COMPREHENSIVE ANALYSIS AND PREDICTION OF INDIA'S CRICKET PERFORMANCE IN MACHINE LEARNING

A Mini Project Report submitted to the Department of Computer Applications, Bharathiar University in the partial fullfillment of the requirements for the award of degree of

MASTER OF SCIENCE IN DATA ANALYTICS

Submitted by

JEER SHRI KARTHICK S

(22CSEG12)

Under the guidance of

Dr.P.Tamilarasi, MCA.,M.Phil.,Ph.D.,SET.,

GUEST FACULTY

DEPARTMENT OF COMPUTER APPLICATIONS

BHARATHIAR UNIVERSITY

COIMBATORE - 641 046

DECEMBER – 2023

DECLARATION

I hereby declare that this Mini-project report titled **"COMPREHENSIVE ANALYSIS AND PREDICTION OF INDIA'S CRICKET PERFORMANCE IN MACHINE LEARNING"** submitted to the Department of Computer Applications, Bharathiar University is a record of original work done by **JEER SHRI KARTHICK S (22CSEG12)** under the guidance of **Dr.P.Tamilarasi MCA.,M.Phil.,Ph.D.,SET,** Guest Faculty**,** Department of Computer Applications, Bharathiar University and this project work has not formed the basis for the award of any Degree/ Diploma/ Associate ship/ Fellowship or similar title to any candidate of any University.

Place: Coimbatore Signature of the Candidate

Date: (JEER SHRI KARTHICK S)

CERTIFICATE

This is to certify that the Mini-Project report titled **"COMPREHENSIVE ANALYSIS AND PREDICTION OF INDIA'S CRICKET PERFORMANCE IN MACHINE LEARNING"** submitted to the Department of Computer Applications, Bharathiar University in partial fulfilment of the requirement for the award of the degree of the Master of Science in Data Analytics is record of the original work done by **JEER SHRI KARTHICK S (22CSEG12)** under my supervision and guidance and this project work has not formed the basis for the award of any Degree/Diploma/Associate ship/ Fellowship or similar title to any candidate of any University.

Place: Coimbatore

Date:

Project Guide Head of the Department

Submitted for the University Viva-Voice Examination held on _____

Internal Examiner External Examiner

ACKNOWLEDGEMENT

I express my respectful thanks to our Professor & Head of the Department **Dr. M. Punithavalli, B.Sc., M.Sc., M.Phil., Ph.D,** Department of Computer Applications, Bharathiar University, for permitting me to carry out my mini project report work in **"COMPREHENSIVE ANALYSIS AND PREDICTION OF INDIA'S CRICKET PERFORMANCE IN MACHINE LEARNING"**

I really deem it a special privilege to convey my prodigious and everlasting thanks to my guide **Dr.P.Tamilarasi MCA.,M.Phil.,Ph.D.,SET,** Guest Faculty, Department of Computer Applications, Bharathiar University, for her valuable guidance and personal interest in this mini project report work.

Last but not least, I also acknowledge the help done by my parents and acknowledge the encouraging support of my friends who were involved in this mini project, in one way or the other.

I thank Almighty for showering the divine grace on me and offer prayers to the lord for everything that was given to me.

JEER SHRI KARTHICK S

(22CSEG12)

ABSTRACT

Cricket, a sport deeply ingrained in the hearts of millions, offers a rich tapestry of data, statistics, and insights that can be harnessed to uncover patterns and forecast outcomes. In this project, the embark on an extensive journey into the world of cricket, with a specific focus on analyzing and predicting the performance of the Indian cricket team.

Two primary datasets were manually collected from the **Statsguru** website to fuel this analysis. The first dataset, "Indian Performance," encapsulates the overall performance of the Indian cricket team. The second dataset, "Indian Team Performance," delves into the individual player statistics, dividing them into batting and bowling performance categories.

This analysis begins with a meticulous examination of the Indian cricket team's performance, taking into account critical metrics such as runs scored, runs per over (RPO), and overs played. These metrics provide invaluable insights into the team's capabilities, strengths, and weaknesses. Subsequently, the prediction of team's result using a combination of Ground, Opponent, and innings played, offering a glimpse into their potential performance in upcoming matches.

Within the principality of player performance, the two vital dimensions are batting and bowling. In batting performance, the analyze made between batting averages, strike rates, highest scores, and total runs scored. In the realm of bowling performance, the analyze made between bowling averages, economy rates, wickets taken, and overs bowled, unveiling the bowlers' effectiveness and impact on the game.

TABLE OF CONTENTS

1. INTRODUCTION

Cricket, often referred to as the "gentleman's game," is not only a sport but a shared passion and fervor for millions of enthusiasts worldwide. Rooted in tradition, yet constantly evolving, cricket offers a unique blend of skill, strategy, and excitement that captivates fans across the globe. With the advent of modern technology and data analytics, the game has transformed beyond the boundaries of mere entertainment, providing an opportunity to dissect, understand, and predict outcomes like never before.

In this project, embark on a compelling journey into the world of cricket, focusing our lens on the performance of one of the most prominent teams in the sport—Team India. Mission is to unveil the hidden patterns and nuances within the sport's rich tapestry, paving the way for a deeper understanding of the game and its strategies

1.1 Project Overview

The central thrust of this project revolves around the analysis and prediction of India's cricket performance. To do so, meticulously curated two invaluable datasets, collectively referred to as "Indian Performance" and "Indian Team Performance." These datasets serve as the bedrock for this exploration and provide a holistic view of both team and player statistics**.**

1.2 Objectives

Team Performance Analysis: To dissect and understand the collective performance of the Indian cricket team, examining key metrics such as runs scored, runs per over (RPO), and overs played. Through data analysis and visualization, the aim to gain insights into the team's strengths and areas for improvement.

Player Performance Analysis: A deep dive into player statistics, categorizing them into batting and bowling performance. the investigate made on batting averages, strike rates, highest scores, total runs for batsmen, and bowling averages, economy rates, wickets taken, and overs bowled for bowlers. This analysis sheds light on individual contributions and effectiveness.

Predictive Modeling: By employing statistical models and machine learning techniques, predicted the team's performance scores based on RPO, overs, and innings played. Additionally, forecast player performance and winning strategies, considering various player metrics.

1.3 Scope and significance

The significance of this project extends beyond cricket enthusiasts and strategists. It demonstrates the potential of data analytics in a sport that cherishes tradition, offering insights into player performance, team dynamics, and predictive modeling. This documentation serves as a testament to the synergy between sports and data, providing a robust framework for cricket analysis that can be adapted and extended to other teams and formats.

As unravel the intricacies of India's cricket journey, hope this project ignites the same passion for data and analytics that cricket does for its fans. Welcome to a world where technology meets tradition, where explore the past, understand the present, and predict the future of Indian cricket.

2. TOOLS AND REQUIREMENTS

2.1 LANGUAGE: Python

Python is a general-purpose, dynamic, high-level, and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures. Python is an easy-to-learn yet powerful and versatile scripting language, which makes it attractive for Application Development. With its interpreted nature, Python's syntax and dynamic typing make it an ideal language for scripting and rapid application development. Python supports multiple programming patterns, including object-oriented, imperative, and functional or procedural programming styles. Python is not intended to work in a particular area, such as web programming. It is a multipurpose programming language.

Python makes development and debugging fast because no compilation step is included in Python development, and the edit-test-debug cycle is very fast. Python has many web-based assets, open-source projects, and a vibrant community. Learning the language, working together on projects, and contributing to the Python ecosystem are all made very easy for developers. Because of its straightforward language framework, Python is easier to understand and write code in. This makes it a fantastic programming language for novices. Additionally, it assists seasoned programmers in writing clearer, error-free code. Python is an open-source, cost-free programming language. It is utilized in several sectors and disciplines as a result.

2.2 TOOL: Jupyter Notebook

Jupyter notebook is a client-server application. The application starts the server on local machine and opens the notebook interface in web browser where it can be edited and run from. The notebook is saved as ipynb file and can be exported as html, pdf and LaTex files. It is an open-source web application that can use to create and share documents that contains live code, equation, visualization and text Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself.

Notebook contains the inputs and outputs of interactive session as well as additional test that accompanies the code but it is not meant for execution. In this way, notebook files can serve as a complete computational record of a session, interleaving

executable code with explanatory text, mathematics and rich representation of resulting objects. These documents are internally json files and are saved with the .ipynb extension. Since json is a plain text format, they can be version-control and shared with the colleagues. Notebooks may be exported to a range of static formats, including HTML (for example, blog post), restructured text, LATEX, pdf and slide shows, via the nbconvert command.

2.3 APPLICATION: Machine Learning

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data. The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Machine learning is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to improve their performance in tasks through experience. These algorithms and models are designed to learn from data and make predictions or decisions without explicit instructions. There are several types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labeled data, while unsupervised learning involves training a model on unlabeled data. Reinforcement learning involves training a model through trial and error. Machine learning is used in a wide variety of applications, including image and speech recognition, natural language processing, and recommender systems.

SUPERVISED LEARNING: Logistic Regression

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the

same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable with the output variable.

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

3. DATA COLLECTION

Data is the lifeblood of any data-driven project, and in the context of this cricket performance analysis, the quality and comprehensiveness of the data are paramount. In this section, to delve into the data collection process, the two primary datasets are curated—namely,

(i) Indian Performance
(ii) Indian Team Performance

and the sources from which we've meticulously obtained this valuable cricket data.

