Cricket Data Analytics

A Minor Project Report
submitted in partial fulfillment of the requirements for
the award of the degree of

Bachelor of Engineering

in

Artificial Intelligence and Data Science

By

J. Monesh (1601-21-771-039)

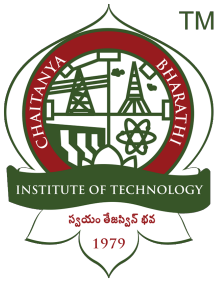
Md. Mushtaq (1601-21-771-050)

M. Gopi Prashanth Raju (1601-21-771-051)

Under the esteemed guidance of

Smt. T. Satya Kiranmai

Assistant Professor



DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075
MAY 2024

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075

INSTITUTE VISION

"To be the center of excellence in technical education and research".

INSTITUTE MISSION

"To address the emerging needs through quality technical education and advanced research".

DEPARTMENT VISION

"To be a globally recognized center of excellence in the field of Artificial Intelligence and Data Science that produces innovative pioneers and research experts capable of addressing complex real-world challenges and contributing to the socio-economic development of the nation."

DEPARTMENT MISSION

1. To provide cutting-edge education in the field of Artificial Intelligence and Data Science that is rooted in ethical and moral values.
2. To establish strong partnerships with industries and research organizations in the field of Artificial Intelligence and Data Science, and to excel in the emerging areas of research by creating innovative solutions.
3. To cultivate a strong sense of social responsibility among students, fostering their inclination to utilize their knowledge and skills for the betterment of society.
4. To motivate and mentor students to become trailblazers in Artificial Intelligence and Data Science, and develop an entrepreneurial mindset that nurtures innovation and creativity.

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

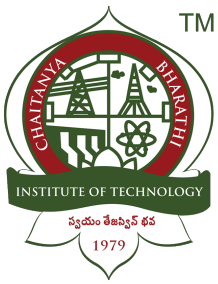
Graduates of AI & DS will be able to:

1. Adapt emerging technologies of Artificial Intelligence & Data Science and develop state-of-the-art solutions in the fields of Manufacturing, Agriculture, Health, Education, and Cyber Security.
2. Exhibit professional leadership qualities to excel in interdisciplinary domains.
3. Possess human values, professional ethics, application-oriented skills, and engage in lifelong learning.
4. Contribute to the research community to meet the needs of public and private sectors.

PROGRAM SPECIFIC OUTCOMES (PSOs)

After successful completion of the program, students will be able to:

1. Exhibit proficiency in Artificial Intelligence and Data Science in providing sustainable solutions by adapting to societal, environmental, and ethical concerns to real-world problems.
2. Develop professional skills in the thrust areas like ANN and Deep learning, Robotics, Internet of Things, and Big Data Analytics.
3. Pursue higher studies in Artificial Intelligence and Data Science in reputed Universities and work in research establishments.



DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
HYDERABAD – 500075

DECLARATION CERTIFICATE

We hereby declare that the project titled **Cricket Data Analytics** submitted by us to the **Artificial Intelligence and Data Science CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY, HYDERABAD** in partial fulfillment of the requirements for the award of **Bachelor of Engineering** is a bona-fide record of the work carried out by us under the supervision of **Smt. T. Satya Kiranmai**. We further declare that the work reported in this project, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Project Associates

J. Monesh (1601-21-771-039)

Md. Mushtaq (1601-21-771-050)

M. Gopi Prashanth Raju (1601-21-771-051)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY

HYDERABAD – 500075

BONAFIDE CERTIFICATE

This is to certify that the project titled **Cricket Data Analytics** is a bonafide record of the work done by

J. Monesh (1601-21-771-039)

Md. Mushtaq (1601-21-771-050)

M. Gopi Prashanth Raju (1601-21-771-051)

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Engineering** in **Artificial Intelligence and Data Science** to the **CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY, HYDERABAD** carried out under my guidance and supervision during the year 2023-24. The results presented in this project report have not been submitted to any other university or Institute for the award of any degree.

Smt. T. Satya Kiranmai
Guide

Dr. Kadiyala Ramana
Head of the Department

Submitted for Semester Mini-Project viva-voce examination held on _____

Examiner-1

Examiner-2

ABSTRACT

Cricket, traditionally reliant on experience and intuition, is undergoing a revolution with the embrace of data analytics. This project delves deeper than basic averages, leveraging advanced statistical methods to unearth hidden insights that can dramatically impact strategic decisions. By analyzing key metrics like strike rate, economy rate, and dismissal patterns, we gain a nuanced understanding of player performance. This allows us to not only identify player strengths and weaknesses but also optimize team selection for specific matches and design targeted training programs to address individual needs, ultimately enhancing both individual capabilities and overall team strategy. Furthermore, the project develops statistical models to predict player performance, empowering teams with the ability to make informed decisions about strategy and in-match adjustments, gaining a crucial edge in the dynamic world of cricket. The incorporation of machine learning techniques allows for the analysis of vast datasets, identifying patterns and trends that may be overlooked by the human eye. This predictive power can also assist in injury prevention by highlighting stress points and potential risk factors for players. As of now, the project uses data exclusively from India vs Australia matches, providing a focused dataset that allows for detailed and specific analysis of these two competitive cricketing nations. Data visualization plays a central role, with charts, graphs, and dashboards transforming complex information into a readily comprehensible format. This empowers coaches, captains, and players to extract actionable insights, leading to informed decisions that translate into winning performances. The interactive nature of these visual tools facilitates real-time analysis and rapid decision-making during matches. Ultimately, this project exemplifies the growing importance of analytics in cricket, demonstrating how data can be transformed into knowledge to elevate player performance, enhance team strategies, and contribute to the evolution of the sport by fostering a more analytical and data-driven approach to cricket management.

ACKNOWLEDGEMENTS

We would like to express our deepest gratitude to the following people for guiding us through this course and without whom this project and the results achieved from it would not have reached completion.

Smt. T. Satya Kiranmai, Assistant Professor, Department of Artificial Intelligence and Data Science, for helping us and guiding us in the course of this project. Without his/her guidance, we would not have been able to successfully complete this project. His/Her patience and genial attitude is and always will be a source of inspiration to us.

Dr. Kadiyala Ramana, the Head of the Department, Department of Artificial Intelligence and Data Science, for allowing us to avail the facilities at the department.

We are also thankful to the faculty and staff members of the Department of Artificial Intelligence and Data Science, our individual parents and our friends for their constant support and help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

In today's cricket landscape, a seismic shift is underway as data analytics takes center stage, reshaping the way teams approach the game. Gone are the days of relying solely on experience and intuition; instead, a data-driven approach is revolutionizing strategic decision-making. This project delves deep into the intricacies of player performance, moving beyond conventional metrics to explore nuanced statistics like strike rate, economy rate, and dismissal patterns. By analyzing these key indicators, teams can gain a comprehensive understanding of each player's strengths and weaknesses, paving the way for optimized team selection tailored to the specifics of each match.

Moreover, this data-centric approach extends beyond team selection to encompass player development and training strategies. By leveraging insights gleaned from statistical analysis, targeted training programs can be devised to address individual player needs and enhance overall performance. Predictive models further augment decision-making, enabling teams to anticipate player performance and adjust strategies accordingly, giving them a crucial advantage in the fast-paced, competitive world of cricket.

Central to this paradigm shift is the role of data visualization, which transforms complex data into actionable insights. Through charts, graphs, and interactive dashboards, coaches, captains, and players alike can easily grasp the implications of the analytics, empowering them to make informed decisions that translate into winning performances on the field. As the influence of analytics continues to grow, this project stands as a testament to the transformative power of data in elevating player performance and refining team strategies, heralding a new era in the sport of cricket.

1.2 Problem Statement

In the realm of cricket, where tradition and experience have long been revered, there's a recognized shortfall in relying solely on intuition and conventional metrics to gauge player performance and formulate effective team strategies. This gap highlights the need for a more sophisticated approach that integrates advanced

statistical methods and data analytics. This project steps into this void with the ambition of developing a comprehensive framework for analyzing cricket data. By transcending traditional averages, it seeks to uncover deeper insights into player capabilities, identifying both strengths and weaknesses, while also shedding light on broader team dynamics that impact strategic decision-making.

At its core, this project seeks to democratize access to actionable insights derived from advanced statistical models and data visualization techniques. By harnessing the power of data analytics, this initiative has the potential to revolutionize the way cricket strategies are formulated and executed, enhancing performance and competitiveness across the board.

It's not just about collecting data; it's about translating that data into meaningful insights that can drive tangible improvements in player performance and team strategies. Through collaboration and innovation, this endeavor aspires to unlock the full potential of data analytics in cricket, ushering in a new era of informed decision-making and elevated standards of play.

1.3 Organization Of The Project

1}Defining the problem: The project begins by precisely outlining the problem statement and objectives, focusing on key questions to be addressed and the scope of analysis. This step ensures clarity and direction, preventing unnecessary deviations during the project.

2}Task Allocation: Responsibilities are assigned to team members based on their expertise, covering data collection, preprocessing, exploratory analysis, model development, and visualization. This allocation optimizes productivity and ensures effective contribution from each team member.

3}Technology Stack and Tools: R programming language is utilized for data analysis, while data is sourced from CSV and JSON files. Tools such as Git for version control and GitHub for collaboration are selected to facilitate seamless communication and workflow management.

4}Data Quality and Model Performance Checks: Continuous evaluation of data quality and model performance is conducted to ensure accuracy and reliability. Checks are implemented to address issues like missing values and outliers, while model performance metrics are monitored and evaluated.

5}Visualization and Reporting: Visualizations, including charts and graphs, are created to present insights clearly. Comprehensive reports summarizing findings, methodologies, and recommendations are prepared to empower stakeholders with informed decision-making capabilities.

CHAPTER 2

SYSTEM REQUIREMENTS AND SPECIFICATIONS

2.1 Functional Requirements

Functional requirements are the requirements that define specific behaviors or functions of the system.

