Cricket Analytics and Predictor

Mr. Suyash Mahajan
Student, Department of Information Technology,
Walchand Institute of Technology, Solapur, India

Ms. Salma Shaikh
Student, Department of Information Technology,
Walchand Institute of Technology, Solapur, India

Mr. Jash Vora
Student, Department of Information Technology,
Walchand Institute of Technology, Solapur, India

Ms. Gunjan Kandhari
Student, Department of Information Technology,
Walchand Institute of Technology, Solapur, India

Ms. Rutuja Pawar
Student, Department of Information Technology,
Walchand Institute of Technology, Solapur, India

Ms. A. R. Deshpande
Assistant Professor, Department of Information Technology.
Walchand Institute of Technology, Solapur, India.
anjali.deshpande10@gmail.com

Abstract—Cricket is one the most watched sport now-a-days. Winning in cricket depends on various factors like home ground advantage, performances in the past matches, experience of the players, performance at the specific venue, performance against the specific team and the current form of the team and the player. In the recent past, a lot of research has been done which measures the player's performance and predicts the winning percentage. This article briefs about the factors that cricket game depends on and discusses various researches which predicted the winning of a team with an advent of statistical modeling in sports. Cricket is one of the most popular team games in the world. With this article, we embark on predicting the outcome of Indian Premier League (IPL) cricket match using a supervised learning approach from a team composition perspective. Our work suggests that the relative team strength between the competing teams forms a distinctive feature for predicting the winner. Modeling the team strength boils down to modeling individual player's batting and bowling performances, forming the basis of our approach. We use statistics and recent performance of a player to model him. Player independent factors have also been considered in order to predict the outcome of a match. Machine learning is used in predicting the outcome of a cricket match before and during a match.

INTRODUCTION

Sports analytics play a major role in various problems associated with sport. Some of these problems are the ranking of individual players and their specialized skills, the composition of teams with an optimal balance of specialized skills, the ranking of teams, the negotiation of contracts, their potential revenue streams, the planning of both physical and mental training, the development of strategies for winning games and tournaments, assessing the effectiveness of coaches and referees, the medical aspects of sports injuries (health and insurance), the analysis and improvisation of rules, the quality of equipment and technology, the determination of awards, historical records and the generation of odds for gambling activities.

Related to above information, the coherent statistical presentation of both raw data and its inference to the decision makers is to facilitate successful planning and implementation. Furthermore, the media and the public have a great appetite for well visualized statistics. New opportunities for sports analytics have arisen due to the advent and availability of detailed and high quality data. For example, in Major League Baseball (MLB), the systems have provided comprehensive data on pitching and fielding. These systems record every play while also tracking the exact movements of all players on the field. Using these data sources, we can make very useful prediction, and various Statistics for improvement purposes.

Today's level of sports analytics has evolved where both the technology which provides data, and the statistical methodologies which provide the tools for analyzing data, improved very rapidly. Though sports analytics has been rapidly developing, it has not been the case with cricket. Due to historical reasons where cricket was perceived as a leisurely gentleman's game played without remuneration to players (until recently), cricket was not subject to large financial transactions.

This has changed in the last few years with the introduction of shorter formats of the game. The shortest and newest format, known as T20, generates intense interest and vast

sums of money, especially in the Indian subcontinent. The demand for cricket analytics has increased accordingly and the main website for cricket information and data is cricinfo.com. Cricket is a sport that originated in England in the 16th century and later spread to her colonies. The first international game however did not feature England but was played between Canada and the United States in 1844 at the grounds of the St George's Cricket Club in New York. In time, in both of these countries, cricket took a back seat to other, faster sports like ice-hockey, basketball, and baseball. International cricket is played today by a number of British Commonwealth countries; the main ones being Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. These teams are members of the International Cricket Council (ICC). A second rung of international teams called Associates includes numerous countries including Canada. Statistical modeling has been used in sports since decades and has contributed significantly to the success on field. Various natural factors affecting the game, enormous media coverage, and a huge betting market have given strong incentives to model the game from various perspectives. However, the complex rules governing the game, the ability of players and their performance on a given day, and various other natural parameters play an integral role in affecting the final outcome of a cricket match. This presents significant challenges in predicting the accurate results of a game. The game of cricket is played in three formats - Test Matches, ODIs and T20s. We focus our research on IPL. To predict the outcome of IPL cricket matches, we propose an approach where we first estimate the batting and bowling potentials of the 22 players using their career statistics and active participation in recent games.