3.1 Indian Performance Dataset

The "Indian Performance" dataset is a comprehensive compilation of historical data reflecting the overall performance of the Indian cricket team. This dataset covers a wide range of information, including match results, runs scored, wickets taken, and key match statistics. The data was manually collected from the renowned **Statsguru** website, which is a reputable source for cricket statistics and match details.

The process of collecting this data involved web scraping and manual curation. To ensure data accuracy, cross-referenced the collected statistics with official cricket scorecards and records. The dataset encompasses a substantial time frame, allowing for a historical perspective on India's cricket journey.

3.2 Indian Team Performance Dataset

Complementing the "Indian Performance" dataset, the "Indian Team Performance" dataset offers a more granular view of individual player statistics. This dataset is divided into two primary categories:

(i) Batting performance
(ii) Bowling performance

It covers key metrics such as batting averages, strike rates, highest scores, total runs for batsmen, and bowling averages, economy rates, wickets taken, and overs bowled for bowlers.

The data collection process for this dataset mirrored that of the "Indian Performance" dataset. It involved meticulous web scraping and manual curation from the same trusted source, Statsguru. The player data was verified and cross-referenced to maintain data integrity.

3.3 Data Sources

The Statsguru website is a pivotal source for cricket enthusiasts and analysts, offering an extensive repository of cricket statistics, historical match details, and player profiles. It is maintained by one of the most renowned cricket portals, Cricinfo, and is widely recognized for its accuracy and comprehensiveness.

In addition to Statsguru, the data collection process involved extracting data from official cricket scorecards, match summaries, and other trusted cricket statistics sources. This multi-faceted approach ensured the robustness of this datasets and the accuracy of the insights generated.

The painstaking process of data collection, cleaning, and validation laid the foundation for this project's success. The datasets that will be curated form the backbone of this cricket analysis and predictive modeling, providing a reliable platform for in-depth exploration and understanding of India's cricket performance. This documentation serves as a testament to the importance of data collection in data-driven projects and its transformative power in the world of cricket analytics.

4. TEAM PERFORMANCE ANALYSIS

Team performance analysis constitutes a fundamental pillar of this project, as aim to unravel the collective strengths, weaknesses, and dynamics of the Indian cricket team. This section provides a comprehensive overview of the approach to team performance analysis, the key metrics under scrutiny, data visualization techniques employed, score prediction model, and the insights derived from this analysis.

4.1 Key metrics for Team Performance

The foundation of team performance analysis rests on a set of key performance metrics, including but not limited to:

Runs Scored: The total number of runs scored by the Indian cricket team across various matches, providing an indication of their batting prowess.

Runs Per Over (RPO): A metric that represents the team's scoring rate per over, reflecting their ability to accumulate runs consistently and maintain a competitive edge.

Overs Played: The number of overs bowled and faced by the Indian team, which influences their run-scoring and wicket-taking strategies.

4.2 Data Visualization

To enhance the understandability and visual appeal of team performance analysis, employed a range of data visualization techniques. This includes Pie charts to depict the trend of runs scored over time, bar charts illustrating RPO in different matches, and scatter plots showcasing the relationship between runs and overs played. These visualizations offer an insightful and accessible representation of the data, aiding in the identification of trends and patterns.

4.3 Score Prediction Model

One of the project's highlights is the development of a predictive model for estimating the team's performance score. This model harnesses machine learning algorithms and statistical techniques to predict the number of runs the Indian cricket team is likely to score based on factors such as RPO, overs played, and innings.

The score prediction model contributes to the project's ability to forecast potential outcomes in upcoming matches and offers a valuable tool for cricket strategists and enthusiasts.

4.4 Insights and Findings

The analysis of team performance has yielded a wealth of insights into the Indian cricket team's dynamics. It unveils the team's ability to adapt to various match scenarios, the impact of batting order changes, and the influence of RPO and overs played on the final score. Additionally, the analysis identifies trends in the team's performance, both historically and in recent matches, offering an invaluable resource for understanding their strengths and areas for improvement.

The team performance analysis component of this project is not only a critical step in the project's overall success but also an illuminating exploration of the Indian cricket team's journey through data. It enriches our understanding of cricket strategy, offering a data-driven perspective that can inform future match tactics and enhance the overall cricket experience.

5. PLAYER PERFORMANCE ANALYSIS

The crux of this cricket performance analysis lies in the individual contributions of the players who make up the Indian cricket team. This section delves deep into the approach to player performance analysis, distinguishing between batting and bowling performance, and elaborates on the specific metrics under consideration for each category.

5.1 Batting Performance

The scrutiny of batting performance focuses on four pivotal metrics that unveil the batsmen's skills and effectiveness:

Batting Average: Batting average is a measure of a batsman's consistency and ability to score runs consistently. It's a reflection of the number of runs scored per dismissal.

Strike Rate: Strike rate represents a batsman's ability to accumulate runs quickly. It's calculated as the number of runs scored per 100 balls faced.

Highest Score: The highest score a batsman has achieved in a match highlights their ability to perform under pressure and their peak performance potential.

Total Runs: The total runs scored by a batsman over a specific period or in their career reflect their overall contribution to the team's success.

This analysis of these metrics provides a nuanced perspective on each batsman's strengths, weaknesses, and their impact on the Indian cricket team's performance

5.2 Bowling Performance Analysis

In the realm of bowling performance, the main focus to four key metrics that evaluate the bowlers' effectiveness and impact on the game:

Bowling Average: Bowling average measures a bowler's ability to take wickets. It is calculated as the number of runs conceded per wicket taken.

Economy Rate: Economy rate reflects a bowler's ability to maintain a tight line and length, minimizing the number of runs conceded per over bowled.

Wickets Taken: The total number of wickets a bowler has taken illustrates their capability to dismiss opposition batsmen and influence match outcomes.

Overs Bowled: The number of overs bowled demonstrates the bowler's stamina and consistency in delivering their allotted quota of overs.

This analysis of these bowling metrics unveils the strengths and areas for improvement of individual bowlers, elucidating their roles within the team's bowling strategy and their influence on match results.

5.3 Player Analysis Findings

The insights derived from this player performance analysis extend beyond individual achievements. They reveal the collective strength of the Indian cricket team, spotlighting standout performers, consistent contributors, and areas where enhancements may be necessary. Furthermore, findings highlight the complementary nature of batting and bowling performance and their crucial roles in shaping the team's outcomes.

By drilling down into the granular details of player performance, to provide a robust framework for player selection, strategy development, and match tactics, empowering the team and its strategists with data-driven insights to improve performance.

Our player performance analysis is an integral component of our cricket analytics project, offering a multifaceted exploration of the individual brilliance and collaborative dynamics that define Indian cricket. It sets the stage for the predictive modeling and strategic insights that culminate in a comprehensive understanding of India's cricket journey.

6. PREDICTIVE MODELLING

Predictive modeling is a pivotal aspect of this cricket performance analysis project, enabling to forecast the outcomes of cricket matches and player performances based on historical data and advanced statistical and machine learning techniques. This section elucidates the approach to predictive modeling, the specific models and methodologies employed, and the insights generated from this predictive analysis.

6.1 Team Performance Score Prediction

The crux of this predictive modeling lies in forecasting the performance of the Indian cricket team. A sophisticated predictive model that takes into account various match-related metrics, including runs per over (RPO), overs played, and innings. The model is designed to predict the number of runs the Indian cricket team is likely to score in a given match.

Score prediction model incorporates statistical techniques and machine learning algorithms, such as **linear regression, decision trees, and ensemble methods**. This comprehensive approach allows to create a robust and accurate model that factors in both historical data and dynamic match conditions. The model is trained on historical matches and their outcomes, making it capable of making predictions for upcoming fixtures.

6.2 Team Player Performance Prediction

In addition to team performance prediction, extended the predictive modeling to individual player performances. The performance of batsmen and bowlers based on their historical statistics and key performance metrics. This models consider factors such as batting averages, strike rates, bowling averages, economy rates, and the number of wickets taken.

These player performance prediction models provide insights into the expected contributions of individual players in upcoming matches. They help selectors and strategists make informed decisions regarding player selection and match tactics.

6.3 Winning Strategies

Predictive modeling also allows to explore potential winning strategies for the Indian cricket team. By considering various player and team performance metrics, easily can identify winning strategies that maximize the team's chances of success. These

strategies may include optimal player combinations, batting orders, and bowling rotations.

