ISM
INTERNATIONAL
SCHOOL OF MANAGEMENT
University of Applied Sciences

Master Thesis at the International School of Management

Sports Analytics: Data-driven Analysis of Team and Player performance for Strategic decision making

Name: **Prajwal Jagadish**

Course: Master of Science Business Intelligence & Data Science

ISM Campus: Hamburg

Supervisors: Prof. Dr. Veith Tiemann
Prof. Dr. Nicole Fabisch

Submission: October 19, 2023

Statutory Declaration

I hereby declare in lieu of oath that I have prepared the present thesis (words) independently and without using other than the stated aids and that the ideas derived directly or indirectly from outside sources are identified as such.

The present paper or thesis has not so far been presented in the same or similar form (in whole or in part) as examination work in any other examination procedure and has also not been published.

Place/Date

Signature
(First & last name)

Further Declaration

I declare that I agree to the present paper or thesis being submitted to a plagiarism test and that I will obtain the consent of the referee before any publication of the paper.

I have been informed that a violation of the statutory declaration entails the threat of disallowance of the examination work as well as deception or fraud proceedings.

Place/Date

Signature
(First & last name)

Abstract

The Indian Premier League (IPL) is a captivating cricket tournament known for its intense competition and thrilling performances. This master's thesis explores the heart of IPL cricket using rigorous data-driven methodologies to uncover performance insights and predictive capabilities, transforming the way we analyze the game.

The research journey begins with comprehensive data collection from Kaggle, including match details, player statistics, and ball-by-ball information. A robust PostgreSQL database serves as the foundation for data storage and schema design, while Python integration enhances data transformation and facilitates advanced machine learning techniques like Support Vector Machines (SVM) and Random Forest (RF) for performance prediction.

Key Performance Indicators (KPIs) such as batting average, strike rate, bowling average, economy rate, boundary percentage, and dot ball percentage are the compass for evaluating individual and team dynamics during IPL matches. The insights are brought to life with Power BI, a dynamic data visualization tool, that creates engaging dashboards and reports.

This thesis transforms the IPL landscape, enabling data-driven decision-making for teams and enthusiasts. It emphasizes the importance of player roles, adaptability, and situational awareness in T20 cricket. By combining PostgreSQL, machine learning, and Power BI, it unlocks the potential to derive valuable insights from structured data and predict player and analyze team performance.

In the highly competitive world of IPL cricket, characterized by narrow margins, the utilization of data-driven insights has emerged as a transformative factor. This thesis offers a scholarly gateway to the forefront of IPL Cricket, transcending the boundaries of sport and narrating a compelling story driven by data and analytics.

Contents

List of Figures

List of Tables

List of Abbreviations

BCCI	The board of control for cricket in india
CSV	comma separated file
DEA	Data Envelopment Analysis
DT	Decision Tree
ICC	International cricket council
IPL	Indian Premier League
JSON	JavaScript Object Notations
KDD	Knowledge Discovery in Databases
KNN	K- Nearest Neighbours
KPI	Key Performance Indicator
LBW	leg before wicket
MAE	Mean Absolute Error
MAPE	Mean absolute percentage error
MSE	Mean Squared Error
ODI	One day Internationals
PCA	Principal Component Analysis

RAA . . . . . Runs Above Average

RF . . . . . Random Forest

RMSE . . . . . Root Mean Squared Error

SQL . . . . . Structured query language

SR . . . . . Strike Rate

SVM . . . . . Support Vector Machine

TRP . . . . . Television Rating Point

WPA . . . . . Win Probability Added

XML . . . . . Extensible Markup Language

1 Introduction

1.1 Sports analytics:

Advancement in measurement technologies has made the movement analysis of various organisms possible. Technologies have enabled an understanding of the ways of wild animals and athletes from the data. Notably, recent advances in sports-related measurement technologies have been explored by many researchers. Based on these advances, it is possible to understand better the principles of real-world biological multi-agent behaviors, which is a significant issue in various engineering and scientific fields. The regulations governing real-world biological multi-agent behaviors still need to be discovered due to the absence of physical connections between the elements involved. Researchers employ mathematical models based on simple rules to gain insight into the multi-agent movements. These models have also found application in more intricate scenarios, including certain sports-related activities. However, when it comes to modeling the overall multi-agent behaviors of living organisms in real-world scenarios like team sports, it becomes mathematically challenging due to the presence of complex higher-order social interactions, cognition, and body dynamics. To better understand these behaviors, there is a necessity for an alternative approach that is both data-driven and model-free. This data-driven modeling proves to be highly powerful, allowing for extracting valuable information and accurate predictions using complex real-world data.

The process of using data and statistical analysis to make informed decisions is known as data-driven decision-making. In sports, it involves collecting and analyzing data from various sources like game footage, player performance measurements, and fan engagement to acquire insights into player performance, team strategy, and fan behavior. In recent years, teams have recognized the potential advantages of utilizing data to inform decision-making, which has led to a rise in the use of sports analytics in talent management and strategy. By using analytics, the teams can gain a more comprehensive understanding of the player's performance, identify the areas of improvement, and make more informed decisions about player acquisition

and development. This thesis aims to analyze team and player performance data for strategic decision-making in the Indian Premier League (IPL).

1.2 Statistics in sports:

Statistics in sports is a developing field that offers a specialized methodology to collect and analyze sports data to make decisions for the successful planning and implementation of new strategies. Before the 21st century, decision-making in sports was made with the data collected by observation. Statistics in sports have altered due to technological advancements, particularly in data collection and the availability of personal computing (Bai and Bai, 2021, 1-11). The term **"Sports Analytics"** has gained more traction than the term "Statistics in Sports," possibly as a result of the expertise's cross-pollination with other disciplines like statistics, computer science, management, and health sciences. Sports analytics is broadly defined as the process of data management, implementation of prediction models, and information systems for decision-making to acquire a competitive advantage in the field of play (Alamar and Mehrotra, 2011, 33-37). Sports analytics have been used in a variety of contexts. For instance, Sports teams evaluate players using statistical analysis to decide on the best game plan; Sports organizations rank players and teams, assess the effectiveness of the current rules, and study the viability of enacting new laws. Sports health professionals utilize statistical methods to determine the physical and mental state of the players. In 1977, Bill James published a book titled "Baseball Abstract," which was considered the beginning of sports analytics. While most of his writing did not use even the essential statistical tools, such as model fitting or graphical displays, the work had a significant impact in contesting long-held beliefs about baseball. "Handbook of Statistical Methods and Analyses in Sports," co-edited by Jim Albert, Mark E. Glickman, Tim B. Swartz, and Ruud H. Koning, is one of the most recent academic works concerning sports analytics. It overviews statistical techniques used in major sports and discusses difficulties encountered in statistical research.

Sports analytics is the study and modeling of sports performance using scientific methods. It is a relatively new scientific discipline used by almost every professional team in every sport. Specifically, it refers to managing structured historical data, applying predictive analytic models that use these data, and using information systems to inform decision-makers better

and enable them to assist their organizations in gaining a competitive advantage. Historical data can be quantitative or qualitative, often collected from various sports-related sources, including medical records, scouting reports, films and videos, box score performance data, and biographical data. The collected data are standardized, integrated, and analyzed using various metrics. A reliable and systematic study of the data is believed to help policymakers, athletes, and coaches improve their decision-making process. 1.1 shows the sample framework of sports analytics.

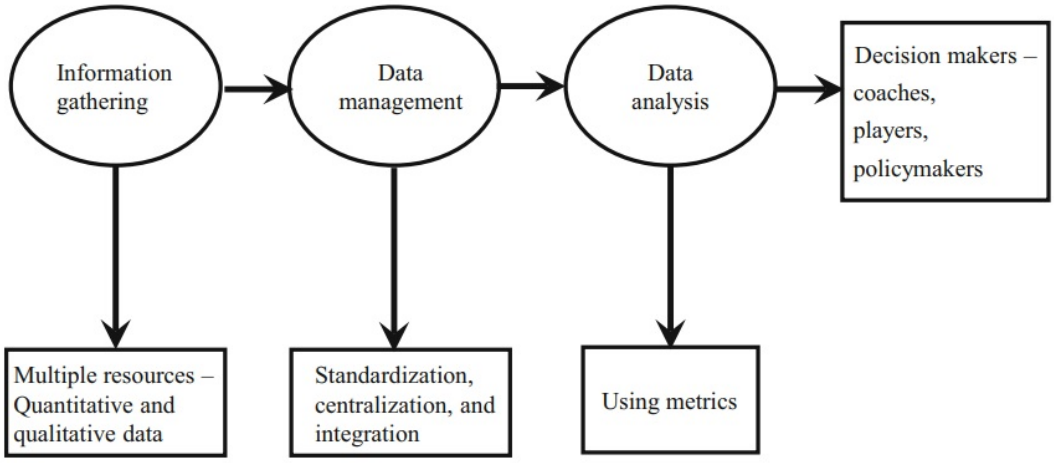

Figure 1.1: Sports Analytics Framework (Elia et al., 2018)

Sports analytics have given managers, players, and selectors a better platform to improve field performance. The second component of the framework is the decision-makers, and analysis is the process of using statistical tools and algorithms to acquire insight into what is likely to occur in the future. Sports analytics has been applied in various sports such as football, basketball, cricket, etc. Every ball movement, player strike rate, run rate, and other statistics are recorded using specialized camera systems and other recording devices. These data were run through various statistical algorithms and tools to gain deeper insights and to provide recommendations to the player or team. Advanced analytics and machine learning approaches are used to create a predictive model for different team sports like cricket with the ease of collecting and storing data.

1.2.1 Role of data analytics in modern sports and its impact on Decision-Making:

Sport is a significant endeavor in the lives of many people. Because many of them are involved in sports to exercise, improve their health, and lead healthier lives. Another reason is that watching and keeping track of professional sports is a popular pastime for young people and adults. Sports have been watched by people all around the world on television. Regular sports sections are found in many newspapers, and the team has its own television channels. The Olympic Games, the World Cup in football, the World Championships in basketball and swimming, and other major sporting events are among the most highly seen occurrences worldwide. The numerous facets of the sports industry cost billions of dollars, from the price of game tickets to the cost of broadcasting licenses, wages of top players, and advertising. Data analytics fundamentally change modern sports, revolutionizing how teams and athletes approach decision-making and performance improvement. Data analytics in sports has substantially advanced over time, changing the sports environment in several ways. Such as;

- **Performance Analysis:** Data analytics permits teams and coaches to collect and analyze much data during training and matches. This data comprises player statistics, fitness levels, movement patterns, Etc. By understanding these patterns and trends, coaches can identify areas for improvement and customize training programs to enhance the performance of individuals and teams (De Silva et al., 2018, 130).

- **Tactical Insights:** Teams can analyze the playing styles and strategies of the opponent team using the data from previous matches. This information enables them to adapt their tactics and game plans to exploit weaknesses and increase their chances of winning.

- **Injury Prevention and Player Health:** Data analytics can track and monitor player health and workload, which helps in reducing the risk of injuries. Wearable devices and sensors can provide real-time data on player movements and physical exertion, enabling coaches to effectively manage player fatigue and recovery (Seshadri et al., 2022, 71).

- **Recruitment and Talent Identification:** Data analytics assists the recruiter in identifying promising talents and potential players. Scouting tools analyze player performance metrics and provide insights into the player's suitability for a particular team or playing style.

- **Fan Engagement and Experience:** Data analytics also impacts the fan experience by providing interactive and engaging content. Fans can access real-time statistics, visualizations, and personalized content, enhancing their connection with the sport and their favorite teams (Lin et al., 2022, 962-971).
- **Referee and Officiating Support:** Data analytics can help referees make critical decisions during games. Video review and advanced tracking systems can provide additional information for fair and accurate decisions.
- **Game Strategy and Decision-Making:** The coaches can use data analytics to make informed decisions during matches. Real-time data can help with substitutions, tactical adjustments, and strategic changes based on player performance and opponent behavior (Ferran Vidal-Codina and Billingham, 2022, 1-15).
- **Player Contract Negotiations:** The sports clubs and agents can use data analytics to objectively examine players' value and contributions, aiding in contract negotiations.

1.3 Cricket

Two teams of eleven players play a bat and ball game called cricket. It is the second-most popular sport in the world. While the other team fields, each team swings at the ball and tries to score runs. Each turn is known as an inning. To achieve more runs than the opposition is the goal. Test match, One Day International (ODI), and Twenty20 are the three cricket match types accepted worldwide. The game's length, as scheduled, is the primary distinction between the three versions. A Test match lasts five days, an ODI is one day or a day/night combination, and T20 is a shorter version where each team plays for twenty overs, hence the term T20. An over is a group of six balls bowled from one end of a cricket pitch. Between two teams, an ODI is a type of limited-overs cricket in which each team plays fifty overs. These forms of cricket are played in numerous contests and championships around the globe(Neeraj and Hardik, 2016, 56). Some commonwealth countries playing the game are India, Pakistan, Sri Lanka, Australia, New Zealand, South Africa, the West Indies, Zimbabwe, Bangladesh, Britain, and recent additions. More than a billion people watch and follow the sport in these countries that play cricket. Millions of spectators follow rival matches between teams like England, Australia, India, and Pakistan (Shashi, 2013, 1-3).

According to available evidence, cricket began in England in 1611, and its official rules were published in 1744. It is one of the oldest sports ever and was introduced to India during the 1700s by the British. The sport spread worldwide through British colonial authority and is still widely practiced in Commonwealth nations, including Australia, South Africa, and the Caribbean West Indies (Ghai and Zipp, 2020, 1-16).

The International Cricket Council (ICC) is the world's cricket governing organization. Australia, England, and South Africa were represented in the 1909 founding of what was then known as the Imperial Cricket Conference. In 1965, the organization changed its name to the International Cricket Conference, and in 1987 it adopted its current name. Dubai, United Arab Emirates, is home to the ICC's headquarters. Examining the ICC and its history of operation on a worldwide scale to comprehend the governance of cricket as a whole. Only 12 of the 104 members of the ICC, which serves as cricket's worldwide governing body, are eligible to participate in traditional test matches. Match-fixing, gambling, and conflicts of interest are just a few of the concerns the ICC's anti-corruption branch addresses (Ghai and Zipp, 2020, 1-16).

The private governing organization for all kinds of cricket in India is the **Board of Control for Cricket in India (BCCI)**. At the state, national, and international levels of cricket, it holds the sole authority for organizing and conducting play. State-level associations oversee cricket-related activities at the state level, and these organizations also provide board members. Board members are frequently well-known former cricket players or prominent people, including ministers and business people who support cricket in India. BCCI represents the national organization, and several state association delegates choose its board members. There are five zones with 27 state associations and three non-playing members. North, South, East, West, and Central are these zones(Ghai and Zipp, 2020, 1-16).

1.3.1 Cricket in India

Indian cricket has a lengthy history. Madras and Calcutta played the first-ever cricket match in India in 1864(Mirjalili, 2019). The sport of cricket was brought to the Indian subcontinent by the British. Around this period, the Indian population developed an interest in cricket. However, like in England, the aristocratic elite mainly dominated the sport. When India gained independence in 1947, the full extent of cricket's infiltration into the country became evi-

dent(Ramachandra, 2003). Consequently, India's leading position in cricketing passion is a relatively recent social phenomenon. Since then, cricket's prominence in India has shaped the global game. During the 1980s, international satellite television networks recognized cricket's vast global audience and its suitability for advertising. Understanding India as the most significant cricket market, these networks directly approached the BCCI to negotiate a broadcasting deal(Gupta, 2003, 1779-1790). This move resulted in a tremendous influx of wealth into the hands of the BCCI, which subsequently wielded significant influence over the global governance of the game. People revere cricket players as gods and consider the game as a festival. The moments when it gained popularity in India are as follows;

- In 1983, the "World Cup" India won the championship by defeating the West Indies, who had previously held the trophy. Kapil Dev was the captain of the team.
- In 2007, the first-ever "Twenty World Cup", India won by defeating Pakistan. MS Dhoni was the captain of the team.
- In 2011, in the "World Cup," India won for the second time after defeating Sri Lanka. MS Dhoni was the captain of the team.

There have also been other moments when people in India celebrated cricket and helped it become more prevalent in India.

1.3.2 Indian Premier League (IPL):

IPL is a cash-rich cricket league conducted by the BCCI. Since 2008, it has had a significant impact globally and has become one of the most renowned cricket leagues (Enderwick and Nagar, 2010, 130-143). BCCI organizes and manages various international and domestic cricket tournaments. The domestic matches encompass a range of prestigious cricket competitions organized by the BCCI. These tournaments include the BCCI Corporate Trophy, Syed Mushtaq Ali Trophy, Irani Trophy, NKP Salve Challenger Trophy, Ranji Trophy, Duleep Trophy, Vijay Hazare Trophy, Deodhar Trophy, and the Indian Premier League. Of all the tournaments mentioned above, IPL is the most financially lucrative. Despite the participation of numerous high-level players, other matches experience low attendance rates. Furthermore, it collaborates with the ICC and facilitates the coordination of international cricket competitions, including test matches, 50-50 games, and 20-20 matches, among various cricket-playing nations. In addition to the test matches, the viewership, sponsorships, and income earned by the other two

formats are considerable. Among India's several sports organizations and groups, BCCI stands out as the most financially lucrative. The organization does not rely on financial support from the Ministry of Sports, instead utilizing internal funding to cover expenses related to match operations, board member compensation, and employee salaries (Shashi, 2013, 1-5). It has been played by teams selected by BCCI selectors. But in tournaments like the IPL, where players perform at their peak in front of a million spectators, the selection committee holds the responsibility of choosing the right players to be played in the national team.

The business model aims to establish a win-win situation based on Major League Baseball games in the USA and the English Premier League football cup in the UK. The first IPL tournament was held in 2008. Mr. Lalit Modi, a former vice-president of the BCCI, launched the IPL in 2008. He was the mastermind behind this idea. On 24 January 2008, the first IPL auction was held with a 400 million dollar base price for each franchise. Bangalore, Kolkata, Chennai, Delhi, Hyderabad, Mumbai, Jaipur, and Mohali are the teams made up in the tournament. It is significant from a financial perspective and a cricket perspective. It also helps in increasing the financial area of India. The business model of the IPL included the loss of intricate processes such as hard-hitting cricket, glamour, pricing, marketing, and above all, entertainment, and it was appreciated by everyone (Shashi, 2013, 2-4). In 2009, the IPL brand was valued at 2 billion USD; in 2010, it was worth 4.13 billion USD. This pattern was particularly encouraging, demonstrating how successful packaging and marketing could draw customers and generate revenue. However, the brand's valuation decreased to 3.67 billion USD in 2011 and 2.92 billion USD in 2012. Featuring 8 teams representing different cities or states, the IPL has evolved into one of the world's most popular and lucrative cricket leagues, boasting an approximate value of 6.8 billion U.S. dollars in 2022. Similar trends can be seen in television's TRP ratings. Television ratings during the 2008 season were 4.81, with 102.2 million viewers. The TVR decreased to 3.27 during the seasons (Shashi, 2013, 2-3). However, the game is top-rated in many countries that play cricket, like the UK and Australia, whose national team members play for various IPL franchises. The inaugural season of the IPL Twenty20 cricket tournament took place in April 2008, including eight teams that various franchisees controlled. The following teams were designated as Chennai Super Kings (CSK), Deccan Chargers (DC), Delhi Daredevils (DD), Kings XI Punjab (KXIP), Kolkata Knight Riders (KKR), Mumbai Indians (MI), Rajasthan Royals (RR), and Royal Challengers Bangalore (RCB). The franchisee assembled their teams by selecting players from various international, national, and local sources. The inclusion of international players in these teams not only enhanced their talent but also

added a sense of prestige and allure. The IPLT20 is widely recognized as the inaugural cricket competition in which clubs and players were subjected to a high-stakes auction, resulting in multi-million dollar transactions (Singh et al., 2015).

