PREDICTIVE ANALYSIS ON IPL - INDIAN PREMIER LEAGUE

**ABSTRACT: -**

Most of the people prefers the Twenty20 format of cricket since no one can guess who will win until the final ball of the match. The Indian Premier League (IPL) has been running since 2008 and is now Asia's most famous T20 league. We used ensemble technique models to predict matches for the outcomes, and we have compared them to other models such as the Random Forest Regressor and other models. Different python libraries and frame works are used to get the required prediction, as well as the tableau platform for creating various analysis and dashboards, and with the help of HTML and CSS languages we have developed user interface for a better user experience.

**INTRODUCTION**: -

Cricket is the most popular sport in Asia [2], a large number of players performs throughout the seas and oceans of different nations every year, and the World Cup is held once in every 4 years for all cricket-playing countries. Cricket is played in a variety of formats, including One Day International (ODI), Twenty20 International (T20I), and Test matches [1]. Aside from that, several matches are played at the state and district levels throughout the nation [11]. A suitable squad of playing 11 and four additional players should be selected to form a team for all the matches of cricket, including World Cup matches [4]. The cricket team consists of both batsmen and bowlers, as well as one wicket keeper who can bat or bowl [5]. In the group, the selectors and team commander must choose the batsmen and bowlers with a wicket keeper. Every batsmen in the team will bat in a different position in the playing eleven, and the bowlers in the team includes spin, pace, and medium pace bowlers [12].

The Indian Premier League (IPL) is a professional Twenty20 cricket league that takes place in India between March and May each year, with eight teams representing eight different cities or states [11]. The BCCI founded the IPL cricket league in 2007 [12]. It is the most popular cricket league in the world [2], ranking sixth in average attendance among all other sports tournaments in 2014, and it was the first sporting event in the world to be streaming live on YouTube in 2010. The IPL league's brand value was 475 billion (US$6.7 billion) in 2019, according to Phelps. The IPL season in 2015 added over $11.5 billion (US$160 million) to the Indian economy's GDP, according to the BCCI. The IPL tournament has a total of 13 seasons where Mumbai Indians are the current IPL champions, having the tournament cup in the 2020 season. Due to the COVID-19 pandemic, the venue for the 2020 season was changed to the United Arab Emirates [8].

The predictions and analysis are based on the viewers, who are unable to predict who will win a certain match in the IPL or to analyse the data from all previous matches of a particular team. As a result, we had established predictions for this issue. As shown below, there are various sections such as User Interface, Data Cleaning, Data Analysis, Predictions, and so on, where we will be forecasting matches, results, and also analysing past IPL data with various charts in each  segment.

**LITERATURE REVIEW: -**

‌In [1] the paper describes a data visualisation and prediction tool that uses HBase, an open source distributed and non-relational database, to store information about IPL cricket matches and players. This knowledge is then used to visualise the players' previous results. Furthermore, the data is used to predict the outcome of a match using various machine learning techniques. The proposed tool could be useful for team management in player auctions to choose the best team.

‌In [2] the data proposed by the author is then used to visualise the previous achievements of players. Additionally, using different machine learning methods, the data is used to predict the result of a match. The proposed tool could be useful for team managing in player auctions when choosing the correct team. To generate prediction models for the problem, the authors used SVM, Random Forest, and Decision Tree classifiers.

‌In [3] Sports analytics e-book tells the how availability of data that transformed the way sports are played, promoted and managed. This book explain how big data evolution is having major influence across the sport industry.

‌In [4] the authors used machine learning models like neural networks, decision trees, random forests and gradient boosting techniques machine accept a feature vector and provide a prediction. The experiment exhibits how much the AI model is fit for synthesizing the required component on its own.

‌In [5] the author proposes a multivariate regression-based solution for calculating points for each player in the league, as well as the total weight of a team based on the past performances of the players who have played the most for the team. Atlast, a dataset is created based on the seven factors that have been established which will be influencing the outcome of an IPL match.

In [6] IPL website provides the information of the players, matches and match schedules and every details that is going to happen.

‌In [7] the blog, author helps to classified how ensemble learning to improve machine learning results and why the ensemble method placed first in many prestigious events and competitions like Netflix competition, KDD2009, and Kaggle. author tells what is ensemble methods.

‌In [8] the history of the ipl and everything that happened in the ipl history is provided by the Wikipedia.

‌In [9] Kaggle is website where we get the datasets of the IPL which is used in the project.

‌In [10] ESPN cricket info used for the gaining the information about the players in every format.

‌In [11] the researcher used various data sets and previous stats which are trained in all dimensions, including toss, home ground, captains, favourite players, opposition battle, and previous stats, among other things and with the aid of the KNIME Tool and the added intelligence of the Naive Bayes network and Eulers power measurement method, each element has a different strength.

‌In [12] the author proposes a supervised learning technique that uses an SVM model with straight, nonlinear poly and RBF kernals to predict the outcome of a game against a specific side by gathering players at different levels in both groups requests for play. The order of groups that contributes to winning probability is determined by comparing different groups of players at the same stage.