6.4 Model Evaluation

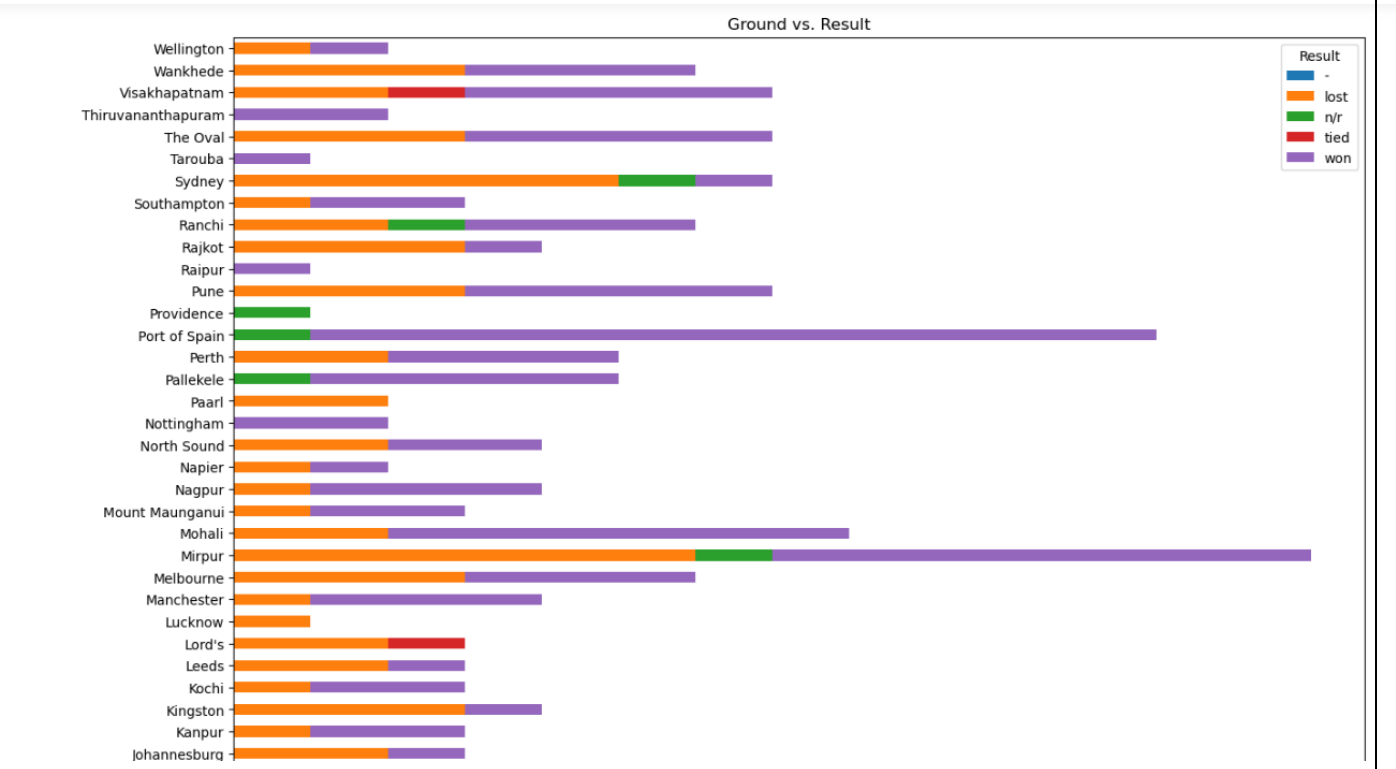
To ensure the accuracy and reliability of the predictive models, subjected them to rigorous evaluation. The metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to assess the models' performance. By comparing model predictions with actual outcomes, gauge their effectiveness and refine them for increased accuracy.

The insights derived from the predictive modeling not only provide valuable information for cricket enthusiasts but also offer practical tools for cricket strategists and selectors. These predictive models empower the cricket community with data-driven insights that can inform match tactics, player selection, and strategic decisions, contributing to a more informed and competitive cricket landscape.

7. RESULTS AND DISCUSSIONS

Bivariate Analysis:

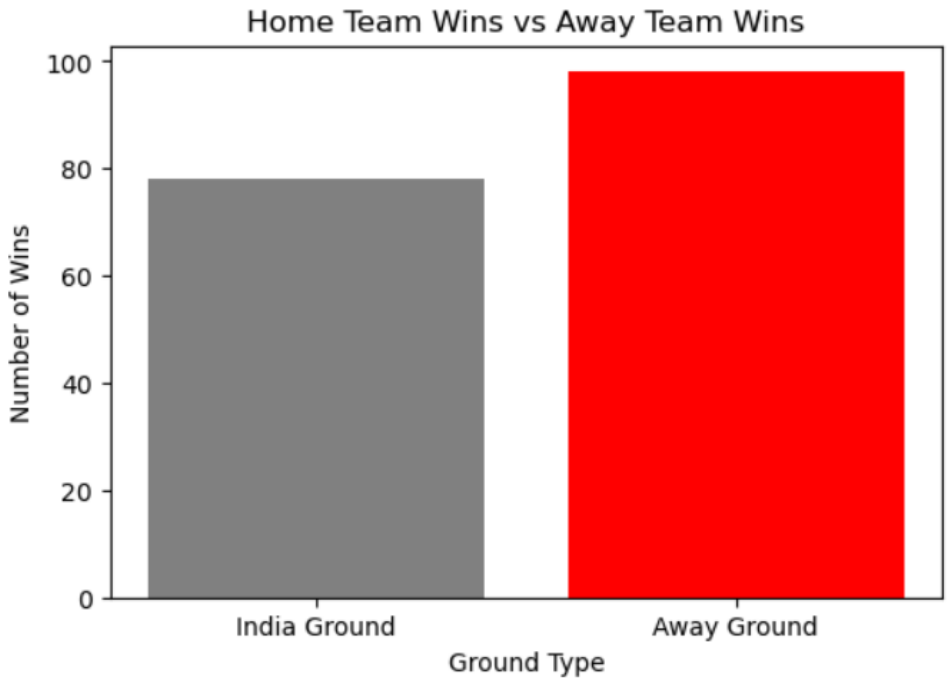
Chart: 1



Insight:

The analysis compares the performance attributes of different grounds in India based on the outcomes. This stacked bar chart indicates that the Port of Spain ground has the highest number of victories, while the Mirpur ground has the highest number of losses.

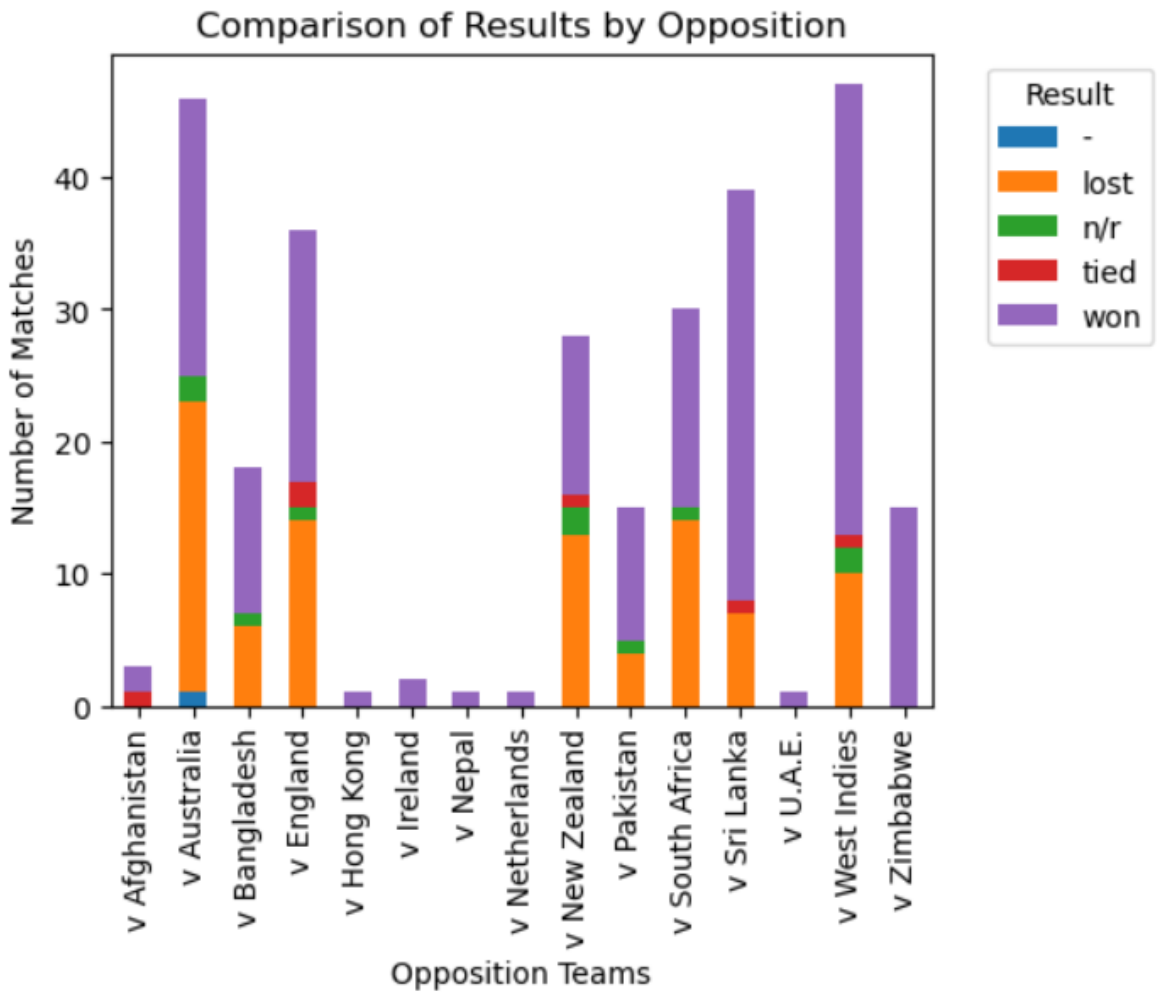
Chart:2



Insight:

The bar chart analysis compares the number of wins for the India team on their home ground and other grounds. It reveals that India has achieved more victories on away grounds compared to their performance on home grounds.

