IPL DATA ANALYSIS AND PREDICTION USING MACHINE LEARNING

Dr. B. Srinivasa Rao[1], G. Lakshman Teja[2], N. Anusree[3], M. Pavan Kumar[4]

[1]*Head of the Department of IT, Lakireddy Bali Reddy College of Engineering, India*
[2-4]*UG Students, Lakireddy Bali Reddy College of Engineering, India*

---***---

Abstract - *When we hear words like sports in India, most people remember cricket. Various formats such as Test, ODI and more recently Cricket T20 have been developed to make this sport more interesting. Recently, the Indian Premier League, known as the IPL, which is played between franchises in various Indian states, has become a popular league not only in India but all over the world. Day by day, the role of data science and machine learning in cricket is increasing due to the huge amount of data generated from a single player to a whole line. We use these available data and statistics to predict things like the team's first inning score and the probability of winning the second team, etc. The project begins with pushing the IPL match data played between 2008 and 2020 using Python modules such as Applications and Beautiful Soup, followed by pre-processing, data analysis and visualization, and finally creates a model that predicts the teams' overall score and probability of winning. When building models, we use machine learning algorithms such as Random Forest, Linear Regression, Logistic Regression and Support Vector Machine.*

***Key Words*: Analysis, IPL, Logistic Regression, Machine Learning, Prediction, Random Forest.**

1.INTRODUCTION

Machine learning is a subset of artificial intelligence, where real-world problems can be solved in the real world. This technique does not require programming but depends only on learning the data if the machine learns from previous data and predicts the result accordingly. Machine learning methods benefit from the use of decision trees, heuristic learning, knowledge acquisition, and mathematical models. Today, the demand for cricket has grown rapidly, with many people focusing on data analysis and data prediction through machine learning technologies. Analyzing and predicting IPL data through machine learning play an important role in player selection. The choice of players depends on various factors. The team board and coach decide the national team selection and the captain also has a bigger role in choosing the squad. Looking at the average scores of all the players on the team against the players of the opposing team. So, this project depends mainly on the success of the winning team of individual players based on the average of the details of previous matches. The team decides on the best batting and the best bowling performances, as well as the analysis of all the rounder performances. So, through these analyzes, fifteen players are selected for the national team. This study addresses prediction and analysis through machine learning algorithms such as linear regression and random forest. Thus, these algorithms predict the player's average score in an efficient way. The results of the analysis show that the prediction of the two-team model is accurate and precise.

2. EXSISTING SYSTEM:

In the existing system there is a formula to calculate the projected score and a win predictor based on the win percentage of a team and polls. These techniques won't give accurate results because they are based on perceptions and predictions based on a particular instant.

Consider calculating projected score, the formula.

Projected score = current run rate * overs in an innings.

The average accuracy obtained by following the above technique is very less.

3. PROPOSED SYSTEM

The proposed system uses Random Forest algorithm for Score Prediction of an innings and Logistic Regression for win prediction. The system scrapes data of all IPL matches from 2008 to 2020 from espncricinfo.com. And then we perform Data Cleaning followed by Data Preprocessing. Data Analysis and Visualization is done to make better understanding of the results and exploring various valuable insights. A model is deployed using one of the salesforce platforms, Heroku.

3.1 Predicting the first innings score:

We tried using multiple Linear regression and Random forest regression to predict the first innings score.X data has the features as said above and Y-Data has the run-rate at which the team has scored in the next overs. Finally we have chosen random forest regression since it gave a lesser RMS error of 1.7 where as Linear Regression gave 2.97.The model inputs all the data at that instance and outputs the final score by calculating from the predicted run-rate.

3.2 Winning probability of the second batting team:

We tried using different classifiers like logistic regression, naive bayes and SVM to predict the winner and find the winning probabilities. X-Data has one extra feature along with the previous features called the remaining runs which has the data of the runs required to reach the target at that instance and the Y-Data has binary values 0 and 1 which represents whether the team has won or not.We have majorly worked on logistic regression and finally chose it since it gave a lesser RMS error of 0.46 and also performed well on custom inputs.