IPL has been compared to the cell phone boom in the telephone industry and the significant changes it brought to the cricket world. Since IPL started, many people have decried it as the future of test cricket. Some feel this new format has made batters impatient and negatively impacted their defensive skills (Allana, 2018). Many cricket fans, commentators, and veteran cricketers blamed the IPL as a reason for the failure of the Indian team in the Twenty-20 World Cups in 2009 and 2010 (Saikia et al., 2012, 96-110). The monopoly of cricket boards has been declining, and this trend is expected to have implications for Test cricket. However, the IPL has not remained unaffected by corruption, as evident from the Spot Fixing scandal during its 2013 season. This scandal created a national outcry, with several prominent cricketers being found guilty and arrested by the Delhi Police. The investigation also revealed the alleged involvement of figures from the Mumbai film industry, franchise owners, and even links to the underworld (Chattopadhyay, 2015, 17-40). Despite controversies surrounding the IPL, it has proven successful in nurturing many exceptional cricketers who have achieved remarkable success on the international stage. The IPL is a source of entertainment and a talent-scouting platform for the Indian National Cricket team. Esteemed players like Yusuf Pathan, Ravindra Jadeja, Ajinkya Rahane, Yuzvendra Chahal, and Hardik Pandya are among those who gained recognition through the IPL and went on to represent the National Cricket team (Sen, 2018).

1.4 Data analytics and its impact on decision-making in IPL

IPL is a Twenty-20 cricket tournament designed to foster the growth of cricket in India and provide a platform for nurturing young and skilled players. The league features teams representing various Indian cities competing against each other. The team formation for the IPL is unique as it involves a player auction, which is not a new concept in sports. The IPL being held in different cities has garnered a massive fan following, attracting significant media attention and engaging various businesses. The involvement of financial aspects, team spirit, city loyalty, and a massive fan following make match outcomes highly significant for all stakeholders. These outcomes rely on complex game rules, the team's luck (toss), the player's abilities,

and their performances on the specific match day. Other key factors include the nature of the pitches, such as flat pitches, those that favor fast bowling, spin bowling, or swing bowling, and whether they offer advantages to batters, non-striker batters, and bowlers and promote a strong partnership between players. Also, various natural parameters, including historical data related to players, play a crucial role in predicting the outcome of cricket matches. In IPL or any sport, the prediction of match outcomes heavily relies on factors such as team strength, the presence of key players, and the home crowd. In cricket history, analytics has emerged as a pivotal element (Christian and Yanyan, 2015, 101-134). However, some uncertainty will inevitably be associated with the average performance of bowlers or batters. Crucial moments like the last overs and power plays often act as turning points in matches. However, it remains a challenging task to identify the right player for crucial overs. The ball movement is changed from one over to another, making it crucial to predict the outcome of each match for every ball.(Shah, 2022, 83-85) developed a model capable of forecasting the match result for every ball played. Predicting match outcomes for different teams can be instrumental in aiding the team selection process (Rory and Fadi, 2019, 27-33). In such challenging situations, analytics can prove to be immensely valuable. Analytics serves as a bridge for team selectors, coaches, and managers, offering a clearer idea of player consistency, fast scoring, and finishing abilities. By leveraging analytics, they can manage risks more effectively and make informed decisions about probable winners on and off the field. However, the diverse range of parameters involved presents significant challenges in the accurate prediction of the outcome of the game. Predictors' accuracy also relies on the volume of data used for analysis. Consequently, several predictive models are developed for predicting match outcomes based on each player's past performance and relevant match-related data (Schumaker et al., 2010, 55-63).

1.5 Research Questions

:

This topic's research question is: What are the primary determinants of team performance in the IPL, and how can these be quantified through data and analytics? Analyzing numerous performance measures and determining the elements that affect team success in the IPL are required to answer this question. Coaches and auctioneers may make data-driven decisions more rapidly by being aware of the essential variables, resulting in league-winning teams.

The IPL's team and individual player performance measures, like batting and bowling av-

erages, run rates, strike rates, and wickets taken, will be examined in the thesis to address this research topic. The characteristics contributing to a team's good performance will be determined using data analysis and data visualization. These metrics will be studied using data from prior IPL seasons.

This research question is vital because the IPL is one of the world's most popular and lucrative professional cricket leagues, and teams are constantly looking for ways to improve their performance and win the league title. By identifying the key factors contributing to team success, this research can help coaches and auctioneers make more informed decisions when selecting players and developing team strategies. Additionally, this research can contribute to the growing field of sports analytics and provide insights into the factors that drive success in professional sports leagues.

1.6 Objectives

The primary objective of this study is to analyze the team and player performance in Cricket and IPL. The various purposes associated with this are as follows;

- To analyze the factors influencing match outcomes and subsequently analyze team and predict player performance based on these features.
- To develop a model to quantify the performance of the various players using past statistics.
- To analyze team performance data to help auctioneers and coaches make strategic decisions.

1.7 Scope and Limitations of the Study

Scope

- In-depth data analysis and performance prediction in the context of the IPL are included in the thesis's scope. It involves gathering, preprocessing, storing, and transforming information about the IPL using PostgreSQL and utilizing various machine learning models to forecast player and team performance.

- **Key Performance Indicators** To assess player performance, the analysis focuses on particular KPIs, including batting average, batting strike rate, bowling average, bowling economy rate, boundary percentage, and dot ball percentage. These KPIs offer insightful data regarding the IPL match dynamics.
- **Data Visualization:** Power BI can be used to visualize data, allowing insights to be presented aesthetically. Data Visualization comprises graphics that show team dynamics, player performance, and match trends.
- **Server Integration:** The integration of PostgreSQL as the database server and its connection to Python for data transformation and manipulation is included in the thesis. In addition, data export to Power BI is necessary for data visualization.
- **Machine Learning Models:** Using historical data to forecast player and team performance, various machine learning models, including SVM and RF, are implemented and evaluated.

Limitations:

- **Data Accessibility:** This thesis's applicability depends on the accessibility of historical IPL data. Any constraints or data gaps may impact the quality and thoroughness of the analysis and forecasts.
- **previous Data:** Future IPL seasons may display different trends and dynamics because the analysis and projections are based on the latest data. The models might not consider unexpected changes in team composition, regulations, or player performance.
- **Model Accuracy:** While machine learning models can offer insightful information, their accuracy depends on the caliber and volume of available data. The models' prediction powers may be limited in highly unpredictable and dynamic cricket matches.
- **Scalability:** As time goes on and more IPL seasons are added, scalability issues may become relevant. The thesis might need to go into better detail about handling massive datasets or real-time data updates.
- **Generalization:** This thesis's conclusions and models were created specifically for the IPL and might not apply to other cricket leagues or formats. The analysis is tailored to the IPL context.

- **Data Privacy:** The thesis admits that there may be inherent dangers involved with managing sensitive player and team data despite data security measures being implemented.
- **Server Performance:** The speed and effectiveness of data processing and analysis may be impacted by the PostgreSQL server's performance and other technological components. Considerations for server performance might need to be more inclusive.

These scope and limitations provide a clear framework for this master's thesis on IPL data analysis and performance prediction, guiding your research and analysis while acknowledging potential constraints and challenges.

1.8 Structure of the thesis:

- The introduction will start with a brief background on sports analytics and its use in cricket, followed by an overview of the IPL, one of the most popular and lucrative professional leagues globally.
- The following section will provide a literature review on data analytics in cricket, focusing on previous team and player performance metrics studies.
- After that, the research question will be presented, focusing on identifying the key factors contributing to team success in the IPL and how these factors can be measured using data analytics.
- The following section will describe the methodology used for the study, including the sources of data, the data cleaning process, and the specific performance metrics used to measure team and player performance.
- The results section will present the analysis findings, highlighting the key factors contributing to team success in the IPL and providing insights into how coaches and auctioneers can use these factors to make data-driven decisions.
- Finally, the conclusion will summarize the study's main findings, discuss the research's implications, and suggest future directions for research in this area.

2 Literature Review

2.1 Introduction

This chapter gives a brief literature review on data analytics and the application of predictive analysis in sports. The various earlier studies on the application of data analytics in Cricket for player and team performance in IPL are also reviewed and presented.

2.2 Formats of Cricket

2.2.1 Test Cricket

Test cricket is often regarded as the most extensive variant of the sport, encompassing a prolonged duration of play. It is commonly acknowledged by coaches, players, and enthusiasts alike as the final assessment of an individual's cricketing prowess. Test cricket is primarily seen as the epitome of the sport of cricket. A study conducted on test match cricket reveals that this format imposes a higher overall physical demand than other formats, primarily due to its longer duration and more cautious approach. Intense planning and careful play are the pillars of a test match(Petersen et al., 2011, 1368-1373).

2.2.2 ODI Cricket

One-day cricket, sometimes called limited overs cricket, exhibits a distinct departure from the conventional form of cricket since it encompasses contests confined to a single day. The birth of limited-overs cricket during the 1960s aimed to address the issue of drawn matches and enhance excitement by promoting a more aggressive batting approach. One-day cricket is characterized by a sequential batting format, where one team bats during their innings, followed by the opposite team batting throughout their innings. The team that accumulates the most number of runs emerges as the victor. The conclusion of a team's innings occurs either

after the completion of 50 overs or the loss of 10 wickets. One exception to this rule is that the second team ceases batting whenever their score surpasses the first team's. Including two specific conditions, namely the requirement of 50 overs and the condition of 10 wickets, introduces complexities. Disregarding certain specifics, an over is comprised of six deliveries executed by a bowler. Hence, the total number of balls that can be bowled in a match is limited to 300, calculated by multiplying 50 by 6 (Swartz Tim and deSilva Basil M, 2006, 1939-1950).

2.2.3 T20-Cricket

The sport changed over time and eventually developed its shortest format, T20. The first official T-20 match between England County sides occurred on June 13, 2003. It was named the Twenty20 Cup (Singh and Bagchi, 2020). The format was officially adopted internationally in 2005 when Australia played New Zealand and won. On February 17th, 2005, the first men's international 20-over game was played. This format gained popularity with the introduction of the 2007 ICC World Twenty20 Cricket. The International Cricket Council (ICC) held its first-ever T-20 Cricket World Cup, which was held in South Africa. The tournament has grown to include 16 teams and three teams making their tournament debuts in its fifth season, which was born in 2014. This is because the competition has been so popular over the years. The BCCI became interested in this aspect of the game after India's stunning success in the 2007 T-20 World Cup, which ultimately inspired the concept of the IPL(Panda, 2018).

2.3 Reviews on Player Performance in T-20 cricket

In all sports, including cricket, the importance of effective team selection cannot be overstated. Consequently, researchers have shown considerable interest in addressing the challenges of optimal team selection. Analyzing the player's performances quantitatively within a team context has been a prominent area of research for scholars, irrespective of the sport being studied. (Kumar et al., 2011) conducted a performance evaluation of fast bowlers and spinners from India across three IPL seasons (2008, 2009, and 2010). They utilized various criteria and parameters like economy rate, bowling average, and bowling strike rate to rank the players using AHP and TOPSIS. The results indicated that Indian bowlers performed exceptionally well in all three seasons being Indian players. In the 2010 session, the Indian spinners have done better than the Indian fast bowlers, whereas in the 2009 session, Indian bowlers have done well in both categories. Still, in 2008, the top bowlers were foreigners of both types.

But overall, the Indian bowlers are outstanding, and the top 7 are Indians, and in the top 10, only two foreigners are there. The performance of 'Zaheer Khan,' India's one of best bowlers, needed improvement during IPL. SK Warne and M. Muralitharan, the best spinners in the world, did not get the rank in the top 5 bowlers.

(Singh et al., 2011, 11-16) introduced a fuzzy logic-based approach for assessing the performance of cricket players. They designed a fuzzy system incorporating eight parameters as linguistic variables that influence the performance of cricket players. These parameters included Runs Scored, Balls Faced, Strike Rate, Out, Fours, Sixes, Team Strength, and Team Against Strength. The input parameters were carefully selected and transformed into linguistic variables to scale their impact on player rankings effectively. A user-friendly and efficient software tool was developed to compute these input parameters' influence on the players' overall order. This fuzzy logic-based technique provided a valuable method for evaluating cricket player performances, offering a comprehensive approach to player assessment.

The study investigated the performance markers in cricket's batting and bowling parts (Moore et al., 2012, 188-207). Pitch-level analysis was employed to uncover micro-level characteristics linked to successful match results. The dataset under examination comprises seven English T20 matches, all contested at a singular venue during the first-class national tournament in the 2010 campaign. The analysis of pitch maps revealed a notable trend wherein bowlers belonging to victorious teams exhibited a higher frequency of wicket-taking through LBW decisions.

A performance evaluation was conducted on the bowlers participating in the IPL season 4 (Bhattacharjee and Pahinkar, 2012, 184-191). Using the multiple regression model, they also identified the factors that influenced the bowler's performance. They found that the bowler's experience and combined bowling rate in Twenty20 internationals were the most significant predictors influencing their performance in IPL season 4.

Using advanced statistical techniques such as factor analysis, a comprehensive investigation was carried out into the interrelationships among various dimensions of batting and bowling capabilities in T20 cricket (Sharma, 2013, 69-76). The dataset comprised accurate data from 85 batters and 85 bowlers participating in the IPL) 2012, considering multiple dimensions of batting and bowling from IPL season 5. The factor analysis grouped five dimensions into factor one (batting) and their respective sizes into factor two (bowling). Notably, the variance explained by factor one (batting) was significantly higher than that of factor two (bowling), clearly indicating the dominant influence of batting capability over bowling capability.

A study explored the performance of IPL teams and players using correlation, association, and classification rules(Prakash et al., 2015, 1-6). They utilized Naive Bayesian classification

to predict team results based on individual player performances. The study also involved an analysis of team performance concerning home and away ground matches. By considering the support and confidence of the players, this study provided valuable insights to selectors for filtering out players for the next IPL season.

A Machine Learning (ML)–based approach was employed to rank the performance of cricket players (Chellapilla et al., 2016, 37-47). They first clustered the players based on their roles, grouping them according to their batting and bowling performances. They introduced a novel "Deep Performance Index" index to formulate the performance ranking. The study focused on players from the IPL season of 2008, and a total of 201 players were analyzed in both T20 and IPL matches. Using machine learning, players were categorized into groups based on their respective batting and bowling performances.

In 2017 (Goyal, 2017, 117-122), an analysis was conducted on the performance of cricket players participating in the IPL from different countries. The study used KPI to gain insights into which country players play well in the league. However, Indian players were excluded from the study due to the league's rule that mandates at least seven out of eleven players in a team to be from India. The dataset of players was taken from the IPL10 and analyzed with cluster analysis. It was revealed that players from England outperformed their counterparts from other countries based on the Key Performance Indicators.

The study employed the Factor Analysis technique to assess the performance of Cricket Players (Shah et al., 2017, 656-660).PCA was utilized to evaluate the soundness of the items and categorize them into coherent clusters. To achieve an ideal amount of variables, the researchers employed a rotation technique called orthogonal varimax rotation. The researchers examined five critical indicators of batting statistics, namely the highest individual score (HS), average batting performance, SR, number of fours (4's), and number of sixes (6's), to analyze the batting and bowling performances of players in both the ninth edition of the IPL and the ICC World Cup in 2015. The study's findings indicate that batting prowess takes precedence over bowling prowess.

In 2020, (Kapadiya et al., 2020, 3406-3409) using various machine classifiers, historical data of team performance, and past individual player performances were utilized to forecast the overall performance of teams in the cricket match. These classifiers included Naive Bayes, Decision Tree, Random Forest, SVM, and Weighted Random Forest. Weighted Random Forest produced the highest accuracy among the different classifiers used, reaching an impressive 93.73 percent. The Naive Bayes classifier achieved 58.12 percent accuracy, the Decision Tree classifier achieved 86.50 percent accuracy, the Random Forest classifier achieved 92.25 per-

cent accuracy, and the SVM classifier achieved 68.78 percent accuracy. The results demonstrated that the Weighted Random Forest classifier outperformed the other classifiers in terms of accuracy, making it a promising method for predicting overall team performance in cricket.

The outcomes of an in-depth analysis of a comprehensive ball-by-ball dataset encompassing all matches played in the entire history of IPL were presented by (Kaviya et al., 2020, 218-228). Processing of datasets, Batsmen performance analysis, Bowler performance analysis, Match analysis, Head-on-head analysis of teams, Team overall performance analysis, Ranking of teams, Match prediction, and User interface creation were related to the metrics of the match were thoroughly examined, and practical visualizations were employed to depict the findings. The authors applied machine learning techniques to rank the players based on the Player Ranking Index. Processing of datasets and Comprehensive Data Analysis on IPL (CDAI), increased prediction accuracy by up to 81 percent. This improvement represented a 12 percent increase compared to the accuracy achieved by the existing system (Deep Mayo Predictor (DMP)).

In their research, (Nasim et al., 2020, 2866-2876) presented a data-driven probabilistic approach developed explicitly to predict the performance of batters in the sport of cricket. This system considers the dependencies between the runs scored by a batsman in consecutive balls. The plan was evaluated using a dataset extracted from the Cricinfo website. The Hidden Markov model (HMM) enables the generation of a prediction model to predict the upcoming performances of the players. The first-order Markov chain assumes that the probability of a batsman scoring runs in the next ball solely depends on the runs scored in the current ball. The parameters of the HMM are learned through a data-driven approach using the available data. A probabilistic matrix is constructed to predict the potential scores a batter may achieve in the upcoming balls. The findings demonstrate that the system can accurately predict the runs scored by a batter in a ball. This study provides insights into the optimal selection of players based on their performances and the selection procedure.

2.4 Review on team Performance in T-20 Cricket

The study undertook a comprehensive and thorough exploratory analysis of the IPL by utilizing a dataset encompassing a time frame of 15 seasons, ranging from 2008 to 2022 (Bhoyar and Agrawal, 2020, 4125-4130). The dataset had a total of 950 matches that were played between 17 distinct teams. The analytical approach discovers the winning percentage of each team, the number of games that have been played, the number of boundaries, 1st inning and 2nd

inning average score, and the number of centuries, half-centuries, and the highest run-getters and wicket-takers as well as which batsman scores the greatest number of 4's and 6's revealed significant insights in the IPL data, presenting possibilities to improve strategies for optimizing team and individual performance.

The study evaluated the data from the 2008-2018 IPL, toss-related analysis, and the extent to which data visualization aids decision-makers in finding suitable players for their teams in the IPL (Vidit and Tulasi, 2019, 4423-4432). Various novel features have been introduced, including the total number of matches played by the team across all eleven seasons, the highest number of Man of the Match awards received, the highest number of centuries scored by batters, the highest number of Player of the Match awards received, the highest count of toss wins, the number of wins by different teams, and the decisions made by each team after winning the toss. The research also encompassed specific KPI;s, such as batsmen with the highest strike rate and batters with the highest number of runs. Through concealed criteria, several patterns and attributes have contributed to the ultimate result of a cricket match, thereby aiding team owners and selectors in their ability to identify more proficient players.

