Indian Premier League (IPL) Win/Loss Prediction using Machine Learning(Logistic Regression)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

Name	**Univ. Roll No.**
Suchanda Banerjee	10800120085
Promit Dey	10800120097
Sanket Bakshi	10800120084
Upasak Sharma Choudhury	10800120123

UNDER THE GUIDANCE OF

Dr. Debasis Chakraborty
(Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ASANSOL ENGINEERING COLLEGE
AFFILIATED TO
MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

June, 2024

Contents

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ASANSOL ENGINEERING COLLEGE
Vivekananda Sarani, Kanyapur, Asansol, West Bengal – 713305

Certificate of Recommendation

I hereby recommend that the thesis entitled, **"Indian Premier League (IPL) Win/Loss Prediction using Machine Learning"** carried out under my supervision by the group of students listed below may be accepted in partial fulfilment of the requirement for the degree of "Bachelor of Technology in Computer Science and Engineering" of Asansol Engineering College under MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY.

Name	**Univ. Roll No.**
Suchanda Banerjee	10800120085
Promit Dey	10800120097
Sanket Bakshi	10800120084
Upasak Sharma Choudhury	10800120123

……

(Dr. Debasis Chakraborty)
Project Supervisor
Dept. of Comp. Sc. & Engg.,
Asansol Engineering College,
Asansol-713305

Countersigned:

……

(Dr. Monish Chatterjee)
Head of the Department
Dept. of Comp. Sc. & Engg.
Asansol Engineering College,
Asansol-713305

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ASANSOL ENGINEERING COLLEGE
Vivekananda Sarani, Kanyapur, Asansol, West Bengal – 713305

Certificate of Approval

The thesis is hereby approved as creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance in the partial fulfilment of the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it is submitted.

……

(**Dr. Debasis Chakraborty**)
Project Supervisor

Dept. of Comp. Sc. & Engg.,
Asansol Engineering College,
Asansol-713305

Acknowledgement

It is our great privilege to express our profound and sincere gratitude to our Project Supervisor, **Prof. (Dr.) Debasis Chakraborty** for providing us a very cooperative and precious guidance at every stage of the present project work being carried out under his/her supervision. His valuable advice and instructions in carrying out the present study has been a very rewarding and pleasurable experience that has greatly benefited us throughout the course of work.

We would like to convey our sincere gratitude towards Dr. Monish Chatterjee, Head of the Department of Computer Science and Engineering of Asansol Engineering College for providing us the requisite support for timely completion of our work. We would also like to pay our heartiest thanks and gratitude to all the teachers of the Department of Computer Science and Engineering, for various suggestions being provided in attaining success in our work.

We would like to express our earnest thanks to Mr. Suman Mallick, of CSE Project Lab for his technical assistance provided during our project work.

Finally, I would like to express my deep sense of gratitude to my parents for their constant motivation and support throughout my work.

……
(Suchanda Banerjee)

……
(Promit Dey)

……
(Sanket Bakshi)

……
(Upasak Sharma Choudhury)

Abstract

The Indian Premier League (IPL) has evolved into one of the most popular and competitive cricket leagues globally, capturing the attention of millions of fans. In this dynamic sporting environment, predicting match outcomes becomes a challenging yet intriguing task. This study explores the application of machine learning techniques to predict the win/loss results of IPL matches. The research leverages historical match data, including team performance metrics, player statistics, match venues, and contextual factors, to develop robust predictive models. Various machine learning algorithms, such as decision trees, random forests, and support vector machines, are employed to analyse the complex interplay of variables influencing match outcomes. Feature engineering techniques are applied to enhance the models' accuracy and interpretability.

The study aims to contribute to the growing body of sports analytics by offering insights into the factors that significantly impact IPL match results. Additionally, the research assesses the performance of different machine learning models in predicting cricket match outcomes, providing valuable information for stakeholders, including team management, analysts, and cricket enthusiasts. The findings of this research not only hold implications for strategic decision-making within the IPL ecosystem but also contribute to the broader domain of sports analytics and machine learning applications in predicting competitive outcomes. Ultimately, the study offers a comprehensive examination of the feasibility and effectiveness of machine learning in forecasting IPL match results, shedding light on the intricate dynamics of T20 cricket.

List of Figures

List of Tables

1. Preface……

1.1 Introduction

The IPL is a dynamic cricketing tournament where teams composed of international and domestic talents battle it out in high stakes matches. The ability to predict team scores and game outcomes is a subject of great interest to fans, sports analysts, and the betting industry. Accurate predictions could have profound implications in enhancing in-game strategies, engaging fans, and influencing the betting market. The topic of the project is related to machine learning methodologies for predicting the outcome of the winning/losing probability of the teams participating in the IPL tournament. Machine learning can be used to build models that can predict the outcome of IPL matches with a high degree of accuracy. Machine learning models are trained on historical data to identify patterns and relationships between the variables that influence match outcomes. These variables can include team performance, player statistics, pitch conditions, weather conditions, past match results, and head-to-head records. Once the model is trained, it can be used to make predictions for future IPL matches by inputting the relevant data. Based on the considered problem and the given dataset, machine learning algorithms can be organized into a taxonomy as: 1) supervised learning (Logistic Regression in [1], Random Forest in [2]); 2) unsupervised learning (K-means clustering in [3]).

The project mainly deals with Logistic Regression Machine Learning methods. Logistic regression can be used for classification tasks. It is a simple but effective algorithm that is often used for binary classification problems, such as predicting whether a cricket team will win or lose a match. To use logistic regression for IPL win/loss prediction, we would first need to collect a dataset of historical IPL match data. This dataset should include features such as team names, match venue, target, and match winner. Once we have collected the dataset, we would need to split it into training and test sets. The training set will be used to train the logistic regression model, and the test set will be used to evaluate the model's performance on unseen data. To train the logistic regression model, we would need to provide it with the training set data. The model will learn the relationship between the features and the target variable (match winner). Once the model is trained, we can use it to predict the winner of a new IPL match. To do this, we would simply provide the model with the features of the new match, and it will output a probability that each team will win.

In addition to developing a robust logistic regression model for predicting IPL match outcomes, our project includes creating an intuitive and user-friendly web application to showcase these predictions. Built with ReactJS [4] and styled using Tailwind CSS, this web application provides users with a seamless experience for accessing match predictions. ReactJS allows for a dynamic and responsive interface, ensuring that users can easily input relevant match details and receive immediate predictions. Tailwind CSS enhances the design with its utility-first approach, enabling rapid and efficient styling that ensures the application is both visually appealing and

easy to navigate. Through this application, fans, analysts, and enthusiasts can interact with the prediction model in real-time, making data-driven insights accessible and engaging for all users.

