iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55793

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Analyzing Key Factors for Second Innings Triumph in IPL Using Machine Learning

Sohon C[1], Yudhajit S[2], Dhrubajyoti G[3], Anita P[4]
[1, 2, 3]*Department of Computer Science & Engineering, OmDayal Group of Institutions, West Bengal*
[4]*Department of Mathematics, National Institute of Technology, Durgapur*

Abstract: ***The Indian Premier League (IPL) has emerged as one of the most captivating and celebrated cricket tournaments worldwide. With its unique blend of sporting excellence, entertainment, and fierce competition, the IPL has captured the imagination of millions of cricket enthusiasts. In this era of data-driven decision-making, harnessing the power of statistical analysis and predictive modelling can provide valuable insights into the dynamics and outcomes of IPL matches. This research paper aims to explore on the analysis of factors contributing to a team's success in the second innings of IPL matches using exploratory data analysis (EDA) and logistic regression modelling. The EDA section will uncover patterns and correlations among variables such as team composition, batting order, run rate, power-play performance, bowling strategies, fielding efficiency, and match conditions. Logistic regression modelling will be applied to develop a predictive model that forecasts the likelihood of a team's victory in the second innings based on the identified factors. The model will be trained, validated, and evaluated using historical data.***

Keywords: Indian Premier League, Exploratory Data Analysis (EDA), Statistical Analysis, Predictive Modelling, Logistic Regression, Victory Prediction

I. INTRODUCTION

The Indian Premier League (IPL) has emerged as one of the most captivating and celebrated cricket tournaments worldwide. With its unique blend of sporting excellence, entertainment, and fierce competition, the IPL has captured the imagination of millions of cricket enthusiasts. In this era of data-driven decision-making, harnessing the power of statistical analysis and predictive modelling can provide valuable insights into the dynamics and outcomes of IPL matches. The paper titled " IPL Predicting IPL Second Innings Score Using Machine Learning Algorithm" aims to delve deep into the vast IPL dataset, applying rigorous statistical techniques and machine learning algorithms to uncover critical patterns and factors influencing match outcomes. By leveraging data preparation, exploratory data analysis, and advanced predictive modelling, this paper seeks to unravel the key drivers behind successful performances, team strategies, and the impact of various factors on match results. Through a comprehensive exploration of the dataset, we gain insights into the most successful IPL teams, the significance of winning the toss, the preference for chasing or defending targets, and the top performers in the league. Armed with these insights, teams, players, and fans can make informed decisions, devise effective strategies, and enhance their understanding of the game. Moreover, this paper goes beyond mere analysis and delves into the realm of predictive modelling. By developing a machine learning model that predicts the winning percentage in the second innings, we aim to offer accurate forecasts and aid decision-making during live matches. This predictive model, coupled with an intuitive web application, brings the power of data-driven predictions to stakeholders, making it accessible and actionable. Ultimately, the paper's findings and methodologies contribute to the growing field of sports analytics, showcasing the immense potential of statistical analysis and predictive modelling in unlocking valuable insights. By shedding light on the intricate dynamics of IPL matches, this paper offers a comprehensive understanding of the game, facilitating informed decision-making and strategic planning for all stakeholders involved.

This paper organised in five sections. Section one contains introduction. Section two describes the literature survey. Section three represents the basic definition of machine learning algorithm. Section four gives the idea of a predictive model that forecasts the likelihood of a team's victory in the second innings of some discussed problem. Finally, Section five draws conclusions based on our study.

II. LITERATURE REVIEW

In a study conducted by Ahmad et al. [1], machine learning techniques were employed to predict emerging players from both batsmen and bowlers. Another research by Song et al.

©IJRASET: All Rights are Reserved

[2] focused on estimating the location of a moving ball using data from a cricket sensor network. Roy et al. [3] developed a ranking system that incorporated social network factors and utilized a distributed framework based on Hadoop and the MapReduce programming model for data processing. Their approach involved evaluating the composite framework. Priyanka et al. [4] predicted the outcome of IPL-2020 by employing Data Mining Algorithms on IPL datasets from 2008-2019, achieving an accuracy of 82.73%. Kansal et al. [5] utilized Data Mining Techniques to predict player evaluation in IPL based on datasets from 2008-2019. They employed data mining algorithms to assess player performance, determine their base price, and aid in player selection for the IPL. Decision tree, Naïve Bayes, and Multilayer Perceptron (MLP) algorithms were used, with MLP demonstrating superior performance compared to other algorithms. Agrawal et al. [6] employed Support Vector Machine (SVM), CTree, and Naïve Bayes classifiers to predict the probability of match winners, achieving accuracies of 95.96%, 97.97%, and 98.98%, respectively. Barot et al. [7] predicted match outcomes based on factors such as the toss and venue. Kaluarachchi et al. [8] predicted match outcomes using the Naïve Bayes classifier, considering home ground, match time, match type, winning the toss, and batting first. Passi et al. [9] addressed two classification problems: predicting player performance based on runs and the number of wickets. They employed machine learning algorithms to classify the runs and wickets into different ranges, with the Random Forest algorithm outperforming other algorithms. Nigel Rodrigues et al. [10] predicted the traits of batsmen and bowlers in current matches, assisting in player selection for upcoming matches by leveraging past performance data and employing Multiple Random Forest Regression. Wright [11] predicted possible cricket match fixtures by considering various factors such as venue, teams, and the number of holidays between matches. They employed a metaheuristic procedure called Subcost-Guided Simulated Annealing (SGSA) to progress from initial to final solutions. Maduranga et al. [12] predicted match outcomes using data mining algorithms and offered solutions based on the approaches used by other authors. Shetty et al. [13] predicted player capabilities based on factors such as the ground, pitch type, and opposition team, utilizing machine learning techniques. The above discussed research focused on evaluating the performance and accuracy of these models in predicting match outcomes. The findings of this study are particularly relevant to this paper as they provide insights into the effectiveness of machine learning techniques in the context of predicting IPL match outcomes.