‌In [13] the discriminant analysis classification functions were established using a training sample of 118 international Twenty20 matches, and their validity was checked using a holdout sample of another 73 matches. In the training and holdout samples, the cross-validation exercise yields a peak of 90.7 percent and 78.4 percent accurate classification, respectively.

**PROPOSED FRAMEWORK: -**

We built a machine learning model using various Python frameworks and libraries, as well as a user interface that allows users to get fast analysis and predictions using HTML and CSS-Bootstrap. And for deployment in backend, we have used python Django Framework. MYSQL was also used to bind the database. As a result, if a user wants to see anything on our website, he must login or register.

**User Interface: -**

The interface is made up of HTML and CSS. There will be various pages with various contents here:

Home Page – The Home Page includes the Navigation Bar, which has a variety of names and will redirect you to another page if you click on it. As shown in fig. 1 (a), the page has different images that change every 2 seconds with the help of a slider.

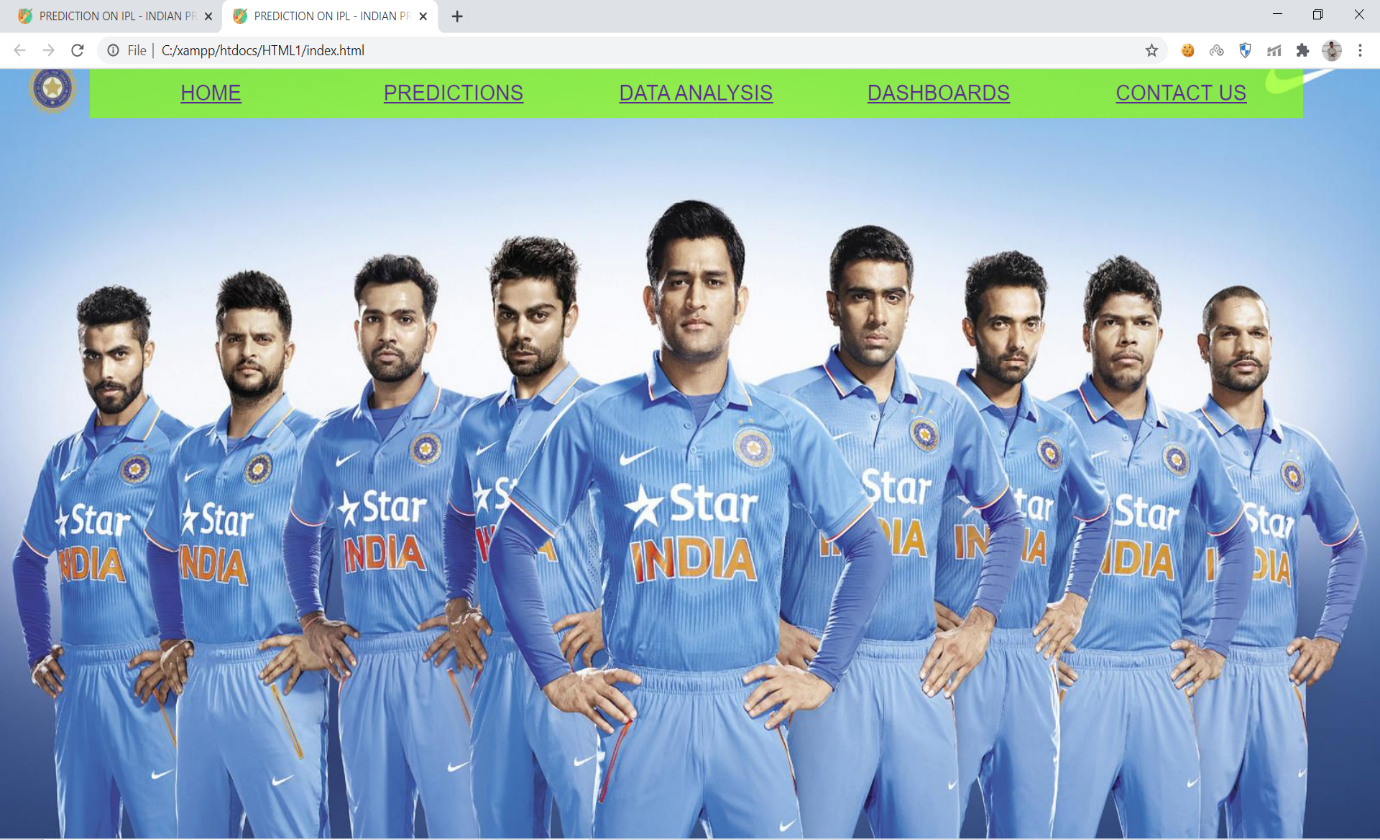


Fig. 1 (a) Home Page

Contact Us – Page As shown in Fig.1 (b), it contains information about all of the team members who worked and helped on the project.