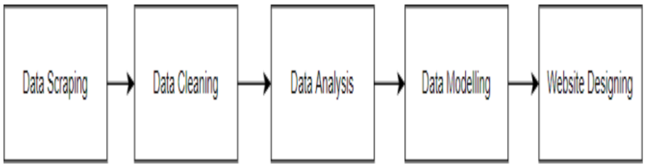


Fig 1 Block diagram of system

3.3 Data set

The dataset is created by scraping ball-by-ball data of all IPL matches from 2008-2020 through espncricinfo website. Every match has a commentary section which consists of detailed ball-by-ball data where that data is requested from a site in which all the data is present in the JSON format. The JSON data consists of ball-by-ball data which includes current score, current overs, details of current batsmen and current bowlers, etc.

3.4 Data Cleaning and Pre-processing:

We have eliminated redundancies from dataset like same team names with different spellings or teams whose names have changed.

3.5 Analysis and Visualization:

From the available data, we can analyse different fields of the game by obtaining the statistics from which we can interpret some useful and interesting results. These visualizations can be helpful for the teams and players to understand the areas of improvement and to plan new strategies against opponents.

3.5.1 Performance of teams in death overs:

We have analysed the performance of teams in death overs by comparing the average number of runs scored per over by teams vs the average number of runs given per over by that team in the death overs. We have used a grouped bar plot to visualize this data. Each teams' average runs scored per over vs average runs given per over are plotted side by side as a group to show a better visualization of all the teams' performances.

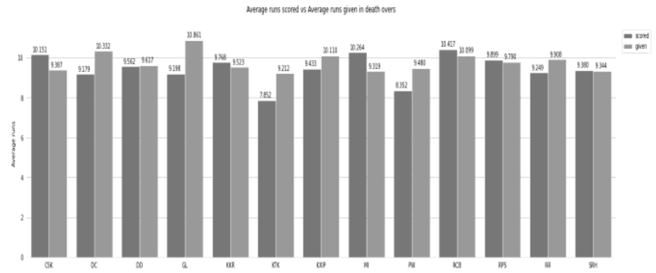


Fig 2 Comparision of all teams performance in death overs

Scoring runs at the death can be a game-changer because it can shift the momentum towards the batting team. From the above statistics we can observe that some of the successful teams like MI, CSK have comparatively more run rate at the death while batting.

3.5.2 Contribution of batsmen to the runs scored by the team in the year 2020:

For a team to perform well in the whole season, all the players need to contribute for the team. Only one or two players cannot win all the matches for the team. We have plotted pie charts for all teams showing the percentage of runs scored by each batsman for his team.