I.   RELATEDWORK

[1]In this paper, a methodology for identifying promising batting orders in one-day cricket was presented. In particular, they suggested some batting orders that have never been tried by the Indian team and contradict prevailing wisdom. As a byproduct of investigation, a simulation procedure was developed for generating first innings runs against an average opponent. The simulation procedure was based on estimates from a Bayesian log-linear model. Finally, methods were developed with the intention of finding optimal or nearly optimal batting orders at the start of a team's innings.

[2]In this paper, two methodologies have been used. MySQL database is used for storing data whereas Java for the GUI. The algorithm used is Clustering Algorithm for prediction. The steps followed are as-
1. Begin with a decision on the value of k being the number of clusters.
2. Put any initial partition that classifies the data into k clusters.
3. Take every sample in the sequence; compute its distance from centroid of each of the clusters. If sample is not in the cluster with the closest centroid currently, switch this sample to that cluster and update the centroid of the cluster accepting the new sample and the cluster losing the sample.

4. Repeat above step until convergence is achieved, that is until a pass through the training sample causes no new assignments.

[3]In this paper, they have featured the various types of possibilities for big data analysis in various fields (Programming Language, Statistical Solutions, and Visualization Tools), also, endeavor to recognize which one of them is more prominent to use than others, and they discovered that R is a normal programming language to use for data scientists. SPSS is great as statistical apparatus for non-analysts clients, and Tableau Public is best suitable visualization instrument to introduce data and break down it in graphical path, yet for web visualization reason D3 will be the best decision.

[4]This paper presents the Usage of the Duckworth-Lewis technique to decide assets staying, toward the end of each finished over; the anticipated run aggregate of the batting group could be refreshed to give a more precise expectation of the match result. Finally, it was discovered that the triumphant probabilities were allocated to the contending groups in ODI matches. With the utilization of D-L approach, this procedure can be promptly adjusted to deliver 'in the run' forecasts.

[5]This paper introduces a model that has three segments which focuses on diverse contemplations developing out of a more profound examination of T20 cricket. The models are made utilizing Data Analytics strategies from machine learning area. In this work 5 highlights of IPL vocation and 5 highlights of International T20 Career have been thought about for both batsmen and bowlers yet in future work more highlights can be made and considered.

[6]In this paper they have clarified the instruments of different techniques utilized for resetting target scores that are interfered with one-day cricket matches. Each of these strategies yields a reasonable focus in a few circumstances. None has demonstrated palatable in inferring a reasonable focus under all conditions. We have introduced a strategy which gives a reasonable reconsidered target score under all conditions.
 A two-factor relationship has been determined which gives the normal numbers of runs which might be scored from any mix of these two assets and henceforth have inferred a table of extents of an innings for any such blend. This empowers the extent of the assets of the innings of which the batting groups are denied when overs are lost because of the stoppage in the play to be computed essentially and subsequently a reasonable revision to the objective score to be made. The parameters of relationship may change, for example, change in principles or conceivably changes in group choice and playing procedure.

[7]In this paper, they have discussed about the Duckworth-Lewis technique for target forecast in the session of cricket and clarified traps in the strategy. During analytics they utilized Correlation based subset assessment technique. As opposed to the conviction of the Duckworth/Lewis strategy, the setting of the diversion and strategic overs influences the expectation. In spite of the inadequate idea of the dataset,

relapse calculations and closest neighbor calculation gave the approximately correct results. Thus a half and half approach of utilizing quadratic relapse display with KNN as a smoothening capacity was utilized as an indicator and in addition the Duckworth/Lewis technique of having 1/1000 of the data was considered. At last the idea of expectation with energy of the amusement as a component was presented.