(Kaviya et al., 2020, 218-228) This study conducted a comprehensive and evaluative exploratory analysis of several aspects of team performance, including matches, batters, bowlers, and other relevant characteristics. Employing a fresh ranking approach and an alternative method that effectively incorporates advanced algorithms is essential to evaluate players. This study demonstrates the need for team owners to possess comprehensive knowledge regarding a player's historical performance and ranking to make informed decisions regarding their suitability for inclusion in the squad. Coaches and team captains thoroughly understand their adversaries, enabling them to devise strategic plans that optimize the composition of their starting lineup at a specific venue. This strategic approach aims to outperform and ultimately defeat their opponents. Individuals that engage in wagering activities on IPL matches. For coaches, it facilitates decision-making processes; it is essential to determine whether a team possesses greater robustness and a higher probability of winning a game, among other factors. For people who make sound investments and optimize financial gains, individuals must select and allocate resources to a competent team carefully.

DEA was employed to assess the technical efficiency of cricket teams participating in the IPL (Singh, 2011, 180-183). For the 2009 season, team inputs were represented by total expenses, encompassing players' salaries, support staff wages, and miscellaneous costs. Outputs were evaluated based on points earned, net run rate, profits, and revenues. The efficiency scores were closely linked to team performance in the league, with a few exceptions. Upon

dissecting inefficiency into technical and scale inefficiency, it was evident that a significant portion could be attributed to suboptimal production scale and inefficient conversion of inputs into outputs.

Advanced statistical methods, specifically factor analysis, were employed to explore the interconnectedness of various batting and bowling skills aspects in T20 cricket (Sharma, 2013, 69-76). The study used data from 85 batsmen and 85 bowlers from the 2012 IPL season. The research findings indicated that batting prowess holds more significance than bowling skills. The study identified five dimensions under factor one (batting) and three under factor two (bowling). Factor one (batting) explained a significantly higher variance than factor two (bowling), highlighting the apparent dominance of batting abilities over bowling capabilities.

A comprehensive analysis was carried out that centered on the role of the toss in cricket matches, and this analysis made extensive use of data visualization techniques to present its findings effectively (Kanungo and Bomatpalli, 2019, 4423-4431). Their objective was to assist decision-makers in identifying inherent talent for their teams. The study concluded that understanding concealed factors, patterns, and attributes influencing cricket match outcomes is vital for team owners and selectors. In the context of IPL cricket, player salaries are determined through an auction process, making it essential for franchises to make strategic decisions. Players' past IPL performances influence this decision-making process. Selectors seek young, dynamic players capable of handling pressure with composure and leading the team to victory.

2.5 Existing sources available for cricket data

A dataset is a compilation of data provided in many formats, such as Excel Spreadsheets, CSV, JSON, and others. The Tabular form is the predominant representation for all of these formats. In this form, each column in the table represents a specific variable, while each row corresponds to a record inside the dataset. In this section, The discussion will encompass websites for discovering cricket datasets, focusing on event data and datasets derived from event data since Tracking Data remains non-publicly accessible.

2.5.1 Cricsheet

Cricsheet is a collection of projects that collectively offer comprehensive cricket data. These projects provide ball-by-ball match data for Men's and Women's Test Matches, One-day inter-

nationals, Twenty20 Internationals, and other international T20 matches. Additionally, they cover various club competitions such as the Afghanistan Premier League, Big Bash League, T20 Blaze, Bangladesh Premier League, Bob Willis Trophy, County Championship, Charlotte Edwards Cup, Caribbean Premier League, CSA T20 Challenge, FairBreak Invitational Tournament, The Hundred, International League T20, Indian Premier League, Cricket Ireland Inter-Provincial Limited Over Cup, Cricket Ireland Inter-Provincial Twenty20 Trophy, Lanka Premier League, Major League Cricket, Mzansi Super League, T20 Blast, Pakistan Super League, Rachael Heyhoe Flint Trophy, Royal London One-Day Cup, SA20, Super 50, Syed Mushtaq Ali Trophy, Sheffield Shield, Super Smash, Women's Big Bash League, Women's Caribbean Premier League, Women's Premier League, Women's Cricket Super League, and Women's T20 Challenge. Cricsheet also offers the convenience of downloading cricket data in various formats like JSON, YAML, CSV, and XML. This comprehensive and accessible collection of cricket data makes it a valuable resource for cricket enthusiasts, analysts, and researchers looking to explore various aspects of sports from different formats and competitions (**?**). Cricsheet is a website referenced by other journal articles such as:(Kaviya et al., 2020) and (Vidit and Tulasi, 2019).

2.5.2 Statsguru in ESPN Cricinfo

ESPN Cricinfo is a cricket news website offering a variety of features, including articles, live match coverage, and a cricket database called Statsguru. While Statsguru does not contain raw data, it presents aggregated data compiled from ball-by-ball information organized into various metrics. Users can submit queries to the database through Statsguru to access the required information. ESPN Cricinfo is a website that is referenced by other journal articles such as (Bhattacharjee and Pahinkar, 2012). The data available in Statsguru is remarkably comprehensive, encompassing nearly every official cricket match ever played. The database's historical records date back to the 1800s, making it a valuable resource for retrieving extensive cricket-related information. Although it does not offer raw ball-by-ball data, the aggregated metrics in Statsguru provide users with valuable insights and statistics for a wide range of cricket matches over the years.

2.5.3 Cricmetric

Cricmetric is a Cricket statistics and analytics website. Like ESPN Cricinfo, Cricmetric does not offer raw ball-by-ball data but provides users with aggregated metrics derived from ball-

by-ball data. These metrics can be filtered based on various criteria, such as competition, club, and country. Users can download the filtered data directly in CSV format if needed. Cricmetric also introduces several innovative in-house metrics, including WPA, RAA, EigenFactor Score, and interactive dashboards, allowing users to analyze and draw insights directly from the website. The work of Cricmetric revolves around three dimensions: data, advanced metrics, and analysis. They have compiled ball-by-ball data for all International Cricket matches played in the last decade. Furthermore, they possess match data for famous T20 tournaments like the Indian Premier League and the Champions T20 tournament. This comprehensive data repository and advanced analytics tools make Cricmetric a valuable platform for cricket enthusiasts, analysts, and researchers seeking detailed statistics and insights from various cricket matches and tournaments.

2.5.4 Cricbuzz

Cricbuzz is an Indian cricket news website that is part of Times Internet. The platform offers a wide range of cricket-related content, including news, articles, and live coverage of matches with multimedia elements such as videos and text commentary. Cricbuzz also provides comprehensive player statistics and team rankings. The website's archives contain aggregated stats like scorecards and point tables from various matches and competitions. Like ESPN Cricinfo, Cricbuzz has historical data for specific matches dating back to the 1800s, making it a valuable resource for accessing cricket statistics and information from past eras (Singh, 2011).

2.6 Introduction to Predictive Analytics

Predictive analytics is a field of data analysis focused on predicting future outcomes by analyzing historical data and patterns. The predictions can be made using statistical algorithms, machine learning approaches, and advanced techniques to analyze data and identify trends, patterns, and correlations to predict future events or behaviors. The figure illustrates the classification of methods used in predictive analytics. These predictive analytics. Forms are grouped into Probabilistic Models, Machine Learning/Data Mining, and Statistical Analysis.

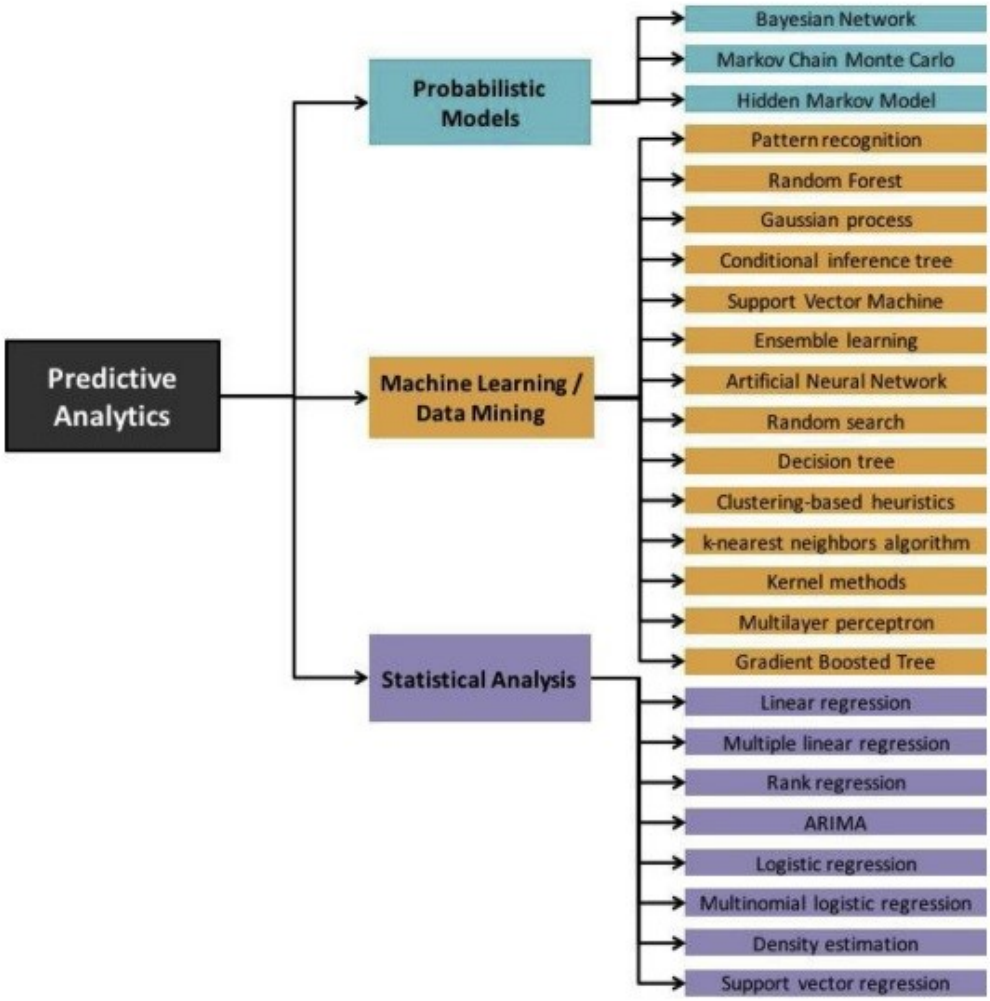

Figure 2.1: Predictive analytics methods (Bousdekis et al., 2020)

2.6.1 Probabilistic models

A probabilistic model quantifies the uncertainty by combining the first principle knowledge with data to represent the dynamics within a distribution of model predictions for state transitions among samples in a batch run (Martínez et al., 2013). This category encompasses representations of uncertain causal relationships between causes and effects. In predictive and prescriptive analytics, probabilistic models are employed to determine the likelihood of specific events occurring rather than simply monitoring actual data to identify events and data points that conform to a set of rules defined by historical analysis (Martínez et al., 2009, 3453-3465).

2.6.2 Machine Learning

Machine Learning refers to algorithms that utilize models and inference to process data without relying on explicit instructions (Nasrabadi, 2007). These algorithms construct a mathematical model from a set of sample data known as "training data" to make predictions or decisions without requiring explicit programming for each specific task. ML is commonly regarded as a subset of artificial intelligence. On the other hand, Data mining involves discovering patterns within large datasets to extract information and transform it into a comprehensive structure for further use. Due to their close interrelation, ML and data mining are considered as one category of methods. By employing these techniques, algorithms can be developed to extract valuable data and reveal essential hidden information. In predictive analytics, the objective is to identify information that can predict future outcomes based on patterns observed in historical data. Conversely, in prescriptive analytics, the focus is on discovering information that can assist in determining the best course of action for a given situation.

2.6.3 Statistical analysis

Statistical analysis is a mathematical field that encompasses data collection, organization, analysis, interpretation, and presentation (Dodge, 2003). It covers various aspects of data, from planning data collection with survey and experiment designs to addressing issues related to statistical populations or models. In predictive analytics, statistical analysis is crucial in extracting valuable information from data and utilizing it to predict trends and behavioral patterns.

2.7 Evolution of Predictive analytics in cricket

Sports analytics originated in the early 20th century when statistical techniques were used to examine sports data. Nevertheless, the fundamental transformation in sports analytics commenced in the late 1970s, propelled by the emergence of computers and the accessibility of more extensive and detailed data sources. Due to technological developments like the snick-o-meter, hawk-eye, Etc., cricket has seen a tremendous shift during the 2000s. Much research has been done on decision-making techniques for cricket team selection, match outcome forecasting, player rankings, and player ratings based on game performance (Borooah and Mangan, 2010).

Summary

Although various studies have been done to develop models for score prediction and winning prediction, many studies are available for analyzing the performance of individual players in cricket. However, the studies on analyzing the performance of individual players and teams could have been more detailed. Hence, an attempt was made to analyze the team and player performance in the IPL and determine the factors influencing match outcomes.

2.8 Review of data visualization and programming tools in sports analytics

Data visualization has a significant role in sports analytics and performance tracking. It aids in transforming raw data into actionable insights, empowering informed decision-making and success both on and off the field. One notable impact of data visualization in sports analytics is its ability to enhance the understanding of player performance. Through visually appealing and easily understandable presentations, coaches and athletes can quickly discern patterns, trends, and areas that need improvement(Kanungo and Bomatpalli, 2019, 4423-4431). Thus, it facilitates targeted training sessions, improves game preparation, and enhances on-field performance. Furthermore, data visualization also plays a crucial role in injury prevention and athlete health management. Sports scientists and medical professionals can identify potential injury risks and implement mitigation strategies by visualizing player workload, fatigue, and recovery data.

2.8.1 Python

The term ”SPYDER” refers to a specific technology or system. Spyder is a robust interactive development environment designed for the Python language. It offers advanced editing capabilities, interactive testing, debugging, and introspection features. Additionally, Spyder serves as a numerical computing environment by leveraging the support of IPython, an enhanced interactive Python interpreter, and well-known Python libraries like NumPy for linear algebra, SciPy for signal and image processing, and matplotlib for interactive 2D/3D plotting. The pandas library is a Python package that offers efficient, adaptable, and expressive data structures specifically designed to facilitate the manipulation of ”relational” or ”labeled” data straightforwardly and naturally. The primary objective of this project is to serve as the foundational component for conducting applied data analysis in Python, focusing on practical

and real-world applications. NumPy is the foundational package for scientific computation in Python programming. The document includes various elements, including linear algebra, Fourier transform, and random number capabilities. Scikit-learn, previously known as scikits-learn, is an open-source machine learning package designed for the Python programming language. The software incorporates various classification, regression, and clustering techniques, such as support vector machines, random forest, gradient boosting, k-means, and DBSCAN. Its design integrates with the Python numerical and scientific libraries, namely NumPy and SciPy (Nimmagadda et al., 2018).

Furthermore, Python is widely used in sports betting and fantasy sports platforms for odds calculation, data analysis, and creating predictive models. Python's versatility and the availability of numerous open-source libraries have made it the preferred choice for sports analysts and organizations, enhancing the overall landscape of sports analytics (Karetnikov, 2019)

2.8.2 SQL

SQL is a fundamental tool in sports analytics that aids in managing and analyzing substantial structured data. Its application in sports analytics is diverse and crucial for various tasks. Here are some critical applications of SQL in sports analytics: SQL is used to create and manage relational databases that store different sports-related data, including player profiles, match statistics, team performance, and historical records. The data can be organized efficiently for easy retrieval and analysis. SQL can aggregate and analyze player performance data, helping coaches and analysts gain insights into individual player strengths, weaknesses, and overall performance trends. SQL enables the comparison of match statistics from different games and seasons, providing valuable insights into team performance, player contributions, and tactical patterns (Pers et al., 2005) SQL queries can help identify patterns and trends related to player injuries, such as injury rates, types of injuries, and their impact on team performance.Data research showed that the outcome of the extra delivery is more influenced by the team's cricket status rather than individual considerations. The obtained dataset is analyzed and categorized into Fact and Dimension tables. Additionally, the identification of range dimensions was conducted to enhance the analysis performed by the end user (Dinesh, 2014).

2.8.3 Power-BI

Power BI, a powerful business analytics tool developed by Microsoft, has found significant application in sports analytics, enabling teams and organizations to derive actionable insights

from their data. One prominent application of Power BI in sports analytics is in performance analysis. Teams collect vast data during games and practices, ranging from player statistics to in-game metrics. Power BI allows analysts to create visually appealing and interactive dashboards that consolidate this data, providing coaches and players real-time insights into their performance. These dashboards can display various metrics, such as shooting accuracy, player movement patterns, and team dynamics, enabling teams to make data-driven decisions for strategy refinement and player improvement.

Researchers and analysts also utilize Power BI for in-depth sports data analysis, uncovering patterns and trends that can inform strategic decisions. This robust analytics tool has become an integral part of sports organizations, allowing them to transform complex data into meaningful insights, leading to improved player performance, enhanced fan engagement, and more informed business strategies (Sherif, 2016).

2.8.4 Microsoft Excel

Microsoft Excel, a widely utilized spreadsheet software, has found diverse applications in sports analytics owing to its user-friendly interface, adaptability, and data manipulation capabilities. Excel is a convenient central repository for storing sports-related data, encompassing player statistics, match outcomes, and team performance metrics. Its organized storage facilitates easy data retrieval for analysis. The software offers fundamental statistical functions, enabling analysts to compute averages, sums, counts, and other summary statistics for player and team performance data. Excel features allow for the creation of straightforward visualizations like bar charts, line graphs, and pie charts, presenting key performance indicators and trends effectively. In sports analytics, Excel proves helpful for fundamental match analysis, facilitating the calculation of run rates, strike rates, and batting averages in cricket, among other relevant metrics (Castellano et al., 2008, 898-905). Also, the software can be employed for budgeting and financial analysis associated with sports teams, encompassing player salaries, revenue projections, and expense tracking.

3 Research methodology

3.1 Introduction

This study aims to examine the performance of players and teams in the IPL by analyzing various performance indicators related to batting, bowling, and fielding statistics. The model uses different machine-learning algorithms. This study explores IPL data and presents the findings through graphical representations and comparative analysis. The insights gained from this study can aid the IPL officials and its fan followers in making decisions about team performance and strategic decisions for the team's better performance. This chapter details further the design and methodology adopted to obtain the study's objectives. It also covers theoretical perspectives on research methodology and design, data collection procedures, tools, and analysis procedures.

3.2 Requirement specifications

All of the necessary conditions for the study are described in depth in this section. Depending on whether particular functionalities are included, particular requirements may change or influence as the project moves forward. The changes' justification will be given.

3.2.1 User Requirements

A requirement lists the features and attributes that software or a product must have to appeal to users. Analysis must be carried out in this investigation. Users should be able to engage with the results displayed on a visualization tool like Power BI. After the project is finished, customers should be able to visit the dashboard, analyze players, choose the greatest player of the year, and evaluate player performance in light of their prior output.

An analysis of Indian T-20 Cricket games will be performed using the data to assess player and team performance. This analysis will look at any potential connections between a player's

performance and the outcome of a game. Websites like espncricinfo.com, Kaggle are used to gather player and match-related dataset, clean the data, and extract pertinent information, including player scores, names, the number of matches won or lost, and team statistics.