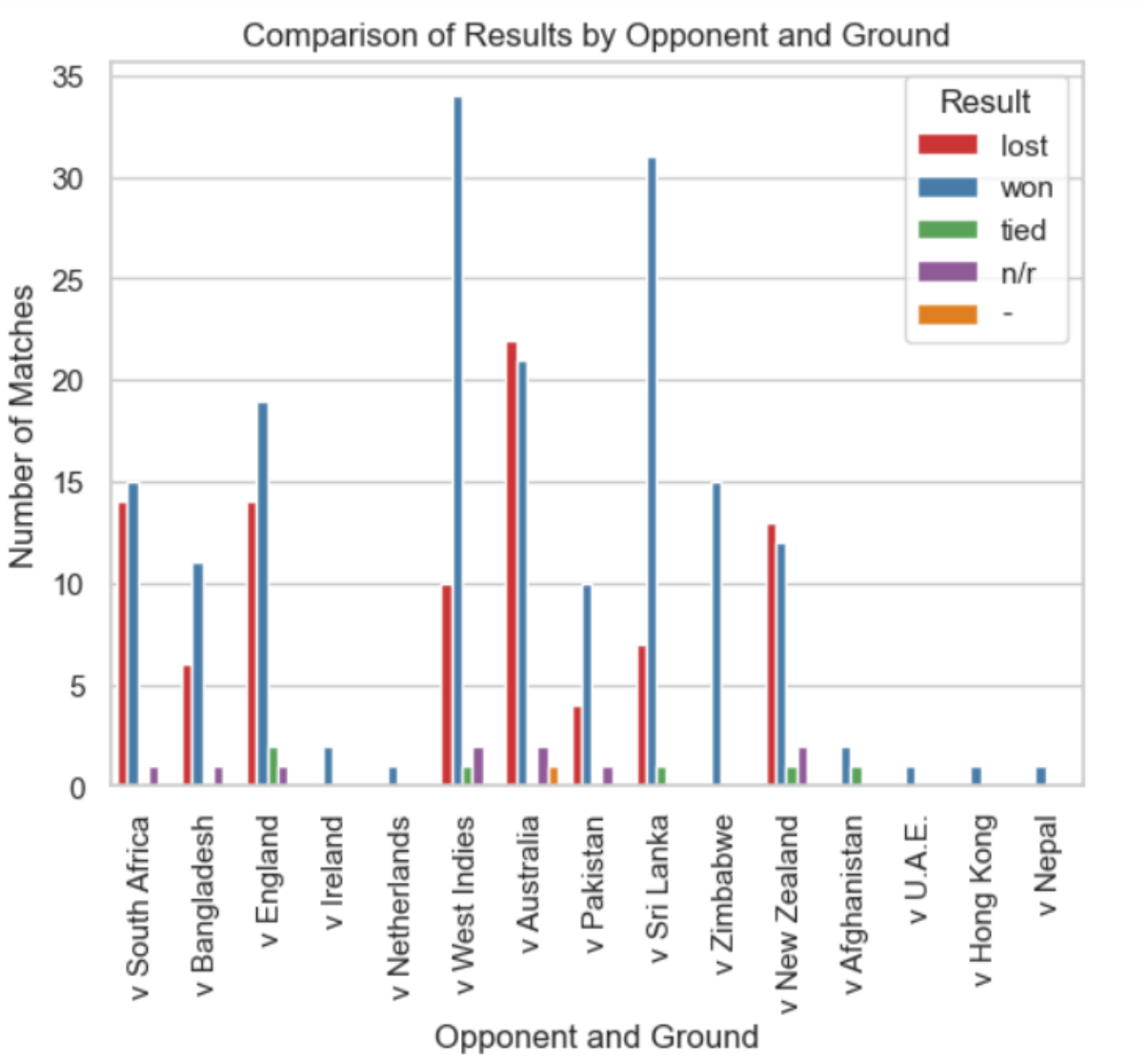
Chart: 3



Insight:

The stacked bar chart analysis compares results against different opponent teams. It indicates that India has the highest number of wins against West Indies. In matches against Australia, India has an equal number of wins and losses. Furthermore, in encounters with teams such as Hong Kong, Ireland, Nepal, Netherlands, and UAE, India has not incurred any losses.

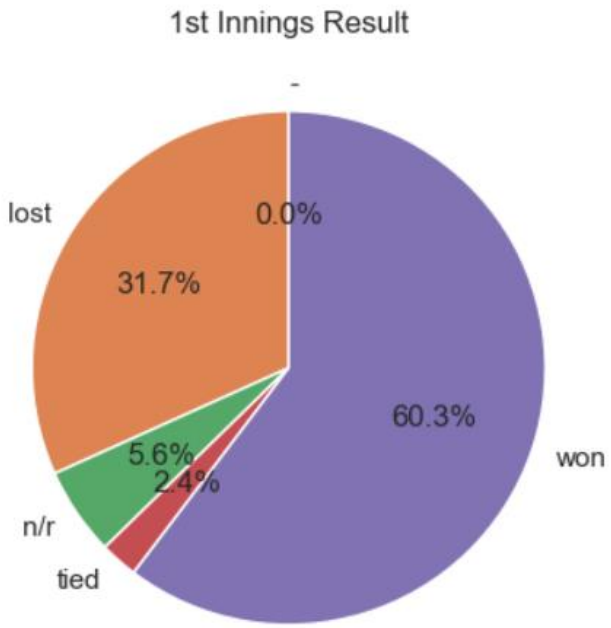
Chart: 4



Insight:

This bar chart indicates that India has the highest number of wins against West Indies and highest number of losses against Australia.

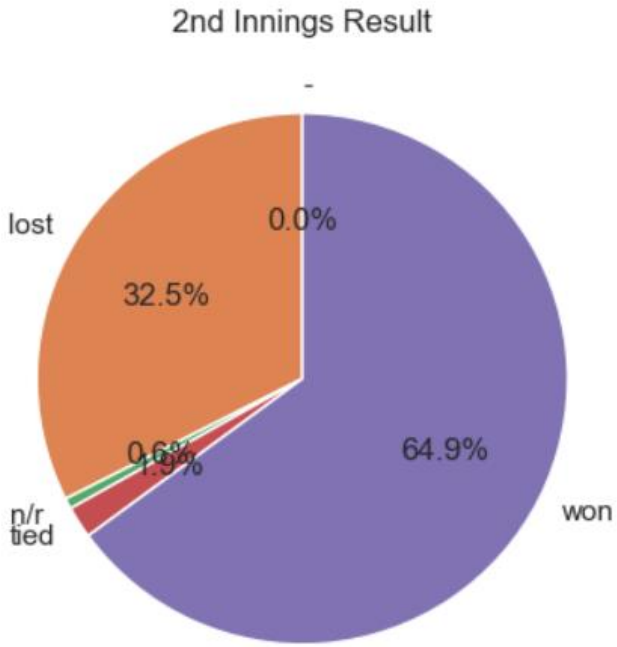
Chart: 5



Insight:

The pie chart analysis reveals that in the first innings, India has the highest winning percentage at 60.3%.

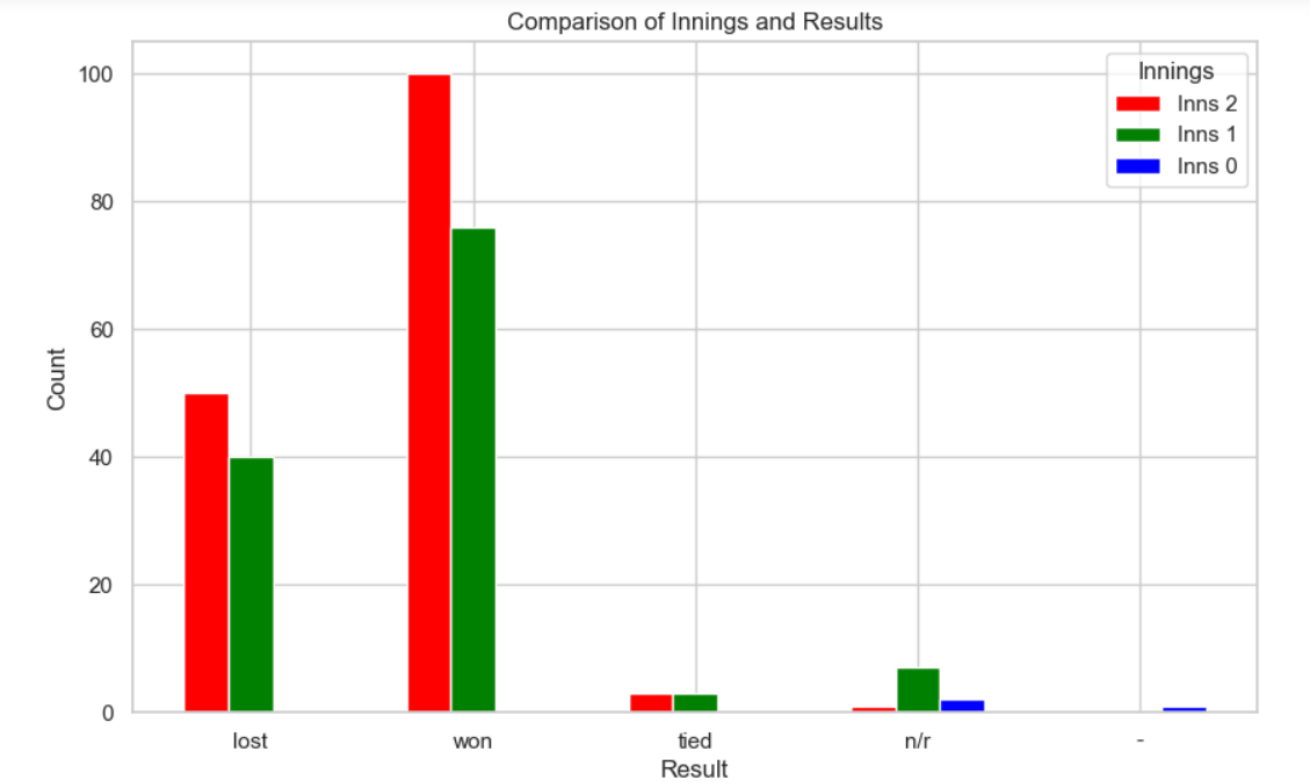
Chart: 6



Insight:

The analysis of pie charts for the first and second innings indicates that India has the highest winning percentage in the second innings, standing at 64.9%. This suggests that whenever Team India plays in the second innings, there is a higher likelihood of winning against the opponent team compared to the first innings.

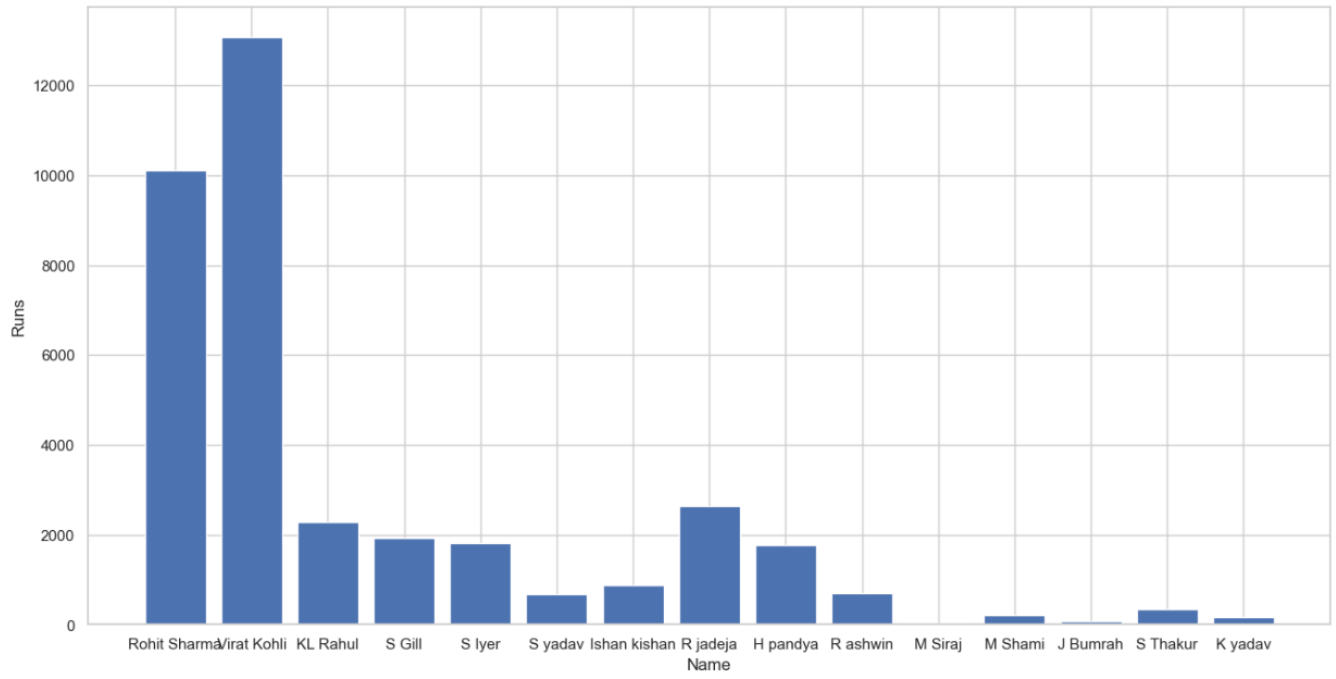
Chart: 7



Insight:

The comparison between innings and results, as depicted in the bar chart, reveals that the Indian team's second innings has the highest number of wins compared to the first innings.

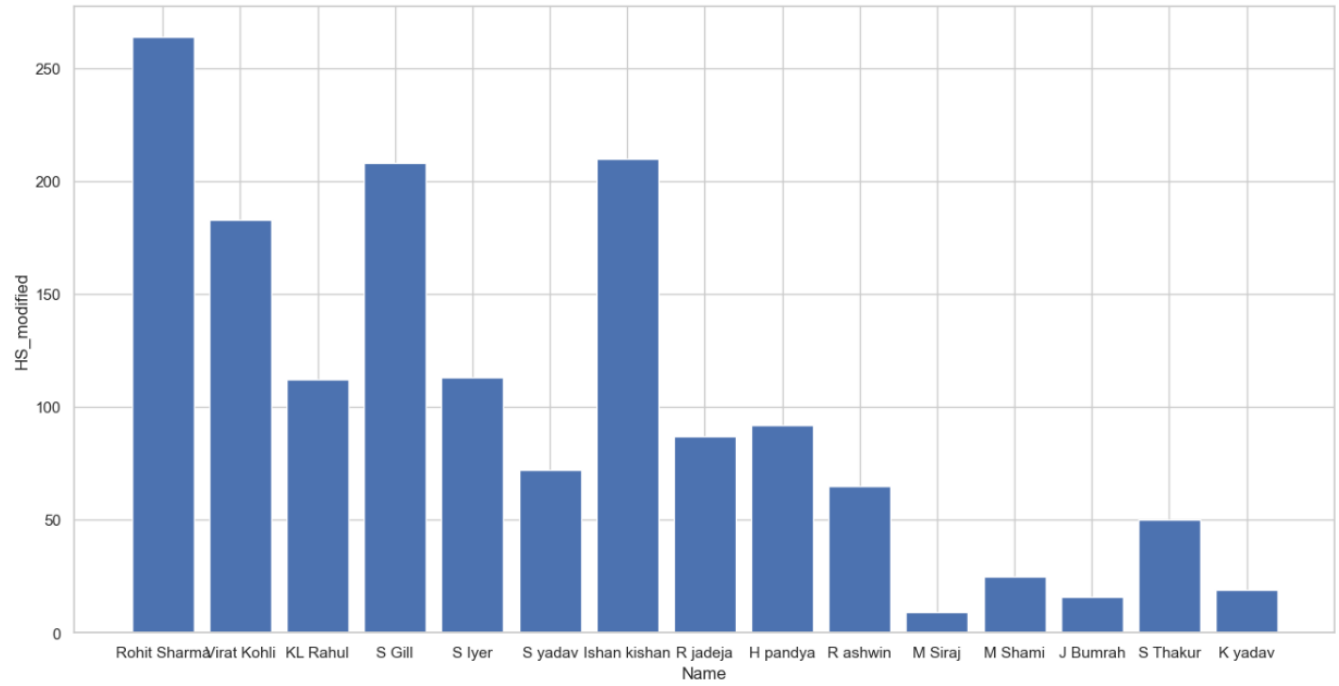
Chart: 8



Insight:

The comparison made between player names and their total scores reveals that Virat Kohli has scored the highest total score for their team, while Rohit Sharma has secured the second-highest run total.