[8]This paper focuses on the execution of players as what number of runs will every batsman score and what number of wicket will every bowler take for both the groups. Both the issues are focused as grouping issues where number of runs and number of wickets are ordered in various reaches. The utilization of Naïve Bayes, random forecast, multiclass SVM and choice tree classifiers produce the expectation models for both the issues were made. Random Forest classifier was observed to be the most precise for both the issues. Four multiclass grouping calculations were utilized and thought about. Random Forest ended up being the most precise classifier for both the datasets with an exactness of 90.74% for foreseeing runs scored by a batsman and 92.25% for anticipating wickets taken by a bowler. Consequences of SVM were amazing as it accomplished an exactness of only 51.45% for foreseeing runs and 68.78% for anticipating wickets.

[9]The paper displays a data visualization and prediction device in which an open source, circulated, and non-social database, H-Base is used to keep the data identified with IPL (Indian Premier League) cricket matches and players. This data is then utilized for picturing the past execution of players' execution. Moreover, the data is utilized to anticipate the result of a match through different machine learning approaches. The proposed instrument can demonstrate the group administrations in the player barters for choosing the correct group. Finally, it was concluded that the novelty of the proposed approach lies in addressing the problem as a dynamic one and using a suitable non-relational database, H-Base for scalability of application. Out of all the machine learning algorithms used, KNN has been observed to be the most accurate.

[10]This Paper specifies the various factors that affect the game, winning in Cricket relies upon different variables like home group advantage, exhibitions before, involvement in the match, execution at the particular setting, execution against the particular group and the present type of the group and the player. Amid the previous couple of years part of work and research papers have been distributed which measure the player execution and their triumphant predictions. This article briefs about the variables that cricket diversion relies upon and focuses on couple of other research papers that anticipated the cricket wining.

[11]With statistical displaying in sports, foreseeing the result of an amusement has been built up as a central issue. Cricket is a standout amongst the most prevalent group recreations on the plane. It is observed that the relative group quality between the contending groups, frames an unmistakable component for foreseeing the victor. The utilization of profession insights and also the ongoing exhibitions of a player are shown. Player autonomous

variables have additionally been considered with a specific end goal to anticipate the result of a match. It was demonstrated that the K Nearest Neighbor (KNN) calculation yields better outcomes when contrasted with different classifier. That is, the paper tends to the issue of anticipating the result of an ODI cricket match utilizing the measurements of 366 matches. The oddity of this approach lies in tending to the issue as a dynamic one and utilizing in taking part of players as key component in anticipating the prediction of the match.

[12]In this paper, they focus on anticipating the best appropriate Team to be lined for a specific match. We propose statistical displaying way to deal with the ideal players for the match to be played. This work recommends that the relative group quality between the contending groups frames an unmistakable component for foreseeing the victory. Demonstrating the group quality comes down to displaying singular player batting and rocking the bowling alley exhibitions, framing the premise of approach utilized. Vocation insights and also the ongoing exhibitions of a player have been utilized to demonstrate. Player free factors have additionally been considered keeping in mind the end goal to foresee the result of a match. Exploratory investigation was performed utilizing Hadoop and Hive for Indian players. Results show up to 91% exactness when contrasted with the genuine outcomes accessible over web. Finally, Making strategies of order of the batting innings or the bowling order can be sorted with these scores.

[13]In this Research paper it is intended to distinguish the variables which assume a key part in anticipating the result of an ODI cricket match and furthermore decide the exactness of the prediction made utilizing the method of data mining. In this examination, statistical hugeness for different factors which could clarify the result of an ODI cricket match is investigated. Home field advantage, winning the hurl, approach (batting first or handling first), match write (day or day and night), contending group, setting commonality and season in which the match is played will be key highlights considered for the examination. For motivations behind model building, three calculations are focused: Logistic Regression, Support Vector Machine and Naïve Bayes. Logistic regression is connected to data as of now acquired from beforehand played matches to distinguish which includes independently or in a mix with different highlights assumed to be a part in the prediction. SVM and Naïve Bayes Classifier are utilized for display preparing and prescient examination. Graphical portrayal and Perplexity frameworks are used to examine the models. An offering situation is likewise considered to clarify the choices that can be taken after the model has been constructed. Impact of this choice on the cost and result of the model is additionally examined.