3.2.2 Functional Requirements

The essential functional requirements for the study include the following:

- The analyst must collect data from the website.
- The data needs to be cleansed by the analyst.
- The analyst is responsible for analyzing the data.
- The analyst is required to develop a predictive data model.
- The analyst must create a dashboard for the end user with data analysis functionalities.

3.2.3 Research Design

Research is a body of knowledge that explores how things currently exist compared to how they might be (Thomas et al., 2015). The study design is considered a significant part of the research process and serves as a guide for the researcher to gather information. The various available methods of research design are discussed below.

1. **Exploratory research design**

   This type of research design focuses on identifying patterns, ideas, or hypotheses rather than attempting to test or confirm hypotheses. Exploratory studies serve the purpose of formulating problems for further investigation, including the formulation of hypotheses. They enhance the researcher's familiarity with the phenomenon under study or the specific situation in which the research will occur (Thomas and Lawal, 2020). Furthermore, exploratory studies help prioritize areas for future research, clarify concepts, and gather information on the feasibility of additional research. They can also provide insight into the most pressing issues identified by experts in the field. Exploratory studies are often considered the initial step in an ongoing research process, aiding researchers in designing structured investigations. Their primary goal is to gain new insights into a phenomenon. Researchers can refine problem statements or develop hypotheses for further research by seeking fresh perspectives and ideas.

2. **Descriptive research design**

   Descriptive research involves exploring the relationships between variables, testing hypotheses, and developing generalizations, principles, or theories with universal validity. Descriptive studies require a clear frame of reference for the questions being addressed (Peniel, 2015). Unlike experimental designs, the researcher does not manipulate variables or arrange specific events. The research process goes beyond simple data collection and tabulation. It adopts a fact-finding approach, primarily focused on the present, drawing inferences from a cross-sectional study of the current situation. The description is complemented by comparison or contrast, involving measurement, classification, interpretation, and evaluation to demonstrate the significance of the findings. In descriptive research studies, the relationships between non-manipulated variables are examined in a natural rather than an artificial one. The researcher selects relevant variables to analyze their relationships, as the events or conditions have already occurred or exist. These studies involve hypothesis formulation and testing, relying on logical methods of inductive-deductive reasoning to derive generalizations.

3. **Experimental Design**

   An experimental design refers to a systematic arrangement for allocating experimental units to different treatment levels and the corresponding statistical analysis conducted following this arrangement. The formulation of an experimental design encompasses a multitude of interconnected tasks. Development of statistical hypotheses that are relevant to the scientific hypothesis. A statistical hypothesis refers to a proposition that pertains to either (a) one or more parameters of a population or (b) the functional representation of a population. Statistical hypotheses are not typically congruent with scientific ideas but serve as verifiable formulations of scientific assumptions. The selection of treatment levels for the independent variable, identifying the dependent variable to be measured, and managing extraneous conditions (nuisance variables) are of utmost importance. Determining the necessary number of experimental units and the population from which they will be sampled is an essential aspect of the research design. The randomization process for allocating the experimental units to the treatment levels is outlined in detail. The selection of the statistical analysis to be conducted will be determined. In essence, an experimental design delineates the independent, dependent, and nuisance variables while specifying the methodology for implementing randomization and statistical procedures within an experiment. The main purpose of an experimental

design is to establish a causal relationship between the independent and dependent variables. A secondary goal is to obtain the maximum of information with the minimum expenditure of resources (Roger and Alberto, 2009).

4. **Survey Method design**

   In social research, surveys entail collecting data from many individuals. Using postal questionnaires, data can be obtained through face-to-face interviews or from a distance. Several interview types exist, including structured interviews with well-defined questions and ethnographic-style interviews that are highly unstructured. The survey questions may be either open-ended or closed-ended in nature.

5. **Cross-sectional or correlation design**

   The prevailing design often employed in survey research is the cross-sectional design. In this approach, measures are collected from at least two groups of people simultaneously, enabling a comparison of the extent to which these two groups differ concerning the dependent variable (e.g., Level of smoking). An exploratory research design was used in this study to analyze the IPL data. It can take two main approaches: quantitative and qualitative. Quantitative research aims to draw conclusions based on generalizations and the acceptance or rejection of theoretical propositions. On the other hand, qualitative research seeks to understand a phenomenon profoundly and develop new theories (Thomas and Lawal, 2020). The difference between quantitative and qualitative research is the expression of information. Quantitative research deals with numerical data, enabling the use of statistical analysis. In contrast, qualitative research involves subjective data in words where individual interpretation is utilized in the analysis (Thomas and Lawal, 2020). This study adopts a quantitative approach with an observational design to investigate the impact of different performance variables on the IPL-T20 league. According to (Thomas and Lawal, 2020), quantitative research is characterized by its concise and logical reasoning, employing deductive methods. It allows for examining cause-and-effect relationships and facilitates understanding associations, thereby determining causality among variables (Burns and Grove, 2001). This study utilizes statistical data from Cricinfo.com to explore the performance indicators influencing cricket analytics in the IPLT20 competition from 2015 to 2017. (Sharma, 2013) Also identified various indicators in analyzing the performance of players, and this study will specifically investigate the prevalence and impact of the following indicators (variables): batting average,

highest total runs per game, strike rate, fours, sixes, bowling average, economy, strike rate and wickets taken.

3.3 Methodology

The scientific methodology employed in this study is based on the KDD methodology, which involves the process of extracting valuable information from a dataset. This methodology followed in this study was illustrated in Figure

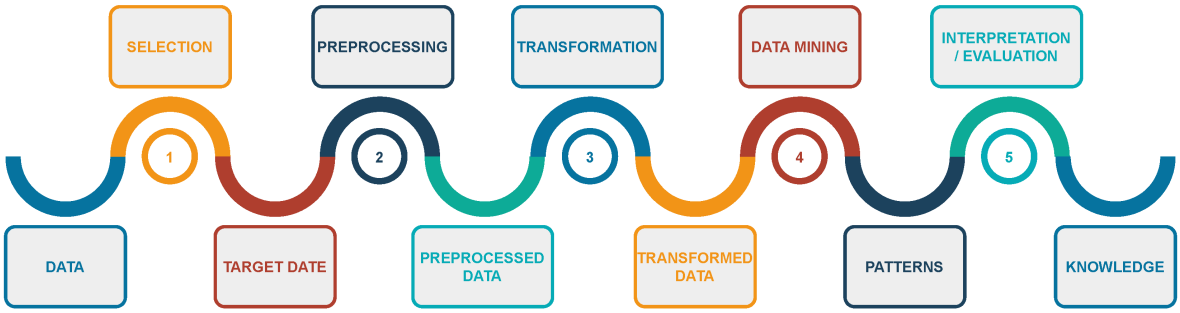

Figure 3.1: Research Methodology (szymon. Kozak Jan and przemyslaw, 2023)

- **Phase-1 (Collection of data):** In this phase of the study methodology, the T-20 cricket league dataset is acquired from the Kaggle website. This step involves carefully selecting and collecting the relevant cricket data that is appropriate for the study.
- **Phase-2 (Pre-processing of data):** In the second phase of the study, the collected cricket data undergoes a cleaning process, which involves eliminating any irrelevant or unnecessary data and leaving only the essential data required for the study.
- **Phase-3 (Transformation):** The third phase involves taking the cleaned dataset produced in the second phase and converting it into a more optimized format. This transformation may include tasks such as dimension reduction and then saved in a CSV file format.
- **Phase-4 (Data mining):** This pivotal phase involves extracting crucial data from vast datasets such as the number of matches won, played and lost by a team.
- **Phase-5 (Interpretation of results):** The results are visualized using the Power-BI

Table 3.1: IPL Winners and Runners-up

Year	Winner	Runner
2008	Rajasthan Royals	Chennai Super Kings
2009	Deccan Chargers	Royal Challengers Bangalore
2010	Chennai Super Kings	Mumbai Indians
2011	Chennai Super Kings	Royal Challengers Bangalore
2012	Kolkata Knight Riders	Chennai Super Kings
2013	Mumbai Indians	Chennai Super Kings
2014	Kolkata Knight Riders	Kings XI Punjab
2015	Mumbai Indians	Chennai Super Kings
2016	Sunrisers Hyderabad	Royal Challengers Bangalore
2017	Mumbai Indians	Rising Pune Supergiant
2018	Chennai Super Kings	Sunrisers Hyderabad
2019	Mumbai Indians	Chennai Super Kings
2020	Mumbai Indians	Delhi Capitals
2021	Chennai Super Kings	Kolkata Knight Riders
2022	Gujarat Titans	Rajasthan Royals
2023	Chennai Super Kings	Gujarat Titans

3.4 Data Collection

This study utilized a quantitative approach with an observational design and retrospective data analyses to determine the batting and bowling performance variables that correlate the highest with the team's success in IPL T20 cricket. The IPL is a well-known and esteemed cricket league that offers aspiring cricketers an excellent platform to showcase their talents and provides an ideal learning environment for young players. The analysis used publicly accessible secondary data from the 2008 to 2022 Indian Premier League T20 tournaments collected from Kaggle. It contains comprehensive ball-by-ball data and matches descriptions (Agrawal Shilpi, 2018). Kaggle is recognized as a Publicly available and dependable dataset, utilized for different purposes and referenced by numerous published authors. The dataset contains an extensive amount of data to be explored. The variables considered for the analysis are total runs scored, maximum runs scored, the number of fours, the number of sixes scored in a cricket match, as well as the wickets lost, wickets lost in the power play, wickets taken in the match, and wickets taken in the power play. The table 3.1 lists out the winner and runner of the IPL from 2008-2022.

3.5 Data Pre-processing and Transformation

Data pre-processing is a critical phase in this study, aimed at ensuring the reliability and accuracy of the subsequent analysis. This phase involves several essential steps: Error and Inconsistency Handling: One of the initial tasks is to address erroneous and inconsistent data. Error handling includes identifying and rectifying any inaccuracies or discrepancies in the datasets. Data accuracy is crucial, as errors could lead to incorrect conclusions (Karthik et al., 2020).

Missing Values:

- **Data Formatting:** The data is formatted to ensure consistency in attributes such as date formats, numerical representations, and naming conventions. This standardization facilitates data integration and analysis.
- **Handling Missing Values:** Detecting and addressing missing values is crucial to data pre-processing. Missing data can significantly affect the quality of machine-learning models. Thorough checks are performed to identify missing values in the datasets, and appropriate strategies are employed to handle them effectively.
- **Duplicate Data Removal:** Duplicate observations or records are identified and removed from the datasets to prevent redundancy and ensure data accuracy.
- **Data Consistency Checks:** Data consistency is examined to ensure that all attributes return valid data and that there are no null entries or inconsistencies within the dataset.

Data Collaboration and Database Creation:

To facilitate comprehensive analysis and modeling, data collaboration and database creation are essential steps:

- **Primary Key Generation:** A primary key is generated by combining the match number and player ID. This primary key is a unique identifier to interlink individual datasets(Barman, 2020, 157-167).
- **Merging Datasets:** All relevant datasets are merged into a unified database based on this primary key. This process combines data from different sources, such as match details, player statistics, and ball-by-ball information, into a single database.
- **Data Frame Creation:** From this unified database, data frames are created to facilitate feature engineering. These data frames incorporate the maximum relevant features required for the analysis.

Server Connection and Feature Selection:

- **Server Connection:** To enable mathematical computations by machine learning models, data attributes such as venue names, seasons, player names, player teams, and opponent team names are transformed into numerical representations. This encoding ensures the data can be effectively utilized in the modeling process.
- **Materialized Views:** Materialized views are created to import necessary data from the database, enhancing the efficiency of data retrieval and analysis.

Feature selection and data encoding are pivotal in preparing the data for machine learning models:

Feature Selection: Specific attributes are selected from the unified dataset to create two new datasets, **"Batsman.CSV"** and **"Bowler.CSV**. These datasets contain relevant information such as match IDs, runs scored, wickets taken, and key match details. Attributes that are not directly related to the analysis, such as umpire names, are excluded to optimize model performance.

Data pre-processing and transformation involve multiple stages of cleaning, integration, and preparation to ensure that the data is in a suitable format for machine learning modeling (Agrawal Shilpi, 2018, 67-71). Once this transformation is complete, the prepared data is ready for testing and training machine learning models. The resulting insights can be exported and visualized through a PostgreSQL server, allowing for informed decision-making and enhanced understanding of player and team performance in the IPL.

3.6 Data Mining

Following data preparation, the dataset is split into training and testing datasets. The training dataset comprises 80 percent of the data, while the remaining 20 percent is used as testing datasets. According to researchers, the most dependable machine learning algorithms, such as SVM, Random Forest, and Decision Tree, are used for the analysis.

3.6.1 K- Nearest Neighbours (KNN)

The K-nearest neighbors (KNN) technique is widely employed as a classification method, utilizing a similarity measure to assign labels to new occurrences or data points. Frequently, it is utilized to categorize data points by classifying their neighboring data points. The parameter K in the KNN algorithm represents the number of nearest neighbors to consider for the majority voting process. The value of K is determined by assessing the degree of similarity between the features of each object. To enhance precision, selecting the appropriate value of K is imperative, a process referred to as parameter tuning. Lower values of K are more vulnerable to errors and will significantly influence the ultimate result. There is a positive correlation between higher values of K and the presence of smoother decision boundaries. This correlation leads to a decrease in variance but an increase in bias. Furthermore, it should be noted that this process is also characterized by its time-consuming nature. In addition to the hyperparameter K, two more hyperparameters in the KNN algorithm require optimization: the distance metric and the distance weight. The KNN diagram is shown in figure 3.2

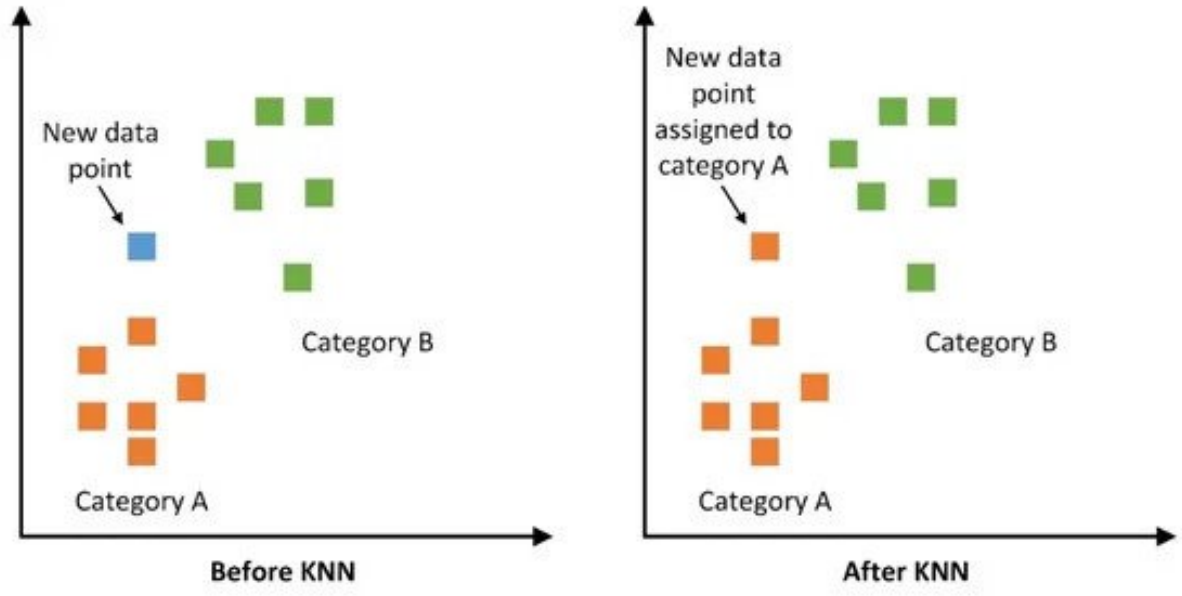

Figure 3.2: K- Nearest Neighbours (Yang et al., 2022)

3.6.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) employ the concepts of margin and support vectors to partition the input into linear and nonlinear structures, generating a highly efficient decision boundary. The margin measures the distance between the boundary and the data points. Within the framework of SVM, the data is divided into multiple groups based on the margin, explicitly focusing on the margin with the highest value. Compared to alternative machine learning methodologies, SVM exhibits superior efficiency and simplicity. Developers often rely on intuition when selecting kernel functions and adjust hyperparameters to establish optimal boundary conditions, facilitating the implementation of their ideas. The hyperparameter in question utilizes data distribution to determine the magnitude of the margin and kernel that collectively form the shape of the border. 3.3

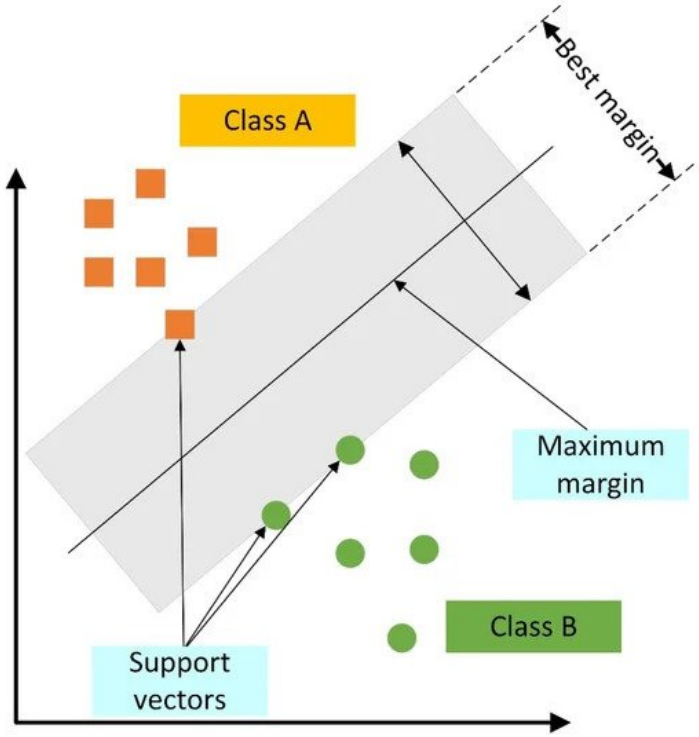

Figure 3.3: Support Vector Machine (Yang et al., 2022)

3.6.3 Decision Tree

The Decision tree is commonly employed for classification tasks. It operates by predefining steps to be followed, earning its label as a supervised algorithm. It facilitates decision-making and problem-solving by offering multiple solutions to a given problem. The tree structure includes a root node at the top, test nodes, and leaf nodes representing the test results. Notably, it boasts speed, efficiency, and minimal time consumption. This method helps prevent errors in problem-solving. The diagram below illustrates how it divides problems into subproblems to reach a decision. The decision tree diagram is shown in figure 3.4

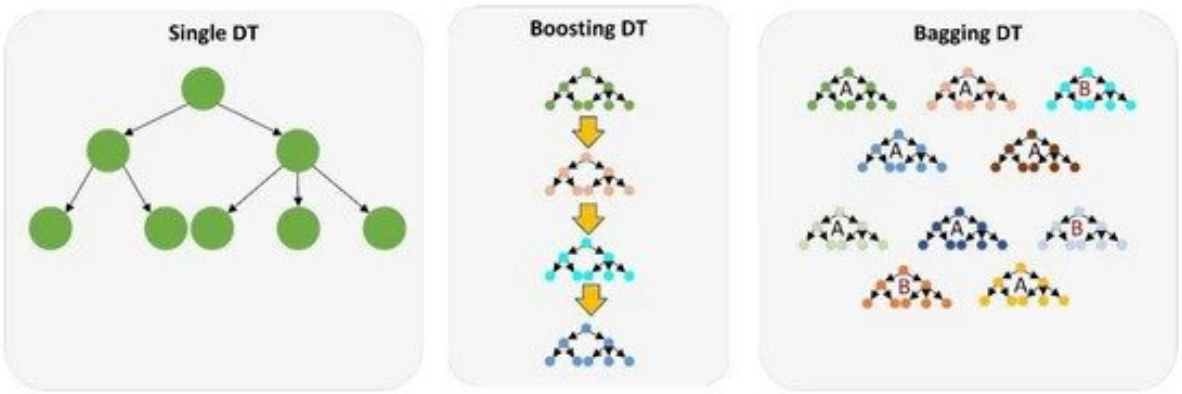

Figure 3.4: Decision Tree (Yang et al., 2022)

3.6.4 Random Forest (RF)

RF is a robust and reliable ML algorithm introduced by Leo Breiman and Adele Cutler (Breiman, 2001, 5-25). It is a classification and regression algorithm that belongs to the bagging algorithm in integrated learning. It was characterized by a Decision Tree (DT), in which a model was constructed based on the randomized training set; the values of various DT's are not correlated and are determined independently, and average results obtained utilizing these decision trees are used in the prediction process. The samples were randomly selected from the training data during the construction of DT instead of using all the data characteristics; some were randomly chosen for training. No proper information is given about which samples are anomalous or which aspects significantly impact classification outcomes during this random feature selection approach. Generally, the accuracy of RF is higher than the DT alone; while solving complex problems, it is widely used in regression and classification contexts. The problem

considered in this paper is a regression problem. The RF regression algorithm outputs the results of all decision trees and then takes the mean value. The regression equation is given as:

$\overline{H}(x) = \frac{1}{K}\sum_{i=1}^{K} h_i(x, \theta_k)$

$\overline{H}(x)$ represents prediction output; h_i represents a single decision tree; θ_k represents an independent distributed random variable that determines the growth process of the single decision tree; K represents the number of decision trees. Each decision tree in the random forest does not capture all the features simultaneously, so each decision tree has uncertainty, which increases the model's generalization ability.

