2	Praveen Kumar (P Kumar) Note: Currently Retired	Lasith Malinga (SL Malinga)
3	Sandeep Sharma	DJ Bravo

Table – 4 : Top 3 bowlers in IPL (2008 - 2019)

We can see from the above analysis that economic bowlers during powerplay and death overs are different. This is the reason bowlers excel in their strength across the overs.

From table - 4, it shows that Bhuvaneshwar Kumar(B Kumar), Praveen Kumar(P Kumar) and Sandeep Sharma are consistent bowlers during powerplay throughout the seasons. Bowlers like JJ Bumrah, SL Malinga, DJ Bravo are performing extremely good during death overs across all seasons.

If IPL team owners want to have best bowling side then should definitely bid for Bhuvaneshwar Kumar(B Kumar),
JJ Bumrah, SL Malinga, DJ Bravo in future auctions.

V. CONCLUSIONS AND FUTURE WORKS

In this work we predict the outcome with machine learning algorithms from the IPL match winners. We predict the winners by using the criteria including bowling average, batting average, teams and the venue. We compared three algorithms, namely logistic regression, k-nearest neighbors and random forest, from all of the tests performed and the simulation results. From the analysis, we can infer that in our case, the best prediction algorithm is random forest, which provides the best results compared with state-of-the-art algorithms. This research also offers an area of scope as part of potential work to forecast real-time match outcomes and consider other specific aspects of the game, such as the Duckworth Lewis Method and Rain Disrupted Games, to make the prediction of results much easier. This research study also sets the IPL franchises as a benchmark in looking for better players year after year, thereby determining the best team for the tournament.

REFERENCES

[1] D. Parker, P. Burns, Natarajan, H. Player, ”valuations in the Indian Premier League”, Frontier Economics, October, 2008, 1-17.

[2] Kansal P, Kumar P, Arya H and Methaila A, ”Player valuation in Indian premier league auction using data mining technique”, International Conference on Contemporary Computing and Informatics (IC3I), 2014, 197-203.

[3] S. Singh, ”Measuring the Performance of Teams in the Indian Premier Leaguee”,American Journal of Operations Research, 2011, vol 1, No 3,pp 180-184

[4] G. Barr, B. Kantor, A criterion for comparing and selecting batsmen in limited overs cricket. Journal of the Operational Research Society, 2004, 55, 1266-1274.

[5] D. Bhattachartjee, D. G. Pahinkar, Analysis of performance of bowlers using combined bowling rate. International Journal of Sports Science and Engineering, 2012, 6, 184-192.

[6] M. M. Islam, J. R. Khan, E. Raheem, Bradley-Terry model for assessing the Performance of ten ODI cricket teams adjusting for home ground effect. Journal of Data Science, 2017, 16, 657-668.

[7] Stylianos Kampakis, William Thomas,”Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches”,Cornell University, 2015

[8] K. A. A. D. Raj and P. Padma, ”Application of Association Rule Mining: A case study on team India”, 2013 International Conference on Computer Communication and Informatics, 2013

[9] S. Singh,P. Kaur, IPL Visualization and Prediction Using HBase. Procedia Computer Science, 2017, 122, 910-915.

[10] Tim B. Swartz, Paramjit S Gill and S. Muthukumarana,”Modelling and simulation for one-day cricket”, Canadian Journal of Statistics, 2009, Vol 37, No 2, pp-143-160

[11] H. H. Lemmer, The combined bowling rate as a measure of bowling performance in cricket. South African Journal for Research in Sport, Physical Education and Recreation, 2002, 24, 37-44.

[12] Veppur Sankaranarayanan, Vignesh and Sattar, Junaed and Lakshmanan,”Auto-play: A Data Mining Approach to ODI Cricket Simulation and Prediction”,SIAM Conference on Data Mining, 2014

[13] K. Passi, N. Pandey,Increased Prediction Accuracy in the Game of Cricket using Machine Learning. 2018. arXiv preprint arXiv:1804.04226.

[14] L. Pin, T. Kunhao, Z. Luo, C. Dunbao, WU Yuntao, "Single-pass Clustering K-nearest-neighbor Algorithm for Sentiment Classification", International Joumal of Advancements in Computing Technology(IJACT), vol. 5, no. 8, pp. 172-180, 2013.

[15] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade, "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016.

GITHUB: https://github.com/Amruthhs/Sports-Analysis