[14]The paper tends to the issue of foreseeing the after effect of an ODI cricket match using the bits of knowledge of 5000 matches. The interest of this approach lies in tending to the issue as a dynamic one, and using the consequences of the past matches as the key component in foreseeing the prediction of the match. It was observed that basic features can yield especially reassuring outcomes. Predicting the winner of the matches utilizing distinctive

administered calculations has been accomplished and now we can anticipate the upcoming matches. There may be some more calculations coming in future which give better outcomes at that point utilized as a part of this paper.

II. METHODOLOGY

The work of our project focuses on two models.
The two models are:
1. Descriptive model
2. Predictive model

DESCRIPTIVE MODEL:
The descriptive model focuses mainly on two aspects:
It describes the data and statistics of the previous information i.e. batting, balling or all-rounder.
It gives the past information of the matches played by the IPL teams.

PREDICTIVE MODEL:
The predictive model focuses on predicting the winning percentage of the team. The ranking of the players is displayed as well.
The user has the liability to choose the two teams playing against each other. The selection of the teams works on the criteria as:
1. If the players are batsmen then, sorting is done according to the strike rate of the batsmen.
2. If the players are bowlers then, sorting is done according to the average rate of the bowler.
3. If the players are all-rounder then, sorting is done considering both strike rate as well as average rate.
The algorithm used for this model is Decision Tree Classifier. A decision tree is built using top-down approach. In this algorithm the root node i.e. the prior factor considered is the 'city' where the match is being played. The tree is built according to the prominent factors (city, venue, teams, toss decision) considered in the match.
Decision Tree Classifier Diagram:
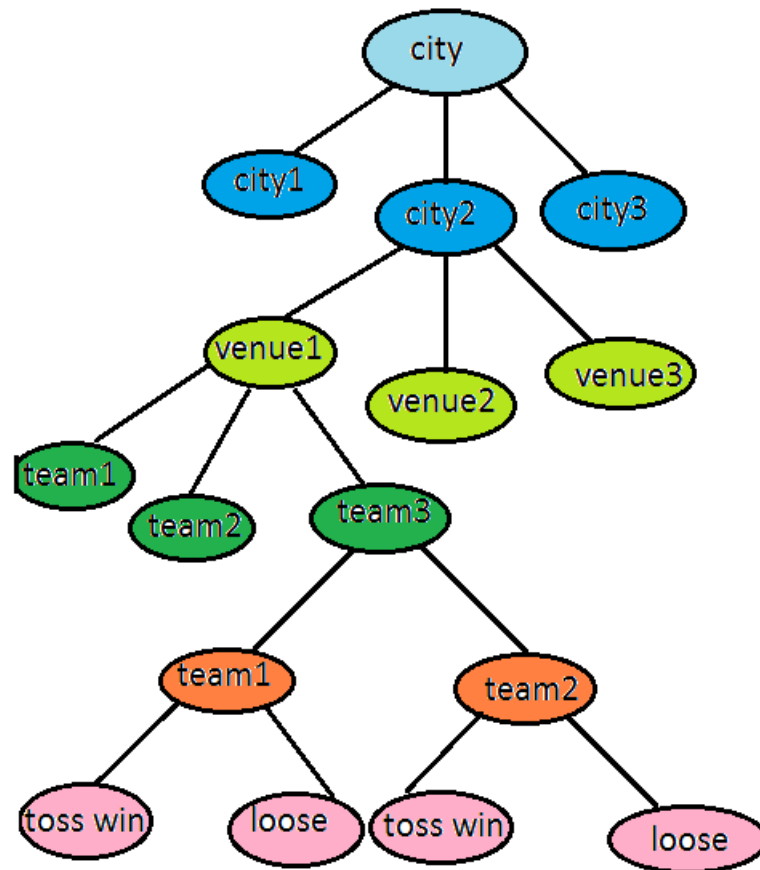
Algorithm:
1. Start
2. Select the root node as 'city'.
3. Choose one of the cities from 'city1','city2','city3', etc.
4. Select the venues among one of the venue ('venue1','venue2','venue3') present in the city.
5. A team is selected and compared against the other teams.
6. Toss decision is made and the result is predicated upon the win/loose criteria.
7. End

III. RESULT

The user has the option of sign up/login. After the successful sign up/ login procedure, the user has liability to access two models that is Descriptive model that shows the statistics of the player and the Predictive model that predicts the winning percentage of the team that the user has selected.
The user can also read the latest tweets and news on the website.