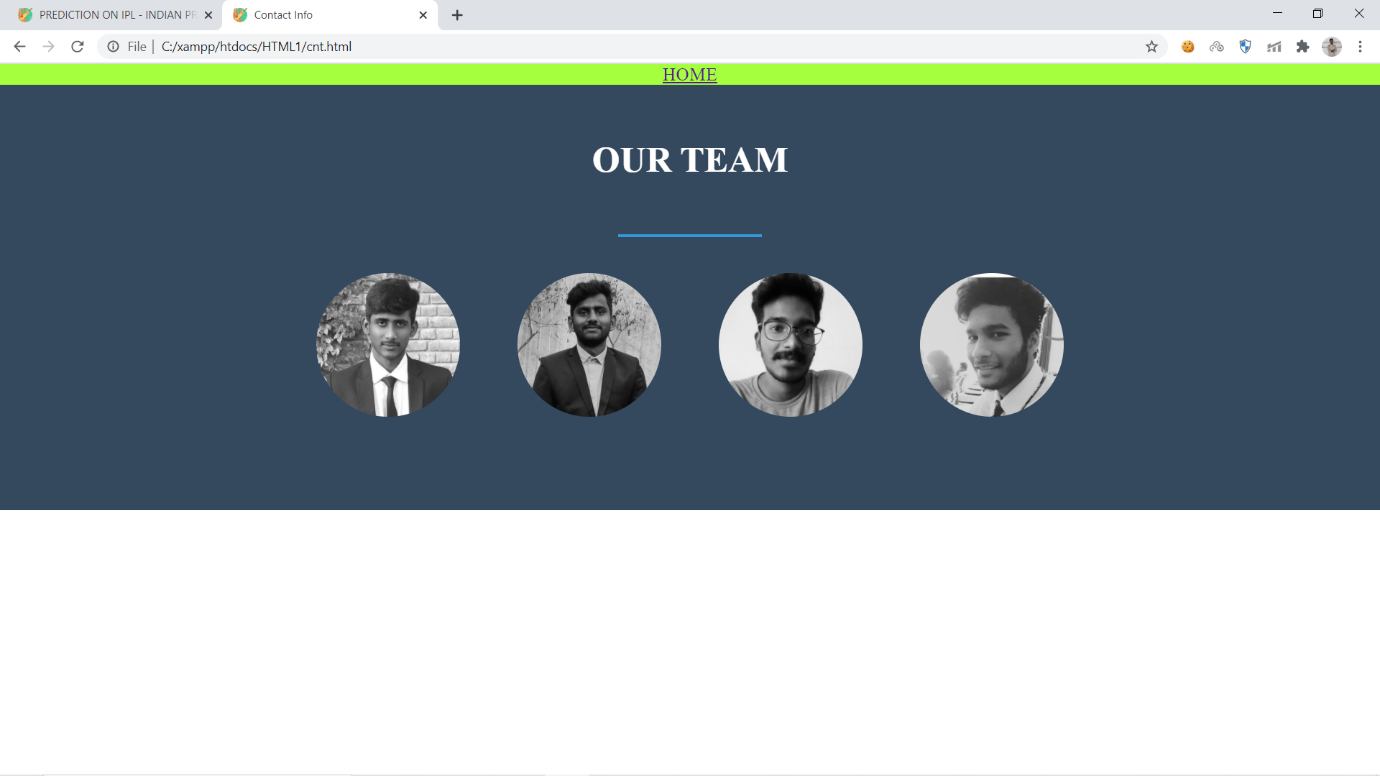


Fig. 1 (b) Contact Us Page

Page Feedback – Users may provide feedback on the interface via the feedback page, whether it is successful website or not.

Data Analysis Pages – There are 5 to 6 pages of data analysis here, each with a different analysis. The user can modify the data by year, and see the results as shown in Fig. 1 (c).



Fig. 1 (c) Teams and their Average runs per Match

Dash Board Pages – A dashboard is created by combining two or more analyses. The user may modify the dashboard to see all of the analyses that are available at once.

Predictions Pages – Users can give their own inputs as shown in Fig. 1 (d) within the website, where the output from the ML model will be reflected through the backend development. As a result, users will be able to get the necessary predictions based on their own inputs.