III. PRELIMINARIES

A. Data Visualisation

Data visualization is a crucial tool used in this paper to simplify and present data effectively. It transforms complex information into clear and visually appealing representations, aiding comprehension and data-driven decisions. Visualizing data through charts, graphs, and interactive dashboards simplifies complex data, facilitating insights. Visual representation helps identify patterns, outliers, and correlations, vital for informed decision-making. In this paper, data visualization conveys key findings and insights, with interactive dashboards enabling user-friendly exploration. It facilitates communication to both technical and non-technical audiences, presenting information concisely. Visualizations highlight trends, comparisons, and outliers, telling a compelling data story. They also help detect data anomalies and errors, ensuring data quality. Overall, data visualization is a vital component, enhancing data understanding, communication, and decision-making, crucial for the paper's success.

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital step in data analysis, involving data examination, visualization, and preprocessing. It begins with dataset structure examination: rows, columns, data types, and missing values. Descriptive statistics, like count, mean, and standard deviation, provide data central tendencies and dispersion. Visualizations like histograms identify numerical variable distributions and outliers, while scatter plots reveal relationships between numerical variables. Bar and pie charts explore categorical variable distributions. EDA includes data cleaning, handling missing values, and treating outliers. Correlation analysis calculates correlations between numerical variables, helping select relevant variables. EDA offers a comprehensive dataset understanding, identifies issues, and reveals patterns, guiding subsequent steps for accurate analysis and decision-making. It ensures analysis reliability and quality.

C. Logistic Regression

Logistic Regression is a popular statistical modelling technique used for predicting binary or categorical outcomes. In this paper, Logistic Regression was employed to analyze the relationship between the independent variables and the binary target variable, and to make predictions based on the trained model.

Logistic Regression begins by formulating a hypothesis about the relationship between the independent variables and the target variable.

The dataset used in this paper consisted of various features, and the objective was to determine how these features influence the likelihood of a particular outcome. The Logistic Regression model was built by estimating the coefficients for each independent variable. These coefficients represent the impact of each variable on the probability of the binary outcome. The model uses a logistic function, also known as the sigmoid function, to convert the linear combination of the independent variables and their coefficients into a probability score between 0and 1.

To evaluate the performance of the Logistic Regression model, various metrics were used. Accuracy, precision, recall, and F1-score were computed to assess the model's ability to correctly predict the positive and negative classes. Additionally, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were used to evaluate the model's discriminatory power and determine an appropriate threshold for classification. The dataset was split into training and testing sets to validate the performance of the Logistic Regression model. The model was trained on the training set, and then the accuracy and other performance metrics were computed on the testing set to assess the model's generalization ability. Feature selection techniques were applied to identify the most influential variables for the Logistic Regression model. This involved analyzing the statistical significance of each variable and considering domain knowledge to select the relevant features.

IV. PROPOSED SYSTEM METHODOLOGY

Design is a meaningful technical representation of something to be built. It is the most important phase in the development of a system. Software design is a process by which the requirements are translated into a representation of the software. Design is a place where design in software development is encouraged. Based on the user requirements and the detailed analysis of the existing system, the new system needs to be designed. This is the system design phase. Design creates a representation or model and provides details of the software data structure, architecture, interfaces, and components required to implement a system. The logical system design obtained as a result of the system analysis is converted into the physical system design. The figure below explains the approach we took in building the predictive model using machine learning algorithms. This system architecture involves five significant steps:

1) Data Acquisition & Pre-processing
2) Training & Testing
3) Exploratory Data Analysis (EDA)
4) Model selection learning algorithm
5) Final model evaluation
6) Fit the model to data and predict the results

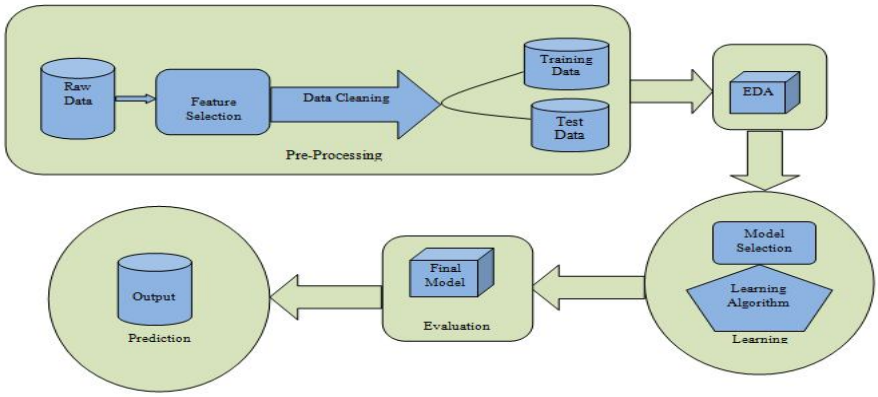


Fig. 1 Proposed System Architecture

A. *Data Collection*

Data collection is the gathering and measurement of information from a myriad of different sources. To use the data, we collect it to develop practical machine learning solutions. By collecting data, you can create a record of past events so that we can use data analysis to find patterns that are returning. From these patterns, you build predictive models using machine learning algorithms that look for trends and predict future changes. To achieve our objectives, we utilize a comprehensive dataset sourced from Kaggle, which provides rich information about IPL matches, teams, players, venues, and umpires.