- **Data Conversion and Preprocessing:** Convert JSON files to CSV format for easier processing and analysis. Extract relevant attributes such as batter, bowler, runs per ball, total score, extras, wickets, winner, toss winner, and decision.
- **Player Performance Analysis:** Extract and analyze player performance metrics, including runs scored by each player and wickets taken by bowlers. Generate separate CSV files containing player details, such as batting and bowling statistics.
- **Player Details Extraction:** Utilize APIs to extract player details, including country, batting hand, and bowling style. Generate a comprehensive list of player details for both Australian and Indian players involved in the matches.

2.2 Non-Functional Requirements

- **Security:** The system should implement appropriate security measures to protect sensitive data and ensure confidentiality, integrity, and availability. Access controls, encryption, and data anonymization techniques should be employed to safeguard the dataset and user privacy.
- **Performance:** The system should be capable of efficiently processing and analyzing a large volume of data contained in 83 JSON files. Data conversion, preprocessing, and analysis tasks should be completed within a reasonable time frame to ensure timely insights and decision-making.
- **Reliability:** The system should accurately handle and process the data without loss or corruption. Data extraction, transformation, and loading processes should be robust and resilient to errors, ensuring the integrity of the dataset throughout the analysis.

- **Scalability:** The system should be scalable to accommodate potential increases in data volume or complexity. It should be able to handle additional cricket matches or datasets without significant performance degradation, allowing for future expansion and analysis.
- **Maintainability:** The system should be easy to maintain and update, allowing for future enhancements or modifications. Codebase documentation, version control practices, and modular design principles should be followed to support ongoing development and maintenance activities.

2.3 Software Requirements

- Operating System: Windows 7 or above, Mac OS, Linux
- Visual Studio Code
- RStudio

2.4 Hardware Requirements

- x86 64-bit CPU (Intel / AMD architecture)
- 4 GB RAM
- 5 GB free disk space

CHAPTER 3

LITERATURE SURVEY

3.1 Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach

Parameter	**Details**
Year of Publication	2016
Conference / Journal Details	IEEE
Dataset Used	ESPN Cricinfo (2010-2014 ODI matches)
Description of the Dataset	Team data, Player data (career statistics and recent performance)
URL of the Dataset	-
Tasks Carried Out	Predicting the winner of an ODI cricket match
Algorithms Used for Each Task	Batsmen Score, Bowler Score, Team Strength
Results Obtained for Each Task	Achieved team-wise winning accuracy between 68% and 70%
Gaps Reported in the Paper	Lacks data on external factors like weather and ground design. Player performance data only considers matches played between 2010-2014. Inability to compare the model against previous models due to different underlying datasets.

Table 3.1: Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach

3.2 Predicting Optimal Cricket Team using Data Analysis

Parameter	**Details**
Year of Publication	2021
Conference / Journal Details	2021 International Conference on Emerging Smart Computing and Informatics (ESCI)
Dataset Used	Not specified
Description of the Dataset	Cricket player and team statistics (batting, bowling, and wicket-keeping parameters)
URL of the Dataset	-
Tasks Carried Out	Data Collection, Data Preprocessing
Algorithms Used for Each Task	Players are selected based on the highest batting average
Results Obtained for Each Task	Assigned weights to different player attributes based on venue and importance. Developed formulae to calculate a resultant score for players on different pitches.
Gaps Reported in the Paper	Specific details and size of the dataset are not available. Algorithms used for weight calculation and resultant scores are not mentioned. Ignores factors like player form, injuries, and team dynamics.

Table 3.2: Predicting Optimal Cricket Team using Data Analysis

3.3 Cricket Player Profiling: Unraveling Strengths and Weaknesses Using Text Commentary Data

Parameter	**Details**
Year of Publication	2023
Conference / Journal Details	IEEE
Dataset Used	Text commentary of T20 International matches (2008-2019)
Description of the Dataset	Text commentary data containing information on batsman names, scores, dismissals (if any), type of delivery (length), and where the ball was hit on the field.
URL of the Dataset	-
Tasks Carried Out	Identify batsman strengths and weaknesses. Calculate region-wise strike rate
Algorithms Used for Each Task	Regular expressions to identify patterns in the commentary text
Results Obtained for Each Task	Specific results related to batsman strengths and weaknesses or strike rates by region aren't provided in the summary.
Gaps Reported in the Paper	The dataset only focuses on T20 Internationals. Extending the analysis to include other cricket formats (Test and ODI). Calculating bowler-specific strike rates for batsmen.

Table 3.3: Cricket Player Profiling: Unraveling Strengths and Weaknesses Using Text Commentary Data

3.4 Cricket Team Prediction Using Machine Learning Techniques

Parameter	**Details**
Year of Publication	2011
Conference / Journal Details	Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, India
Dataset Used	Not explicitly mentioned, but authors suggest scraping data from websites like ESPN Cricinfo.
Description of the Dataset	Cricket player statistics including batting and bowling averages, strike rates, etc.
URL of the Dataset	-
Tasks Carried Out	Predict the best 11 players for a team
Algorithms Used for Each Task	Data scraping (web scraping), Random Forest Classifier
Results Obtained for Each Task	The system outputs a recommended team of 11 players based on their historical data and factors like the opposition team and venue.
Gaps Reported in the Paper	Data limitations: Authors acknowledge the lack of data on factors like weather and ground design. Dataset used only considers historical data up to 2017. Model improvements: Incorporating additional parameters like player fitness and location-based performance. Experimenting with other machine learning algorithms like XGBoost.

Table 3.4: Cricket Team Prediction Using Machine Learning Techniques

3.5 Analysis and Winner Prediction of Cricket Match

Parameter	**Details**
Year of Publication	2022
Conference / Journal Details	International Journal of Research Publication and Reviews (Volume 3, Issue 4, pp 1626-1632)
Dataset Used	Not specified in the paper
Description of the Dataset	IPL match data including details like teams, toss winners, scores, and winners.
URL of the Dataset	-
Tasks Carried Out	Exploratory Data Analysis (EDA), Winner Prediction - Score based method
Algorithms Used for Each Task	Linear Regression, Random Forest Regression, Decision Tree
Results Obtained for Each Task	Random Forest achieved the highest R-squared score (0.7516) for score-based prediction. Random Forest achieved the highest accuracy (88.23%) for toss-based prediction.
Gaps Reported in the Paper	Specific features employed for model training are not mentioned. Performance metrics (MAE) are only reported for score-based prediction using linear regression. The paper doesn't discuss limitations of the models or potential improvements.

Table 3.5: Analysis and Winner Prediction of Cricket Match

3.6 Enhancing Cricket Performance Analysis with Human Pose Estimation and Machine Learning

Parameter	**Details**
Year of Publication	2023
Conference / Journal Details	Sensors (MDPI)
Dataset Used	Cricket video dataset
Description of the Dataset	Videos featuring eight different batsman strokes: pull, cut, cover drive, straight drive, backfoot punch, on drive, flick, and sweep. Each video frame has 17 key points extracted related to the batsman's body pose using MediaPipe library.
URL of the Dataset	-
Tasks Carried Out	Stroke prediction for batsman
Algorithms Used for Each Task	Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Logistic Regression (LR), Long Short-Term Memory (LSTM)
Results Obtained for Each Task	Random Forest achieved the highest accuracy (99.77%) for stroke prediction. k-fold validation of the RF model yielded 95.0% accuracy with a standard deviation of 0.07.
Gaps Reported in the Paper	The paper doesn't mention the size of the video dataset. The dataset itself is not publicly available.

Table 3.6: Enhancing Cricket Performance Analysis with Human Pose Estimation and Machine Learning

3.7 Cricket Analytics and Prediction Using Machine Learning

Parameter	**Details**
Year of Publication	2019
Conference / Journal Details	International Journal of Scientific and Engineering Research (Volume 10, Issue 12, pp. 158-165)
Dataset Used	2008-2019 IPL seasons
Description of the Dataset	The dataset is split into two main categories: player data and match data. Player data includes runs scored, wickets taken, strike rates, etc. Match data includes match outcomes, toss results, and venue details.
URL of the Dataset	-
Tasks Carried Out	Player Analysis: Identifying key players and their contribution to match outcomes. Team Analysis: Analysis of team strategies (e.g., how often a team wins after winning the toss) and identification of successful patterns (e.g., batting first or second).
Algorithms Used for Each Task	Player Analysis: Statistical functions (mean, median, mode). Team Analysis: Statistical functions, Correlation analysis
Results Obtained for Each Task	Player Analysis: Identification of key players and their contribution to match outcomes. Team Analysis: Analysis of team strategies (e.g., how often a team wins after winning the toss) and identification of successful patterns (e.g., batting first or second).
Gaps Reported in the Paper	Limitations of the Data Collection Methodology: Possible data inaccuracies due to scraping errors or inconsistencies in the source website's data format. Lack of advanced analytics: Basic statistical functions used for player and team analysis, with no advanced machine learning models employed. Dataset constraints: Analysis limited to 2008-2019 seasons, and does not account for recent player performances or changes in team strategies beyond this period.

Table 3.7: Cricket Analytics and Prediction Using Machine Learning

3.8 Artificial Intelligence in Cricket

Parameter	**Details**
Year of Publication	2023
Conference / Journal Details	2023 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)
Dataset Used	Cricbuzz API
Description of the Dataset	Player statistics including batting averages, bowling averages, strike rates, etc.
URL of the Dataset	`https://www.cricbuzz.com`
Tasks Carried Out	Team Selection Prediction
Algorithms Used for Each Task	Machine Learning: Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN)
Results Obtained for Each Task	Random Forest achieved the highest accuracy (95.3%) for team selection prediction.
Gaps Reported in the Paper	The dataset doesn't include external factors such as weather conditions, ground conditions, and player injuries. Team selection algorithm only considers player statistics. Future work should integrate more real-time data and improve feature engineering.