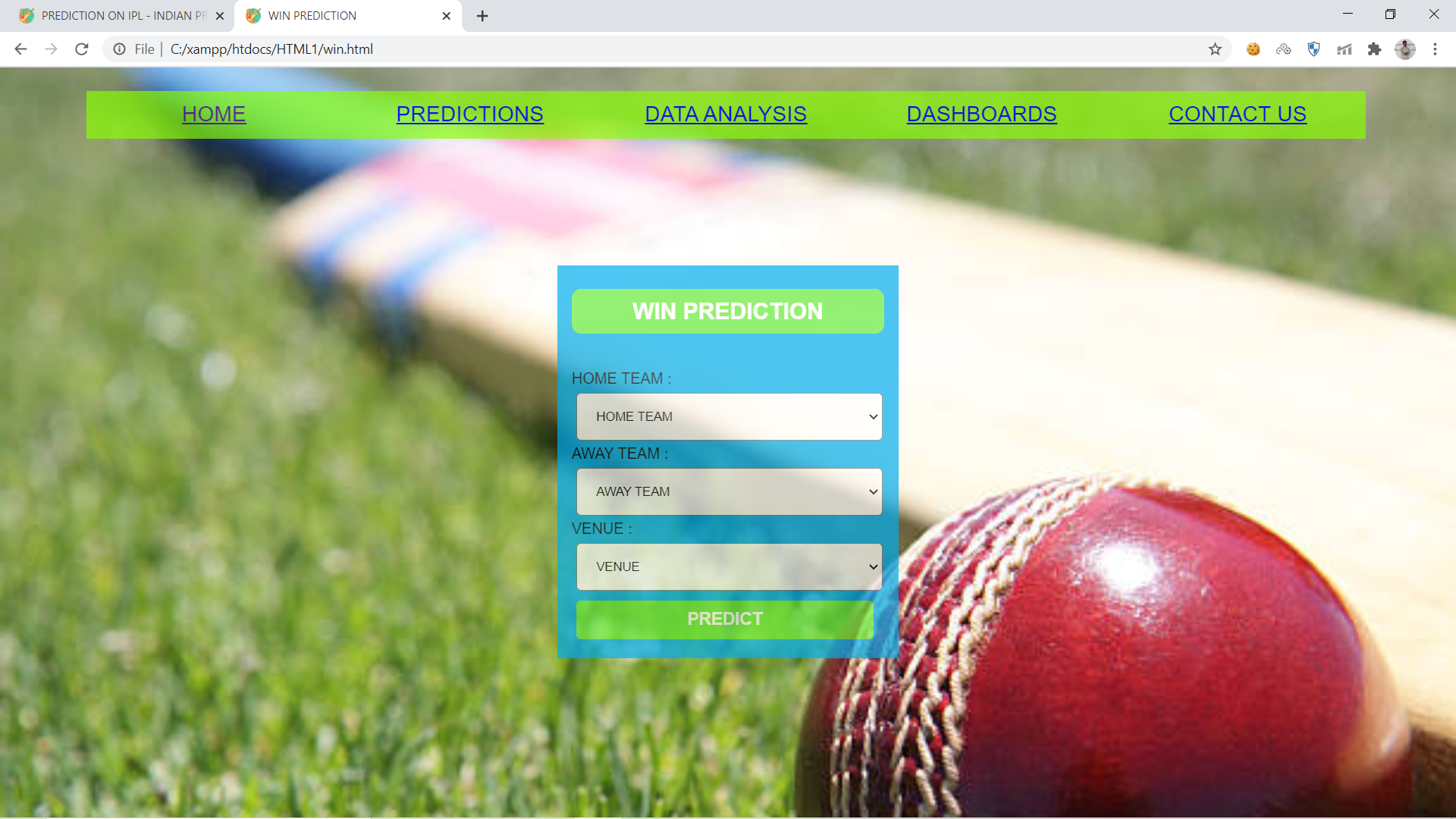


Fig. 1 (d) Win Prediction

**Data Collection**

The information was gathered from the Kaggle website. It has to do with all of the players and their results, as well as the matches that took place in the winning or losing positions. The data was cleaned using the tableau platform, and the current teams SunRisers Hyderabad and Delhi Capitals were combined with the Deccan Charges and Delhi Daredevils. Also, for score prediction, a specific set of data was gathered from Kaggle [9].

**Data Visualisation**

Tableau was used to perform data analysis and visualisations, and different views of analysis were performed, some of which are mentioned below.

**Analysis:**

Average runs per match, - As shown in fig. 2 (a), the visualisation was created using the runs scored by each team in each match and calculating an average runs per match for each team.