3.6.5 Evaluation metrics for the model

The performance of the model was assessed using metrics such as RMSE, MAE, MSE, and MAPE (Davide Chicco and Jurman, 2021). The Y_i element in the following formulas represents the actual i-th value, whereas X_i represents the expected i-th value. The ground truth dataset's associated Y_i element is predicted by the regression approach for the X_i element. Set two constants: the genuine values median.

- **RMSE** is a mathematical measure derived from the average squared differences between predicted and observed outcomes. This metric is primarily employed in regression analysis and forecasting, where precision is paramount. A lower RMSE value indicates a higher accuracy in the model's predictions, whereas a higher RMSE suggests a more significant disparity between the model's predictions and the actual outcomes.The two numbers, MSE and RMSE, are monotonically connected through the square root. The ordering of models based on RMSE and regression models based on MSE will be the same.

  RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((\hat{y}_i - y_i)^2\right)}$

  n is the number of observations; y_i represents the observed value; $\hat{y}_i$ represents the predicted values.

  (best value = 0; worst value = +∞)

- **MAE**

  If outliers reflect contaminated portions of the data, MAE may be applied. Since the L1 norm in some way smoothes out all the errors of potential outliers, MAE does not penalize training outliers as harshly as it could. MAE gives the model a general and

bounded performance measure. On the other hand, the model's performance will be subpar if the test set contains many outliers.

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i|$$

(best value = 0; worst value = +∞)

- **MSE**

  If any outliers need to be found, MSE can be employed. Due to the L2 norm, MSE is excellent at assigning higher weights to such points. If the model ultimately produces a single inferior prediction, the squaring portion of the function magnifies the error.

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2$$

(best value = 0; worst value = +∞)

- **MAPE**

  Another performance indicator for regression models is called MAPE, which has a relatively simple definition regarding relative error. Because of this, its application is advised for tasks where it is more crucial to be sensitive to relative fluctuations than absolute variations.

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{Y_i - X_i}{Y_i}\right|$$

(best value = 0; worst value = +∞)

3.7 Analysis and Interpretation

3.7.1 Player Performance

The visual representation depicts a comprehensive outline of the architectural framework and methodology employed in analyzing and predicting player performance in the IPL through diverse tools and methods. The following are the primary procedures entailed:

Figure 3.5: Player Performance Architecture

1. The IPL data from 2008 to 2022 has been kept in a PostgreSQL database. PostgreSQL is a robust relational database management system that facilitates SQL queries and data processing operations.

2. The PostgreSQL database is linked to a Jupyter Notebook, an interactive web-based platform that enables users to write and execute Python code and visualize and analyze data.

3. The Jupyter Notebook employs a range of libraries, including pandas, numpy, sci-kit-learn, and Plotly, to analyze data and machine learning operations on the IPL dataset. Several tasks are included:

    - Cleaning and prepping the data involves several tasks, including addressing miss-

ing numbers, outliers, duplicates, and other similar issues.

- Exploratory data analysis encompasses various techniques, including identifying descriptive statistics, relationships, and distributions.
- The feature engineering and selection process involves various techniques, such as creating new variables, transforming existing variables, and the reduction of dimensionality, among others. Machine learning modeling involves several key steps, including selecting suitable algorithms, partitioning the data into separate training and testing sets, fitting and assessing the models, and optimizing the hyperparameters, among other considerations.
- Machine learning prediction involves using learned models to predict or generate probability for fresh or unexplored data.
- The Jupyter Notebook utilizes the PostgreSQL library for establishing a connection with the PostgreSQL database. This connection enables the storage and updating of data analysis and machine learning outcomes within the same database.

4. The PostgreSQL database is integrated with Power BI, a business intelligence and data visualization solution that enables users to generate interactive dashboards and reports from diverse data sources.

5. Power BI utilizes the data obtained from the PostgreSQL database to generate a diverse range of visual representations, such as charts and graphs, which effectively illustrate the patterns and observations about players' performance in the IPL.

   The key four charts and graphs include:

   - Most Valuable batter (batting average):
   - Most Valuable bowler (batting average):
   - Most impactful batter (batting strike rate):
   - Most impactful bowler(bowling economy rate):

   Power BI also allows users to draw visual insights, strategic decisions, and key takeaways from the visualization for all the KPIs

3.7.2 Team Performance

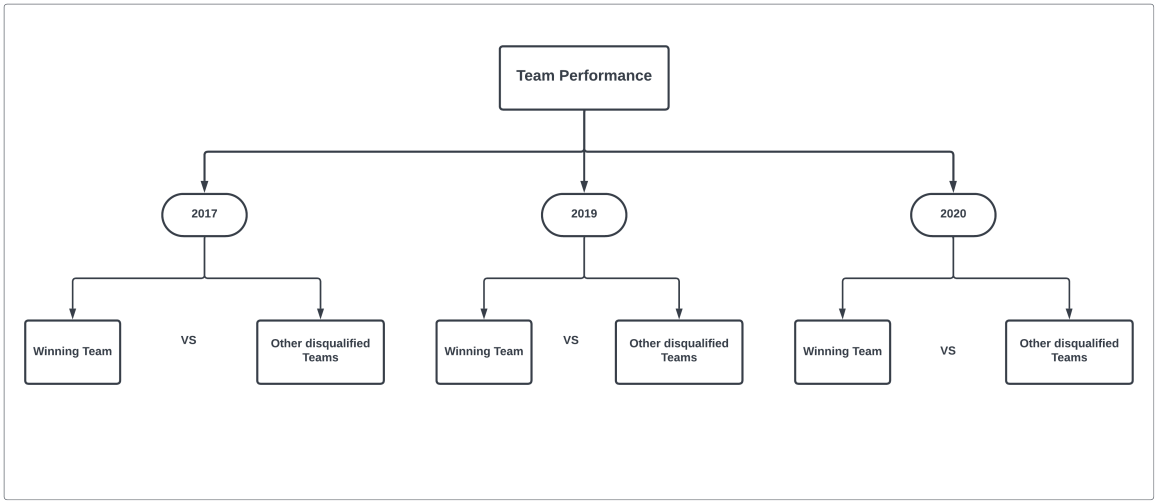

Figure 3.6: Team Performance Overview

1. **Data Selection:**

   A comprehensive examination was undertaken to analyze teams' performance in the IPL, with particular emphasis on three years of data spanning 2017, 2019, and 2020. This period provides a thorough perspective on recent seasons of the IPL, as it encompasses the Mumbai Indians'(winning team) triumphant performances on three occasions during this span.

2. **Research Objective**

   The research objective of this study is to investigate and analyze the phenomenon under investigation to get a deeper understanding and contribute to the existing body of knowledge in the field. The main research topic is evaluating and contrasting the performance of the Mumbai Indians with other teams that did not secure a spot in the IPL playoffs during the specified years of 2017, 2019, and 2020. This comparison research focuses on the KPI associated with batting and bowling statistics. These KPIs include batting average, strike rate, boundary percentage, bowling average, bowling economy rate, and dot ball percentage.

3. **Source of Data and Data Preparation**

   The data utilized in this study was obtained from a PostgreSQL database, which served as a central repository for IPL-related data from multiple sources. These sources encompass match records, player information, ball-by-ball specifics, and other relevant

data. Materialized views were designed to enhance the efficiency of data retrieval and processing.

4. **The segmentation of data**

   The data was categorized based on the outcomes of matches, distinguishing between games won and matches lost for the specific years of 2017, 2019, and 2020. This approach allowed for a more detailed and nuanced analysis of the data. The implementation of this division facilitated the acquisition of strategic insights, visual insights, and critical takeaways about team performance.

5. **Comparative analysis**

   This study aims to conduct a comparative analysis of the selected variables. The research involved doing a thorough comparison analysis, explicitly examining the performance of the Mumbai Indians to other clubs that did not qualify for the playoffs in the IPL. This analysis aimed to investigate performance patterns, areas of proficiency, and areas for improvement by applying the designated key performance indicators (KPIs) for both batting and bowling elements.

6. **The Process of Selecting KPI's**

   Key performance indicators were selected with great attention to detail to assess team performance comprehensively. The KPI offer valuable insights into how much a team's batters and bowlers contribute to achieving success or facing elimination in the IPL). KPI related to batting, such as batting average, strike rate, and boundary percentage, provide valuable insights into a team's capacity to score runs. In bowling, KPI's like average, economy rate, and dot ball percentage serve as metrics that gauge a team's proficiency in bowling.

7. **Data visualization using Power-BI**

   Data visualization refers to the graphical representation of data to facilitate understanding and analysis. It uses many visual elements, such as charts, graphs, and maps. Utilizing Power BI Power BI, a robust data visualization tool, was applied to enhance the display and interpretation of research findings. This process enabled the generation of dynamic and perceptive visual depictions of data and analysis findings, thereby streamlining intricate data for a broader range of viewers.

8. **The Implications and Significance of the Research**

   Examining team performance, explicitly considering the Mumbai Indians' significant success and the implementation of other teams, provides vital insights for stakeholders, enthusiasts, and decision-makers involved in the IPL. Comprehending the factors that contribute to the success or disqualification of a team provides valuable insights for team management plans. It stimulates conversations regarding the competitive dynamics of T20 cricket.

The chosen methodological approach for this study is described in this section. The research methodology employs a systematic strategy for gathering, converting, and analyzing data. The PostgreSQL database is widely recognized for its robust capabilities in data management, while materialized views offer an efficient means of extracting data. The selected KPI provide precise and targeted information regarding the team's performance. The application of Power BI for data visualization serves to improve the accessibility and interpretability of the data.

3.8 Various performance indicators used for Player and Team Performance

The players' performance is assessed using various metrics as discussed below.

3.8.1 Measures for Batters

Innings: This parameter reflects the total count of innings in which the batsman has participated up to the current match date. It indicates the batsman's experience, where a higher number of innings suggests a more experienced player.

Batting Average: The batting average represents the mean number of runs scored by the player per inning. This metric indicates the player's ability to achieve runs (Gaur and Bhattacharjee, 2016).

$$\text{Average} = \frac{\text{Runs scored}}{\text{Number of match played}}$$

Strike Rate (SR): The strike rate is the average number of runs scored by the player per 100 balls faced. In limited-overs cricket, scoring runs quickly is crucial because teams have

limited overs to play. A lower strike rate, meaning fewer runs scored per 100 balls, can harm the team's performance. This metric indicates how quickly the batter can score runs.

$$\text{Batting Strike Rate} = \left(\frac{\text{Total Runs Scored}}{\text{Total Balls Faced}}\right) \times 100$$

Centuries: This parameter denotes the number of innings in which the batsman has scored more than 100 runs. Fifties: This measure represents the number of innings in which the batsman scored more than 50 runs but less than 100. Zeros: This parameter indicates the number of innings in which the batsman is dismissed without scoring any runs. Highest Score (HS): This refers to the maximum number of runs scored by a batsman in a single innings throughout their career.

Batting Strike Rate in Powerplay, Middle Overs, and Death Overs:

Batting strike rate is a crucial indicator that assesses a batsman's capacity to score runs fast in the Powerplay, middle overs, and death overs. It is stated as the number of runs a batsman scores for every 100 balls faced. The ability to analyze batting strike rates in various game situations, including the Powerplay, Middle Overs, and Death Overs, offers important clues about a player's adaptability and aggression.

- **Powerplay:** A high strike rate is preferred during the Powerplay (overs 1-6) as it aids the team in laying a solid foundation. Batters that are successful in this stage are frequently aggressive and look to take advantage of field limitations.
- **Middle Overs:** Batsmen balance aggression with forming partnerships in the Middle Overs (overs 7–15). In this stage, a high strike rate demonstrates the capacity to keep up a scoring pace without taking unwarranted risks.
- **Death Overs:** The Death Overs (overs 16-20) demand aggressive batting to maximize runs. Finishers who can speed up the team's score are batters with a high strike rate in this phase.

Boundary Percentage:

Boundary Percentage: In the IPL, boundary percentage is a crucial batting KPI. It shows the percentage of a batsman's runs (fours and sixes included) from boundaries. A player who consistently finds openings in the field and clears the boundary ropes has a high boundary percentage. Batters who often hit boundaries are essential members of their

teams because they may swiftly boost the run rate and maintain the scoreboard.

3.8.2 Measures for Bowlers

Innings: This parameter reflects the total count of innings in which the bowler has participated up to the current match date. It serves as an indicator of the experience of the bowler, where a higher number of innings suggests a more experienced player.

Overs: This parameter represents a bowler's total number of overs bowled. It indicates the bowler's experience, where a higher number of overs bowled suggests a more experienced bowler.

Bowling Average: The bowling average refers to the number of runs a bowler concedes per wicket taken. This metric reflects the bowler's ability to limit opponents' scoring while taking wickets. A lower bowling average indicates higher capabilities in restricting runs and achieving wicket-taking success (Bhattacharjee and Pahinkar, 2012).

$$\text{Bowling average} = \frac{\text{Total runs conceded}}{\text{total wickets taken}}$$

Bowling Economy Rate: This metric expresses the average number of runs the bowler gives up per over. This rate is crucial for evaluating a bowler's performance in limited-over cricket. A bowler with a lower economy rate is more effective in limiting the opponent's scoring. Maintaining a low economy rate is particularly difficult in the Powerplay, where fielding limits are in place, making it a valuable indicator to gauge a bowler's efficacy.

$$\text{Bowling economy rate} = \frac{\text{Total runs conceded}}{\text{Total balls bowled}} \times 6$$

Bowling Strike Rate: The bowling strike rate is the number of balls bowled by a bowler per wicket taken. This metric signifies the bowler's ability to take wickets effectively. A lower bowling strike rate indicates that the bowler can take wickets quickly.

$$\text{Strike rate} = \frac{\text{Number of balls bowled}}{\text{Number of wickets taken}}$$

Dot Ball Percentage:

Dot Ball Percentage: The percentage of bowled deliveries (dot balls) that do not result in any runs being scored by the batsman is known as the dot ball percentage. A bowler can put the batter under strain by continuously bowling difficult-to-score-off deliveries, according

to a high dot ball percentage. Dot ball pressure bowlers frequently induce errors in batters, resulting in wickets and fewer possibilities to score runs. These batting and bowling KPI's offer a thorough picture of player performance in the IPL, assisting teams in making defensible choices regarding player selection, tactics, and team makeup.

4 Analysis and findings

4.1 Data Architecture:

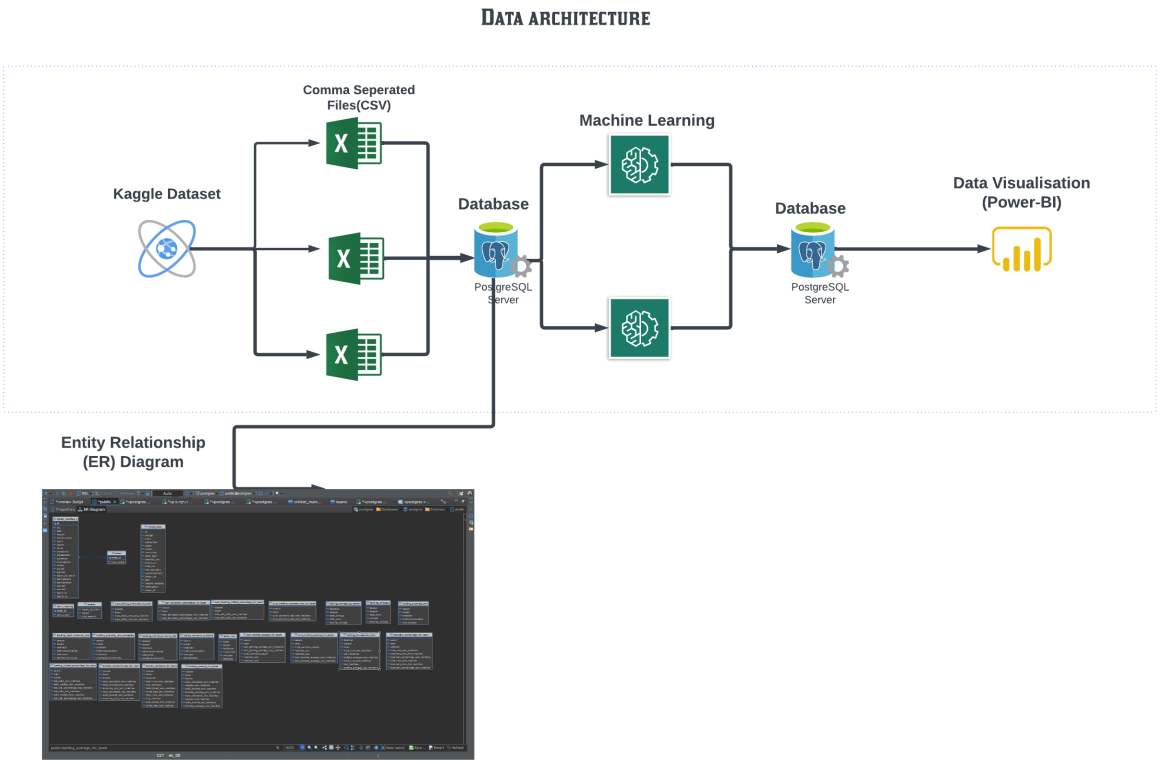

Figure 4.1: Data Architecture

4.1.1 Data Collection:

Downloading data from Kaggle was the first stage in the data collection process. The dataset ”Matches.csv” provides a comprehensive collection about IPL matches. It includes facts such as the venue, location, season, participating teams, toss-related information, match outcomes,

winning margins, players of the game, umpire details, and the victorious team. The dataset titled ”Deliveries.csv” provides a comprehensive and detailed record of every ball bowled in all Indian Premier League (IPL) matches spanning from 2008 to 2022. The dataset encompasses essential statistical information about each delivery, encompassing comprehensive details regarding the match, bowling team, batting team, the batters engaged, the bowler, non-striker, the nature of runs achieved (singles, doubles, fours, sixes), supplementary runs (such as no-balls and penalty runs), over specifics, and the cumulative runs obtained from the specific delivery.