1.2 Motivation of the project

The motivation for this project comes from the growing importance of data-driven decision-making in sport. As the IPL is a highly dynamic and competitive cricket tournament, the ability to predict match results can provide valuable insights into team strategies, player management and fan engagement how it meets the need. The project aims to bridge the gap between traditional cricket analytics and modern data science, offering a new approach to understanding and predicting the complex dynamics of the IPL.

1.3 Basic description of the project

This project focuses on employing machine learning techniques for the prediction of Indian Premier League (IPL) match outcomes. By leveraging historical match data, including team and player statistics, match venues, and contextual factors, the study aims to develop predictive models capable of forecasting whether a team will win or lose a match. The project involves key steps such as data preprocessing, feature engineering, and the application of various machine learning algorithms [2]. The ultimate goal is to provide a comprehensive analysis of the factors influencing IPL match results and to assess the efficacy of machine learning in enhancing predictive accuracy. The findings are expected to have implications for strategic decision-making within the IPL ecosystem and contribute to the broader field of sports analytics.

Fig 1.3.1 Web Application home screen

2. Literature Review…….

2.1 General

The use of machine learning techniques to predict match outcomes is very common in the growing field of sports research in this context, cricket, and especially T20 tournaments like the Indian Premier League (IPL), have proven to be interesting topics for predictive modelling.

Research in sport analysis has ranged in various fields, including player performance analysis, team strategies and the impact of contextual information on the outcome of the tournament Researchers have examined the relevance of machine learning algorithms to predict outcomes in games.

Many projects have shown that machine learning is effective in predicting cricket tournament results considering factors like player statistics, team dynamics and tournament conditions but the nuanced understanding of these factors has a different IPL format of the 19th century remains a relatively unexplored area. This literature review aims to synthesize the current state of knowledge, identify gaps and opportunities in existing research, and provide a basis for specific applications of machine learning to incorporate IPL seminar outcomes.

2.2 Review of related works

As of today, there are many fantasy gaming platforms that are using this tech like,

Dream11:

Dream11 is one of the pioneers in the fantasy sports industry and has played a significant role in popularizing fantasy gaming in India. It offers users the opportunity to create fantasy teams across various sports, including cricket, football, basketball, and more. Dream11 has been particularly associated with the IPL, allowing users to assemble virtual teams based on real players' performances in IPL matches. The platform is known for its user-friendly interface, extensive player statistics, and a range of contests catering to different skill levels and preferences. However, user reviews have occasionally highlighted concerns about the app's performance during peak times, and there have been discussions around the legal status of fantasy sports in certain regions [5].

Fig 2.2.1 Dream11 App Home Page

My11Circle:

My11Circle is another fantasy sports platform that gained popularity in India. Sponsored by renowned cricketer Sourav Ganguly, My11Circle focuses on cricket fantasy leagues, providing users with the opportunity to create teams and earn points based on players' actual performances. The platform emphasizes ease of use and offers various contests with different entry fees, allowing users to compete at their preferred levels. Like Dream11, My11Circle has been associated with the IPL, leveraging the tournament's immense popularity to attract users. Users have appreciated the platform for its engaging contests and user-friendly interface. However, it's crucial to consider the evolving landscape of fantasy sports regulations and user experiences [5].

Connection to our Project:

Given our project's focus on IPL win/loss prediction using machine learning, exploring user engagement and data patterns on platforms like Dream11 and My11Circle could provide valuable insights. Analysing user behaviours, team compositions, and strategies adopted on these platforms may contribute to understanding the factors influencing users' virtual team selections and, by extension, contribute to the broader field of sports analytics. Keep in mind that obtaining relevant data from these platforms may have legal and privacy implications, so ensure compliance with applicable regulations.

3. Related theories and Algorithms…….……

3.1 Fundamental theories underlying the work

Logistic Regression for Binary Classification:
Logistic regression [6] for binary classification is a statistical method used to model the probability of a binary outcome based on one or more predictor variables.
It provides a clear explanation of logistic regression as a machine learning algorithm specifically designed for binary classification tasks.
Discuss how logistic regression models the probability of an event occurring, making it suitable for predicting binary outcomes like win or loss in IPL matches.

Supervised Learning Concept:
Supervised learning [6] is a type of machine learning where a model is trained on labelled data to make predictions or decisions based on new, unseen data.
It elaborates on the concept of supervised learning, the overarching paradigm in which our project operates.
It explains how the model is trained on labelled historical data, where the algorithm learns patterns and relationships between input features and the target variable (match outcomes).

Feature Selection and Importance:
Feature selection [7] is the process of identifying and selecting a subset of relevant features for use in model construction to improve performance and reduce overfitting.
These features may include team performance metrics, player statistics, pitch conditions, and more.
It explains the process of selecting relevant features and their importance in influencing the model's predictive accuracy.

Training and Testing Sets:
Detail the rationale behind splitting the dataset into training and test sets.
Emphasize the role of the training set in teaching the model and the importance of the test set in assessing the model's generalization to new, unseen data. You can learn more about Training and Testing sets in [8].

Probability and Decision Thresholds:
Explore how logistic regression outputs probabilities for each class (team winning or losing).
It discusses the concept of decision thresholds [1-3] and how it influences the classification decision based on predicted probabilities.

Evaluation Metrics:
The evaluation metrics that will be used to assess the performance of the logistic regression model.
Common metrics for binary classification include accuracy, precision, recall, and F1 score [1].

Historical Data and Pattern Recognition:
It highlights the role of historical IPL match data in training the model.
Emphasize how the algorithm leverages this data to recognize patterns [2] and relationships between features and match outcomes.

Relevance to IPL Win/Loss Prediction:
It explicitly connects these theoretical concepts to the specific task of predicting IPL match outcomes.
Discuss how the principles of logistic regression [1] and supervised learning [3] align with the dynamics of cricket matches and the factors influencing team performance.

About NPM (Node Package Manager):
NPM stands for node packet manager [9]. It functions as a package manager for programming language mainly speaking -JavaScript is a product of
GitHub or we can say it's GitHub subsidiary, which gives a host like service i.e server for development of software and control of Version by using Git as a version control system. node packet manager is the default package manage of the programming language which is JavaScript. Interestingly enough, node packet manager is the world's largest software registry. The developers which contribute to open source from every place in this world use node packet manager to give and take data in form of packets, however many organizations use it for private uses also which are not available to the general population.