IV. CONCLUSION

The website developed is an authorized website. This website is beneficial for the coach as he can rank the

players on their priority from the previous data, it is beneficial to the owner to get the details of the IPL match played and the users who predict the winning percentage of the team and get the statistics of the player.

V. REFERENCES

[1] T. B. Swartz, P. S. Gill, D. Beaudoin, and B. M. Desilva, "Optimal batting orders in one-day cricket," Computers & Operations Research, vol.33, no. 7, pp. 1939–1950, 2006.

[2] Preeti Satao, "Cricket Score Prediction System (CSPS) Using Clustering Algorithm." Technical Research Organization India, Vol. 3, no. Issue 4, 2016, pp. 2394–0697., troindia.in/journal/ijcesr/vol3iss4/43-46.pdf.

[3] Tamanna Siddiqui, Mohammad Alkadri, Najeeb Ahmad Khan, "Review of Programming Languages and Tools for Big Data Analytics", International Journal of Advanced Research in Computer Science, vol.8,no.5,May-June 2017.

[4] "Predicting The Match Outcome in One Day - jssm.org." [Online]. Available: http://www.jssm.org/volume05/iss4/cap/jssm-05-480.pdf&p=DevEx.LB.1,5063.1. [Accessed: 11-Aug-2018].

[5] C. Deep, C. Patvardhan, and C. Vasantha, "Data Analytics based Deep Mayo Predictor for IPL-9," International Journal of Computer Applications, vol. 152, no. 6, pp. 6–11, 2016.

[6] F. C. Duckworth and A. J. Lewis, "A Fair Method for Resetting the Target in Interrupted One-Day Cricket Matches," Operational Research Applied to Sports, pp. 128–143, 2015

[7] Vijay Ramakrishnan, Sethuraman K, and Parameswaran R, "Target Score Prediction in the game of Cricket." [Online]. Available: https://people.ucsc.edu/~praman1/static/pub/ML_Project_CS7641_report.pdf. [Accessed: 11-Aug-2018]

[8] K. Passi and N. Pandey, "Increased Prediction Accuracy in the Game of Cricket Using Machine Learning," International Journal of Data Mining & Knowledge Management Process, vol. 8, no. 2, pp. 19–36, 2018.

[9] S. Singh and P. Kaur, "IPL Visualization and Prediction Using H-Base," Procedia Computer Science, vol. 122, pp. 910–915, 2017.

[10] "Analysis on Attributes Deciding Cricket Winning," Scribd. [Online]. Available: https://www.scribd.com/document/357690109/Analysis-on-Attributes-Deciding-CricketWinning. [Accessed: 11-Aug-2018].

[11] Madan Gopal Jhawar and Vikram Pudi, "European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases."

[12]S. Agarwal, L. Yadav, and S. Mehta, "Cricket Team Prediction with Hadoop: Statistical Modeling Approach," Procedia Computer Science, vol. 122, pp. 525–532, 2017.

[13] Mehvish Khan and Riddhi Shah, "Role of External Factors on Outcome of a One Day International Cricket (ODI) Match and Predictive Analysis" International Journal of Advanced Research in Computer and Communication Engineering, vol. Vol. 4, no. Issue 6, pp. 192–197, Jun. 2015.

[14]Geddam Jaishankar Harshit and Rajkumar S, "A Review Paper on Cricket Predictions Using Various Machine Learning Algorithms and Comparisons Among Them," International Journal for Research in Applied Science & Engineering Technology (IJRASET) , vol. Vol 45, no. 98, pp. 27–32.