Table 3.8: Artificial Intelligence in Cricket

CHAPTER 4

SYSTEM DESIGN OR METHODOLOGY

The system follows a structured conceptual model that encompasses the system's structure, behavior, and views. Here is an outline of the process:

- **Data Collection and Preprocessing:** Acquire cricket match data from reliable sources such as official cricket databases or APIs. Preprocess the data to ensure consistency and usability, including tasks such as data cleaning, handling missing values, and formatting data into a suitable structure for analysis.
- **Feature Engineering:** Identify relevant features from the dataset that can contribute to the analysis and decision-making process. Engineer new features if necessary to enhance the predictive power of the models, such as calculating batting averages, strike rates, bowling averages, and economy rates.
- **Model Development:** Utilize statistical and machine learning techniques to develop models for various tasks, such as predicting player performance, team outcomes, and match results. Select appropriate algorithms based on the nature of the prediction task, such as regression for continuous outcomes or classification for categorical outcomes. Train the models using historical cricket data, validating their performance through cross-validation techniques.
- **Data Visualization and Interpretation:** Visualize the insights derived from the data using plots, charts, and dashboards to make them understandable to stakeholders. Interpret the visualizations to extract actionable insights regarding player performance, team strategies, and match dynamics. Provide interactive visualization tools where necessary to allow users to explore the data and gain deeper insights.
- **Validation and Testing:** Validate the models and analytical findings using appropriate evaluation metrics and testing methodologies. Conduct sensitivity analysis to assess the robustness of the models and their performance under different scenarios. Iterate on the model development and refinement process based on the validation results to improve accuracy and reliability.

CHAPTER 5

IMPLEMENTATION

5.1 Introduction

The aim of this project is to predict cricket scores based on historical match data. We utilize machine learning techniques, specifically linear regression and hybrid models, to achieve this goal. The process involves reading and preprocessing the data, training models, evaluating their performance, and making predictions.

5.2 Data Collection and Preprocessing

1. **Data Collection:**
    - We begin by acquiring cricket match data from reliable sources such as official cricket databases or APIs. This data includes detailed records of player statistics, match outcomes, and various performance metrics.
    - Ensuring the data covers multiple aspects of the game is crucial for comprehensive analysis.
2. **Data Cleaning:**
    - Next, we handle missing values through appropriate imputation methods or by removing incomplete entries if necessary.
    - We standardize the format of data entries to ensure consistency across the dataset, which is essential for accurate analysis.
3. **Data Transformation:**
    - We convert categorical variables into numerical formats where necessary, facilitating their use in statistical models.
    - Additionally, we normalize or scale numerical data to prepare it for analysis, ensuring that all features contribute equally to the model.

5.3 Feature Engineering

1. **Feature Identification:**

- We identify key features that influence cricket performance. These include batting averages, strike rates, bowling averages, and economy rates.
- Selecting the right features is critical for building effective predictive models.

2. **Feature Creation:**

- We engineer new features to enhance the predictive power of our models. For example, we calculate complex metrics from raw data, such as moving averages or player consistency scores.
- These new features help capture underlying patterns and trends that might not be apparent from basic statistics alone.

5.4 Model Development

1. **Model Selection:**

- We choose appropriate algorithms based on the nature of the prediction task. For example, we use regression for continuous outcomes and classification for categorical outcomes.
- The choice of algorithm is guided by the specific goals of our analysis and the characteristics of the data.
- We select relevant features and the target variable for the prediction task. In this section, we train a linear regression model using the 'caret' package.

2. **Hybrid Models:**

- We implement hybrid models that combine different machine learning techniques to improve predictive accuracy. For example, we use ensemble methods like Random Forests, which combine multiple decision trees, and Gradient Boosting, which builds models sequentially to correct errors from previous models.
- By leveraging the strengths of various algorithms, hybrid models help in capturing complex relationships within the data.

3. **Training:**

- We split the data into training and testing sets to develop our models.
- The models are trained using historical cricket data, which allows them to learn patterns and relationships within the data.

4. **Validation:**

   - After training the model, we evaluate its performance using the test dataset. We calculate the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
   - We validate our models using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics help us assess the performance and reliability of our models.
   - Based on the validation results, we iterate on the model development process to refine and improve accuracy and reliability.

5.5 Data Visualization and Interpretation

1. **Visualization Tools:**

   - We used a variety of visualization tools to present our findings, including matplotlib and seaborn libraries in Python for creating static visualizations. For interactive visualizations, we utilized Plotly and Tableau.
   - These tools were chosen for their ability to handle large datasets and provide clear, insightful representations of the data.

2. **Kernel Density Estimation (KDE):**

   - We used KDE plots to estimate the probability density function of continuous variables like batting averages and strike rates. KDE helped us smooth out the distribution and identify the underlying patterns in player performance.
   - By overlaying KDE plots for different players or teams, we were able to compare their performance distributions and identify key differences.

3. **Violin Plots:**

   - Violin plots were used to visualize the distribution of performance metrics across different groups. For example, we compared the distribution of strike rates for different players or teams.
   - These plots combined the benefits of box plots and KDE, showing both the distribution and the summary statistics (median, quartiles) in a single visualization.

4. **Clustered Bar Charts:**

- Clustered bar charts were employed to compare categorical data across multiple groups. For example, we used them to compare the number of runs scored or wickets taken by different players in different matches.
- These charts helped in visualizing the differences and similarities between groups, making it easier to identify standout performances.

5. **Joint Distribution Plots:**

   - Joint distribution plots were used to examine the relationship between two continuous variables, such as strike rate and batting average. These plots helped us understand how two performance metrics were related and identify any underlying trends.
   - By plotting the joint distribution, we could also identify clusters and outliers, providing deeper insights into player performance.

6. **Interactive Bubble Charts:**

   - Interactive bubble charts were used to visualize multi-dimensional data. For example, we plotted players' performances with bubble sizes representing the number of matches played and colors representing different teams.
   - These interactive charts allowed stakeholders to explore the data dynamically, filter by different criteria, and gain insights into player performances and team dynamics in a more engaging way.

7. **Interpretation:**

   - The visualizations were interpreted to extract actionable insights regarding player performance, team strategies, and match dynamics. For example, KDE plots helped in identifying consistent performers, while joint distribution plots highlighted the correlation between different performance metrics.
   - Interactive dashboards allowed stakeholders to explore different scenarios and outcomes, enhancing their ability to make informed decisions. For example, a coach could use the dashboard to simulate different team compositions and predict their potential performance.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Introduction

This chapter presents the results obtained from the analysis of cricket data between India and Australia. Various visualization techniques were employed to gain insights into the performance metrics of players from both teams.

6.1.1 Linear Regression Model

The scatter plot in Figure 6.1 shows the actual values plotted against the predicted values for the Linear Regression model. The red dashed line represents the ideal case where the predicted values exactly match the actual values. As observed, there is a significant deviation from the ideal line, indicating that the linear regression model does not capture the relationship between the variables very well.

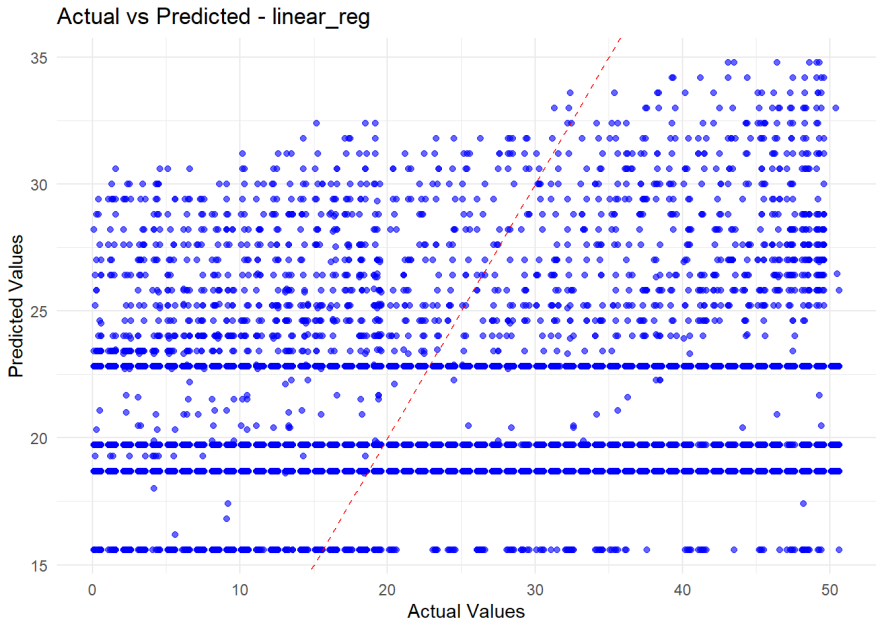


Figure 6.1: Actual vs Predicted values for the Linear Regression model

6.1.2 Random Forest Model

The scatter plot in Figure 6.2 shows the actual values plotted against the predicted values for the Random Forest model. Similar to the previous plot, the red dashed line represents the ideal case. The Random Forest model appears to have a slightly

better alignment with the ideal line compared to the Linear Regression model, although there are still notable deviations.