4.1.2 Database Creation in PostgreSQL:

Data tables were created within a PostgreSQL database using CSV files downloaded from Kaggle. The two datasets, Matches.csv and Deliveries.csv, were imported into the PostgreSQL database to consolidate and structure the data for analytical purposes.

4.1.3 Data Transformation:

Refinement and Organization: Within the PostgreSQL environment, data transformation activities were carried out. These included processes like data aggregation, adjustments of data types to suit analysis requirements, and the creation of tables that optimized the data for subsequent analysis.

Team and Player ID Creation: Unique Team IDs and player IDs were generated during this stage to ensure data consistency and efficient referencing in the subsequent analysis.

4.1.4 ER Diagram Utilization: An Entity-Relationship (ER):

An Entity-Relationship (ER) diagram visually represented the database structure. This diagram was pivotal in elucidating the relationships between different data entities, thereby assisting in database design and comprehension. The ER diagram for this analysis is shown in 4.2

4.1.5 Python Integration and Data Transformation:

Python libraries and PostgreSQL connectors connect the database and the Python environment. This connection facilitated data extraction from the database and its transformation into Python data frames.

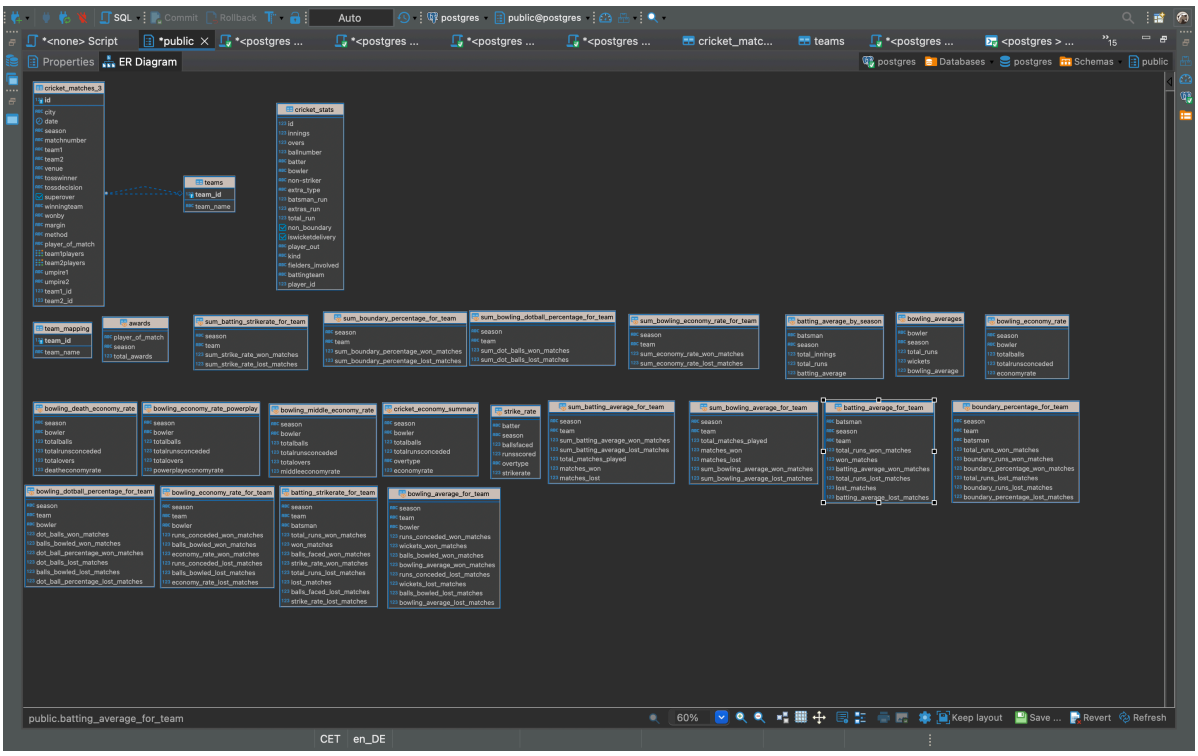

Figure 4.2: ER Diagram

4.1.6 Machine Learning Model Implementation:

Various machine learning models, including Support Vector Machines (SVM) and Random Forest (RM) were employed to analyze the data. These models were instrumental in identifying the most valuable batters and bowlers based on distinct criteria within the dataset.

4.1.7 Server for Visualization:

The same server was employed for both data analysis and visualization. This server facilitated the creation of visualizations in Power BI. These visualizations allowed for effectively presenting insights regarding player and team performance.

4.2 Player Performance

Player performance is crucial in deciding a team's performance in professional T20 cricket leagues like the IPL. Understanding and assessing player performance is a complex process beyond simple statistics. It dives deeply into the subtleties of bowling and batting, shedding

light on players' contributions during different game phases.

The KPI used in this thesis section provides in-depth insights for player performance analysis. Prediction and data analysis are two separate domains that are investigated. The study uses sophisticated machine learning algorithms to analyze KPI like batting and bowling averages in the prediction domain and determine the most valuable batter and bowler.

The analysis additionally explores the dynamic elements of player performance throughout various stages of a T20 cricket match. Batting strike and bowling economy rates emerge as vital metrics, shedding light on the most impactful players during the Powerplay, Middle Overs, and Death Overs. Understanding player contributions during these periods can reveal strategies that go beyond the realm of simple statistics. These phases represent crucial turning points in a T20 game.

This part investigates the intersection of data science and cricket, using analytical methods and modeling techniques to decipher the mysterious realm of player performance. The lessons gained from this investigation provide a profound understanding of the people who determine the outcome of their teams on the cricket pitch. The entire data architecture for player performance is also explained in 3.7.

4.2.1 Most Valuable Batter:

1. **Introduction**:

   The batting average is a crucial cricket statistic that is a fundamental measure of a Most Consistent Batter. The calculation involves dividing the aggregate amount of runs accumulated by the batter by the frequency at which the batter is dismissed. A more excellent batting average indicates a batter who has achieved better skill and success.

2. **Overview of the Data:**

   - The dataset comprises comprehensive information regarding the batters who have participated in IPL cricket matches from 2008 to 2022. The dataset has been filtered to encompass just those batters who have played over ten innings in a given season and have achieved a batting average of over 30. The Random Forest machine learning model generates predictions for batting averages by leveraging certain variables. These forecasts are then graphically shown on a scatter plot juxtaposed with the corresponding values.

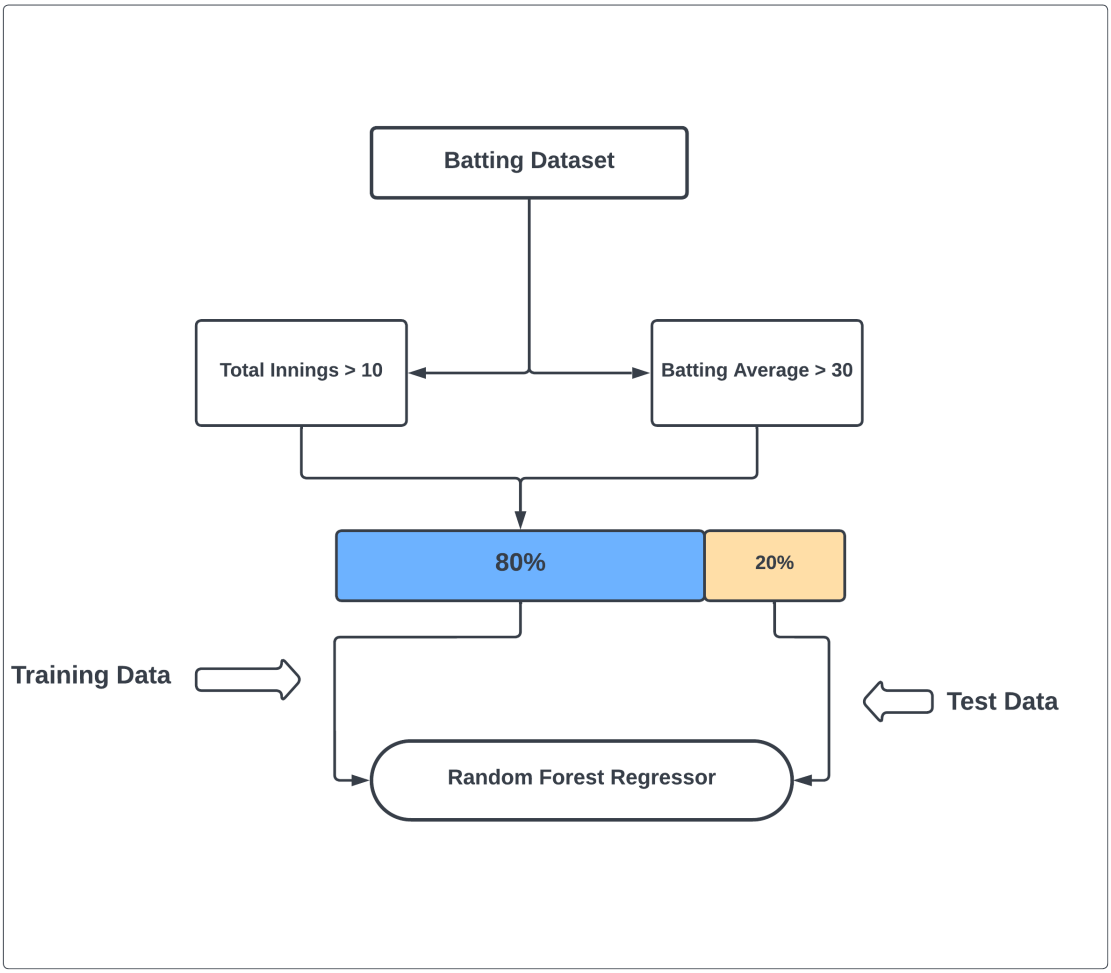

Figure 4.3: Most valuable Batter Architecture (Random Forest)

- The graph depicts the relationship between the anticipated and actual batting averages of the cricketers engaged in the IPL from 2016 to 2022.
- The graph displays three colors for its data points: blue, green, and red. The data points represented by the color blue relate to hitters whose batting average is less than 30. The green data points correspond to hitters whose batting average ranges from 40 to 50. The data points represented in red color indicate hitters who have attained a batting average of 50 or above.

3. **Visual Insights**

- The graph displays a diagonal linear regression line, demonstrating a positive slope. The graphic suggests a good correlation between the projected and observed bat-

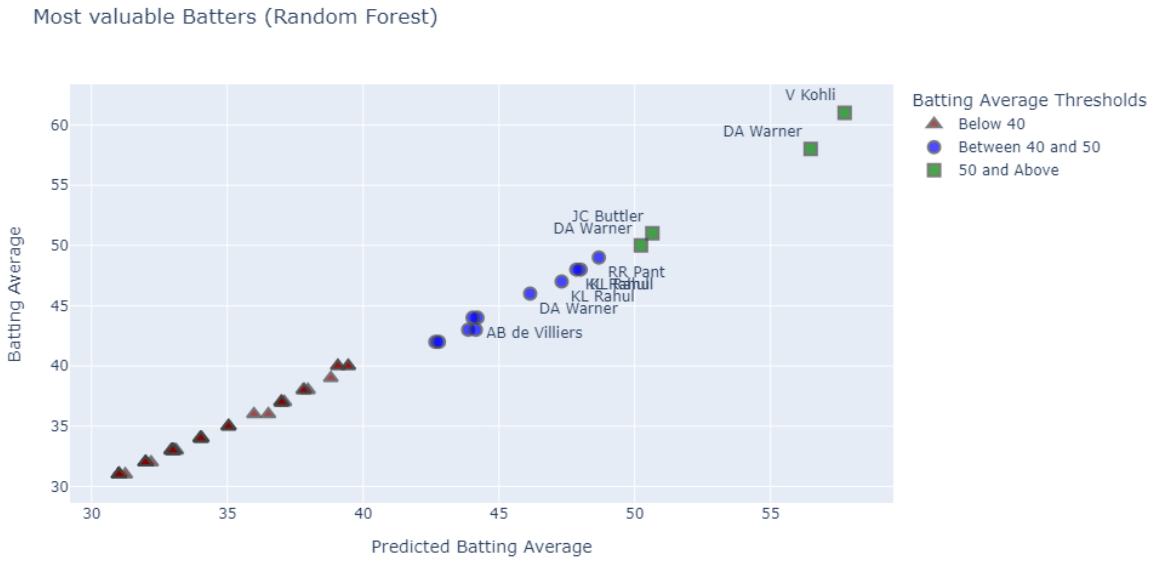

Figure 4.4: Most valuable Batter (Random Forest)

ting averages. The precision of a forecast improves as the number of data points converges toward the regression line.

- Several notable data points include the names of prominent cricket players, namely ”V Kohli,” ”JC Buttler,” ”DA Warner,” ”AB de Villiers,” and ”KL Rahul.” The individuals mentioned herein are regarded as highly appreciated batters in the IPL since their outstanding batting statistics closely correspond to their expected performance.
- The random forest model demonstrates a robust predictive capability for the batting average of batters in the IPL, as most of the data points closely align with the line of best fit. Batters who exhibit a high predicted batting average also show a correspondingly high actual batting average, indicating their consistent performance and reliable contributions to their respective teams. Conversely, batters with a low predicted batting average also exhibit a low actual batting average, suggesting their limited effectiveness and efficiency in scoring runs. Batters demonstrating a substantial disparity between their observed and projected batting averages can be classified as outliers. This discrepancy suggests that these individuals have performed extraordinarily in specific matches while underperforming in others. The reliability of these entities is still being determined and poses a significant risk.

In the IPL, players who maintain batting averages beyond 40 significantly impact the outcome of matches. These individuals can independently guide their teams toward triumph, rendering them invaluable assets inside the league.

4. **Strategic decisions:**

   - Teams can deliberately deploy resources and investments by favoring individuals projected to have high batting averages. The use of this focused strategy has the potential to enhance the efficiency and effectiveness of team formation. The provided insights aid in achieving a harmonious equilibrium between star athletes who possess remarkably high averages and consistent performers who consistently exceed a score of 40, guaranteeing the formation of a comprehensive and formidable team.
   - Teams should allocate resources towards acquiring batters who possess both high projected and observed batting averages, as this is likely to result in an increased number of runs scored and a higher probability of winning matches.
   - Conversely, teams should exercise caution when considering batters with low projected and observed batting averages, as they are more prone to squandering balls and losing wickets, which can negatively impact team performance.
   - Teams should exercise caution when facing batters who exhibit high actual but low expected batting averages, as these individuals are more likely to demonstrate inconsistency and experience a lack of success.

5. **Key Takeaways:**

   The random forest model is a valuable technique for assessing the worth of hitters in the IPL) by considering their batting averages.

   The batting average is critical in assessing the performance and potential of hitters participating in the IPL. The worth of batters in the IPL is contingent upon their ability to consistently, reliably, efficiently, and effectively accumulate runs. By leveraging data-driven insights to optimize player performance, clubs participating in the IPL have the potential to gain a competitive advantage. The comprehension of distinct characteristics that contribute to achievement can be utilized to enhance individual and collective performances, hence augmenting the likelihood of triumph.

6. **Summary:**

The RMSE score remains constant at 0.4699 and is more than just a statistical measure. IPL clubs approach the auction round and player selection meetings with solid assurance, given the shallow RMSE achieved. The individuals possess a high level of confidence in the dependability of their predictive tool, enabling them to plan their actions with meticulous accuracy to achieve success strategically.

SVM Model for IPL Most Valuable Batter

1. **Introduction:**

   The present section serves as an introduction to the topic at hand. SVM is a powerful machine-learning approach commonly used for regression or classification tasks. SVM strives to discover the best hyperplane that splits data points into discrete groups or predicts their values. A hyperplane is a geometric construct visualized as a line in two dimensions, a plane in three, or a surface in higher-dimensional spaces.

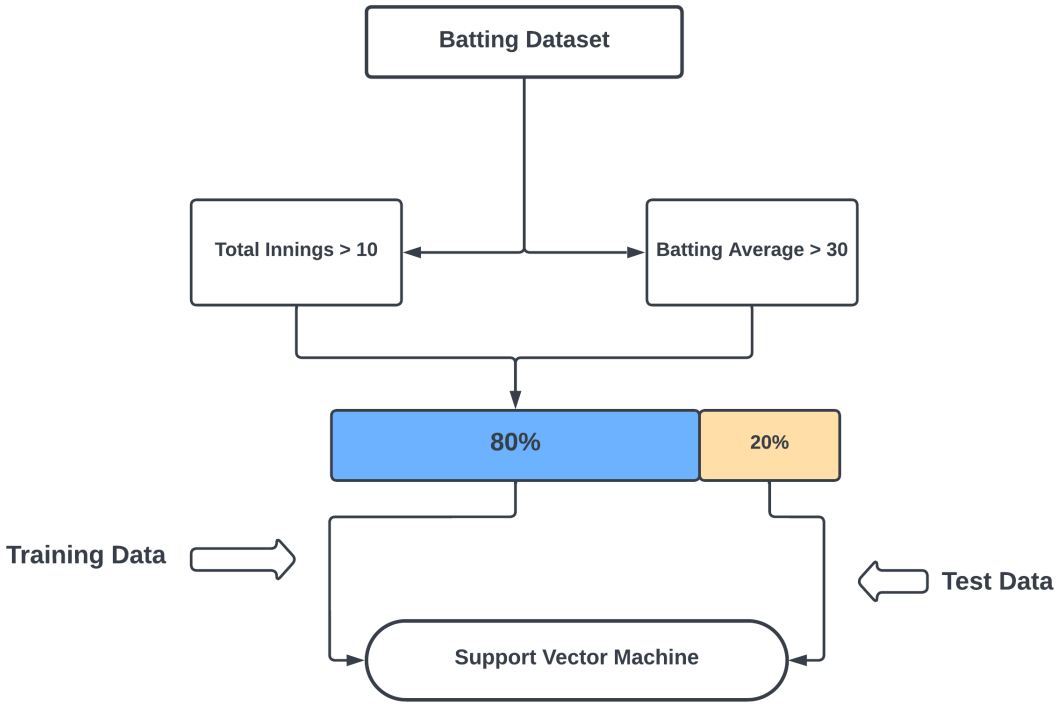

Figure 4.5: Most valuable Batter Architecture (Support Vector Machine)

2. **Data Overview**.