We perform data pre-processing, exploratory data analysis, and develop predictive models using appropriate statistical techniques and machine learning algorithms. Additionally, we develop a user-friendly web application that enables users to interactively explore the analysis results, visualize insights, and access match predictions.

```
In [366…
match = pd.read_csv('./Dataset/matches.csv')
delivery = pd.read_csv('./Dataset/deliveries.csv')

In [367…
match.head()
```

Out[367…

	id	Season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_runs	wi
0	1	IPL-2017	Hyderabad	05-04-2017	Sunrisers Hyderabad	Royal Challengers Bangalore	Royal Challengers Bangalore	field	normal	0	Sunrisers Hyderabad	35	
1	2	IPL-2017	Pune	06-04-2017	Mumbai Indians	Rising Pune Supergiant	Rising Pune Supergiant	field	normal	0	Rising Pune Supergiant	0	
2	3	IPL-2017	Rajkot	07-04-2017	Gujarat Lions	Kolkata Knight Riders	Kolkata Knight Riders	field	normal	0	Kolkata Knight Riders	0	
3	4	IPL-2017	Indore	08-04-2017	Rising Pune Supergiant	Kings XI Punjab	Kings XI Punjab	field	normal	0	Kings XI Punjab	0	
4	5	IPL-2017	Bangalore	08-04-2017	Royal Challengers Bangalore	Delhi Daredevils	Royal Challengers Bangalore	bat	normal	0	Royal Challengers Bangalore	15	

Fig. 2 Data collection of matches

```
In [369…
delivery.head(6)
```

Out[369…

	match_id	inning	batting_team	bowling_team	over	ball	batsman	non_striker	bowler	is_super_over	…	bye_runs	legbye_runs
0	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	1	DA Warner	S Dhawan	TS Mills	0	…	0	0
1	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	2	DA Warner	S Dhawan	TS Mills	0	…	0	0
2	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	3	DA Warner	S Dhawan	TS Mills	0	…	0	0
3	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	4	DA Warner	S Dhawan	TS Mills	0	…	0	0
4	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	5	DA Warner	S Dhawan	TS Mills	0	…	0	0
5	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	6	S Dhawan	DA Warner	TS Mills	0	…	0	0

6 rows × 21 columns

Fig. 3 Data collection of Bowl Delivery

B. Data Pre-Processing

1) 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 performed accurately. So we tried replacing the null values in different columns with dummy values.

2) Choosing Required Attributes: This step is the main part where we can remove some columns of the dataset that are not useful for estimating the winning team. This is estimated based on feature importance. The attributes considered have the following feature meaning:

id: The IPL match id.

season: The IPL season

city: The city where the IPL match was held. date: The date on which the match was held. team1: One of the teams of the IPL match team2: The other team of the IPL match toss_winner: The team that won the toss

toss_decision: The decision taken by the team that won the toss to ‘bat’ or ‘field’

result: The result(‘normal’, ‘tie’, ‘no result’) of the match.

dl_applied: (1 or 0)indicates whether the Duckworth-Lewis rule was applied or not.

winner: The winner of the match.

win_by_runs: Provides the runs by which the team batting first won

win_by_runs: Provides the number of wickets by which the team batting second won.

player_of_match: The outstanding player of the match.

venue: The venue where the match was hosted.

Umpier1& Umpire2 : on-field umpires who officiate the match.

Umpire3: The off-field umpire who officiates the match.

```
match.columns
✓ 0.0s
Index(['id', 'Season', 'city', 'date', 'team1', 'team2', 'toss_winner',
       'toss_decision', 'result', 'dl_applied', 'winner', 'win_by_runs',
       'win_by_wickets', 'player_of_match', 'venue', 'umpire1', 'umpire2',
       'umpire3'],
      dtype='object')
```

Fig. 4 Dataset Columns Overview

```
print(match['team1'].unique())
print(len(match['team1'].unique()))

print(match['team2'].unique())
print(len(match['team2'].unique()))
✓ 0.0s
['Sunrisers Hyderabad' 'Mumbai Indians' 'Gujarat Lions'
 'Rising Pune Supergiant' 'Royal Challengers Bangalore'
 'Kolkata Knight Riders' 'Delhi Daredevils' 'Kings XI Punjab'
 'Chennai Super Kings' 'Rajasthan Royals' 'Deccan Chargers'
 'Kochi Tuskers Kerala' 'Pune Warriors' 'Rising Pune Supergiants'
 'Delhi Capitals']
15
['Royal Challengers Bangalore' 'Rising Pune Supergiant'
 'Kolkata Knight Riders' 'Kings XI Punjab' 'Delhi Daredevils'
 'Sunrisers Hyderabad' 'Mumbai Indians' 'Gujarat Lions' 'Rajasthan Royals'
 'Chennai Super Kings' 'Deccan Chargers' 'Pune Warriors'
 'Kochi Tuskers Kerala' 'Rising Pune Supergiants' 'Delhi Capitals']
15
```