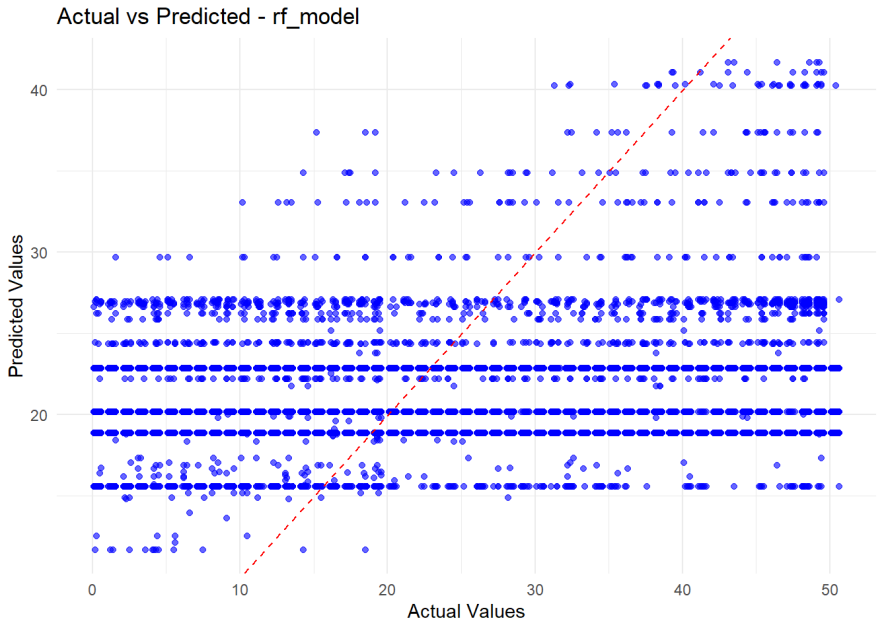


Figure 6.2: Actual vs Predicted values for the Random Forest model

6.1.3 XGBoost Model

The scatter plot in Figure 6.3 shows the actual values plotted against the predicted values for the XGBoost model. The red dashed line represents the ideal case where the predicted values exactly match the actual values. As observed, the XGBoost model also deviates significantly from the ideal line, indicating that it struggles to capture the relationship between the variables accurately.

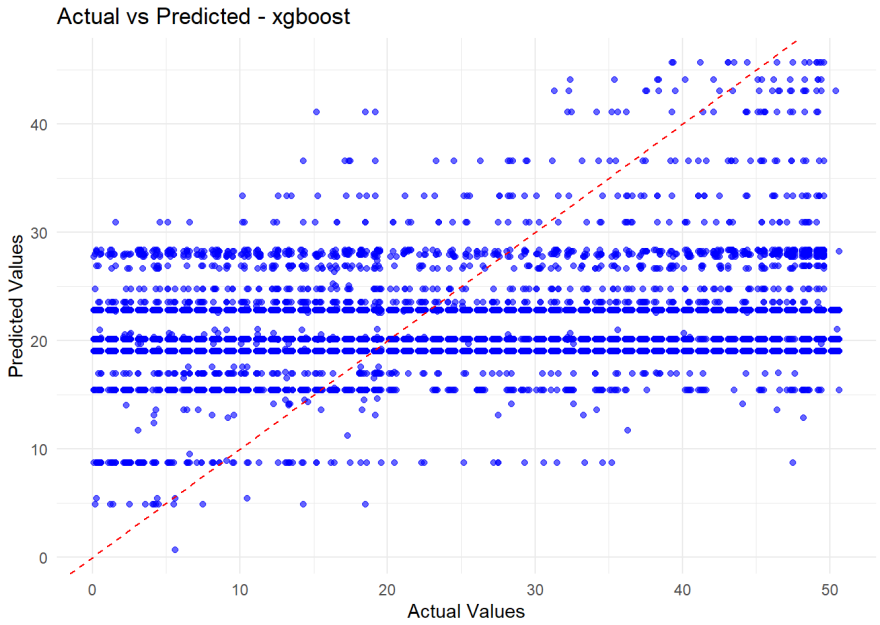


Figure 6.3: Actual vs Predicted values for the XGBoost model

6.1.4 Comparison of Models

Comparing Figures 6.1, 6.2, and 6.3, we can observe the following:

- **Linear Regression Model:** The scatter plot shows a significant deviation from the ideal line, indicating that the linear regression model does not capture the relationship between the variables well.
- **Random Forest Model:** The scatter plot for the Random Forest model shows a slightly better alignment with the ideal line compared to the Linear Regression model. However, there are still notable deviations, suggesting limitations in its predictive power.
- **XGBoost Model:** The scatter plot for the XGBoost model also shows significant deviations from the ideal line. While it exhibits some clustering of predicted values, it does not perform substantially better than the other models.

6.1.5 Linear Regression Model Residuals

The histogram in Figure 6.4 shows the distribution of residuals (errors) for the Linear Regression model. Residuals are the differences between the actual values and the predicted values. The residual distribution provides insight into the performance of the model.

As observed, the residuals for the Linear Regression model are roughly centered around zero, with a relatively symmetrical distribution. This suggests that the model does not exhibit a significant bias in its predictions. However, the spread of the

residuals indicates that there is considerable variability, implying that the model's predictions are not consistently accurate. The presence of large residuals on both the positive and negative sides indicates that the model struggles to accurately capture the relationship between the variables.

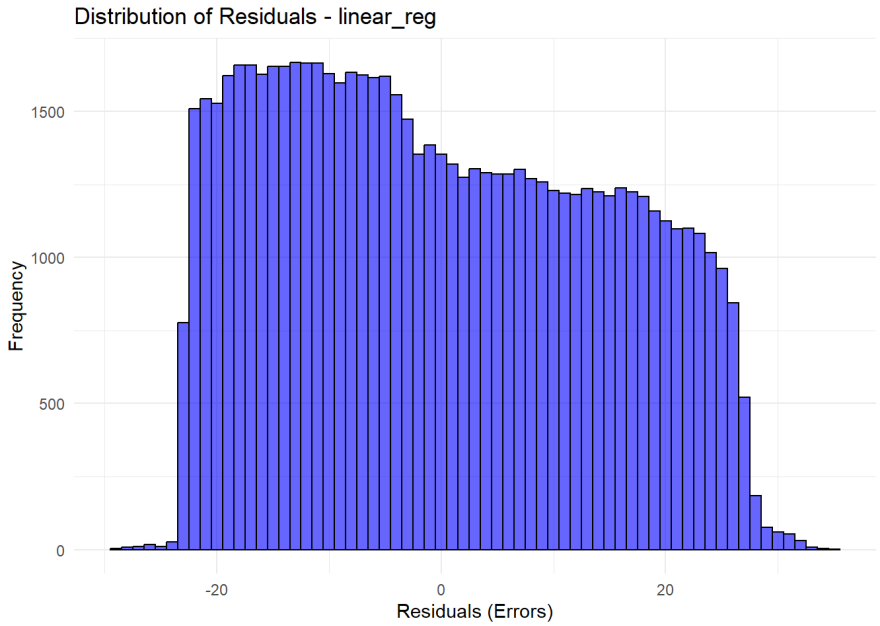


Figure 6.4: Distribution of Residuals for the Linear Regression model

6.1.6 Random Forest Model Residuals

The histogram in Figure 6.5 shows the distribution of residuals for the Random Forest model. Similar to the Linear Regression model, the residuals are the differences between the actual values and the predicted values.

The residual distribution for the Random Forest model is also centered around zero, indicating no significant bias in the predictions. The shape of the distribution suggests that the residuals are more tightly clustered around zero compared to the Linear Regression model, which implies that the Random Forest model produces more accurate predictions. However, there is still a noticeable spread, indicating some variability in the model's predictions. The presence of residuals with larger magnitudes on both sides suggests that while the Random Forest model performs better than the Linear Regression model, it still has limitations in accurately capturing the underlying relationships in the data.

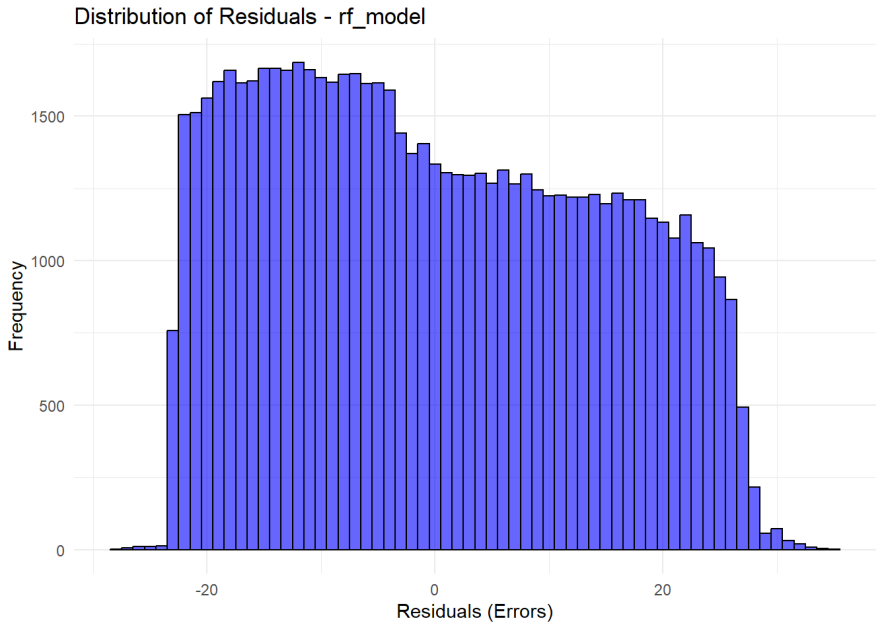


Figure 6.5: Distribution of Residuals for the Random Forest model

6.1.7 XGBoost Model Residuals

The histogram in Figure 6.6 shows the distribution of residuals for the XGBoost model. As with the previous models, the residuals are the differences between the actual values and the predicted values.

The residual distribution for the XGBoost model is centered around zero, indicating no significant bias in the predictions. The shape of the distribution suggests that the residuals are more tightly clustered around zero compared to both the Linear Regression and Random Forest models. This implies that the XGBoost model produces more accurate predictions. However, there is still some spread, indicating variability in the model's predictions. The presence of residuals with larger magnitudes on both sides suggests that, despite its overall superior performance, the XGBoost model still has some limitations in capturing the underlying relationships in the data.