NPM has 3 major parts:
- website
- registry
- Command Line Interface (CLI)

We use NPM for package discovery, setting up a profile, and managing various other things of NPM usage. Let's clear it with an example, you can identify organizations to manage the grant to public or private packages. The CLI functions from the terminal generally cmd or git, and is how most developers interact with NPM. The NPM registry acts as large open-source database of JavaScript and it's meta information.

About ReactJS:
React [4] is a framework which makes creating interactive UI's a lot less time consuming and makes it painless to create interactive UIs. React lets us design easy to create views for every stage in our application development, React has a great efficient update and it works on the components that need the change not on the whole application. The views which are declarative in nature makes the code easy to analyse and even easier to optimize and debug it. Building components which are encapsulated and manage their own state, we combine then to make better UI's. The logic of the components is written in JavaScript language and

not as templates, we can pass data with ease through our app. No assumptions are made about what technologies a person might be using, we can develop new features in React without having to again rewrite the existing code. ReactJs can also make changes on the server with the help of Node and can make powerful apps via React Native. To understand react in more depth, let us discuss the functioning of React in the background i.e background processes. The best and most important benefit of ReactJs is performance. The speed of React is a thing of beauty, and it works on low memory principles that are achieved by abstraction of the DOM (Document Object Model) with a virtual DOM in action. To implement data binding most of the front-end technologies use one of the two i.e Key-Value Observation like (Meteor, Ember) or the Dirty Checking like (AngularJS). React acts in a different manner and acts on a JavaScript approach. Let's see how React works on this implementation:
First happens the DOM Abstraction and a virtual representation of the DOM that is stored in the memory. Now when this happens, the data model shifts/changes, React then goes through the process of re-rendering and only the components that use the data that has changed "Diffs" the previous version of virtual DOM with the new version of virtual DOM updates to the actual DOM, and leads to only modifying those components which require this change and will be directly beneficial to them.

```
import React from "react";
import ReactDOM from "react-dom/client";
import { createBrowserRouter, RouterProvider, Outlet } from "react-router-dom";
import Header from "./components/Header";
import Predict from "./components/Predict";
import Body from "./components/Body";
import Footer from "./components/Footer";
import AboutUs from "./components/AboutUs";
import Faq from "./components/Faq";

const AppLayout = () => {
  return (
    <div className="body-custom p-2 bg-cover">
      <Header />
      <Outlet />
      <Footer />
    </div>
  );
};

//React Router
const appRouter = createBrowserRouter([
  {
    path: "/",
    element: <AppLayout />,
    children: [
      {
        path: "/",
        element: <Body />,
      },
      {
        path: "/about",
        element: <AboutUs />,
      },
      {
        path: "/predict",
        element: <Predict />,
      },
      {
        path: "/faq",
        element: <Faq />,
      },
    ],
  },
]);

const root = ReactDOM.createRoot(document.getElementById("root"));
root.render(<RouterProvider router={appRouter} />);
```

Fig 3.1.1 App.js File

3.2 Fundamental Algorithms

Logistic Regression:

Logistic Regression serves as a foundational algorithm in our project's predictive modelling efforts. By leveraging the principles of logistic regression, our model effectively estimates the probability of a binary outcome—in this case, whether a team will win or lose an IPL match. This algorithm is particularly chosen for its simplicity, interpretability, and proven suitability for binary classification tasks.

The choice of logistic regression is driven by several key factors. Firstly, its simplicity ensures that the model can be easily implemented and understood. Logistic regression provides a clear understanding of the relationship between the dependent variable (match outcome) and the independent variables (historical data on team and player performance). This interpretability allows us to draw meaningful insights from the model, making it easier to communicate findings to stakeholders and make informed decisions based on the predictions.

Moreover, logistic regression's suitability for binary classification makes it an ideal choice for predicting IPL match outcomes, which are inherently binary (win or lose). The algorithm calculates the probability of a team winning a match based on various predictors such as previous match performances, player statistics, and other relevant factors. These probabilities can then be used to classify the match outcome, providing a straightforward yet powerful predictive capability.

To develop a robust predictive framework, we analyse extensive historical data on team and player performance. This data includes metrics such as batting averages, bowling statistics, head-to-head records, and other performance indicators. By incorporating these variables into our logistic regression model, we capture the nuances of team dynamics and player contributions, enhancing the model's accuracy and reliability.

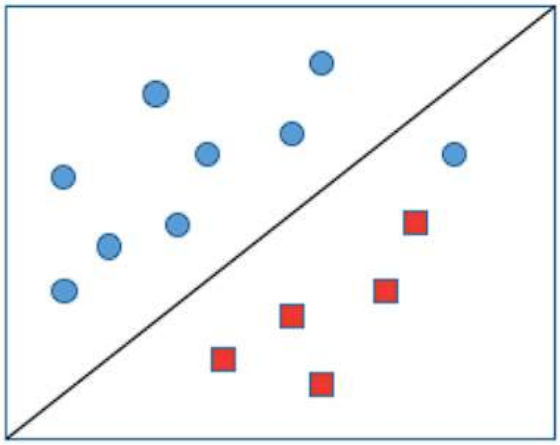

Fig 3.2.1 Logistic Regression representing classification

Random Forests:

Random Forests [2], an ensemble learning method, are utilized to enhance predictive accuracy and robustness. By constructing multiple decision trees and aggregating their predictions, Random Forests mitigate overfitting and improve generalization. This algorithm is crucial in handling the intricacies of IPL match prediction, where diverse factors can impact outcomes, and a collective decision-making approach proves advantageous.

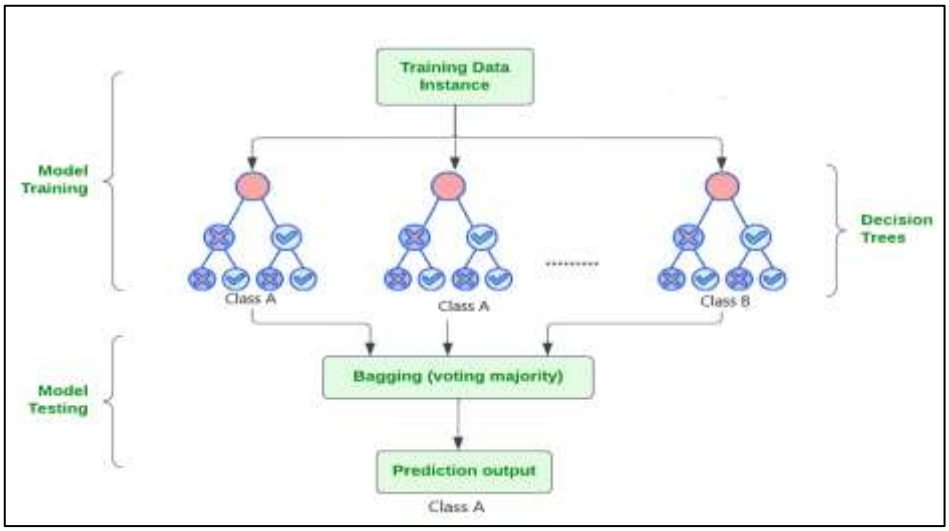

Fig 3.2.2 Random Forest

Use of Random Forest in our Project:

The difference in accuracy between the Random Forest model (0.78) and the Logistic Regression model (0.80).