Fig. 5 Dataset of Franchise Names

The dataset analyzed in this paper consists of records for 15 distinct teams that participated in the Indian Premier League (IPL). During the course of the IPL history, several teams underwent changes, including renaming under new managing bodies, disbandment, and rebranding while remaining under the same management. These transformations highlight the dynamic nature of franchise teams in the IPL and their evolution over time. In our analysis, we have taken into account the changes related to the Delhi Daredevils and Rising Pune Supergiants teams. It is important to note that these changes were only a rebranding of the team names, without any impact on the players or management structure. Therefore, these changes do not have any significant effect on the data and its interpretation. The Rising Pune Supergiants (RPS) and Delhi Daredevils (DD) are two teams that experienced name changes during their participation in IPL. The team was renamed from "Rising Pune Supergiants" to "Rising Pune Supergiant" to align with the singular form of the team name. Similarly, the Delhi Daredevils went through a name change to become the Delhi Capitals (DC) ahead of the 2019 season. The rebranding was undertaken to give the team a fresh identity and to better represent the spirit of the city it represents. On accounting name changes the unique team count reduced to 13 from the originally declared 15.

```
In [422…  final_df.sample()
```

Out[422…

	batting_team	bowling_team	city	runs_left	balls_left	wickets	total_runs_x	crr	rrr	result
48626	Deccan Chargers	Rajasthan Royals	Nagpur	155	106	10	159	1.714286	8.773585	0

Fig. 6 Final Data after Cleaning and pre processing

C. Training & Testing

In order to build an important prepared set, the problem being resolved must be grasped. The preparation set, which is the larger of the two, is used. When you run a series of preparations through an AI framework, it informs the network how to optimally weight certain features and translates them into coefficients based on their likelihood of limiting error in the results. These coefficients, called bounds in each case, are stored in tensors, collectively called a model because they encapsulate a model of the data being trained on. These are the key takeaways from creating an AI framework. Then comes the test set. It serves as an insurance seal and is only used last. The neural network can be tested against this final subjective assessment after it has been prepared and the data set.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In [426...
```
X_train
```
Out[426...

	batting_team	bowling_team	city	runs_left	balls_left	wickets	total_runs_x	crr	rrr
41116	Chennai Super Kings	Deccan Chargers	Chennai	189	112	10	190	0.750000	10.125000
56101	Rajasthan Royals	Delhi Daredevils	Jaipur	30	22	6	151	7.408163	8.181818
122680	Mumbai Indians	Kings XI Punjab	Mumbai	160	87	8	177	3.090909	11.034483
33384	Delhi Daredevils	Kolkata Knight Riders	Durban	45	29	9	154	7.186813	9.310345
131158	Delhi Daredevils	Sunrisers Hyderabad	Raipur	113	76	9	163	6.818182	8.921053
...	...	...	...	...	...	...	...	...	...
72266	Deccan Chargers	Chennai Super Kings	Visakhapatnam	108	46	7	193	6.891892	14.086957
130220	Kings XI Punjab	Royal Challengers Bangalore	Bangalore	191	76	5	226	4.772727	15.078947
148655	Sunrisers Hyderabad	Rajasthan Royals	Hyderabad	103	92	9	133	6.428571	6.717391
42819	Kings XI Punjab	Deccan Chargers	Cuttack	105	47	5	170	5.342466	13.404255
133897	Rajasthan Royals	Royal Challengers Bangalore	Pune	77	8	1	180	5.517857	57.750000

67015 rows × 9 columns

Fig. 7 Trained data

1) The most successful IPL team: In a game of sports, every team competes for victory. Hence, the team that has registered the most number of victories is the most successful

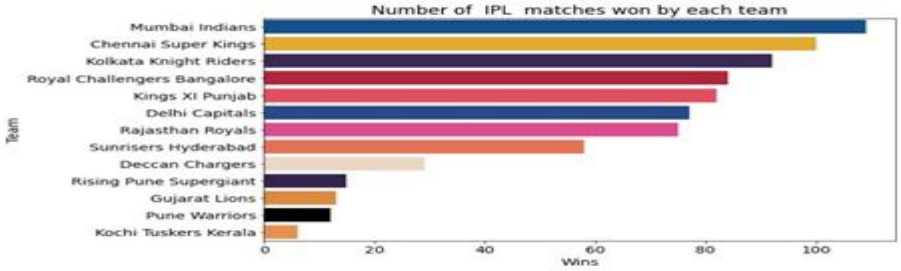


Fig. 8 Most successful IPL teams

Observations:
Mumbai Indians is the most successful team (as they have won the maximum number of IPL matches -109) followed by Chennai Super Kings and Kolkata Knight Riders.