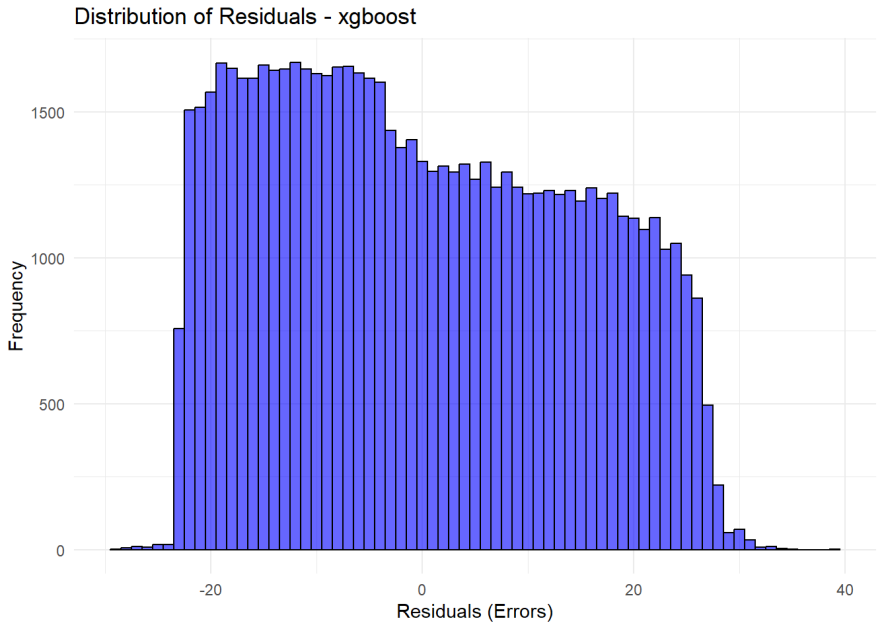


Figure 6.6: Distribution of Residuals for the XGBoost model

6.1.8 Comparison of Residual Distributions

Comparing the residual distributions of the three models, we can observe the following:

- **Linear Regression Model:** The residuals are centered around zero and exhibit a relatively symmetrical distribution. However, there is considerable variability in the residuals, indicating that the model's predictions are not consistently accurate. The presence of large residuals on both sides suggests difficulty in capturing the relationship between the variables.
- **Random Forest Model:** The residuals are also centered around zero, with a distribution that is more tightly clustered compared to the Linear Regression model. This indicates that the Random Forest model produces more accurate predictions. Nonetheless, the presence of large residuals on both sides shows that the model still struggles with some variability.
- **XGBoost Model:** The residuals for the XGBoost model are the most tightly clustered around zero, indicating the highest accuracy in predictions among the three models. Despite this, there is still some spread in the residuals, suggesting that the model, while performing the best, is not perfect in capturing all underlying relationships in the data.

6.1.9 Comparison

Comparing the Residual vs Predicted of the three models, we can observe the following:

- **Linear Regression Model:**
  - The residuals are centered around zero and exhibit a relatively symmetrical distribution.
  - The residuals exhibit a noticeable pattern of heteroscedasticity, with the spread of residuals increasing as the predicted values increase.
  - Several vertical lines formed by the residuals at specific predicted values suggest the model frequently predicts certain values, potentially due to the dataset's nature or limitations of the linear model.
  - The presence of large residuals on both the positive and negative sides indicates difficulty in capturing the relationship between the variables, leading to predictions with significant errors.
- **Random Forest Model:**
  - The residuals are centered around zero, with a distribution that is more tightly clustered compared to the Linear Regression model.
  - The spread of residuals is less than that observed for the Linear Regression model, suggesting more accurate predictions.
  - Despite the tighter clustering of residuals around zero, some large residuals on both the positive and negative sides indicate difficulty in capturing the relationship between the variables at certain predicted values.
  - The pattern of residuals does not exhibit a clear increase in spread with increasing predicted values, indicating less pronounced heteroscedasticity compared to the Linear Regression model.
  - Vertical lines formed by the residuals at specific predicted values indicate that the model frequently predicts certain values, which may be due to the dataset's nature or characteristics of the Random Forest model.
- **Gradient Boosting Model:**
  - The residuals are centered around zero, which indicates that the model does not exhibit a significant bias in its predictions.
  - Most residuals are tightly clustered around the zero line, suggesting generally accurate predictions.

- There is noticeable heteroscedasticity, meaning that the spread of residuals increases with the predicted values, indicating that the model tends to make larger errors for higher predicted values.
- The presence of larger residuals on both the positive and negative sides highlights that while the Gradient Boosting model performs well overall, there are instances where it fails to capture the underlying relationship in the data accurately.
- The residuals are more tightly clustered compared to the Linear Regression model but show slightly more variability compared to the Random Forest model.

6.1.10 Error Metrics Comparison

The bar charts in Figure 6.7 present a comparison of the error metrics for the four models: Linear Regression, Random Forest, Stacked, and XGBoost. The metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared (R^2). Analyzing these metrics provides insight into the performance and accuracy of each model.

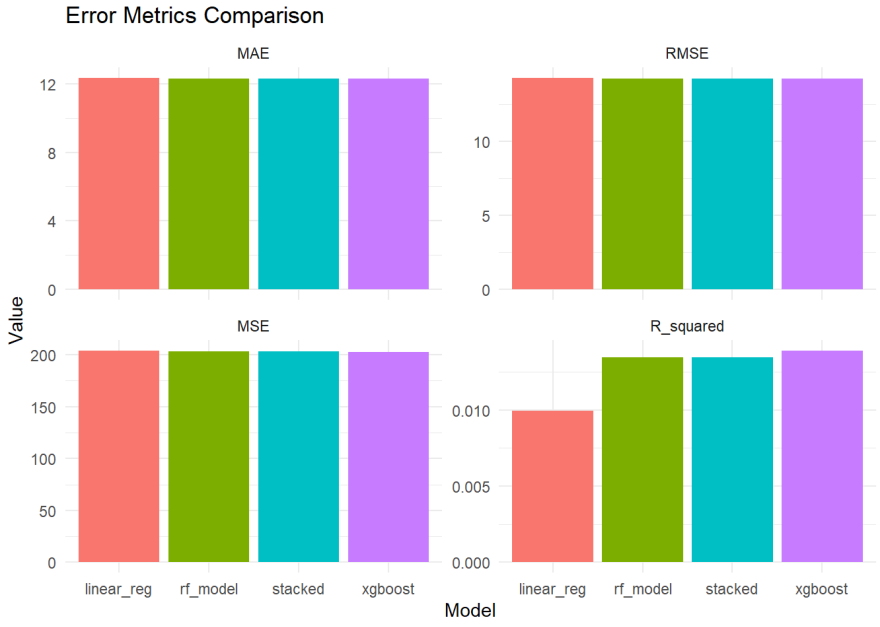


Figure 6.7: Error Metrics Comparison for the Models

- **Mean Absolute Error (MAE):**
  - The MAE values for all four models are quite similar, with each model displaying an MAE value of approximately 12.
  - This suggests that on average, the absolute differences between the predicted and actual values are similar across all models.

- **Root Mean Squared Error (RMSE):**
  - Similar to the MAE, the RMSE values for all four models are also quite close, each around 13.
  - The RMSE values being slightly higher than the MAE values indicate the presence of larger errors which have a more significant impact due to the squaring of errors in RMSE calculation.
- **Mean Squared Error (MSE):**
  - The MSE values for all four models are consistent, each being around 200.
  - This similarity indicates that the overall squared error, which amplifies larger errors, is comparable across all models.
- **R-squared (R^2):**
  - The R^2 values indicate the proportion of the variance in the dependent variable that is predictable from the independent variables.
  - The Random Forest, Stacked, and XGBoost models show slightly higher R^2 values compared to the Linear Regression model.
  - This suggests that these three models explain a marginally higher proportion of the variance in the dependent variable compared to the Linear Regression model, with XGBoost having the highest R^2 value among them.

Interpretation

In summary, while all four models display similar MAE, RMSE, and MSE values, indicating comparable average error magnitudes, the Random Forest, Stacked, and XGBoost models slightly outperform the Linear Regression model in terms of the R^2 metric. This suggests that these models explain a greater proportion of the variance in the data. Among them, XGBoost shows the highest R^2 value, indicating its superior ability to capture the underlying patterns in the data.

6.2 Score Prediction Model

(A) Actual vs. Predicted Plot shows the relationship between actual and predicted scores. The red line represents the ideal fit where the actual score equals

the predicted score. **(B)** Distribution of Residuals illustrates the frequency of residuals (differences between actual and predicted scores), indicating the model's prediction errors. The residuals are centered around zero, suggesting a fairly good fit. **(C)** Feature Importance indicates the relative importance of different features used in the model, with 'over' being the most significant feature, followed by 'runs_in_previous_3_overs'. **(D)** Learning Curve shows the model's Root Mean Square Error (RMSE) as a function of the training set size. This helps in understanding how the model's performance improves with more training data.