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,y_pred)
```

```
0.9990188520569065
```

Fig 3.2.3 Random Forest Classifier Accuracy Score in our project

4. Proposed Model/ Algorithm……

4.1 Proposed Model

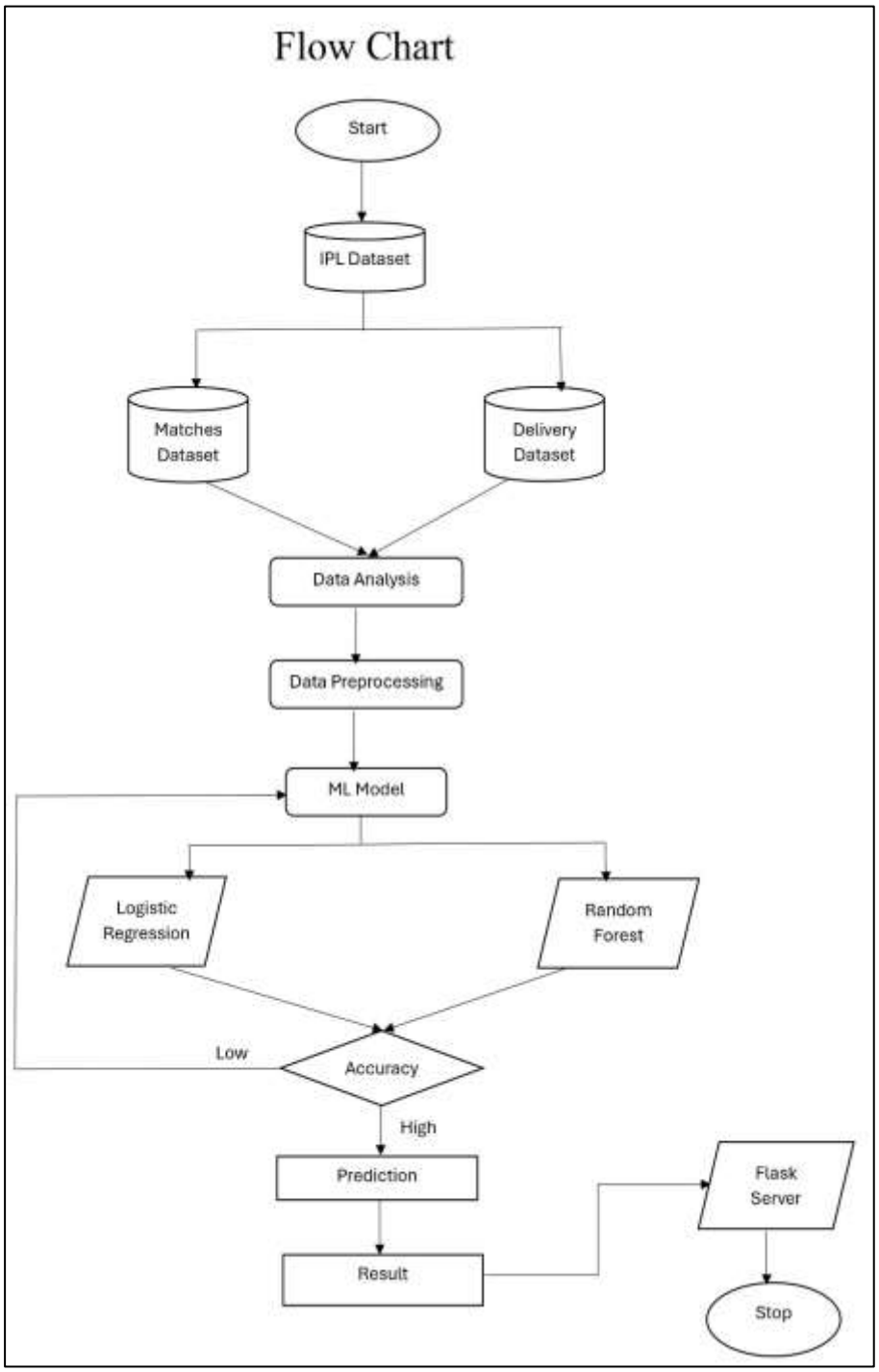

Fig 4.1 Flow diagram illustrating the sequential application of data preprocessing, feature engineering, and machine learning algorithms for IPL win/loss prediction, providing a systematic overview of the predictive modelling process.

4.2 Proposed Algorithm

Logistic Regression:

The Logistic Regression algorithm [10] is a foundational component of the IPL win/loss prediction project, serving as a robust statistical method for binary classification tasks. The fundamental theory underlying Logistic Regression is based on modelling the probability of a binary outcome, such as whether a team will win or lose an IPL match.

Model Formulation: In the context of the project, the Logistic Regression model is formulated [11] as follows:

$$P(\text{Win}) = \frac{1}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Here :

- **P(Win)** is the probability of wining.
- **e** is the base of the natural logarithm.
- $\boldsymbol{\beta_0, \beta_1, \beta_2, \ldots., \beta_n}$ are the coefficients.
- $\boldsymbol{x_1, x_2, \ldots, x_n}$ are the input features.

Model Interpretation: The Logistic Regression model [6] provides interpretable coefficients (β), allowing us to understand the impact of each feature on the likelihood of winning an IPL match. Positive coefficients indicate a positive correlation with the win probability, while negative coefficients suggest a negative correlation.

Feature Selection: Feature selection [7] plays a crucial role in enhancing the model's predictive power. Relevant features, such as team performance metrics, player statistics, and contextual factors, are carefully chosen to capture the intricate dynamics of IPL matches.

Training and Evaluation: The Logistic Regression model is trained on historical IPL match data, optimizing the coefficients to maximize the likelihood of observed outcomes. The model's performance is evaluated using metrics such as accuracy, precision, recall, and the area under the Receiver Operating Characteristic (ROC) curve [1].

Iterative Refinement: Given the dynamic nature of cricket and the IPL, the Logistic Regression model undergoes iterative refinement [3]. Continuous evaluation allows for adaptation to evolving player strategies, team dynamics, and changing match conditions, ensuring its relevance and accuracy over time.

The application of Logistic Regression in this project reflects a commitment to a transparent, interpretable, and well-established algorithm, contributing to the comprehensive analysis of IPL match outcomes.

5. Simulation Results……

5.1 Experimental Setup

The experimental setup for the IPL Win/Loss Prediction project involves a systematic approach to data preparation, model training, and evaluation to ensure the robustness and generalizability of the predictive models.

Data Collection: Historical IPL match data is collected, encompassing a diverse range of seasons to capture variations in team performance, player form, and contextual factors. The dataset includes information on team statistics, player metrics, match venues, and other relevant features.

	id	Season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_runs	win_by_wickets
0	1	IPL-2017	Hyderabad	05-04-2017	Sunrisers Hyderabad	Royal Challengers Bangalore	Royal Challengers Bangalore	field	normal	0	Sunrisers Hyderabad	35	0
1	2	IPL-2017	Pune	06-04-2017	Mumbai Indians	Rising Pune Supergiant	Rising Pune Supergiant	field	normal	0	Rising Pune Supergiant	0	7
2	3	IPL-2017	Rajkot	07-04-2017	Gujarat Lions	Kolkata Knight Riders	Kolkata Knight Riders	field	normal	0	Kolkata Knight Riders	0	10
3	4	IPL-2017	Indore	08-04-2017	Rising Pune Supergiant	Kings XI Punjab	Kings XI Punjab	field	normal	0	Kings XI Punjab	0	6
4	5	IPL-2017	Bangalore	08-04-2017	Royal Challengers Bangalore	Delhi Daredevils	Royal Challengers Bangalore	bat	normal	0	Royal Challengers Bangalore	15	0