   The dataset comprises data on batters who have participated in IPL cricket matches from 2008 through 2022. The dataset has been subjected to a filter that explicitly includes

batters who have played over ten innings in a given season and have achieved a batting average of over 30. The SVM machine learning model generates predictions for batting averages by leveraging certain variables. These predictions are then graphically shown on a scatter plot with the corresponding data.

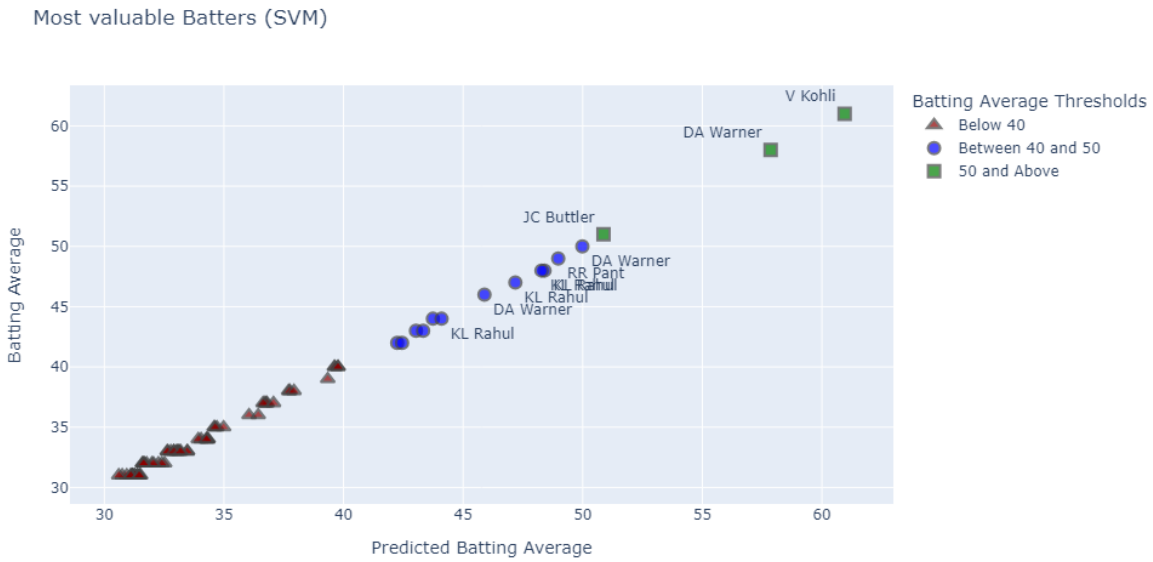

Figure 4.6: Most valuable Batter (Support Vector Machine)

3. **Visual Insights:**

   - Examining the visual depiction yields significant observations. The scatter plot graph depicts the correlation between each batter's observed and projected batting averages. The degree of proximity to the diagonal line indicates the accuracy level in making predictions.
   - The data points depicted on the graph are categorized into three groups based on color: Red, representing batting averages below 40; Blue, representing batting averages between 40 and 50; and Orange, representing batting averages of 50 and above.
   - The graph illustrates that most of the data points are below the regression line, suggesting that the SVM model frequently tends to underestimate batting averages. Possible explanations for this disparity may involve:
     - The presence of strange data points.

  - The presence of random fluctuations in the data.
  - The absence of pertinent and meaningful variables in the predictive model.

- Batters whose batting averages over 30 may not consistently garner prominent media attention. Yet, they play a crucial role in establishing stability and serving as the bedrock for their respective teams, making substantial contributions to the overall performance. Players who maintain batting averages exceeding 30 are a dependable support system for star players, enabling them to exhibit their innate skills and abilities with a sense of assurance.
- The graph highlights several batsmen, namely V Kohli, DA Warner, JC Buttler, and KL Rahul, with considerably higher accurate batting averages than their forecasted averages. These players probably possess distinctive skills or traits not adequately captured by the SVM model, rendering them important assets to their respective teams.

4. **Strategic decisions:**

- Strategic decisions should prioritize the selection of batters with high actual and expected batting averages since this approach is likely to yield consistent and high-scoring performances.
- Additionally, this research emphasizes the significance of players with batting averages exceeding 40, hence emphasizing their pivotal contribution to the club's overall performance. The regular and reliable contributions of individuals frequently serve as the foundation for the overall effectiveness of a team. Players who possess batting averages exceeding 40 in the IPL are not merely statistical frontrunners but individuals who can secure victories for their teams. Their exceptional performances substantially influence the overall results of IPL matches.
- It is advisable to exercise prudence when evaluating batters with low actual and expected batting averages, as their performance may be subpar.
- One should assess batters who exhibit high actual batting averages but low predicted batting averages, as they may contain latent abilities not accounted for by the SVM model, hence offering potential avenues for enhancement.
- It is advisable to use caution when dealing with hitters with low actual batting averages but high predicted batting averages since they may possess undiscovered

restrictions or obstacles not identified by the SVM model.

5. **Key Takeaways:**

   - The SVM is a valuable tool for predicting batting averages by considering multiple characteristics. However, it is essential to note that SVM has flaws and some limits.
   - Discrepancies between observed and projected batting averages can be attributed to various variables, including outliers, noise in the data, the absence or irrelevance of specific attributes, or outstanding individual abilities.
   - The study provides valuable support for strategic decision-making processes involved in the selection or evaluation of batters, considering their actual and expected performances.

6. **Comparative Analysis: SVM vs Random Forest**

   - In this analysis, comparing the SVM) model and its counterpart, the Random Forest model. The SVM model exhibits superior prediction accuracy, as evidenced by its reduced root mean square error (RMSE) compared to the Random Forest model, establishing its superiority. Due to this characteristic, SVM model emerges as the preferred choice for IPL teams seeking precise forecasts of player batting averages.
   - To assess the effectiveness of the SVM model, initially examining its RMSE metric. This pivotal measure functions as a benchmark for assessing the precision of predictions. The SVM model demonstrates a noteworthy RMSE value of 0.3078, indicating its exceptional ability to forecast batting averages with a high level of precision accurately. The decreased RMSE seen in the SVM model, in comparison to the Random Forest model, highlights the enhanced dependability of the SVM model. Consequently, this renders the SVM model an essential resource for decision-makers in the context of IPL.
   - The categorization precision of the SVM model is notable due to its ability to capture a more detailed range, specifically encompassing players with average scores ranging from 40 to 50. Both models classify players according to their projected averages. Due to enhanced precision, IPL teams possess a more comprehensive understanding of player performance and potential.
   - The interpretability of the result interpretation is enhanced by the SVM model's distinctive color-coding technique for visual representation. Using graphic repre-

sentation has improved the accessibility of decision-making processes, particularly in the context of team ownership and player selection. This visual tool facilitates the rapid identification of players who align with the strategic objectives and plans of team owners and selectors.

4.2.2 Most Valuable bowler:

1. **Introduction:**

   Bowling averages, a fundamental metric in the realm of cricket, provide valuable insights into the overall performance of bowlers. As mentioned above, the standards are derived by dividing the aggregate amount of runs a bowler concede by the corresponding number of wickets they have successfully captured. A lower bowling average is indicative of a greater degree of bowling proficiency.

   The Random Forest algorithm, which falls within the domain of machine learning, utilizes the combined capabilities of decision trees to generate predictions. Every decision tree is trained using a randomly selected sample of both data and features. The ultimate prediction outcomes are derived by taking the average of the predictions made by all the trees within the ensemble.

2. **Data Overview:**

   The dataset encompasses the IPL, containing seasons spanning from 2008 to 2022. The filtration criteria encompass bowlers with more than 15 wickets and a bowling average below 25, specifically from 2016 onwards.

3. **Visual Insights:**

   - The analysis of visual representations provides valuable insights. The scatter plot employs a random forest model to compare bowlers' actual and anticipated bowling averages. The x-axis denotes the anticipated bowling average, while the y-axis signifies the observed average. The graph displays three blue, green, and red lines, corresponding to bowling averages of 16, 20, and 24, respectively.
   - The random forest model has a high degree of conformity with most bowlers, as evidenced by the near alignment of data points with the blue line, confirming precise predictions.

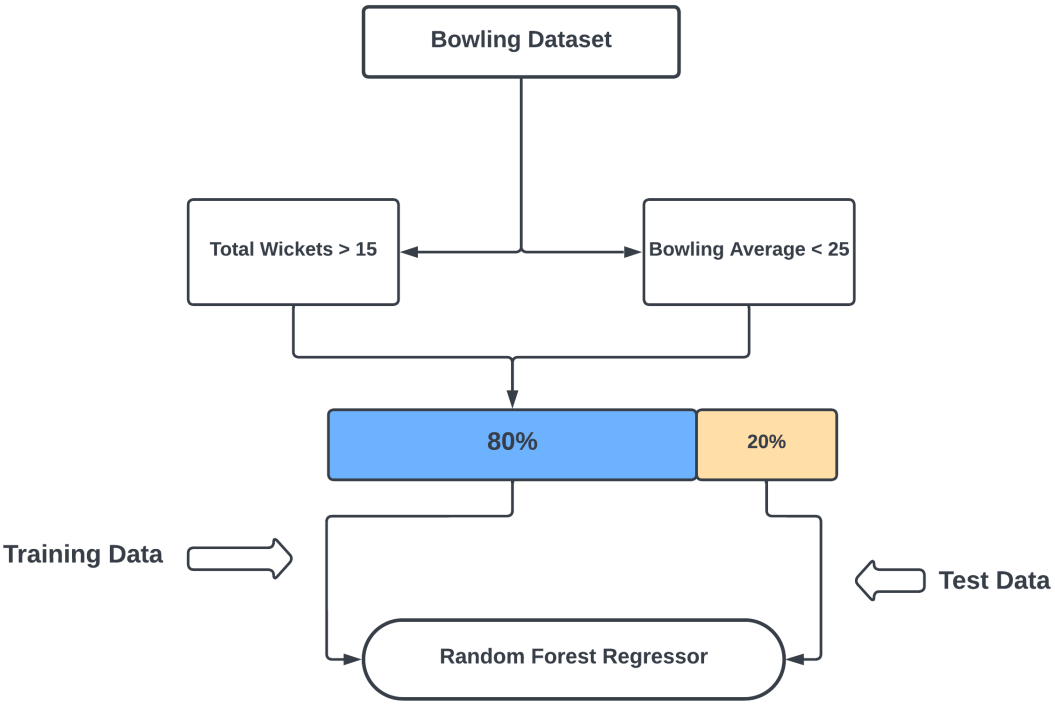

Figure 4.7: Most valuable Bowler Architecture (Random Forest)

- The scatter plot is a graphical representation illustrating the relationship between projected and observed bowling averages. Bowlers with high performance tend to concentrate on the upper sections, underscoring their sustained excellence.
- The analysis recognizes the significant contribution of bowlers with averages below 20, highlighting their pivotal significance in earning triumphs for their respective teams. The individuals who consistently take wickets play a crucial role as the foundation of a team's bowling strategy, making essential contributions to their progress in the Indian Premier League.
- Players who possess bowling skills of exceptional quality and maintain an average below 16 have the potential to significantly influence the outcome of an IPL match, altering the course of the game. The performances of individuals have a substantial impact on the final results of matches.
- The empirical data reveals that A.D. Russell's observed bowling average stands at 21.05, while the predictive model expects it to be 16.52, thus implying a tendency to underestimate his performance. Conversely, T. Perera's average is 23.76,

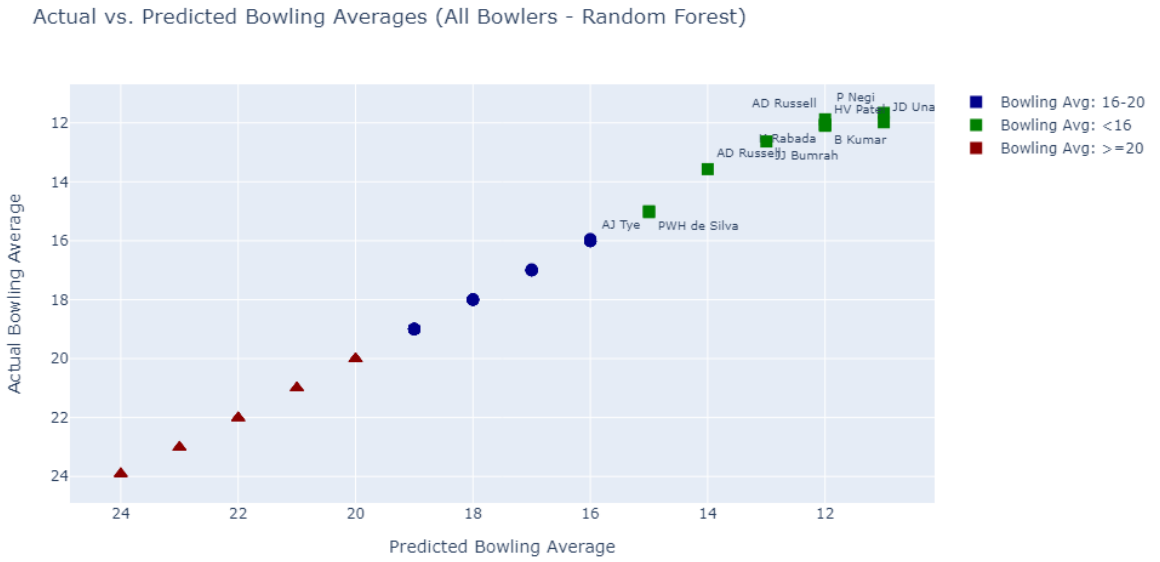

Figure 4.8: Most valuable Bowler (Random Forest)

although the model predicts it to be 24.38, showing an overestimation.

- P. de Silva poses a challenge by demonstrating a significant discrepancy between the actual average of 16.88 and the expected average of 20.71, indicating a considerable prediction error.

4. **Strategic decisions:**.

   - Teams can effectively allocate resources and assets by employing a strategic approach that prioritizes bowlers projected to have high bowling averages. The aforementioned focused strategy aims to maximize the effectiveness of team composition by achieving a harmonious combination of exceptional bowlers and reliable performers under the age of 20.
   - A considerable value and cost-effectiveness bowler, A.D. Russell merits retention and further opportunities.
   - T. Perera, a bowler with a high cost and a relatively low number of wickets, may benefit from substituting a more cost-effective alternative or being utilized strategically.
   - P. de Silva demonstrates potential despite occasional inaccuracies in predictions. Additional examination is necessary to reveal the strengths and shortcomings and

improve coherence. • These highly skilled bowlers, comparable to Kagiso Rabada and Rashid Khan, frequently guide their teams to victories that significantly impact their future.

- The analysis above highlights the importance of bowling averages to assess bowlers' performance. The Random Forest algorithm has been identified as a reliable and effective approach for accurately predicting bowling averages. Additionally, this analysis provides insights into the discrepancies between observed and projected means, informing strategic decision-making.

5. **RMSE as a Measure of Model Success:**

   Within this captivating exploration of statistical analysis, the RMSE indicates achievement. IPL clubs enter the auction round and player selection meetings with a high confidence level, as noted in an impressively low RMSE value of 0.2402. The individuals exhibit a high level of assurance in the precision of their forecasting technologies, hence facilitating their ability to strategize their path toward success through careful and precise calculations.

SVM for Most Valuable Bowler

1. **Introduction:**

   Bowling averages play a crucial role as a performance indicator in the sport of cricket, providing valuable insights into the effectiveness of cricketers throughout their time on the field. The calculation involves dividing the aggregate amount of runs conceded by the bowler by the cumulative number of wickets taken. A lower bowling average indicates a bowler's higher level of proficiency. Bowling averages play a crucial role in facilitating the comparison and rating of bowlers across different formats and groups within the game.

   The SVM is a powerful machine-learning technique commonly used for regression or classification applications. The SVMalgorithm aims to determine the best hyperplane that effectively separates data points into various groups or predicts their values. The algorithm demonstrates proficiency in managing non-linear connections and high-dimensional data, showcasing its adaptability by utilizing different kernels and parameters.

2. **Data Overview**

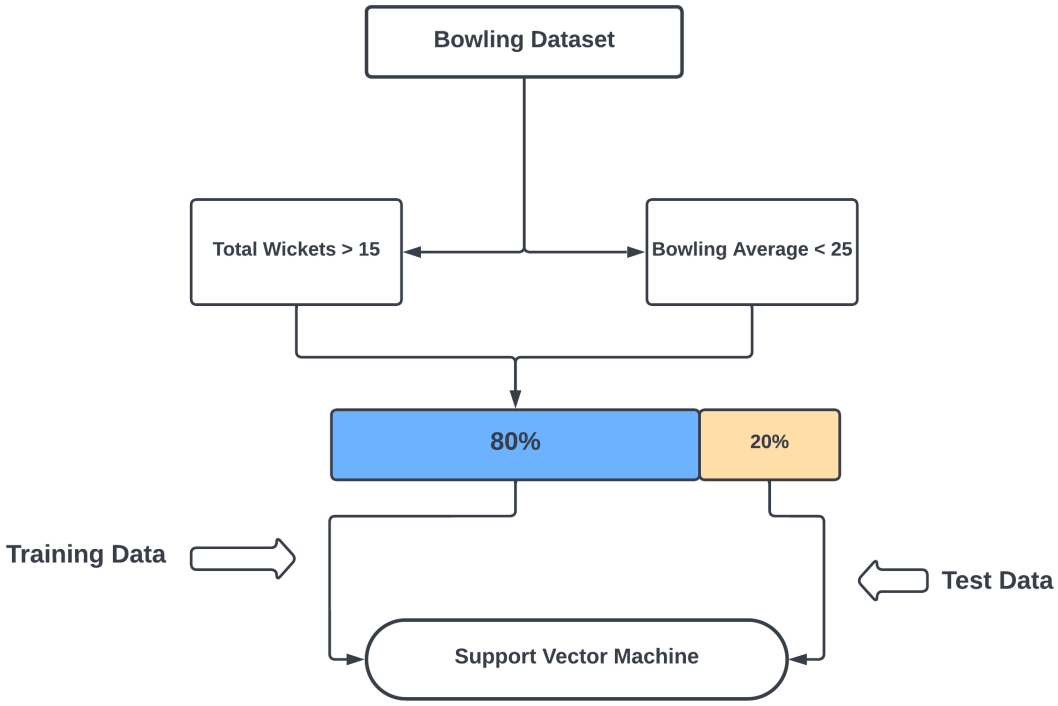

Figure 4.9: Most valuable Bowler Architecture (Support Vector Machine)

The IPL is a highly regarded Twenty20 cricket tournament in India every year since 2008. The visual information consists of a scatter plot graph labeled ”Actual vs. Predicted Bowling Averages (All Bowlers - Support Vector Regression).” The presented diagram depicts the bowling averages of eight IPL bowlers who actively participated in the tournament between 2016 and 2022. Additionally, the graph compares the observed and projected values created by the SVM model. The x-axis is designated as ”Predicted Bowling Average,” while the y-axis is denoted as ”Actual Bowling Average.” The graph comprises two discrete data point categories: blue circles and red triangles. The bowling averages of A. Russell, J. Unadkat, B. Kumar, A. Tye, J. Bumrah, and T. Perera are represented by blue circles, but red triangles denote the bowling averages of K. Rabada and D. Silva. The graph displays two reference lines: a green line indicating a bowling average of 16 and a purple line marking a bowling average of 20.