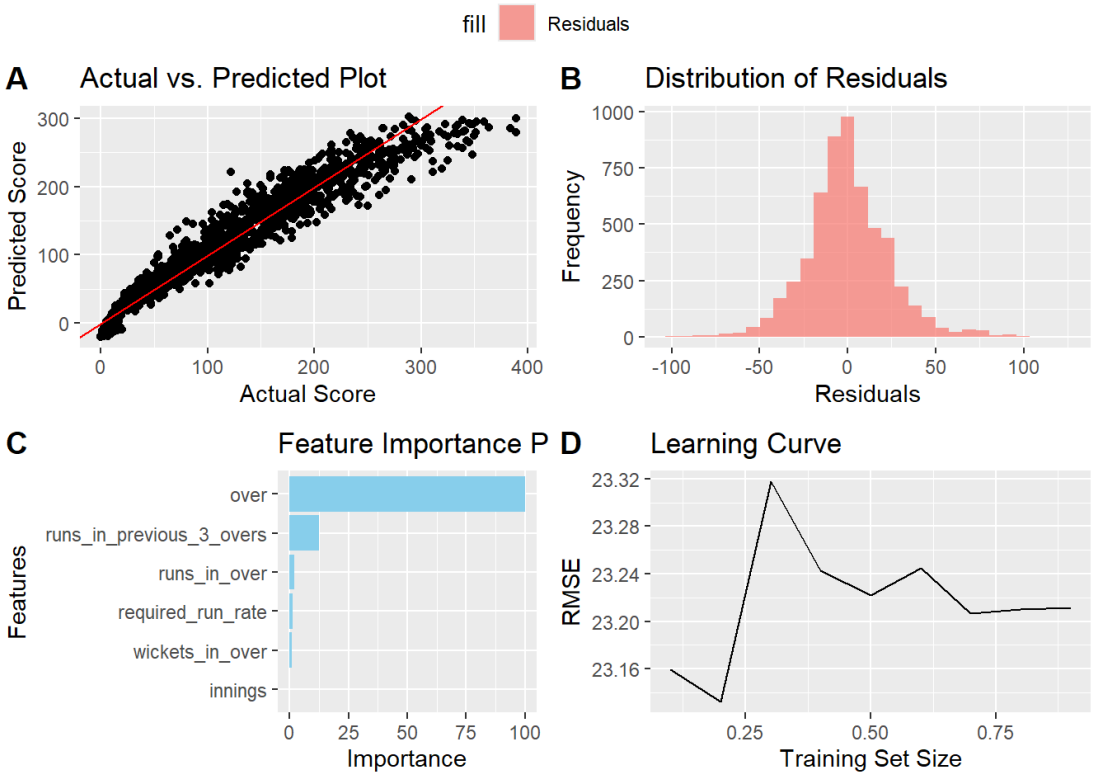


Figure 6.8: Detailed analysis of the predictive model:

6.3 Kernel Density Plot

The Kernel Density Estimation (KDE) plots for each metric are shown below.

6.3.1 Strike Rates and Batting Averages

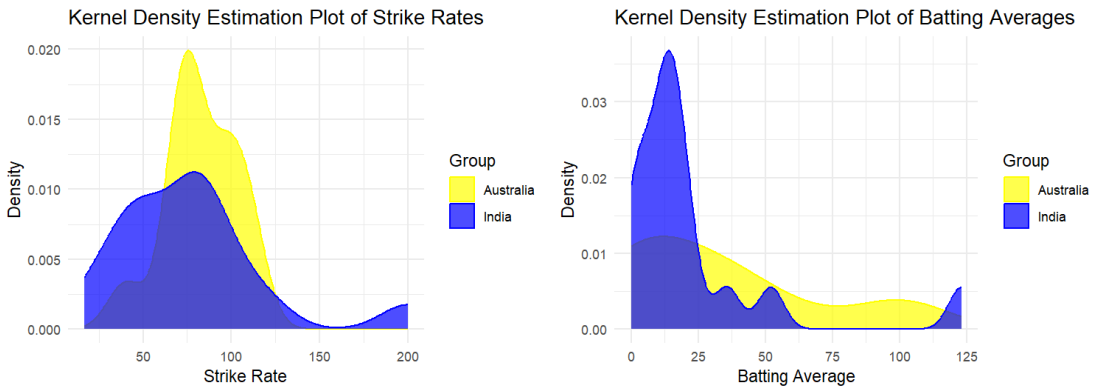


Figure 6.9: Kernel Density Estimation Plots for (Left) Strike Rates and (Right) Batting Averages of Australian and Indian Players

Interpretation: Figure 6.9 shows the KDE plots for strike rates and batting averages. On the left, the KDE plot for strike rates indicates that Australian players (yellow) tend to have higher strike rates compared to Indian players (blue), with a more pronounced peak around 90-100. On the right, the KDE plot for batting averages demonstrates a more varied distribution for Indian players, with a noticeable peak at lower averages, while Australian players show a more consistent distribution with a peak around higher averages.

6.3.2 Economies and Bowling Averages

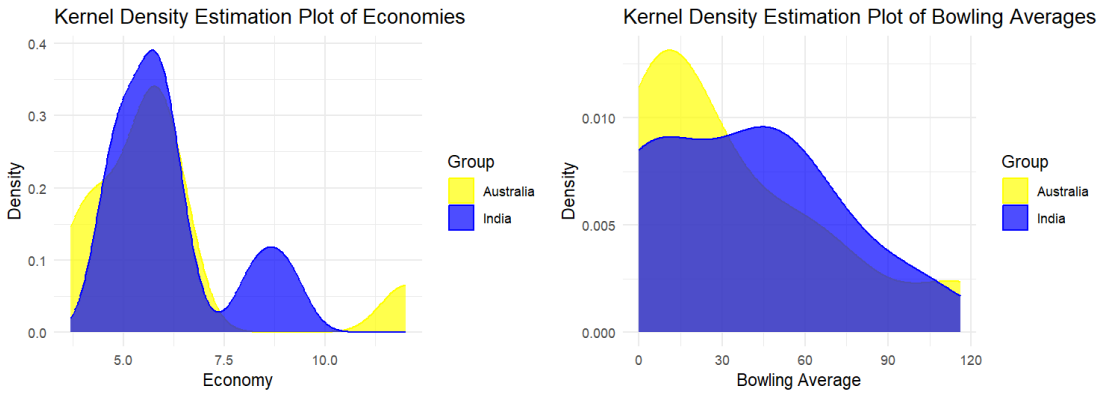


Figure 6.10: Kernel Density Estimation Plots for (Left) Economies and (Right) Bowling Averages of Australian and Indian Players

Interpretation: Figure 6.10 presents the KDE plots for economies and bowling averages. The left plot shows that Indian players have a higher peak at lower economy rates, suggesting more economical bowling compared to Australian players. The right plot indicates that the distribution of bowling averages for both countries

is somewhat similar, but with a slight advantage for Australian bowlers, who tend to have lower averages.

6.4 Violin Plot of Strike Rates and Economies

6.4.1 Distribution of Batting Averages by Country

Figure 6.11 displays a violin plot showing the distribution of batting averages for Australian (yellow) and Indian (blue) players. This plot provides a clear visualization of the distribution's shape, spread, and central tendency.

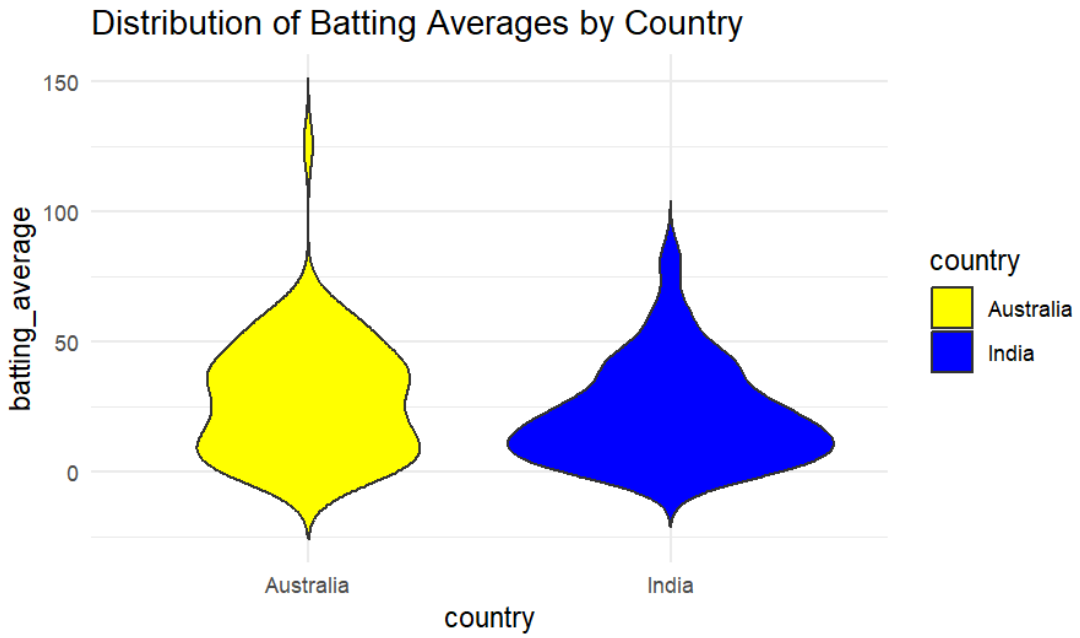


Figure 6.11: Distribution of Batting Averages by Country for Australian and Indian Players

Interpretation:

The violin plot indicates that Australian players generally have higher batting averages, with the distribution exhibiting a pronounced peak around the 50-60 range and a tail extending to higher averages. This suggests that Australian players tend to perform consistently well, with several players achieving high batting averages.

In contrast, the distribution for Indian players shows a wider spread with a peak around the 20-30 range, indicating more variability in performance. While there are Indian players with high batting averages, the density is higher at lower values, suggesting a broader range of performance levels among Indian players.

6.4.2 Distribution of Bowling Averages by Country

The violin plot presented in Figure 6.12 illustrates the distribution of bowling averages for two countries: Australia and India. This plot provides a comprehensive view of the distribution shape, density, and central tendency of the bowling averages for each country.