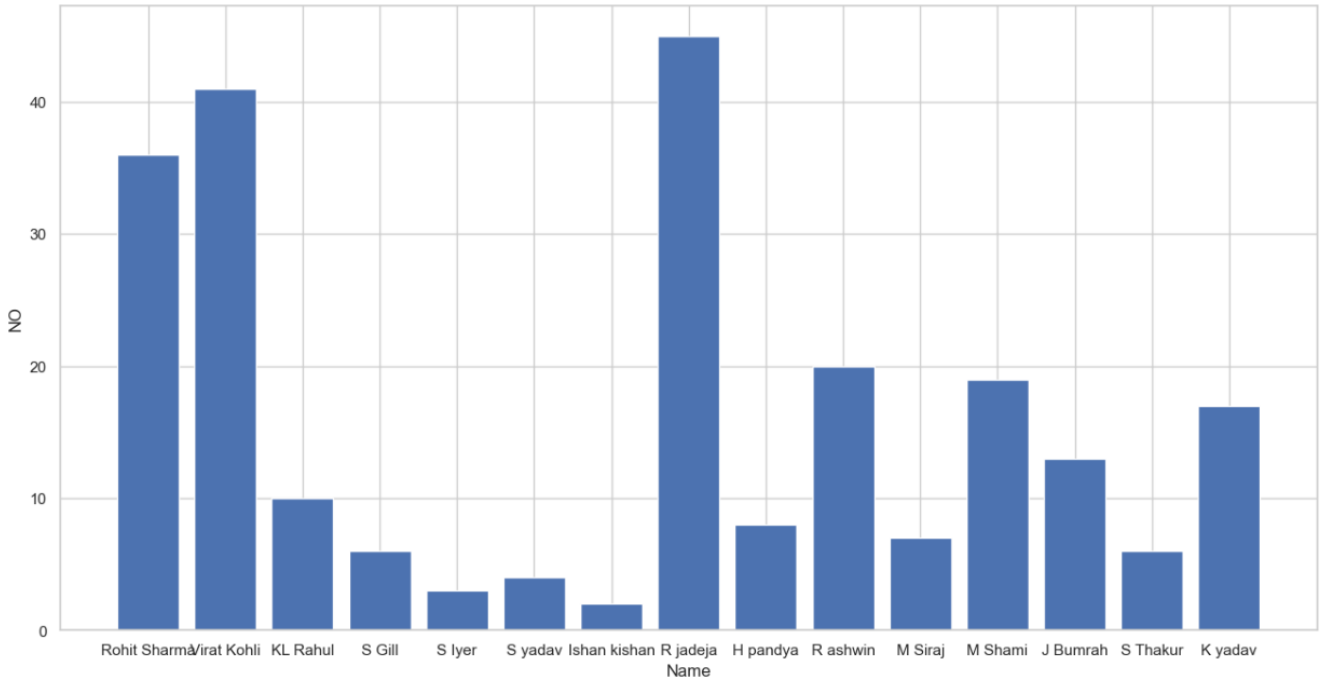
Chart: 9



Insight:

The comparison made between player names and their highest scores in the tournament shows that Rohit Sharma achieved the highest score, while Ishan Kishan and Gill share the second-highest score, holding an equal position.

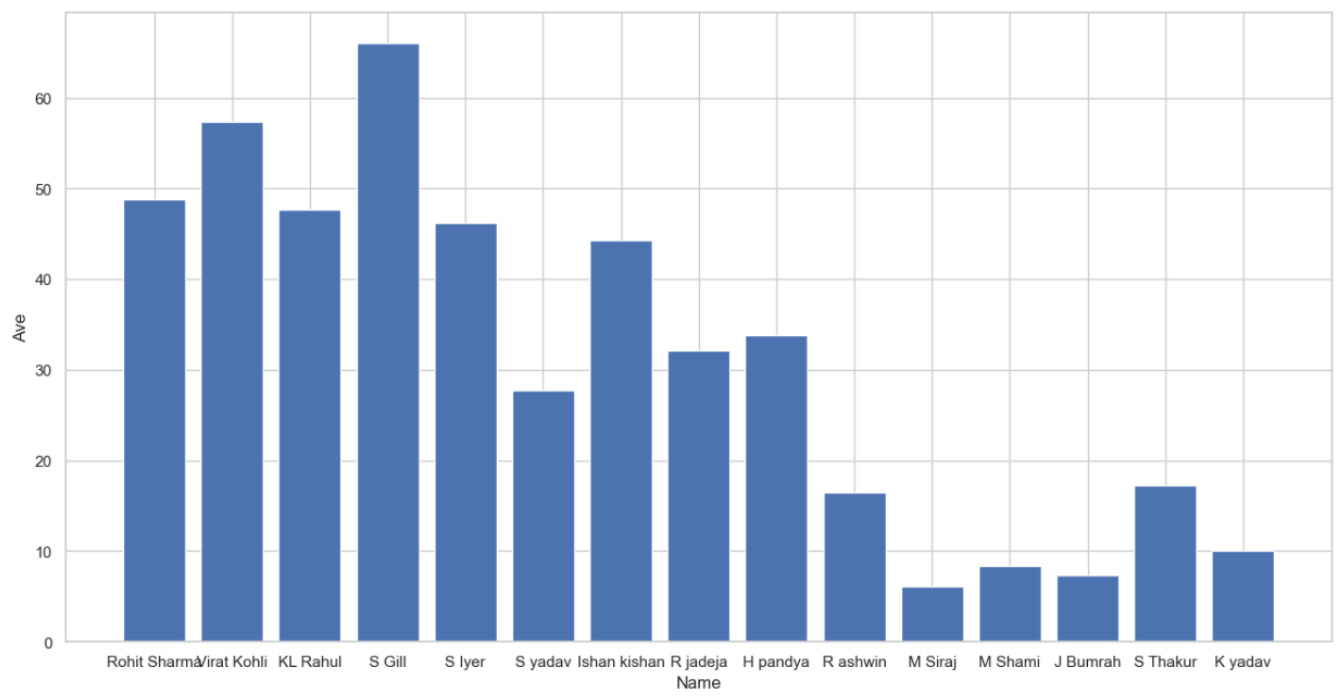
Chart: 10



Insight:

The bar chart analysis, focusing on team players and their "not out" statistics, indicates that Jadeja has the highest number of not outs, implying that he has remained unbeaten in the majority of matches.

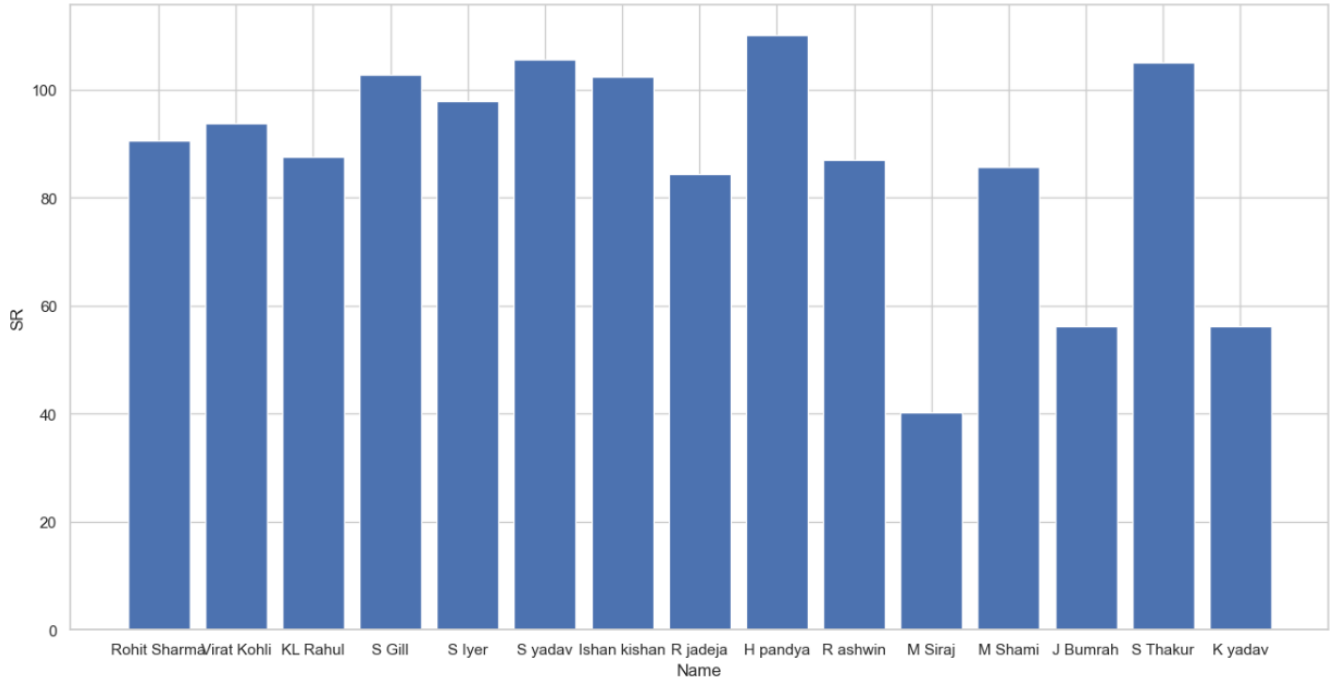
Chart: 11



Insight:

The bar chart analysis compares player names with their average runs, representing the average number of runs a particular player scores in previous matches. It reveals that Gill has the highest average, exceeding 60, while Virat Kohli has the second-highest average, which is close to 60.