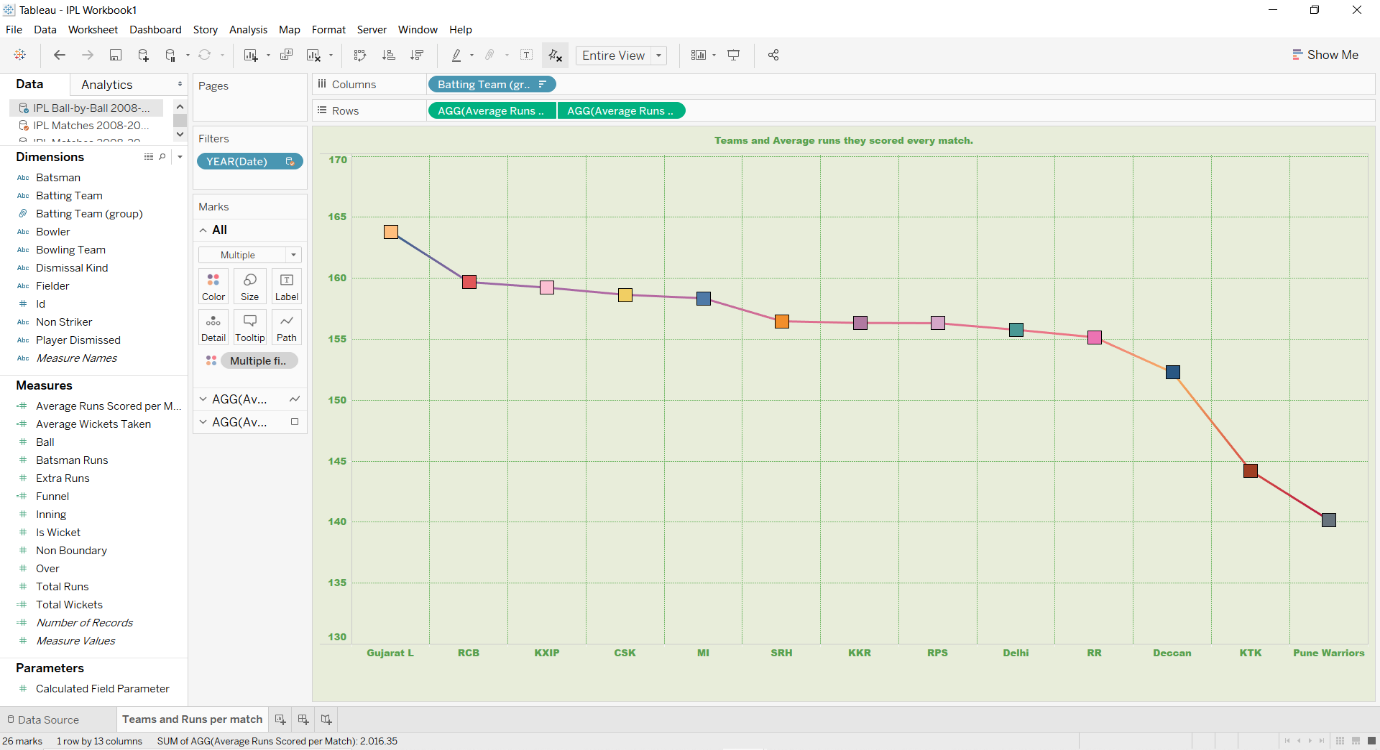


Fig. 2. (a) Teams and their Run per Match

Most Wickets in IPL/Player, - The most wickets taken per player in each IPL session, with the year which is added in the filters as shown in Fig. 2 (b).