2) Top 10 Stadiums:

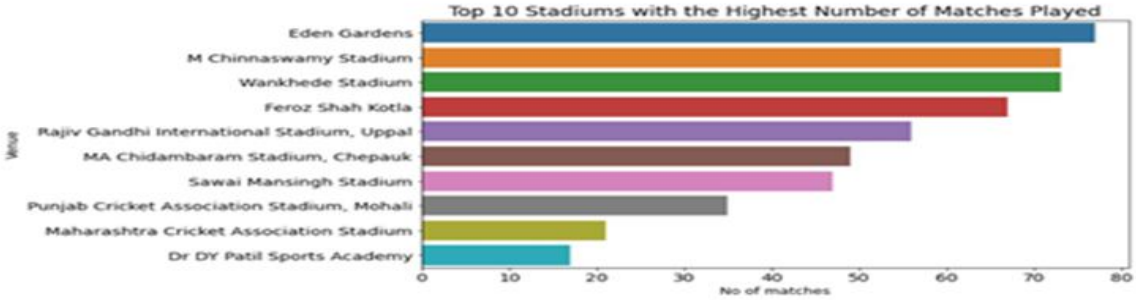


Fig. 8 Top 10 stadium

Observations:
Eden Gardens has hosted the maximum number of IPL matches followed by Wankhede Stadium and M Chinnaswamy Stadium.

3) IPL Teams fared in toss:

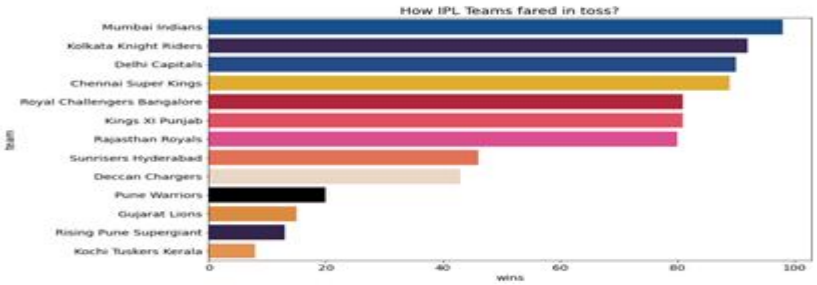


Fig. 10 IPL team fared in toss

Observations:

Mumbai Indians has won the most toss (till 2019) in IPL history. All the top teams in IPL are successful in winning the toss as well

4) Top 10 IPL Players: In a game of sports, every team competes for victory. Hence, the team that has registered the most number of victories is the most successful

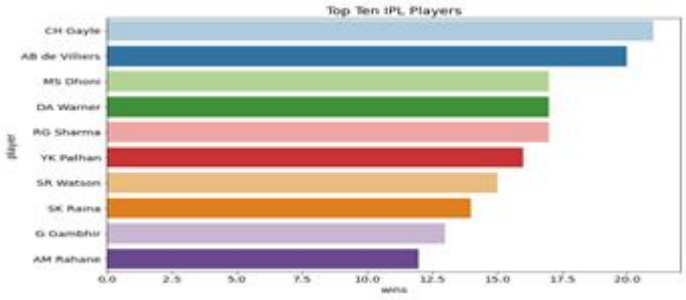


Fig. 11 Top Most valuable players

Observations:

Chris Gayle is the player who won the most player of the match awards and hence is the most valuable player. Six Indian players have figured in the top ten IPL players list.

5) Most Wins per season by an IPL Team:

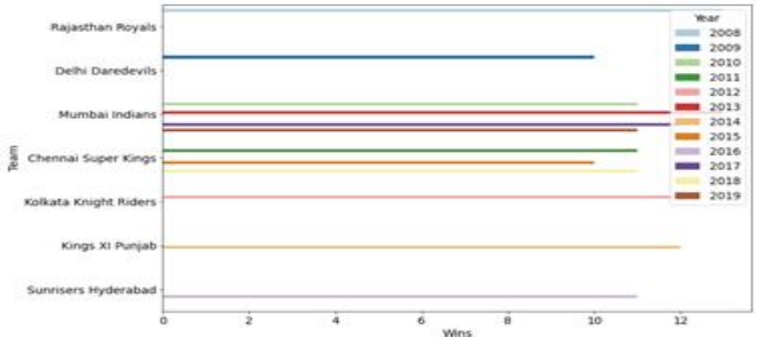


Fig. 11 Most Wins per Season by an IPL Team

Observations:

Mumbai Indians stand out as the most successful team in IPL history, topping the chart with the highest number of wins in four seasons, showcasing their dominant performance. Chennai Super Kings follow closely with three seasons as the second-highest achievers, highlighting their consistent presence among the top teams in the league.