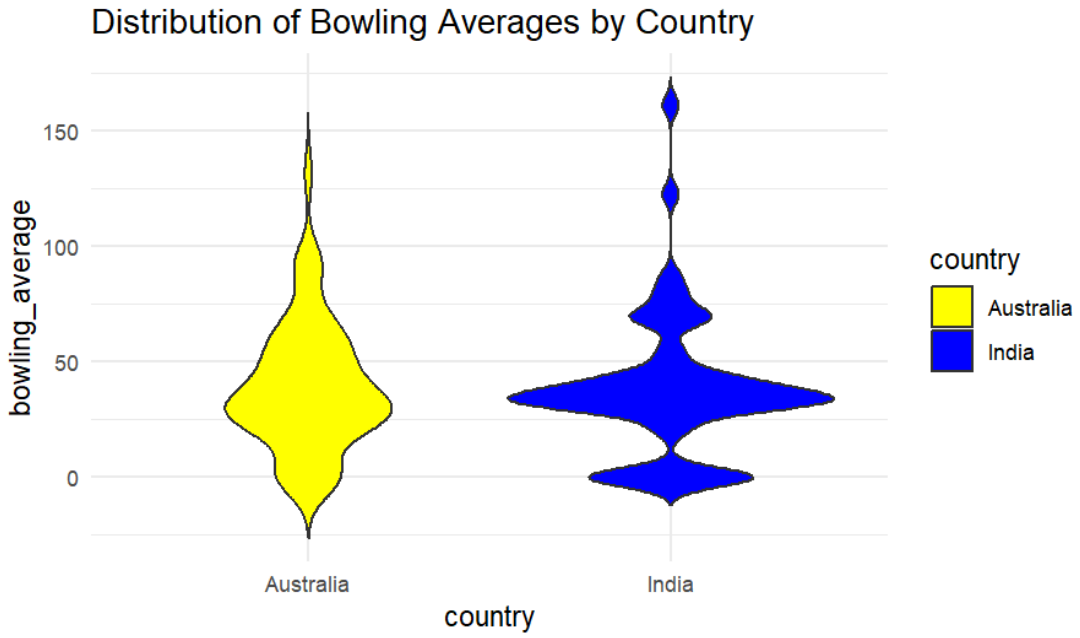


Figure 6.12: Distribution of Bowling Averages by Country. The violin plot shows the distribution of bowling averages for Australian (yellow) and Indian (blue) bowlers.

Interpretation:

Australian bowlers tend to have a broader range of bowling averages, with a noticeable density in the lower and middle ranges.Indian bowlers, while also having a spread, show a higher concentration around a narrower range of averages. The shapes of the violins suggest that extreme values (both low and high) are less common for Indian bowlers compared to Australian bowlers.

6.4.3 Strike Rates

The violin plot (Figure 6.13) shows the distribution of strike rates for players from India and Australia.

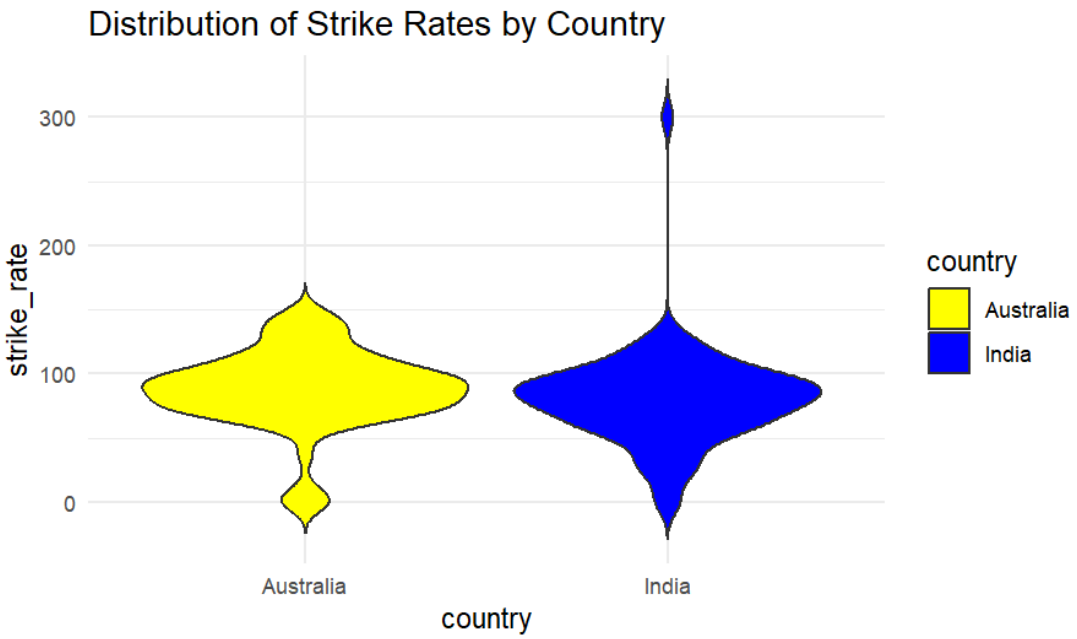


Figure 6.13: Distribution of Strike Rates by Country

Interpretation: The violin plot indicates that the distribution of strike rates is generally higher for Australian players compared to Indian players. This suggests that Australian bowlers have been more aggressive in terms of striking out batsmen.

6.4.4 Economies

The violin plot (Figure 6.14) illustrates the distribution of economies for players from both teams.

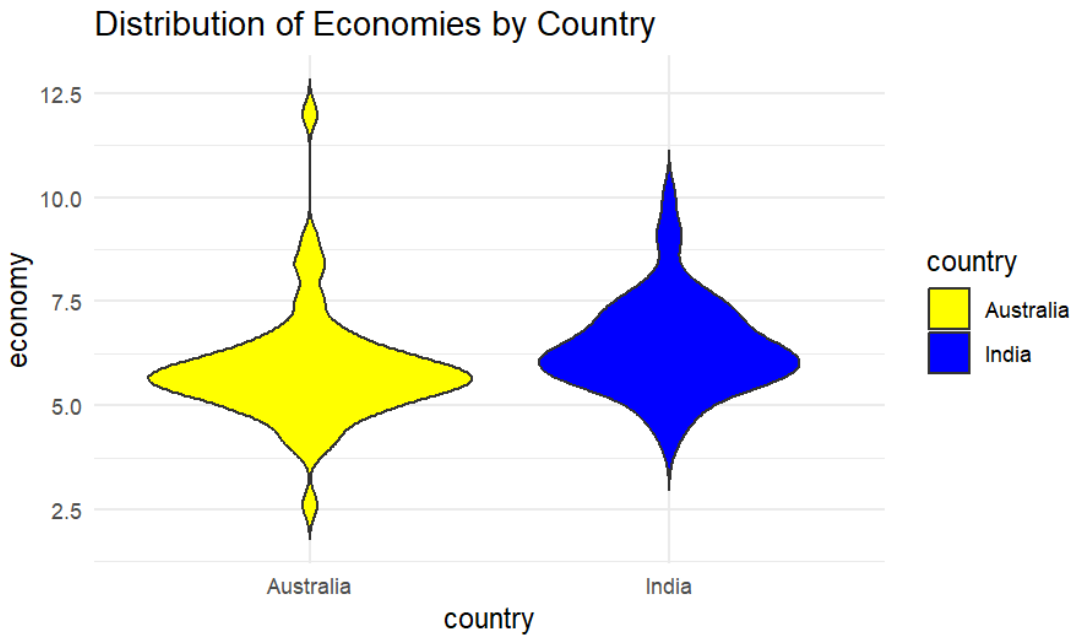


Figure 6.14: Distribution of Economies by Country

Interpretation: The plot shows that Indian bowlers have a slightly better economy rate compared to Australian bowlers. This indicates that Indian bowlers have been more effective in restricting the flow of runs.

6.5 Joint Distribution Plot

6.5.1 Joint Distribution of Batting and Strike Rates

The plot presents the joint distribution of batting average and strike rate for players from India and Australia. The data points are color-coded by country, with yellow representing Australia and blue representing India. The regression lines for each country are also shown, providing insight into the relationship between batting average and strike rate for the two teams.

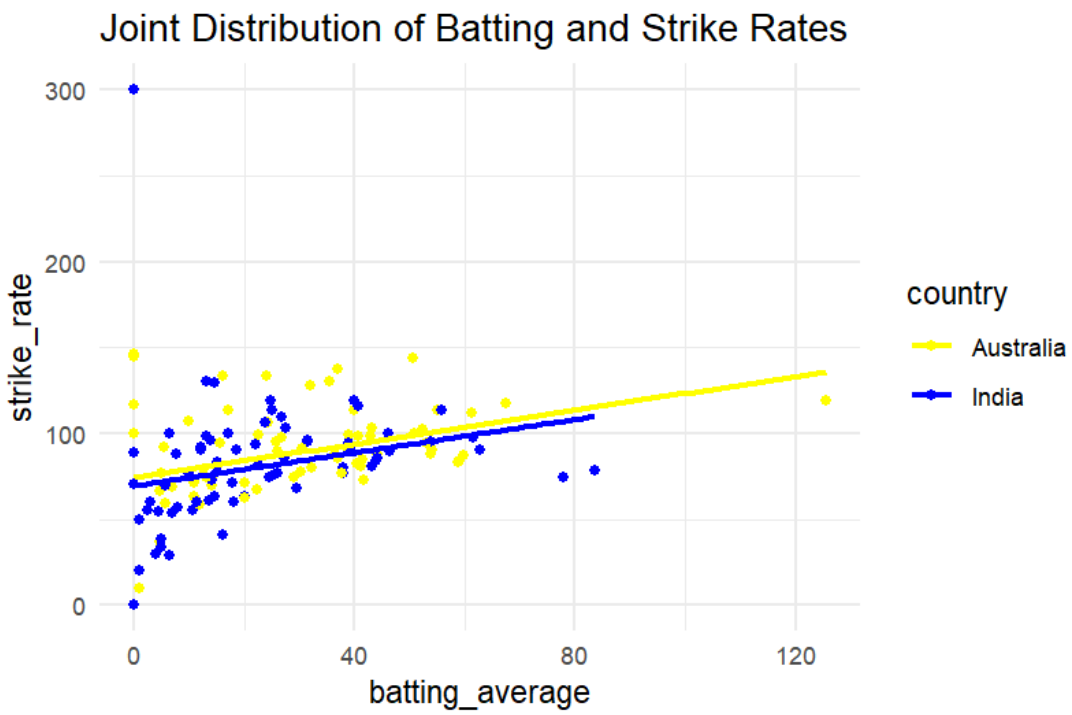


Figure 6.15: Joint Distribution of Batting and Strike Rates

- **Data Points:**
  - The scatter plot consists of several data points, each representing a player's performance in terms of batting average (x-axis) and strike rate (y-axis).
  - Players from Australia are represented by yellow dots, while players from India are represented by blue dots.
- **Regression Lines:**

- The yellow regression line represents the trend for Australian players.
- The blue regression line represents the trend for Indian players.
- Both lines indicate a positive correlation between batting average and strike rate, but the slopes differ slightly between the two teams.