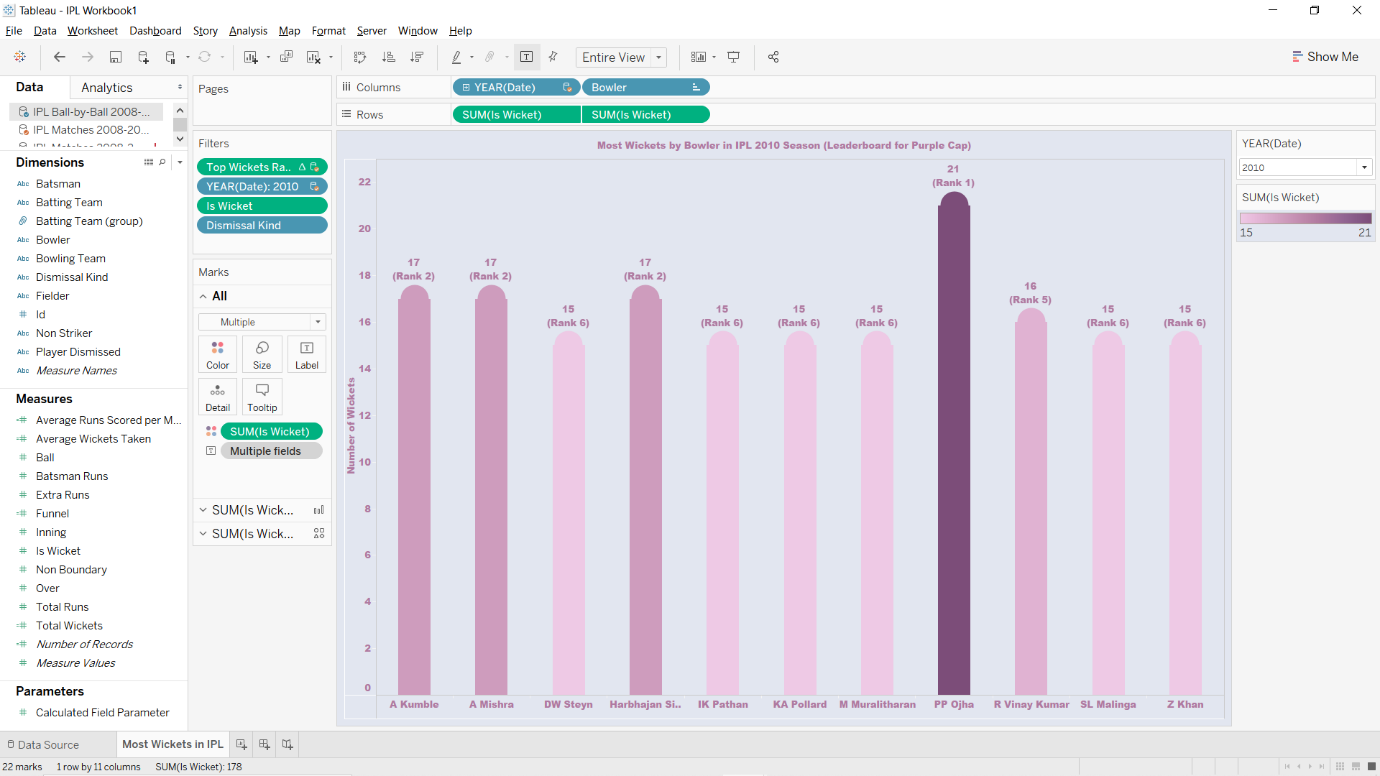


Fig. 2 (b) Most Wickets Taken in IPL

Status of Teams, - Every team’s winning and losing statistics have been analysed and shown in a table with total matches played, wins and losses (See Fig. 2 (c)).

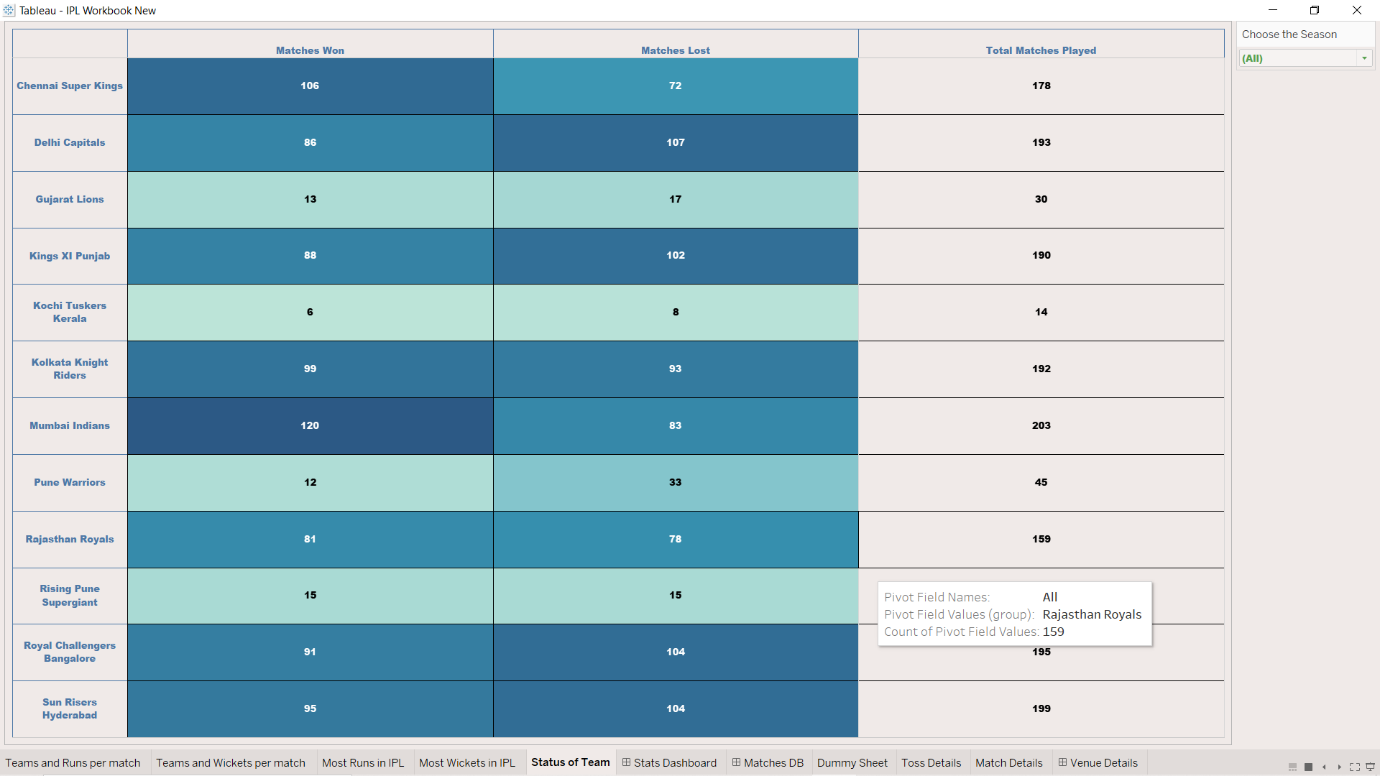


Fig. 2 (c) Status of Team

Most Runs in IPL/Player, - The most runs scored per player in each IPL session, with the year which is added in the filters as shown in Fig. 2 (d).



Fig. 2 (d) Most Runs Scored in IPL

**Dashboards of Data analysis**

Dash boards are made by merging two or more tableau sheets into one new layer.

Stats Dashboards, - As shown in Fig. 3 (a), this dashboard includes both orange cap and purple cap players, as well as a year in filter.