6) Impacts of Team in IPL in winning toss: In a game of sports, every team competes for victory. Hence, the team that has registered the most number of victories is the most successful.

```
The number of times the team winning toss have won: 393
The percentage of winning if won the toss: 52 %
```

Fig. 12 Impact on winning toss

Observations:

The probability of winning when the team had won the toss is 52%. So winning the toss gives a slight edge over the opponent. However, it would be naive to term winning the toss as a greater advantage as there were 363 instances when the team losing the toss has won the game

7) Optimal Strategy in IPL Matches: Chase or Defend: The Based on the analysis of 756 IPL matches played between 2008 and 2019, teams batting second and chasing a target emerged victorious in 55% (419 Matches) of the matches. This indicates a higher success rate compared to teams defending a total. These findings suggest that chasing is a more favourable strategy in the IPL, highlighting the importance of adaptability and flexibility in match tactics for teams vying for victory.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

id	44
Season	IPL-2017
city	Delhi
date	06-05-2017
team1	Mumbai Indians
team2	Delhi Capitals
toss_winner	Delhi Capitals
toss_decision	field
result	normal
dl_applied	0
winner	Mumbai Indians
win_by_runs	146
win_by_wickets	0
player_of_match	LMP Simmons
venue	Feroz Shah Kotla
umpire1	Nitin Menon
umpire2	CK Nandan
umpire3	NaN

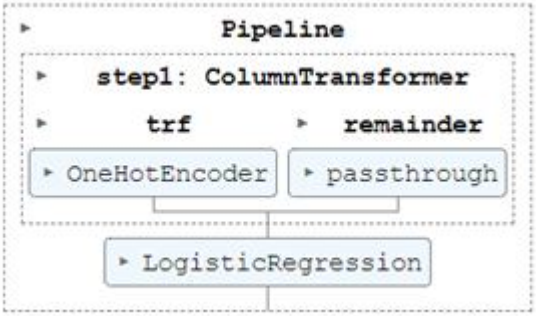
Name: 43, dtype: object

Fig. 12 Match details

Observations:
The greatest victory in IPL on defending a total is for Mumbai Indians when they defeated Delhi Daredevils by 146 runs on 06 May 2017 at Feroz Shah Kotla stadium, Delhi.

D. Model selection learning algorithm

In this paper, logistic regression was employed as a binary classification model to predict the outcome variable in a given dataset. The logistic regression model was implemented using scikit-learn's LogisticRegression () function with the liblinear solver. The dataset was pre-processed using a pipeline, which included a ColumnTransformer for one-hot encoding of categorical variables. This pipeline facilitated the transformation of the dataset by encoding the 'city' feature into a set of binary features. representing each unique category. The resulting transformed dataset was then used as input for the logistic regression model. The dataset was further split into training and testing sets using train_test_split, ensuring the model's evaluation on unseen data. The logistic regression model was trained on the training data using the pipeline's fit method, enabling the model to learn the underlying patterns and relationships between the input features and the binary outcome. Predictions were made on the test set using the predict method, and the accuracy of the model was evaluated using the accuracy_score function. Logistic regression, a widely-used algorithm for binary classification, proved to be a suitable approach for this task, demonstrating its effectiveness in predicting the outcome based on the input features of 'team' and 'winner'.



E. Rationale for Choosing the liblinear Solve

The liblinear solver was the top choice for predicting IPL win percentages in the second innings due to its distinct advantages over other solvers like Newton's method, gradient descent, and stochastic gradient descent. It offered exceptional computational efficiency, critical for handling large IPL datasets with many matches, teams, and variables. Its optimization algorithms were designed for large-scale datasets, enabling faster model training and evaluation. Furthermore, the liblinear solver's ability to handle both L1-regularization (Lasso) and L2-regularization (Ridge) set it apart. L1-regularization facilitated automatic feature selection, helping identify influential factors for a team's second-inning win percentage. Additionally, the solver showed robustness against multicollinearity, a common issue in sports datasets with highly correlated features.

©IJRASET: All Rights are Reserved

This ensured model stability and prevented overfitting. In summary, the liblinear solver excelled due to its computational efficiency, support for regularization techniques, and robustness against multicollinearity. It was the optimal choice for our IPL win percentage prediction paper.

F. Final model evaluation

The aim of this paper was to perform exploratory data analysis (EDA) and develop a predictive model to estimate win percentage in the second innings of cricket matches. Initially, the focus was on conducting a comprehensive EDA to gain insights into the dataset and identify key factors influencing team performance. Exploratory analyses included descriptive statistics, data visualization, and correlation analysis. These EDA findings served as the foundation for subsequent machine learning (ML) model development. If we used Random Forest model, then we suffered from overfitting. To address the overfitting issue, logistic regression with the liblinear solver was selected as an alternative approach. The liblinear solver exhibited exceptional performance and stability in predicting win percentages. Leveraging regularization techniques, such as L1 and L2 regularization, the logistic regression model effectively handled multicollinearity and feature selection, leading to improved model generalization and interpretability. Once the logistic regression model was trained and evaluated, it was serialized into a pickle file format. This allowed for convenient storage and reusability of the model.



Fig. 13 Final accuracy score

G. Fit the model to data and predict the results

Our model demonstrates the capability to prognosticate match outcomes, specifically gauging the probability of victory for each respective team during the second innings of a given match, predicated upon the target set during the initial innings. This machine learning paradigm, having undergone rigorous development and optimization, has attained a commendable predictive accuracy rate of 81 percent. This technological advancement underscores our commitment to fostering data-driven insights and facilitating informed decision-making within the context of competitive sporting events