Table 5.1.1 “matches.csv” dataset [12]

	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

Table 5.1.2 “deliveries.csv” dataset [12]

Data Preprocessing: Missing values, if any, are addressed through imputation or removal, maintaining the integrity of the dataset. Categorical variables are encoded, and numerical features are scaled to standardize the input data for modelling.

	match_id	inning	total_runs
0	1	1	207
2	2	1	184
4	3	1	183
6	4	1	163
8	5	1	157
...	...	...	...
1518	11347	1	143
1520	11412	1	136
1522	11413	1	171
1524	11414	1	155
1526	11415	1	152

Table 5.1.3 Processed data of teams after 1st Innings

batsman	...	noball_runs	penalty_runs	batsman_runs	extra_runs	total_runs_y	player_dismissed	dismissal_kind	fielder	current_score	runs_left
CH Gayle	..	0	0	1	0	1	NaN	NaN	NaN	1	206
Mandeep Singh	..	0	0	0	0	0	NaN	NaN	NaN	1	206
Mandeep Singh	..	0	0	0	0	0	NaN	NaN	NaN	1	206
Mandeep Singh	..	0	0	2	0	2	NaN	NaN	NaN	3	204
Mandeep Singh	..	0	0	4	0	4	NaN	NaN	NaN	7	200
...	..	...	...	...	...	...	...	...	...	...	...
RA Jadeja	..	0	0	1	0	1	NaN	NaN	NaN	152	0
SR Watson	..	0	0	2	0	2	NaN	NaN	NaN	154	-2
SR Watson	..	0	0	1	0	1	SR Watson	run out	KH Pandya	155	-3
SN Thakur	..	0	0	2	0	2	NaN	NaN	NaN	157	-5
SN Thakur	..	0	0	0	0	0	SN Thakur	lbw	NaN	157	-5

Table 5.1.4 Processed data of teams after adding current_score and runs_left colums

penalty_runs	batsman_runs	extra_runs	total_runs_y	player_dismissed	dismissal_kind	fielder	current_score	runs_left	balls_left
0	1	0	1	NaN	NaN	NaN	1	206	119
0	0	0	0	NaN	NaN	NaN	1	206	118
0	0	0	0	NaN	NaN	NaN	1	206	117
0	2	0	2	NaN	NaN	NaN	3	204	116
0	4	0	4	NaN	NaN	NaN	7	200	115
...	...	...	...	...	...	...	...	...	...
0	1	0	1	NaN	NaN	NaN	152	0	4
0	2	0	2	NaN	NaN	NaN	154	-2	3
0	1	0	1	SR Watson	run out	KH Pandya	155	-3	2
0	2	0	2	NaN	NaN	NaN	157	-5	1
0	0	0	0	SN Thakur	lbw	NaN	157	-5	0

Table 5.1.5 Processed data of teams after adding balls_left column

batsman	...	batsman_runs	extra_runs	total_runs_y	player_dismissed	dismissal_kind	fielder	current_score	runs_left	balls_left	wickets
CH Gayle	...	1	0	1	0	NaN	NaN	1	206	119	10
Mandeep Singh	...	0	0	0	0	NaN	NaN	1	206	118	10
Mandeep Singh	...	0	0	0	0	NaN	NaN	1	206	117	10
Mandeep Singh	...	2	0	2	0	NaN	NaN	3	204	116	10
Mandeep Singh	...	4	0	4	0	NaN	NaN	7	200	115	10

Table 5.1.6 Processed data of teams after adding wickets columns

Feature Engineering: Relevant features are selected based on their potential impact on match outcomes, including team composition, player form, and historical performance. Additional derived features may be created to capture complex relationships within the data.

	batting_team	bowling_team	city	runs_left	balls_left	wickets	total_runs_x	crr	rrr	result
125	Royal Challengers Bangalore	Sunrisers Hyderabad	Hyderabad	206	119	10	207	6.000000	10.386555	0
126	Royal Challengers Bangalore	Sunrisers Hyderabad	Hyderabad	206	118	10	207	3.000000	10.474576	0
127	Royal Challengers Bangalore	Sunrisers Hyderabad	Hyderabad	206	117	10	207	2.000000	10.564103	0
128	Royal Challengers Bangalore	Sunrisers Hyderabad	Hyderabad	204	116	10	207	4.500000	10.551724	0
129	Royal Challengers Bangalore	Sunrisers Hyderabad	Hyderabad	200	115	10	207	8.400000	10.434783	0
...	...	...	...	...	...	...	...	...	...	...
149573	Chennai Super Kings	Mumbai Indians	Hyderabad	0	4	5	152	7.862069	0.000000	0
149574	Chennai Super Kings	Mumbai Indians	Hyderabad	-2	3	5	152	7.897436	-4.000000	0
149575	Chennai Super Kings	Mumbai Indians	Hyderabad	-3	2	4	152	7.881356	-9.000000	0
149576	Chennai Super Kings	Mumbai Indians	Hyderabad	-5	1	4	152	7.915966	-30.000000	0
149577	Chennai Super Kings	Mumbai Indians	Hyderabad	-5	0	3	152	7.850000	-inf	0

Table 5.1.7 Processed data showing Current Run Rate and Required run Rate by the batting team and the result column which help in predicting

Model Selection: The choice of machine learning algorithms, including Logistic Regression, is made based on their suitability for binary classification tasks and their interpretability. Multiple algorithms may be explored to identify the most effective model for IPL win/loss prediction.