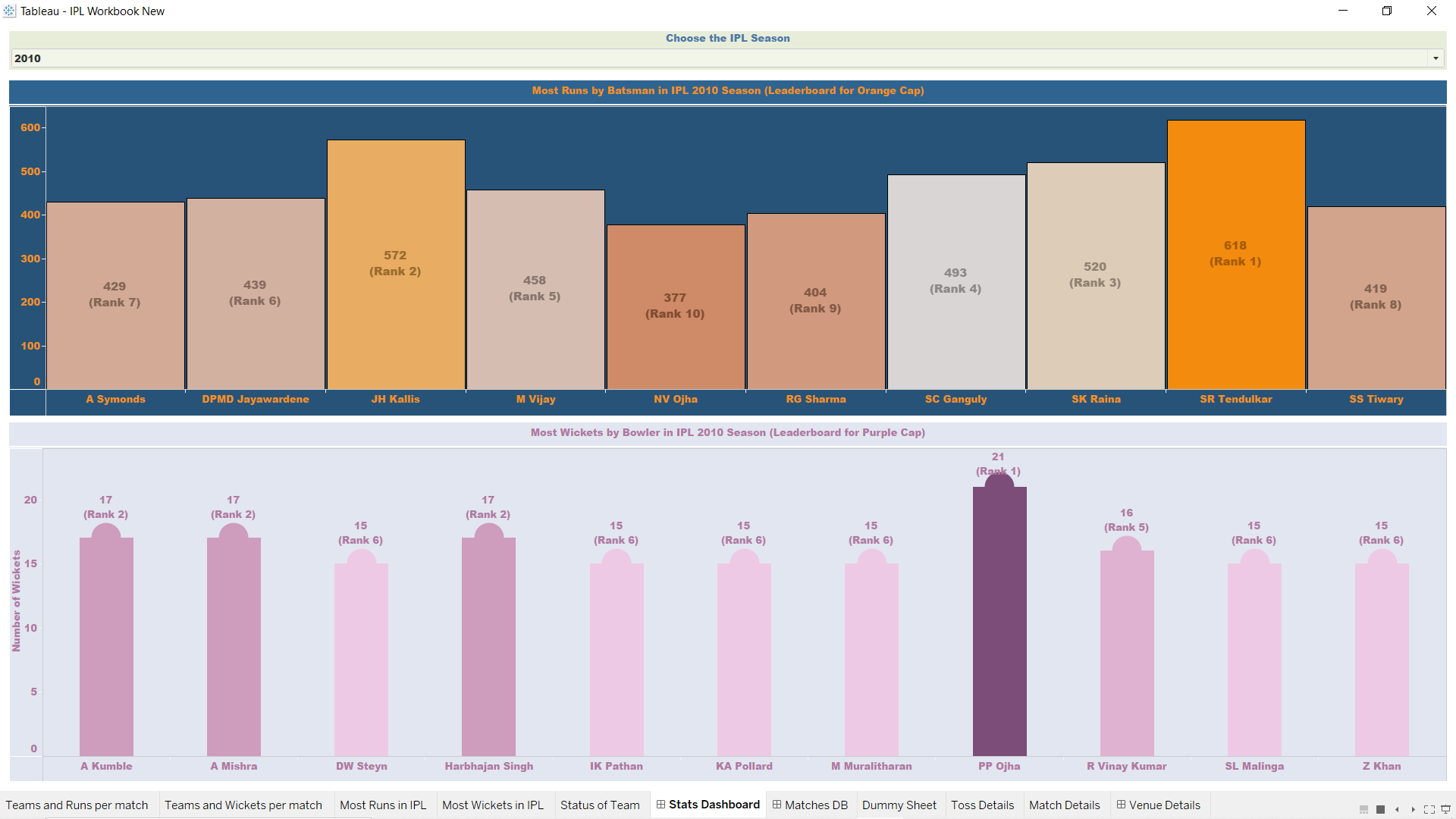


Fig. 3 (a) Stats Dashboard

There are a few other dash boards that are currently being created.

**PREDICTIONS: -**

Predictions are made using the Ensemble Techniques Algorithm, which has a higher prediction accuracy than other prediction algorithms.

**Ensemble Techniques Algorithm: -**

By integrating several models, it aids in improving machine learning predictions. As compared to a single model, the method produces the best predictive results. These are meta algorithms that combine multiple Machine Learning Algorithms or techniques into a single predictive Algorithm to reduce variance (bagging), bias (boosting), or increase prediction accuracy (stacking) [7].

These can be separated into two categories. They are: -

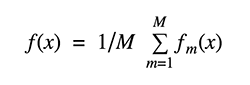
1. Sequential ensemble techniques – It is generated sequentially and used to take advantage of the base learner’s interdependence [7].

* AdaBoost Algorithm, for example.

1. Parallel ensemble techniques – It is generated in parallel and used to take advantage of the base learner’s independence [7].

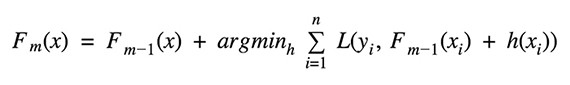
* Random Forest Algorithm, for example.

Bagging: - Bagging refers to bootstrap aggregation.



Boosting: - It refers to a group of algorithms that can transform weak learners into strong learners.