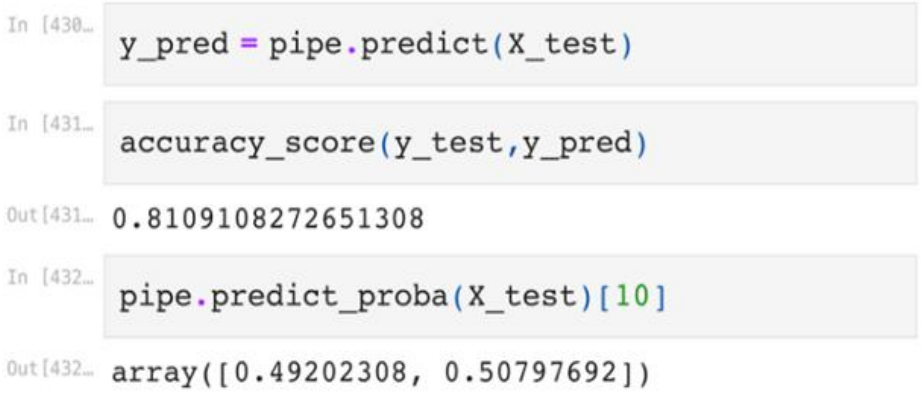


Fig. 14 Result Prediction

H. Web application

We Developed a web application using Streamlit for the research work. The web app is hosted at sportsguru.tech. Users can enter the batting team, bowling team, stadium, target, current score, overs completed, and wickets down. After entering the details, users can press the "Predict" button to obtain the win percentage predictions for both teams. The web app provides predictions based on the input parameters using the ML model. The predictions help users assess the win probabilities for the teams involved in the match.

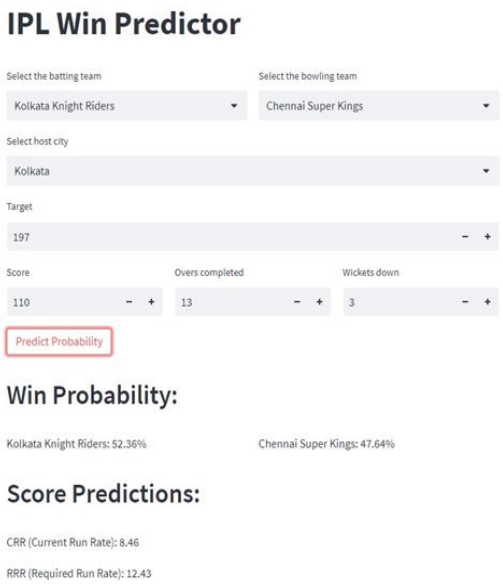


Fig. 13 Web Application in action

V. CONCLUSIONS

This paper has provided valuable insights of Indian Premier League (IPL) matches. Through the utilization of various statistical analytics techniques and data visualization tools, we have gained a deeper understanding of the game and its dynamics. Our objective was to explore the vast IPL dataset, perform exploratory data analysis, and draw meaningful insights. We have successfully identified the most successful IPL teams, analyzed the impact of winning the toss, examined the preferences for chasing or defending, and highlighted the top performers in the league. These insights can be valuable for team management, players, and fans alike, enabling them to make informed decisions and predictions. Furthermore, we have developed a machine learning model that predicts the winning percentage of teams in the second innings of a match. This model showcases the potential of data-driven approaches in making accurate predictions. By leveraging historical data and relevant features, our model can assist in forecasting match outcomes and aiding in strategic decision-making. This paper also emphasized the importance of effective data visualization using tools like Seaborn and Matplotlib. By creating attractive graphs and charts.

.

REFERENCES

[1] H. Ahmad, A. Daud, L. Wang, H. Hong, H. Dawood and Y. Yang, “Prediction of Rising Stars in the Game of Cricket”, vol. 5, pp. 4104 – 4124, Mar. 2017
[2] H. Song, V. Shin and M. Jeon, “Mobile Node Localization Using Fusion Prediction-Based Interacting Multiple Model in Cricket Sensor Network”, IEEE Transactions on Industrial Electronics, vol.59, Nov. 2012.
[3] S. Roy, P. Dey and D. Kundu, “Social Network Analysis of Cricket Community Using a Composite Distributed Framework: From Implementation Viewpoint”, IEEE Transactions on Computational Social Systems, vol. 5, pp. 64-81, Mar. 2018.
[4] P. Kansal, P. Kumar, H. Arya, A. Methaila, “Player valuation in Indian premier league auction using data mining technique”, International Conference on Contemporary Computing and Informatics (IC3I), pp. 27-29, Nov. 2014.
[5] S. Agrawal, S. P. Singh, J. K. Sharma, “Predicting results of IPL T-20 Match using Machine Learning”, 8th International Conference on Communication Systems and Network Technologies (CSNT), pp. 24-26, Nov. 2018.
[6] H. Barot, A. Kothari, P. Bide, B. Ahir, R. Kankaria, “Analysis and Prediction of Indian Premier League”, 2020 International Conference for Emerging Technology (INCET), pp.5-7, June 2020.
[7] A. Kaluarachchi, S. Varde Aparna, “CricAI: A classification-based tool to predict the outcome in ODI cricket”, 2010 Fifth International Conference on Information and Automation for Sustainability, pp.17-19, Dec. 2010.
[8] N. Rodrigues, N. Sequeira, S. Rodrigues, V. Shrivastava, “Cricket Squad Analysis Using Multiple Random Forest Regression”, IEEE Xplore, 1st International Conference on Advances in Information Technology, 2019
[9] M. B. Wright, “Scheduling fixtures for New Zealand Cricket”, IMA Journal of Management Mathematics 16, pp. 99−112, 2005
[10] M. Maduranga, H. Singhe, G. Poravi, “Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links”, 5th International Conference for Convergence in Technology (I2CT), 2019
[11] M. B. Wright, “Scheduling fixtures for New Zealand Cricket”, IMA Journal of Management Mathematics 16, pp. 99−112, 2005.
[12] M. Maduranga, Hatharasinghe, G. Poravi, “Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links”, 5th International Conference for Convergence in Technology (I2CT), Mar 2019.