```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder

trf = ColumnTransformer([
    ('trf',OneHotEncoder(sparse=False,drop='first'),['batting_team','bowling_team','city'])
]
,remainder='passthrough')
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline
```

```
pipe = Pipeline(steps=[
    ('step1',trf),
    ('step2',LogisticRegression(solver='liblinear'))
])
```

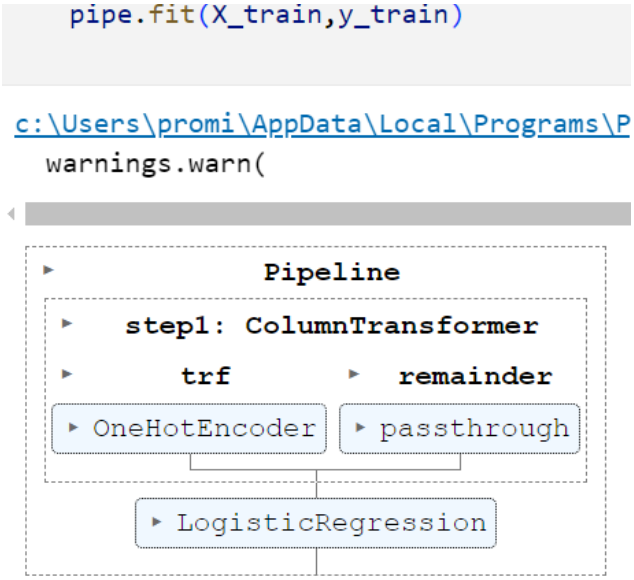

Fig 5.1.8 Machine Learning Model – Logistic Regression

Training and Validation: The dataset is split into training and validation sets to facilitate model training and assessment. Cross-validation techniques, such as k-fold cross-validation, may be employed to ensure the model's stability and prevent overfitting.

Model Evaluation: The trained models are evaluated using a comprehensive set of performance metrics, including accuracy, precision, recall, and the F1 score. Evaluation metrics provide insights into the model's ability to correctly predict IPL match outcomes.

Fig 5.1.9 Testing Accuracy of our model

Results Analysis: The final models' results are analysed to extract meaningful insights into the factors influencing IPL match outcomes. Model interpretations, including feature importance, contribute to a deeper understanding of the predictive capabilities.

File Structure:

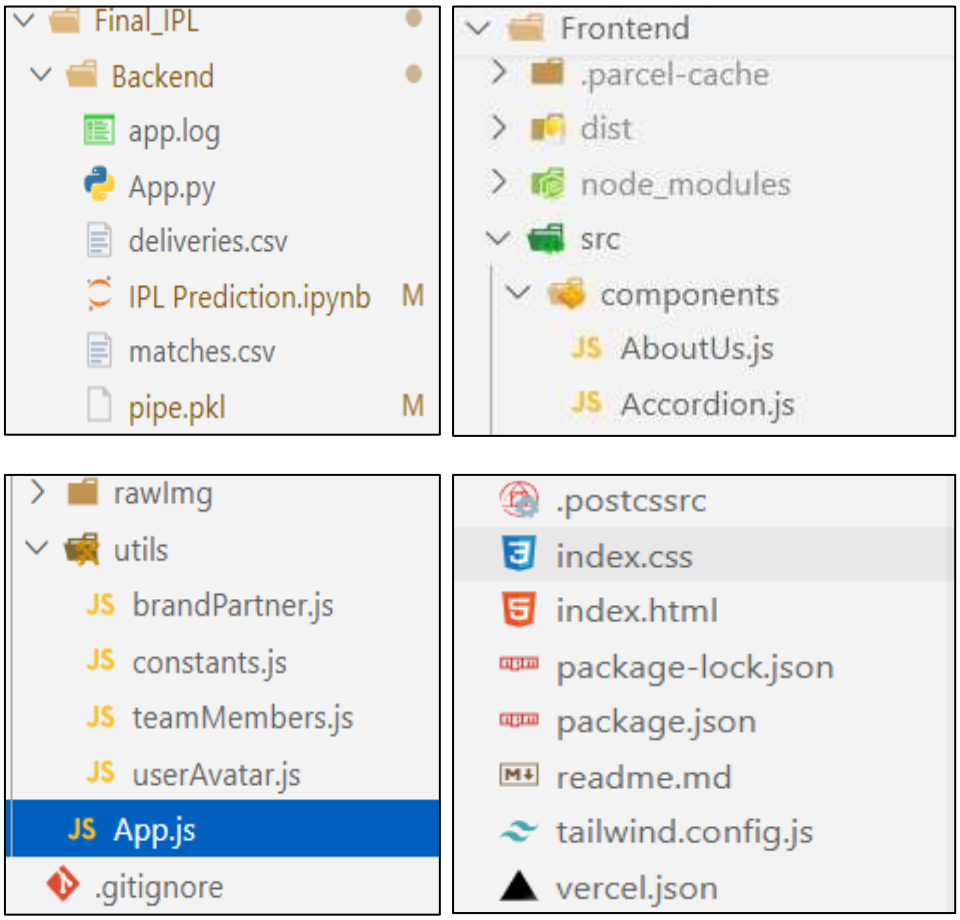

Fig 5.1.10 File Structure

By following this structured experimental setup, the project aims to build robust and adaptive models for IPL win/loss prediction, contributing significantly to the evolving field of sports analytics. This approach involves meticulously gathering and preprocessing historical IPL match data, carefully selecting features that impact match outcomes, and employing logistic regression as the primary machine learning technique for prediction. The systematic training and validation process ensures that the model is both accurate and generalizable, capable of making reliable predictions on new, unseen data.

5.2 Experimental Results

Target- 178

	end_of_over	runs_after_over	wickets_in_over	lose	win
10459	1	4	0	57.0	43.0
10467	2	8	0	51.6	48.4
10473	3	1	0	58.3	41.7
10479	4	7	1	69.9	30.1
10485	5	12	0	59.9	40.1
10491	6	13	0	47.5	52.5
10497	7	9	0	41.6	58.4
10505	8	15	0	27.7	72.3
10511	9	7	0	25.5	74.5
10518	10	17	0	14.0	86.0
10524	11	9	1	19.5	80.5
10530	12	9	0	16.0	84.0
10536	13	8	0	13.7	86.3
10542	14	8	0	11.8	88.2
10548	15	5	1	20.5	79.5
10555	16	8	1	29.2	70.8
10561	17	8	2	55.7	44.3
10567	18	6	1	70.7	29.3
10573	19	8	2	89.4	10.6

Table 5.2.1 Experimental setup to test our model for a particular match

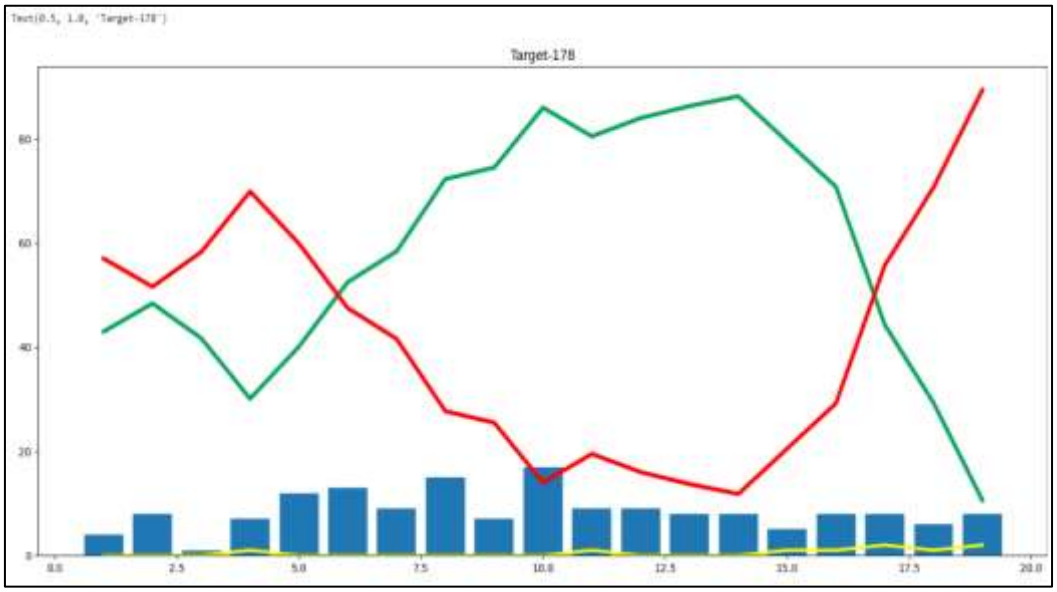

Fig 5.2.2 Green, Red line depicts the winning/losing probability, blue bars depict increase/decrease in runs, yellow lines depict the fall of wicket after every over.

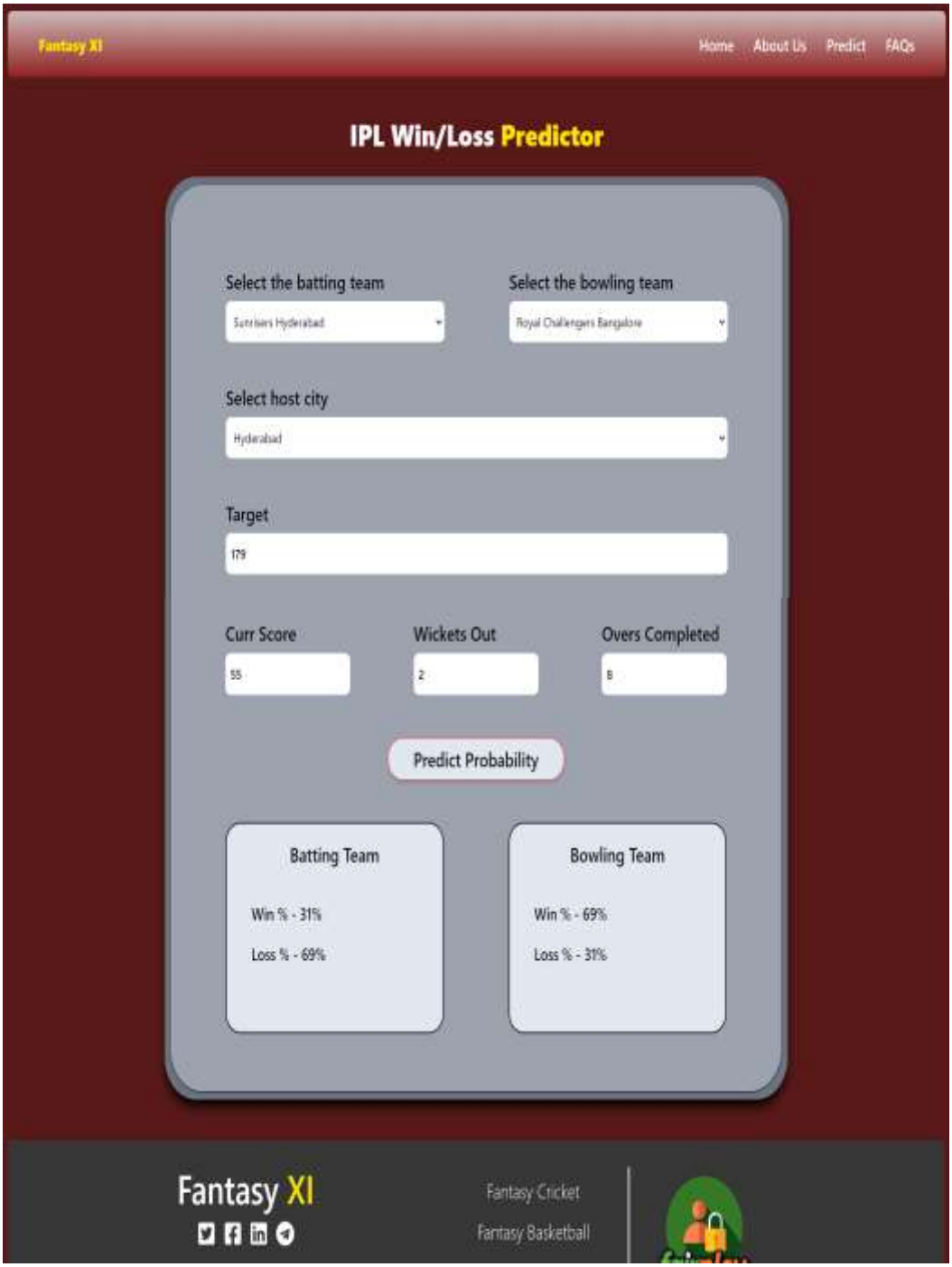

Fig 5.2.3 App Predictor – By giving inputs of the batting team, bowling team, city, target, wickets out, current score, overs completed. It shows the win/loose probability of each team.

6. Discussion and Conclusion……...……

6.1 Discussion

It provides a platform to critically analyse the findings, implications, and limitations of the IPL Win/Loss Prediction project. Firstly, the project's success in leveraging machine learning, specifically Logistic Regression, to predict IPL match outcomes is evident [1-2]. The interpretability of the Logistic Regression model allows for a clear understanding of the impact of various features on the likelihood of a team winning [1][5]. This transparency is essential for cricket stakeholders, including team management, analysts, and fans, as it facilitates informed decision-making based on identifiable factors. The feature engineering process played a pivotal role in enhancing the model's predictive performance [7]. Selecting and transforming relevant features, including team statistics and player metrics, proved crucial in capturing the nuanced dynamics of IPL matches. The incorporation of contextual factors, such as match venues and weather conditions, further enriched the model's ability to adapt to the unique challenges posed by the T20 format [7].