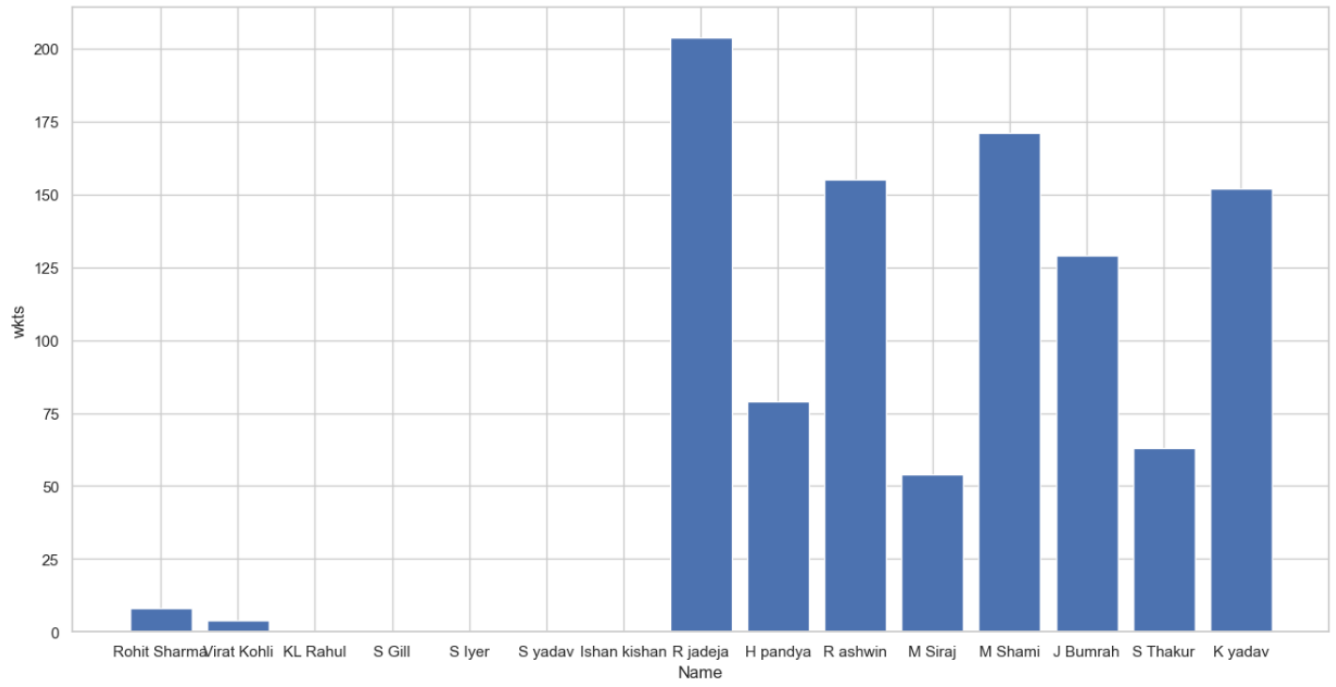
Chart: 12



Insight:

The comparison is made between Indian players and their strike rates, and the bar chart shows that Hardik Pandya has the highest strike rate compared to other players.

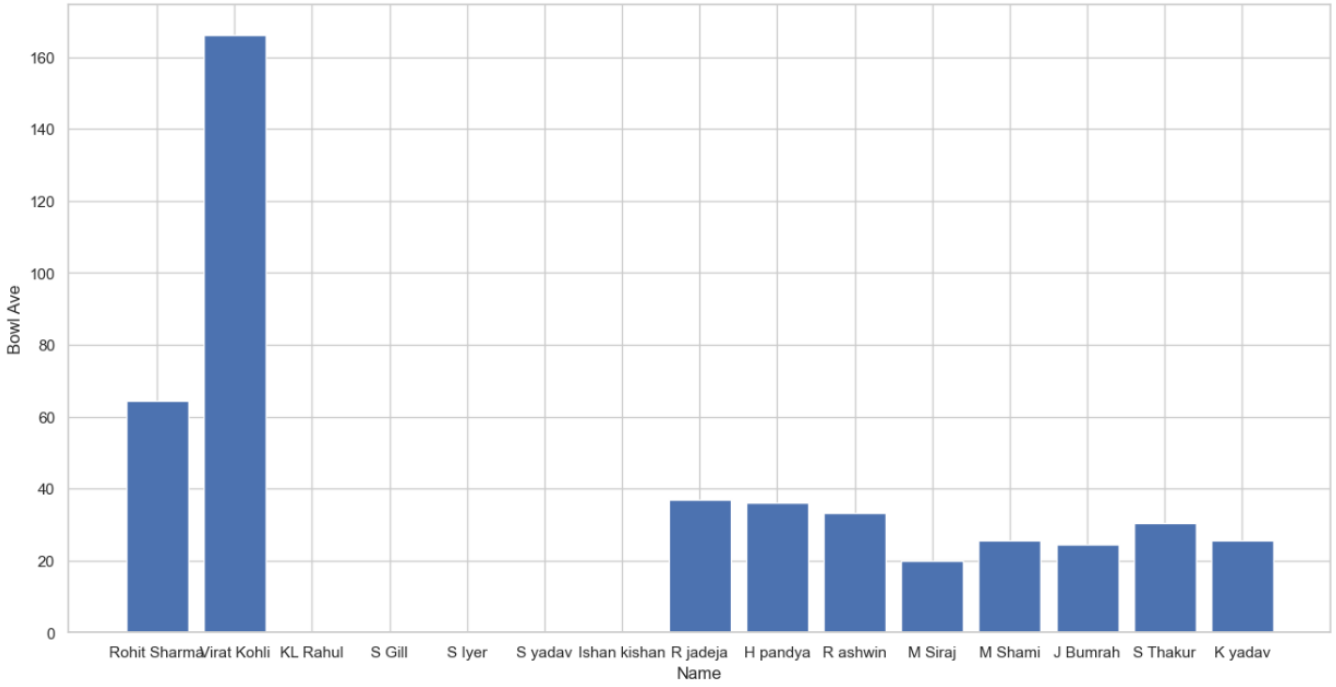
Chart: 13



Insight:

The comparison made between Indian bowlers and the number of wickets taken reveals that Jadeja has the highest number of wickets, surpassing other bowlers.

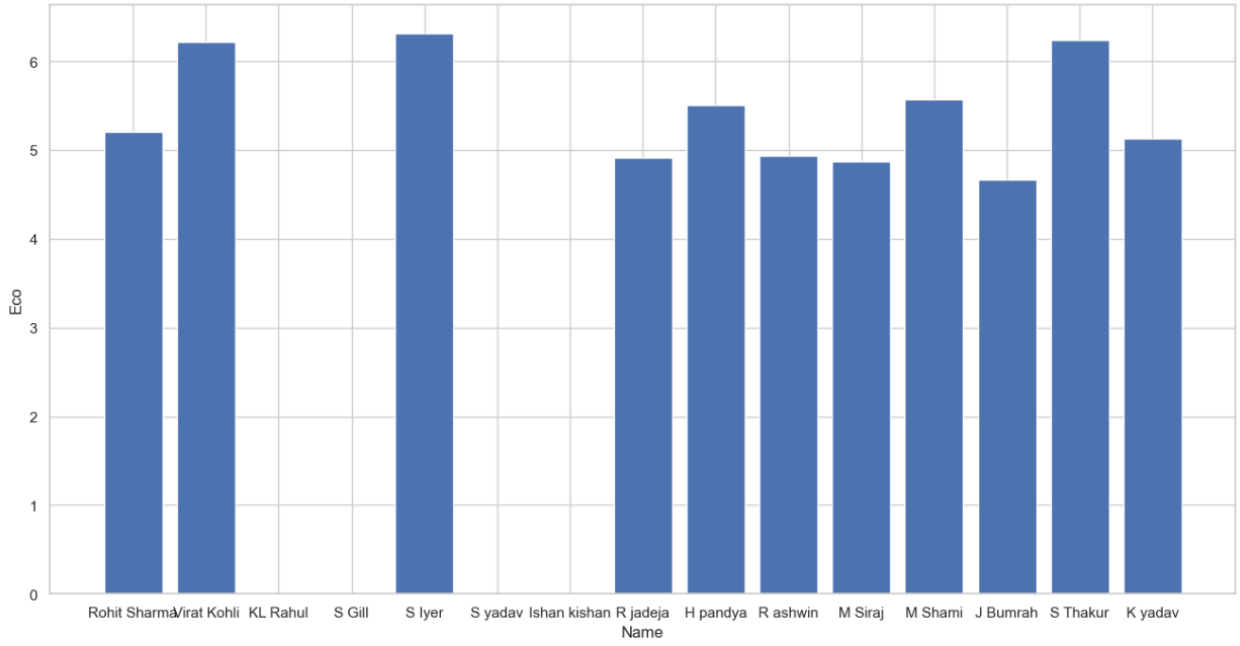
Chart: 14



Insight:

The comparison made between Indian bowlers and their bowling averages reveals that Virat Kohli has the highest bowling average, indicating that they have conceded the most runs per wicket. On the other hand, Siraj has the lowest bowling average, suggesting that he has been more economical and effective in restricting runs during his bowling spells. This performance indicates that Siraj has bowled many dot balls and is considered a reliable bowler for not giving away high runs to the opponent team.

Chart: 15



Insight:

The comparison made between Indian bowlers and their bowling economy indicates that Bumrah has the lowest bowling economy, making him the most economical bowler.

Prediction analysis:

The prediction process involved partitioning the dataset into training and testing sets. Logistic regression was utilized to predict the outcome, taking into account features such as the opponent team, innings, and venue. The target variable for this project is the 'result,' indicating the likelihood of India winning the match based on the opponent team, innings, and venue.

In summary, the primary goal of this project is to predict the probability of India winning a match by considering key factors such as the opponent team, innings, and venue.

8. CONCLUSION

Summary of Findings

In conclusion, the comprehensive analysis reveals significant patterns in India's cricket performance. Notably, India has demonstrated remarkable success against Port of Spain ground, securing the highest number of wins, while Mirpur ground poses a challenge with the highest number of losses. The team has exhibited a propensity for triumph on away grounds compared to their home ground. Against West Indies, India boasts the highest number of victories, whereas encounters with Australia result in an even balance of wins and losses. Furthermore, the data suggests that Team India's second-innings performances often lead to a higher likelihood of victory compared to the first innings.