Stacking: It’s an ensemble technique that combines several classification or regression models using a meta classifier or a meta regressor. The meta model is trained as a feature on the outputs of the base level model, and the base level models are then trained on a complete training set [7].

Predictions are made for things like winning a match, predicting a score, and more. The batting team, bowling team, venue, current over, current score, wickets dropped, runs scored in previous 5 overs, and wickets taken in previous 5 overs, all these parameters are taken into score prediction.

**RESULTS: -**

The Ensemble Techniques Algorithm is used to make the win prediction, and the accuracy is 81 percent. Users provide inputs for all of the columns in Table 1.1 except Winner, which is expected as output. We predicted over 30 matches, where some of them are mentioned in below Table 1.

Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S. No | Home team | Away team | Venue | Batting Info | Predicted Winner |
| 1 | CSK | DC | Home\_DC | Home\_DC | CSK |
| 2 | SRH | KIXP | Home\_KXIP | Home\_KXIP | KXIP |
| 3 | RCB | RR | Home\_RR | Home\_RCB | RR |
| 4 | MI | KKR | Home\_MI | Home\_MI | KKR |
| 5 | KKR | KXIP | Home\_KKR | Home\_KXIP | KXIP |
| 6 | MI | CSK | Home\_CSK | Home\_MI | CSK |

Score Prediction is also performed by using Ensemble Techniques Algorithm, which has an accuracy of 80%, and it is compared to other models that have a lower accuracy as compared with Ensemble Techniques Algorithm. Users provide inputs for all of the columns in Table 1.2 except Predicted Score, which is expected as output. We predicted over 20 games, where some of them are shown in Table 2 below.

Table 2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Batting Team1 | Bowling Team2 | Venue | Current Over | Current Score | Wickets Fallen | Runs Scored Last 5 overs | Wickets fallen Last 5 overs | Predicted Score |
| CSK | KXIP | 1 | 14 | 170 | 2 | 45 | 0 | 218 |
| MI | RCB | 2 | 16 | 120 | 4 | 15 | 2 | 176 |
| SRH | DC | 1 | 14 | 180 | 1 | 55 | 1 | 230 |
| KKR | RR | 1 | 10 | 90 | 7 | 10 | 3 | 135 |

These are our project's predictions, a few of them are not being stated in the report.

**CONCLUSION: -**

Our work can be expanded further by extending our dataset to include not only games directed in the IPL, but also coordinates from other well-known Cricketing events such as the Big Bash League and matches from International Cricket. We can predict that the algorithms we used in our models will provide accurate results and that user-website interaction will be faster, such as data analysis and forecasts, which take longer in Tableau and Anaconda Jupiter to load but takes less time in web using backend and frontend.

**REFERENCES: -**

[1] Shubhra Singh and Parmeet Kaur, Using HBase IPL Prediction and Visualization, 2017.

<https://www.sciencedirect.com/science/article/pii/S1877050917327023>

[2] Dr. GR Anantharaman, Ganesh Kumar Macha, and Kasukurti Raviteja, Predicting and Analyzing the Performance of the Indian Premier League Cricket Using Regression Models, 2019.

<http://cij.org.in/pdf/CIJ-23-03-0033.pdf>

[3] Ceyda Mumcu, Gil Fried, Sport analytics is an information approach to sports marketing and management, 2016.

<https://doi.org/10.4324/9781315619088>.

[4] J. Heaton, An evaluation into the use of feature engineering in predictive modelling, 2016.

<https://doi.org/10.1109/SECON.2016.7506650>.

[5] Lamsal R, Choudhary A, To Predict the Future Outcomes of Indian Premier League (IPL) Matches Using ML models, 2018.

<https://arxiv.org/abs/1809.09813>

[6] IPL player to player performances and so on...

Available from: <https://www.iplt20.com>

[7] Ensemble Techniques Algorithm plays an important role in Machine Learning,

Available from: <https://blog.statsbot.co/ensemble-learning-d1dcd548e936>

[8] Indian Premier League history and stats from Wikipedia,

Available form: <https://en.wikipedia.org/wiki/Indian_Premier_League>

[9] Kaggle for Datasets...

Available from: <https://www.kaggle.com/manasgarg/ipl>

[10] ESPN cricinfo is a sports news website dedicated solely to the game of cricket...

Available from: [https:/stats.espncricinfo.com/](https://stats.espncricinfo.com/)

[11] Anurag Sinha, IPL 2020 Predictive Analytics and Machine Learning Modeling, 2020.

<https://www.preprints.org/manuscript/202010.0436>

[12] Sandesh Bananki, Anthony Akas, Abhilasha Gududuru, Jayanth, Shaik Noorni, and Srinivasa Gowri, A squad recommendation process and outcome prediction for the game of cricket, 2018.

<https://content.iospress.com/articles/journal-of-sports-analytics/jsa196>

[13] Priyanka Talukdar and Dibyojyoti Bhattacharjee, Using the pressure index to forecast match results: facts from 2020 cricket, Communications in Statistics - Simulation and Computation, 2019.

<https://doi.org/10.1080/03610918.2018.1532003>.