An integral part of the project's strength lies in its iterative approach. The continuous evaluation and refinement process ensures that the models remain adaptive to the ever-changing landscape of the IPL [1][2]. This adaptability is paramount in a sport where player forms evolve, team strategies shift, and unforeseen circumstances impact match outcomes. However, certain limitations merit consideration. The project's predictive power relies heavily on historical data, assuming that past trends will persist [2]. While this is a common constraint in sports analytics, the inherent unpredictability of cricket introduces an element of uncertainty [2]. Additionally, the model's performance may be influenced by external factors not accounted for in the dataset, such as player injuries or unexpected tactical innovations [3].

Furthermore, the project underscores the importance of contextual analysis but recognizes that the IPL's intricacies extend beyond the captured features [6]. Team dynamics during auctions, player combinations, and the influence of international players are dimensions that, while acknowledged, may warrant more in-depth exploration in future iterations of the project [6].

In conclusion, the IPL Win/Loss Prediction project demonstrates the efficacy of machine learning, particularly Logistic Regression, in forecasting cricket match outcomes [5][6]. The discussion highlights the project's successes, acknowledges its limitations, and emphasizes the necessity of ongoing refinement to ensure continued relevance in the dynamic landscape of the IPL. The insights gained contribute not only to cricket analytics but also to the broader discourse on the intersection of machine learning and sports prediction [1-4].

6.2 Future work

The exploration of machine learning for IPL win/loss prediction using Logistic Regression has unveiled a promising avenue ripe for further advancements and applications. As we look to the future, several areas beckon for deeper exploration and refinement:

1. Real-Time Prediction Platforms: The development of real-time prediction platforms could be the next frontier. Integrating the predictive model into live match scenarios could provide instantaneous insights for coaches, teams, and fans, elevating the in-game decision-making process.

2. Expanding to T20 Leagues Globally: While IPL served as the focal point, extending the methodology to other T20 cricket leagues globally presents an exciting avenue. Each league introduces unique dynamics, and adapting the model to different contexts could be a compelling area for future research.

3. Dynamic Model Updating: In the dynamic world of cricket, where player form and team dynamics evolve, there is potential for dynamic model updating. Research could explore methodologies to adapt the predictive model in real-time, ensuring it remains relevant and effective throughout a cricketing season.

6.3 Conclusion