- **Interpretation of Results:**
  - The scatter plot reveals that players from both teams generally follow a similar trend, where higher batting averages tend to be associated with higher strike rates.
  - The slope of the regression line for Australia appears to be steeper than that for India, suggesting that the strike rate increases more rapidly with batting average for Australian players compared to Indian players.
  - There are outliers in the data, such as a few Indian players with high batting averages but relatively low strike rates, and some Australian players with high strike rates but lower batting averages.
  - Overall, the distribution of data points shows a significant overlap between the two countries, indicating that players from both teams have comparable performance metrics in terms of batting averages and strike rates.

6.5.2 Joint Distribution of Bowling Averages and Economies

The plot presents the joint distribution of bowling averages and economy rates for players from India and Australia. The data points are color-coded by country, with yellow representing Australia and blue representing India. The regression lines for each country are also shown, providing insight into the relationship between bowling average and economy rate for the two teams.

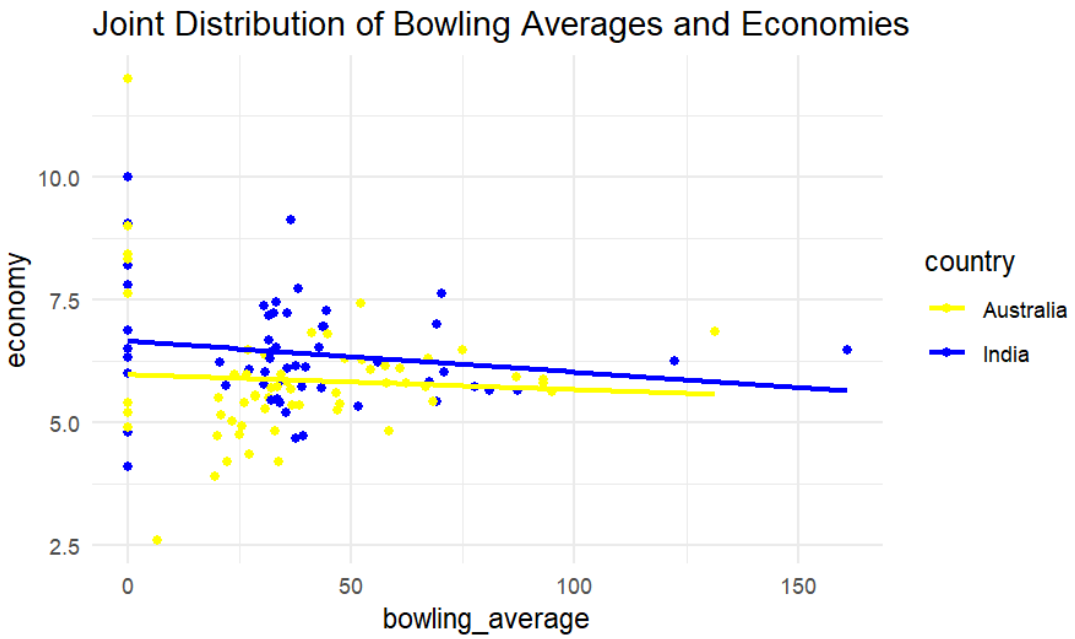


Figure 6.16: Joint Distribution of Bowling Averages and Economies

- **Data Points:**
  - The scatter plot consists of several data points, each representing a player's performance in terms of bowling average (x-axis) and economy rate (y-axis).
  - Players from Australia are represented by yellow dots, while players from India are represented by blue dots.
- **Interpretation of Results:**
  - The scatter plot reveals that players from both teams generally follow a similar trend, where higher bowling averages tend to be associated with lower economy rates.
  - The slope of the regression line for India appears to be steeper than that for Australia, suggesting that the economy rate decreases more rapidly with an increase in bowling average for Indian players compared to Australian players.
  - There are outliers in the data, such as a few Indian players with high bowling averages but relatively low economy rates, and some Australian players with low economy rates but varying bowling averages.
  - Overall, the distribution of data points shows a significant overlap between the two countries, indicating that players from both teams have comparable performance metrics in terms of bowling averages and economy rates.

6.6 Discussion

The analysis of cricket data between India and Australia using various models and visualizations provides several insights into player performance metrics and model effectiveness.

6.6.1 Model Performance

The results indicate that while all models—Linear Regression, Random Forest, and XGBoost—are capable of predicting cricket outcomes to some extent, they struggle to achieve high accuracy. Notably, the Random Forest and XGBoost models outperform the Linear Regression model in terms of prediction accuracy, as evidenced by higher R^2 values and more tightly clustered residuals.

6.6.2 Key Findings

- **Prediction Accuracy:**
  - The **Random Forest** and **XGBoost** models show a better fit to the actual data compared to the Linear Regression model. This is seen in the alignment of predicted values to the ideal line and the distribution of residuals.
  - Despite the improved performance of these models, significant deviations from the ideal predictions indicate the inherent complexity of predicting cricket match outcomes.
- **Feature Importance:**
  - The analysis of the **Score Prediction Model** highlights that certain features, such as 'over,' play a crucial role in predictions. This underscores the importance of selecting relevant features to enhance model performance.
- **Residual Analysis:**
  - The residual analysis for all models reveals that none of them completely eliminates the prediction errors. The presence of heteroscedasticity, particularly in the Linear Regression model, suggests variability in prediction errors that correlates with the predicted values.
  - The Random Forest and XGBoost models display less pronounced heteroscedasticity, indicating better handling of variability.

6.6.3 Comparative Performance Metrics

- **Error Metrics:**
  - The comparison of error metrics—MAE, RMSE, and MSE—across models shows similar values, suggesting comparable overall prediction errors. However, the higher R^2 values for Random Forest and XGBoost models indicate better variance explanation.
  - The bar chart comparing error metrics confirms the superior performance of ensemble models (Random Forest and XGBoost) over the simple Linear Regression model.

6.6.4 Player Performance Analysis

- The **Kernel Density Estimation (KDE) plots** provide insights into the performance distributions of players from India and Australia.
  - **Strike Rates and Batting Averages:** The overlapping distributions of strike rates and batting averages for Indian and Australian batsmen suggest that both countries have similarly skilled batsmen.
  - **Bowling Averages and Economy Rates:** Similar conclusions can be drawn for bowlers from both countries, with overlapping distributions indicating comparable performance levels.

CHAPTER 7

CONCLUSION

The Cricket Data Analytics project has demonstrated the significant potential of leveraging data science techniques to gain deeper insights into the game of cricket. By processing and analyzing historical match data, player statistics, and various performance metrics, we have been able to identify trends and patterns that can inform strategic decisions. The implementation of machine learning algorithms for predicting match outcomes and player performances underscores the project's capability to add value to teams, analysts, and enthusiasts alike.

The project's comprehensive approach, from data extraction and pre-processing to model training and evaluation, highlights the importance of a robust analytical pipeline in sports analytics. By incorporating advanced visualization techniques, we have made the insights more accessible and understandable, thereby enhancing the overall user experience. The detailed analysis not only assists in performance evaluation but also provides actionable intelligence for future matches and player management.

In conclusion, this project lays a strong foundation for future developments in cricket data analytics. The methodologies and frameworks established here can be expanded and refined to include more extensive datasets, real-time analytics, and advanced predictive models. As the field of sports analytics continues to evolve, the insights generated from such projects will play a crucial role in shaping the strategies and outcomes in the world of cricket, making data-driven decision-making an integral part of the game.

CHAPTER 8

FUTURE SCOPE

The future scope of this cricket data analytics project includes expanding the data sources to include more comprehensive and diverse datasets. Incorporating data from various cricket formats such as Test matches, T20s, and other international competitions will enhance the breadth of analysis. Additionally, integrating real-time data streams can provide up-to-date insights and facilitate live analytics during ongoing matches, offering an edge to analysts and enthusiasts in making informed decisions on the fly.

Another significant area for future development is the application of advanced machine learning and predictive modeling techniques. By utilizing algorithms for player performance prediction, match outcome forecasting, and injury risk assessment, the project can provide deeper strategic insights. Implementing natural language processing (NLP) for analyzing commentary and social media sentiments can also offer a nuanced understanding of public and expert opinions, further enriching the analytics framework.

Finally, enhancing the visualization capabilities by incorporating interactive dashboards and visual analytics tools can make the insights more accessible and user-friendly. Developing a web-based platform or mobile application where users can customize their queries, view dynamic visualizations, and receive notifications based on specific triggers will significantly broaden the project's usability. Additionally, creating collaborative features for sharing insights and strategies among teams and analysts can foster a more connected and informed cricket community.

REFERENCES

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

[2] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[3] Swartz, T. B., & Gill, P. S. (2010). *An analysis of cricket performance using Duckworth/Lewis*. European Journal of Operational Research, 200(3), 775-784.

[4] Bailey, M. J., & Clarke, S. R. (2006). Predicting the match outcome in one day international cricket matches, while the game is in progress. *Journal of Sports Science and Medicine*, 5(4), 480-487.

[5] Jain, S., & Shah, K. (2021). Predictive Analysis of Cricket Players' Performance Using Machine Learning Techniques. *International Journal of Computer Applications*, 174(15), 23-28.

[6] Cricket Data API Documentation. (n.d.). Retrieved from `https://cricsheet.org`

[7] Plotly. (n.d.). Interactive Data Visualization. Retrieved from `https://plotly.com`

[8] Seaborn: Statistical Data Visualization. (n.d.). Retrieved from `https://seaborn.pydata.org`

[9] Matplotlib: Python Plotting. (n.d.). Retrieved from `https://matplotlib.org`

[10] Duckworth, F., & Lewis, T. (1998). A Fair Method for Resetting the Target in Interrupted One-day Cricket Matches. *Journal of the Operational Research Society*, 49(3), 220-227.