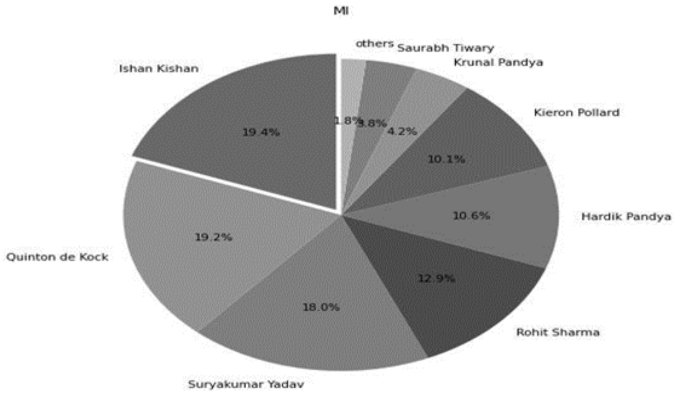


Fig 3 Pie chart of runs scored by MI batsmen in IPL 2020

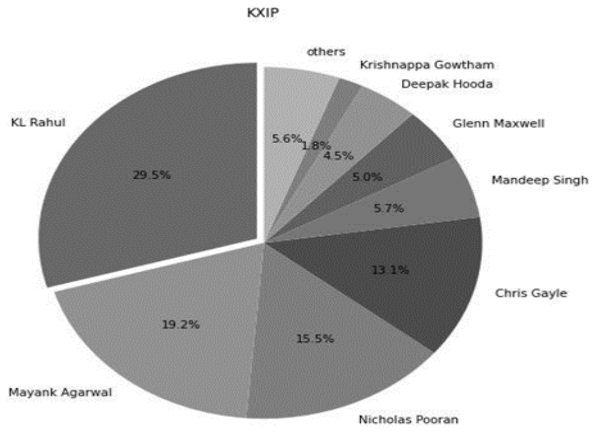


Fig 4 Pie chart of runs scored by KXIP batsmen in IPL 2020

In some teams like MI, six players have scored more than 10% of the team's runs, which means almost all the batsmen have contributed good enough for the team. Whereas in some teams like KXIP, only four players have scored more than 10% of the team's runs out of which only two players have scored around 50% of the runs. This shows the dependency of the team on only two of its batsmen which is a bad sign for any team.

3.5.3 Comparison of players' strike rate in different phases:

We have made a grouped bar plot which compares the strike rates of players at each phase of the game i.e., power play, 7-10 overs, 11-15 overs and death overs. We can observe how the players' strike rates change as the innings progresses.

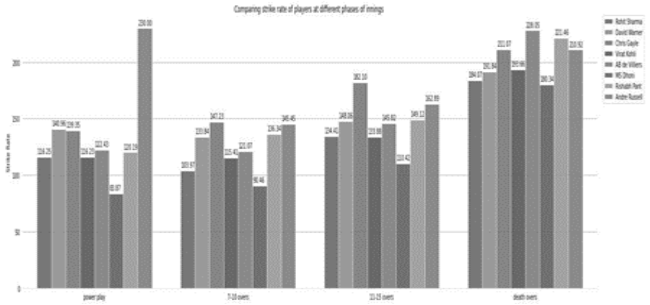


Fig 5 Strike rate comparision of players in different phases

4. EXPERIMENTAL RESULTS

The main objective of the project is to analyse the data to interpret the results and make some conclusions regarding the performances of teams and players, which we have done in the above sections. Also, we have developed a web app for our score predicting and winning predicting models using flask and deployed it on Heroku.

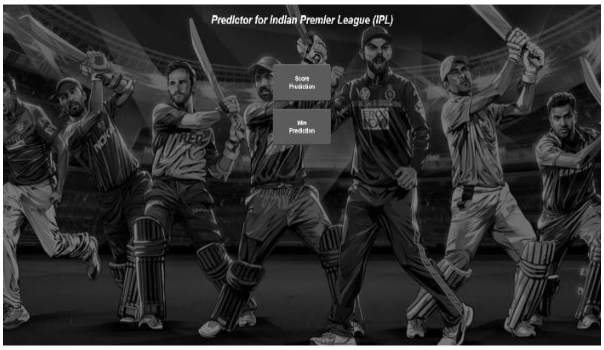
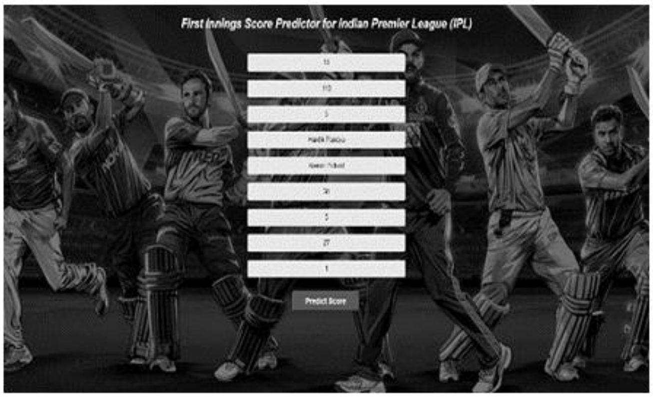