In closing, the journey through the realms of the Indian Premier League (IPL) and the intricate world of machine learning for match predictions culminates in a synthesis of insights and contributions. The introduction illuminated the fervour surrounding IPL matches, capturing the interest of fans, sports analysts, and the betting industry. Against this backdrop, the project articulated a clear objective: to leverage Logistic Regression and unravel the complexities of IPL outcomes through meticulous analysis of historical data and diverse match features. The overarching contribution of the project lies in its dual impact on both academic and practical fronts. It propels the field of sports analytics forward by introducing a novel application of machine learning to cricket, particularly in the context of the IPL. The commitment to advancing knowledge is mirrored in the nuanced exploration of feature importance, offering a deeper understanding of the intricate variables that shape match results. Practically, the project doesn't merely dwell in theoretical realms but extends its influence to the cricketing arena and beyond. The developed Logistic Regression model holds the promise of not only enhancing in-game strategies but also engaging fans and influencing the betting market. This acknowledgment of the broader implications reinforces the project's commitment to translating academic rigor into tangible, real-world applications. The overview of the project serves as a cohesive narrative thread, weaving together the contextual richness of IPL, the well-defined objective of employing Logistic Regression, and the multifaceted contributions made along the research journey. From the methodological intricacies of data collection and model training to the development of a user-friendly interface, the project stands as a testament to a comprehensive and meticulous approach. As a parting note, the comparative analysis with other machine learning models adds depth and richness to the research, offering a nuanced understanding of the relative strengths and weaknesses of algorithms. This comparative lens enhances the robustness of the project's findings and provides a springboard for future research endeavours in the dynamic field of sports analytics. In essence, the project not only serves as a scholarly exploration but also as a practical guide for those seeking to navigate the confluence of cricket, machine learning, and predictive analytics. Through its contributions and insights, this research endeavours to leave an indelible mark on the landscape of IPL predictions, inviting further exploration and refinement in the fascinating intersection of sports and technology.

References…….

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "The Elements of Statistical Learning" Chapter on logistic regression and its applications.

[2] Christopher M. Bishop. "Pattern Recognition and Machine Learning"

[3] Kevin P. Murphy. "Machine Learning: A Probabilistic Perspective"

[4] Greg Sidelnikov. React.js Book: Learning React JavaScript Library From Scratch

[5] ChatGPT - https://chat.openai.com/

[6] Fred C. Pampel. "An Introduction to Logistic Regression Analysis and Reporting". Overview of logistic regression analysis and interpretation of results.

[7] Alice Zheng, Amanda Casari. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists

[8] O Theobald. Machine Learning For Absolute Beginners: A Plain English Introduction (Second Edition) (AI, Data Science, Python & Statistics for Beginners)

[9] Greg Lim. Beginning Node.js, Express & MongoDB Development

[10] Pulkit Sharma. "Understanding Logistic Regression" on Analytics Vidhya.

[11] V. Sindhwani, P. K. Gopalan, and S. S. Keerthi. "Large-Scale Logistic Regression for Text Categorization", IEEE Transactions on Knowledge and Data Engineering in 2009

[12] Kaggle for datasets - https://www.kaggle.com/