[13] M. Shetty, S. Rane, C. Pandita, S. Salvi, "Machine learning-based Atlantis Highlights in Computer Sciences", Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020), IEEE Conference Record, ISBN: 978-1-7281-5371-1, 2020

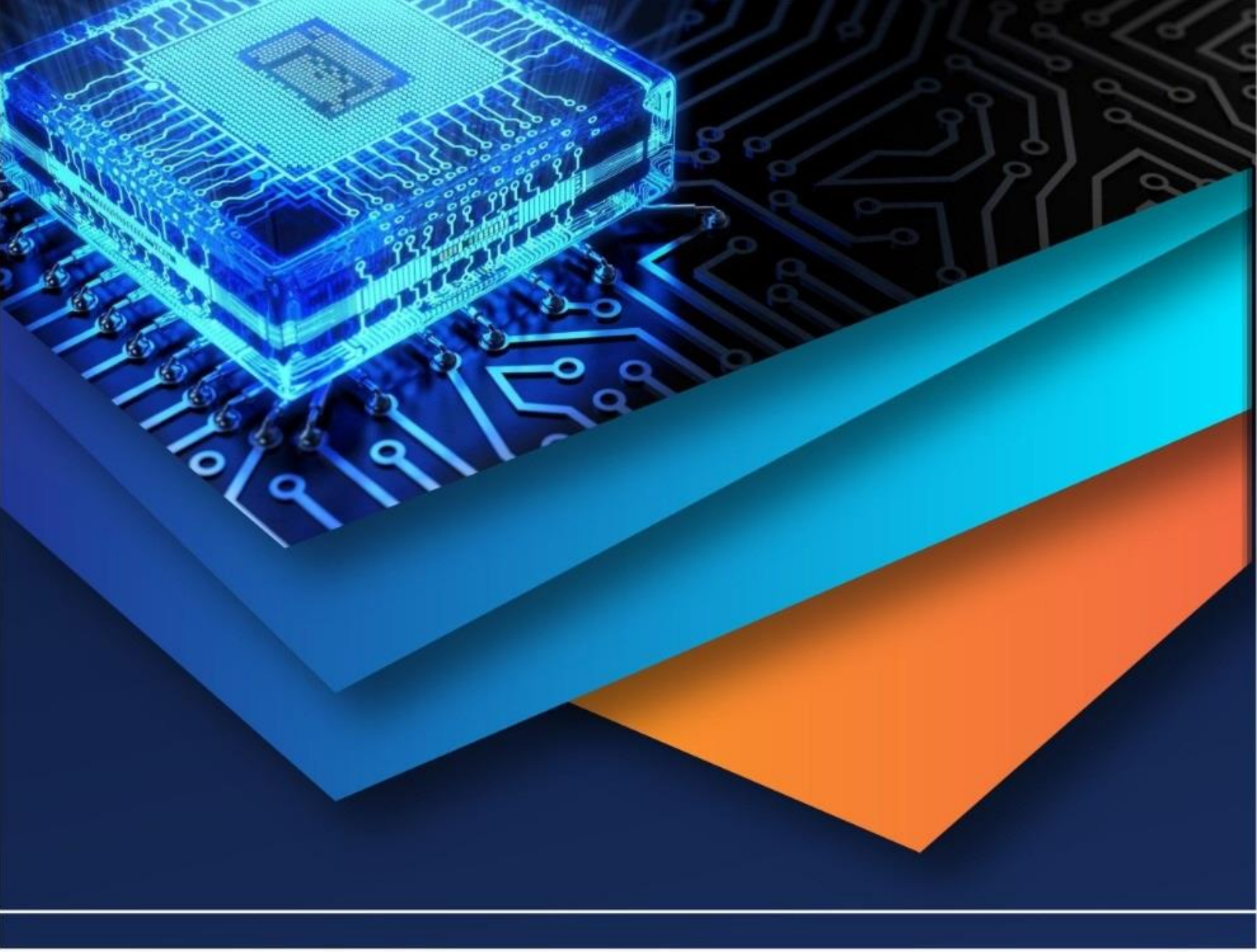
[14] A. Balasundaram, S. Ashokkkumar, D. Jayashree, K. Magesh S, "Data Mining based Classification of Players in Game of Cricket", proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020), IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

[15] J. Kumar, R. Kumar, P. Kumar, "Outcome Prediction of ODI Cricket Matches Using Decision Trees and MLP Networks", 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)

[16] S. Prabu., V. Balamurugan, F. V. Jayasudha, P. Visu, and K. Janarthanan. "Mobile technologies for contact tracing and prevention of COVID-19 positive cases: a cross-sectional study." International Journal of Pervasive Computing and Communications (2020)

[17] Subramani, Prabu, K. Srinivas, R. Sujatha, and B. D. Parameshachari. "Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients." Personal and Ubiquitous Computing, 2021

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)