In terms of individual player performances, Virat Kohli emerges as the leading run-scorer for the team, followed by Rohit Sharma. The highest score accolade goes to Rohit Sharma, while Ishan Kishan and Gill share the second-highest score with equal standing. Jadeja's frequent 'not out' status highlights his resilience on the field. Gill maintains the highest batting average, exceeding 60, while Virat Kohli closely follows with the second-highest average.

Moving to the bowling department, Jadeja excels as the highest wicket-taker, showcasing his proficiency. Siraj stands out with the lowest bowling average, indicating his economic and effective bowling strategy. Bumrah emerges as the most economical bowler with the lowest bowling economy. In summary, these findings offer a nuanced understanding of India's cricket performance, highlighting key strengths and areas for improvement across various aspects of the game.

9. APPENDIX

9.1 Data Extraction

The Indian Performance Analysis dataset was manually collected in the website Statsguru. The dataset used in this project can be accessed at the following link:

https://stats.espncricinfo.com/ci/engine/stats/analysis.html?search=india+performance;template=analysis

The Website screenshot to this project was illustrated below for easy reference:

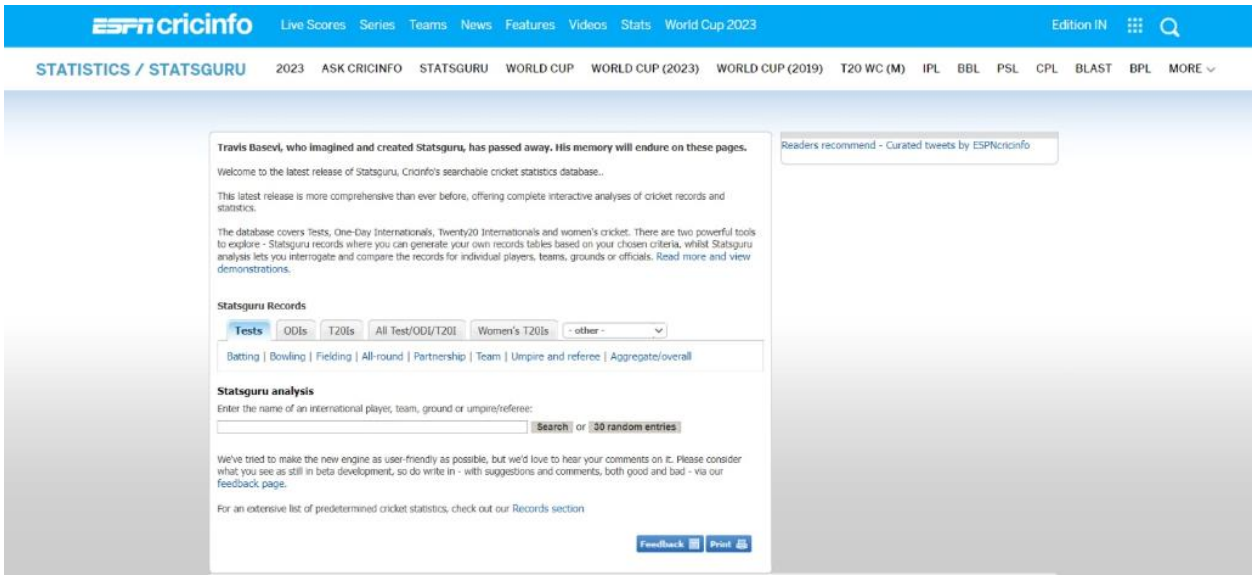


Fig 8.1 Statsguru website for data extraction

9.2Data storage

The Indian Performance dataset was extracted and stored it in CSV format within local library for easy access and analysis

Fig 8.2 Data storage in local directory

9.3Source code

The Python code is used for implementing Linear Regression to analyzed and predict the Indian Player Performance scores.

Importing the datasets

```
Indian_perf = pd.read_csv("Indian performnces.csv")
Indian_team= pd.read_csv("Indian Team.csv")
```

Fig 8.3 Data set importing

Displaying the imported datasets

	Score	Overs	RPO	Target	Inns	Result	Opponent	Venue	Start Date	Year	Scorecard	Win
0	154	35.4	4.31	290.0	2	lost	v South Africa	Durban	12-01-2011	2011	ODI # 3079	0
1	190	47.2	4.01	NaN	1	won	v South Africa	Johannesburg	15-01-2011	2011	ODI # 3080	1
2	223	48.2	4.61	221.0	2	won	v South Africa	Cape Town	18-01-2011	2011	ODI # 3082	1
3	142	32.5	4.32	191.0	2	lost	v South Africa	Gqeberha	21-01-2011	2011	ODI # 3084	0
4	234	40.2	5.80	268.0	2	lost	v South Africa	Centurion	23-01-2011	2011	ODI # 3087	0
5	370	50.0	7.40	NaN	1	won	v Bangladesh	Mirpur	19-02-2011	2011	ODI # 3100	1
6	338	49.5	6.78	NaN	1	tied	v England	Bengaluru	27-02-2011	2011	ODI # 3110	0
7	210	46.0	4.56	208.0	2	won	v Ireland	Bengaluru	06-03-2011	2011	ODI # 3121	1
8	191	36.3	5.23	190.0	2	won	v Netherlands	Delhi	09-03-2011	2011	ODI # 3124	1
9	296	48.4	6.08	NaN	1	lost	v South Africa	Nagpur	12-03-2011	2011	ODI # 3128	0

Fig 8.4 Displaying the dataset

	Span	Mat	Bat Inns	Bowl Inns	NO	Runs	HS	HS_modified	HS_NOT OUT	Name	...	4s	6s	Overs	Mdns	Bow runs	wkts	BBI	Bowl Ave	Eco	Wining_strategy
0	2007-23	251	243	38.0	36	10112	264	264	0	Rohit Sharma	...	928	292	98.5	2	515	8	2\27	64.37	5.21	Select
1	2008-23	281	269	48.0	41	13083	183	183	0	Virat Kohli	...	1226	142	106.5	1	665	4	1\15	166.25	6.22	Select
2	2016-23	61	58	NaN	10	2291	112	112	0	KL Rahul	...	177	52	0.0	0	0	0	0	0.00	0.00	Select
3	2019-23	35	35	NaN	6	1917	208	208	0	S Gill	...	213	40	0.0	0	0	0	0	0.00	0.00	Select
4	2017-23	47	42	5.0	3	1801	113*	113	1	S Iyer	...	176	37	6.1	0	39	0	0	0.00	6.32	Unselect

Fig 8.5 Player performance dataset

Contigency table creation for EDA analysis

```
#Visualize
# Set custom width and height for the figure
fig, ax = plt.subplots(figsize=(15, 20))
contingency_table.plot(kind='barh', stacked=True, ax=ax)
plt.title("Ground vs. Result")
plt.ylabel("Ground")
plt.xlabel("Count")
plt.legend(title="Result")

plt.show()
```

Fig 8.6 Contingency table

Comparing home town wins and away town wins

```
# Create a bar chart to compare the results
plt.figure(figsize=(6,4))
plt.bar(['India Ground', 'Away Ground'], [len(india_home_wins), len(away_home_wins)], color=['grey', 'red'])
plt.title('Home Team Wins vs Away Team Wins')
plt.xlabel('Ground Type')
plt.ylabel('Number of Wins')
plt.show()
```

Fig 8.7 Comparison

Defining feature and target variable

```
# Select features and target
X = Indian_perf[['venue_encoded', 'opposition_encoded', 'Inns']]
y = Indian_perf['Result']
```

Splitting the dataset into training and testing

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Standardize the features

```
# Standardize the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Making the predictions

```
# Make predictions on the test set
y_pred = logistic_regression_model.predict(X_test)
```

Evaluating the model

```
# Evaluate the model
accuracy = logistic_regression_model.score(X_test, y_test)
print(f"Accuracy: {accuracy:.2f}")
```

Result:

```
Accuracy: 0.63
```

Data types of input features

```
# Check data types of user input
print(f"User Input Data Types: {type(user_opponent)}, {type(user_inns)}, {type(user_venue)}")
```

Result:

```
User Input Data Types: <class 'str'>, <class 'int'>, <class 'str'>
```

Standardized input

```
# Check the standardized user input
print(f"Standardized User Input: {user_input_standardized}")
```

Result:

```
Standardized User Input: [[-1.07182738 -1.39007737 12.52549953]]
```

Prediction

```
# Make probability predictions for the user input
probability_prediction = logistic_regression_model.predict_proba(user_input_standardized)
```

Extract the probability prediction

```
# Extract the probability of India winning (class 1)
percentage_chance = probability_prediction[0][1] * 100
```

Prediction output

```
print(f"The predicted percentage chance of India winning is: {percentage_chance:.2f}%")
```

Output:

```
The predicted percentage chance of India winning is: 30.06%
```

10 BIBLIOGRAPHY

(i) https://www.researchgate.net/publication/332429100_Analyzing_the_performance_of_the_Indian_Cricket_Team_using_Weighted_Association_Rule_Mining

(ii) https://www.projectpro.io/article/big-data-analytics-the-new-player-in-icc-world-cup-cricket-2015/89

(iii) https://www.geeksforgeeks.org/ml-linear-regression/

(iv) https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

(v) https://www.statology.org/rmse-vs-r-squared/