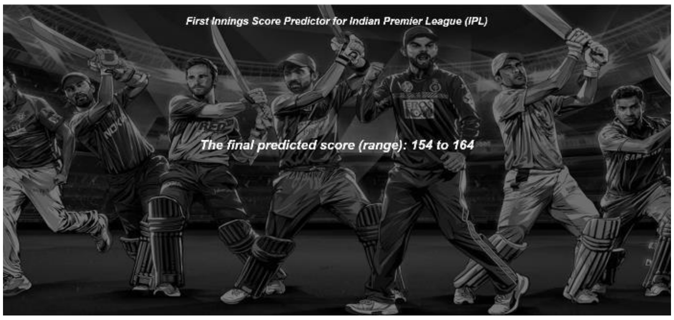
Fig 6 Interface of deployed model


First Innings Score Predictor for Indian Premier League (IPL)
Predict Score

15
110
5
Hardik Pandya
Kieron Pollard
26
5
27
1

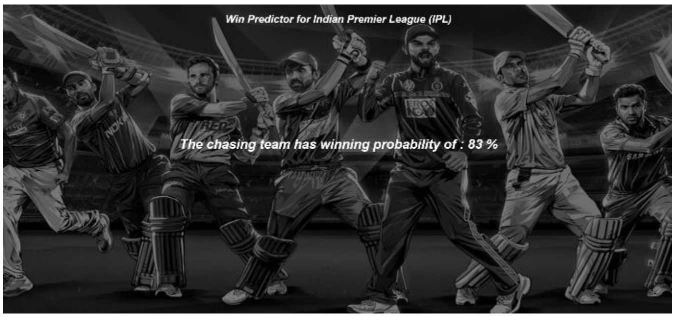
Fig 7 Interface of score predictor

We need to fill the data of the current situation of the match as shown in the above image. To predict the first innings score, we need to provide the current over, current score, current wickets, names of striker and non-striker, current scores of striker and non-striker, runs and wickets in the last 3 overs. The page displays the range of projected scores which the team is expected to score. To predict the winner, we need to provide the runs remaining along with the data mentioned av. The page displays the winning percentage of the chasing team in that situation.


First Innings Score Predictor for Indian Premier League (IPL)

The final predicted score (range): 154 to 164

Fig 8 Predictor score for an innings



The chasing team has winning probability of : 83 %

Figure 10 Wining Probability for a team

5. CONCLUSION

In the IPL we observe that only the current run-rate is used to predict the final score which is not an efficient way, since there are many other factors which can affect the projected score, we use some tools and libraries of python like 'Requests','Beautifulsoup' and analyze the conditions to predict the total score. With this project we can able to analyze and predict the score and winning probability of a team.

6. ACKNOWLEDEMENT

The authors gratefully acknowledge the support from corresponding guide and concerned faculty of IT department, Lakireddy Bali Reddy College of Engineering, Mylavaram.

7. FUTURE ENHANCEMENT

Much more analysis can be made if we could extract information like Nature of the pitch (hard, grassy, etc),Ball pitching (full length, short length, pitched outside off, etc), Speed of the delivery,Bowler type (off spinner, leg spinner, fast bowler, medium pacer, etc) and Whether the bowler and batsman are right handed or left handed

Models can be improved by considering the features like Batsmen who are yet to come, Bowlers in the opponent team , Performance of batsmen in that season (runs, average, strike-rate, etc], Performance of bowlers in that season (wickets, economy, etc)and Nature of the pitch.

8. REFERENCES

[1] K. Abbas and S. Haider, “Winner Prediction in Cricket Using Machine Learning” , 2017.

[2] Aminul Islam Anik et al and Amitabha Chakarbarty, “Player’s Performance Prediction in ODI Cricket Using Machine Learning Algorithms”, IEEE, 2018.

[3]Michal Sipko and William Knottenbelt, “Machine Learning for the Prediction of Professional Tennis Matches”, Imperial College London, 2015.

[4] P. Jhansi Rani and Aadith Menon, “Selection of Players and Team For An Indian Premier League Cricket Match Using Ensembles of Classifiers”, IEEE, 2020.

[5] Rohith Kade and Nikhil Bankar, “Cricket Score Prediction using Machine Learning Algorithms”, Grd Journal, 2020.

[6] Arjun Singhvi and Ashish Shenoy, “Prediction of the outcome of a Twenty-20 Cricket Match”, 2015.

[7] P.Jhansi Rani and D. Rishabh, “Prediction of Player Price in IPL Auction Using Machine Learning Regression Algorithms” , IEEE, 2020.

[8] R. Rajender and V. SivaRama Raju, “A Review of Data Analytic Schemes for Prediction of Vivid Aspects in International Cricket Matches”, IEEE, 2019.