3. **Key Visual Insights:**

   - The SVM model consistently underestimates the bowling averages of most bowlers, as evidenced by data points that fall above the diagonal line representing perfect

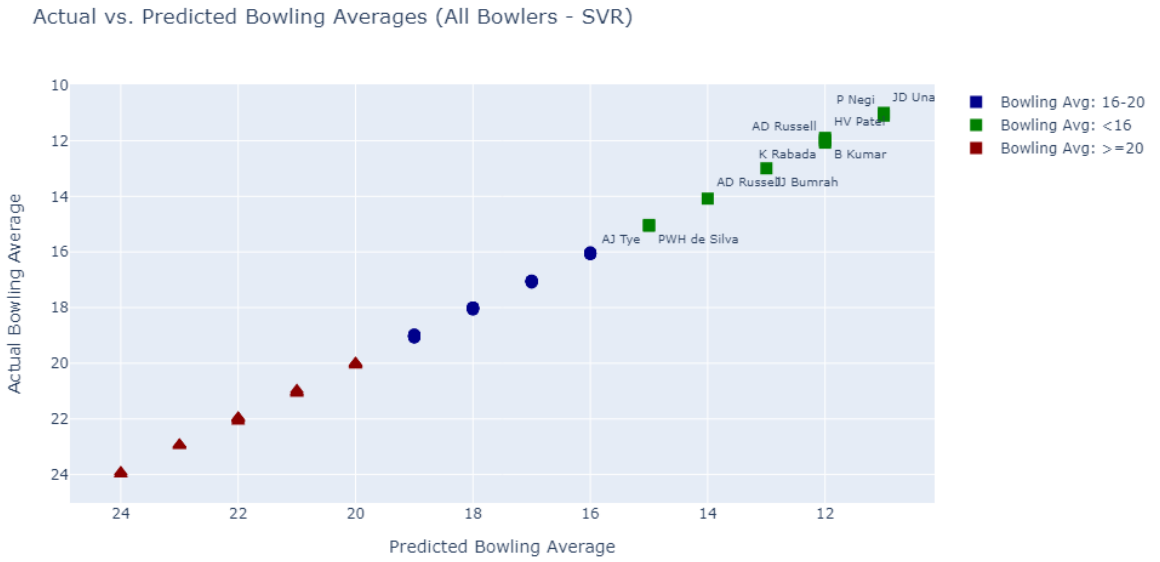

Figure 4.10: Most valuable Bowler (Support Vector Machine)

predictions. Notably, the model exhibits more significant prediction errors for bowlers with lower actual bowling averages, specifically K. Rabada and D. Silva, who are outliers within the dataset.

- On the other hand, the SVM model performs better in predicting the bowling averages of bowlers with higher actual averages, such as A. Russell and J. Unadkat, as their predicted values align closely with the diagonal line. However, the model needs help to accurately capture the variance in bowling averages for certain bowlers, such as B. Kumar and A. Tye, who display a wide range of actual values.
- It is worth mentioning that within the dataset, there are no bowlers whose basic or forecast bowling averages fall below 16, widely regarded as an excellent benchmark in the context of T20 cricket. Furthermore, it is worth noting that among the bowlers in question, namely K. Rabada and D. Silva, only these two individuals own bowling averages that fall below the threshold of 20. It is essential to acknowledge that such figures are widely considered to be praiseworthy within the context of this particular format.

4. **Strategic Decisions:**

- The topic of discussion pertains to strategic decisions. By analyzing the visual representation, one can derive valuable information that can inform the development of strategic decisions. Teams should consider retaining K. Rabada and D. Silva as their primary bowlers, as they have consistently performed well in the Indian Premier League (IPL). On the other hand, teams with A. Due to their higher bowling averages, Russell and J. Unadkat may want to explore alternative roles, such as all-rounders or batters. This could result in them conceding more runs. Bowlers like B. Kumar and A. Tye should be closely monitored regarding their form and fitness, as their bowling averages have shown fluctuations, which could impact their reliability.
- Teams that heavily depend on the bowling abilities of J. Bumrah and T. Perera may consider investigating alternative options or backup bowlers, given that their bowling averages are relatively average within the context of this highly competitive tournament.
- The main points to remember are: This analysis highlights the importance of bowling averages as a crucial performance metric in cricket, particularly in the T20 format, where the value of each run is significant. The SVM has demonstrated its efficacy as a powerful tool in predicting and analyzing cricket data. However, it is crucial to recognize and address its use's inherent limitations and constraints. Moreover, the IPL's dynamic and highly competitive environment requires continuous evaluations and improvements of players and teams.

5. **Comparison of RMSE:**

   - Random Forest RMSE: 0.2402
   - SVM RMSE: 0.0549

   The Support Vector Machine (SVM) model is a powerful tool used for identifying proficient bowlers in the IIPL and predicting bowling averages. The remarkable forecasting accuracy, comprehensive categorization, and visual depiction of IPL teams enable them to access crucial information for making strategic decisions. The Support Vector Machine (SVM) model provides precise insights into bowling performance more effectively than the Random Forest model. Using data-driven analysis and machine learning models, such as Support Vector Machines (SVM), will play a pivotal role in shaping the future of cricket and enhancing team performance as the IPL progresses.

Comprehensive Analysis of the Most Impactful Players in IPL

The present analysis serves as a bridge between the PostgreSQL data store and the graphical functionalities of Power BI, offering a complete perspective on the performance of both players and teams. The investigation largely centers on two KPI: Batting Strike Rate and Bowling Economy Rate.

The primary framework utilized in this examination is the Batting Strike Rate, which seeks to assess several facets of IPL matches during the Powerplay, Middle Overs, and Death Overs. This analysis presents insights into identifying individuals who exhibit outstanding performance in several aspects, providing valuable perspectives for teams seeking to optimize their batting lineups and leverage their players' skills.

Moving on to the second facet of our research, specifically the Bowling Economy Rate, Evaluating the competence of bowlers in limiting the number of runs conceded per over. As the cricket matches progress from the Powerplay phase to the Middle Overs and subsequently to the Death Overs, it becomes evident that certain players demonstrate exceptional performance in particular stages of the game. These insights allow IPL clubs the capacity to make pivotal judgments concerning their bowling selections and formulate strategies that have the potential to exert a substantial influence on the game's outcome.

The ability to make decisions in real-time is of utmost importance in defining the level of success achieved within the ever-changing domain of T20 cricket. The study offers significant findings that can give teams a competitive edge. These insights enable teams to deploy players strategically, adapt to the game's dynamics, and make educated choices by considering the situational context.

Our research integrates data-driven expertise with the excitement associated with IPL cricket. By integrating PostgreSQL and Power BI, Delving into the fundamental aspects of player and team performance and uncovering the key contributors who shape the outcomes of their respective teams. The Batting Strike Rates and Bowling Economy Rates are essential for understanding the intricate dynamics of IPL cricket. These metrics provide valuable insights into the contribution of each run scored, and every dot ball bowled towards attaining success.

4.2.3 Most Impactful Batter

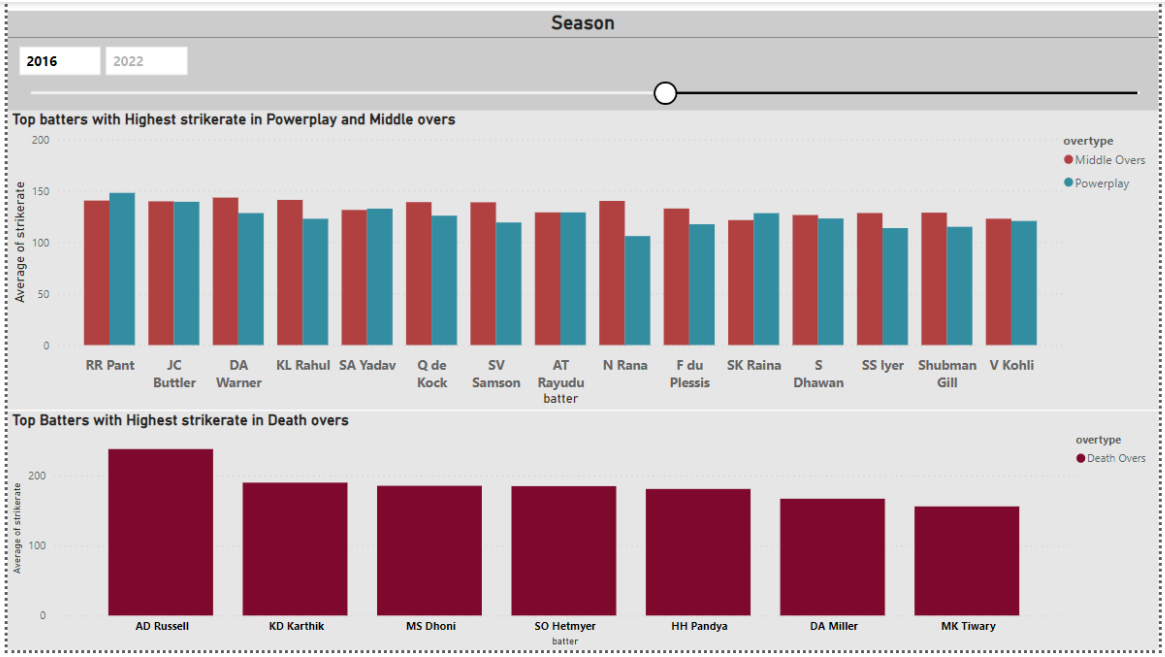

Figure 4.11: Most Impactful Batter (Batting Strike Rate)

1. **Introduction**

   The strike rate in cricket is a metric used to quantify the speed at which a batsman accumulates runs. The calculation involves dividing the total number of runs scored by the batsman by the total number of balls faced, followed by the multiplication of the quotient by 100. A greater strike rate indicates a hitter with a more aggressive and practical approach, whereas a lower strike rate suggests a more cautious and defensive batting style. The strike rate holds particular significance in limited-overs cricket, wherein both teams are allotted a certain number of deliveries to accumulate the maximum amount of runs. The powerplay phase refers to the initial six overs of an inning, during which the fielding team is restricted to having a maximum of two fielders positioned beyond the 30-yard circle. The middle overs refer to the period of play in cricket between the conclusion of the Powerplay and the commencement of the death overs, encompassing the final five overs of an innings.

2. **Data Overview:**

The dataset offers an overview of the average strike rate of different IPL batters across three distinct game phases: the Powerplay, Middle Overs, and Death Overs. The strike rates provide a concise representation of the batters' efficiency in scoring runs throughout these pivotal stages of the game.

3. **Seasonal Slicer:**

   A slicer has been implemented to filter the dataset, commencing from the 2016 season. This decision follows the contemporary period of IPL cricket, which has observed the development of strategies and the presence of dynamic gaming.

4. **Powerplay - Setting the Tone:**

   The Powerplay phase, which encompasses the first six overs of a cricket match, is strategically utilized by teams to exploit the limited fielding restrictions and establish a solid base for their batting effort. The following are several significant discoveries:

   - **Rishabh Pant (148):** With a notable strike rate of 148, Rishabh Pant showcases his prowess as a formidable force during the Powerplay phase. The player's proficiency in rapidly accumulating runs throughout the initial stages of the innings offers Delhi Capitals a dynamic commencement.
   - **David Warner (143.4):** David Warner has shown to be a reliable asset for the Sunrisers Hyderabad team because of his continuous performance in the Powerplay, where he maintains an impressive strike rate of 143.4. The individual's capacity to identify and exploit lacunae and delineate limits is paramount.
   - **KL Rahul (141.17):** KL Rahul, with a strike rate of 141.17 during the Powerplay, plays a crucial role in the success of the Punjab Kings. The team benefits from his adeptness in blending assertiveness with strategic precision, resulting in a formidable beginning.

5. **Middle Overs - Building Momentum:**

   During the middle overs of a cricket match, which typically spans from the 7th to the 15th over, the batting team must adopt a well-rounded strategy. This phase of the game necessitates a focus on consolidating the innings and establishing partnerships between batters. The following are the results of our analysis:

   - **Andre Russell (167.3):** Andre Russell, with a strike rate of 167.3, has an outstanding ability to score runs rapidly throughout the Middle Overs, establishing

himself as a formidable presence in the game. The player's capacity to rapidly increase speed and effectively confront opposing bowlers during this period proves advantageous for the Kolkata Knight Riders.

- **Suryakumar Yadav (132.03):**Suryakumar Yadav, with a strike rate of 132.03, offers Mumbai Indians the necessary steadiness in the Middle Overs to construct significant scores. The constant performance exhibited by the individual in question is deserving of recognition.
- **Shubman Gill (130.05):** Shubman Gill, with a strike rate of 130.05, demonstrates his proficiency in maintaining a high scoring rate during the Middle Overs, thereby establishing himself as a dependable pillar for the Kolkata Knight Riders. The individual effectively employs a strategy of alternating striking rotations while also strategically timing boundary shots.

6. **Death Overs - Finishing with a Flourish:**

   The Final Stages of a Cricket Match - Concluding with a Display of Skill and Precision. The final phase of a cricket match, known as the Death Overs (expressly overs 16-20), necessitates a strategic and assertive approach to batting to optimize the scoring of runs. The following are significant findings:

   - **Andre Russell (238):** Andre Russell has exceptional performance in the Death Overs, exhibiting an impressive strike rate of 238. The consistent capacity of the individual in question to effectively surpass established limitations renders him a transformative force within the context of the Kolkata Knight Riders.
   - **Dinesh Karthik (190):** Dinesh Karthik's strike rate of 190 in the Death Overs showcases his adeptness in concluding innings for the Kolkata Knight Riders. His ability to identify openings and score boundaries is quite helpful.
   - **MS Dhoni (185.5):** MS Dhoni's strike rate of 185.5 during the Death Overs for Chennai Super Kings indicates his prowess as one of the most accomplished finishers in the history of the IPL.
   - **Shimron Hetmyer (185):** Shimron Hetmyer's notable strike rate during the latter stages of the innings contributes significantly to the overall batting prowess of the Royal Challengers Bangalore team. The individual's assertive approach proves advantageous in high-pressure circumstances.

- **Hardik Pandya (181):** Hardik Pandya's impressive strike rate of 181 during the Death Overs renders him an indispensable asset to the Mumbai Indians team. The individual's exceptional ability to consistently execute clean hits and demonstrate innovative thinking in high-pressure situations is noteworthy.
- **David Miller (167):** David Miller's remarkable performance of scoring 167 runs during the Death Overs exemplifies his exceptional skills as a finisher for the Kings XI Punjab (formerly known as Punjab Kings).

  The analysis above emphasizes various participants' importance during different stages of an IPL match. While specific individuals demonstrate proficiency in initiating dynamic beginnings, others succeed in cultivating collaborative relationships, and a limited number exhibit great aptitude in concluding tasks. Teams deliberately employ these insights to maximize their batting order and exploit the strengths of their players.

7. **Key takeaways:**
   - **Diverse Player Roles:** The composition of IPL teams is characterized by a broad array of players, each with distinct skill sets that enable them to thrive in various aspects of the game. The recognition and utilization of these positions are crucial for the achievement of team success.
   - **Dominance in the Powerplay:** Powerplay Dominance: Notable players such as Rishabh Pant, David Warner, and KL Rahul exhibit exceptional performance during the Powerplay phase, establishing the momentum for their respective teams. The ability to score rapidly in the initial six overs of a cricket match frequently leads to the accumulation of substantial final scores.
   - **Middle Overs Anchors:** During the Middle Overs, batters such as Suryakumar Yadav and Shubman Gill play a crucial role in providing steadiness to the team. The players' capacity to alternate strikes and locate sporadic boundaries contributes to the continuous progression of the scoreboard.
   - **Death Overs Finishers:** The Death Overs are a period in cricket where some players, such as Andre Russell, MS Dhoni, Dinesh Karthik, Shimron Hetmyer, and Hardik Pandya, excel. The players' remarkable rates of successful strikes and adeptness in finishing plays can significantly alter the game's outcome.

- **Team Strategy:** The formulation of team strategy in the IPL is contingent upon delineating player responsibilities. They guarantee that the appropriate individuals are assigned to suitable places to optimize their influence on the game.
- **Adaptability:** The ability to adapt one's gameplay to various scenarios and stages of a match is highly esteemed among players. The attribute of versatility holds significant value in the context of T20 cricket.
- **Game Situation Awareness:** IPL teams who achieve success demonstrate a comprehensive understanding of the importance associated with situational awareness. Players can discern the appropriate moments to increase their pace, solidify their position, or conclude the game with a strong performance, all contingent upon the specific circumstances of the match.

4.2.4 Most Impactful Bowler

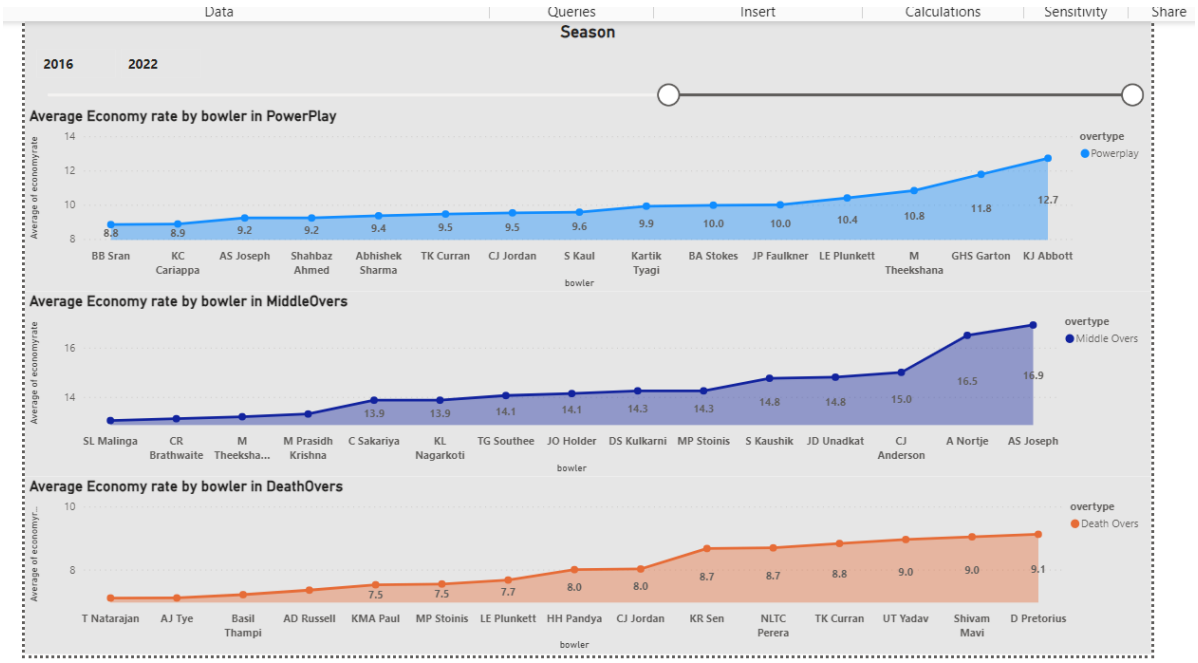

Figure 4.12: Most Impactful Bolwer (Bowling Economy Rate)

1. **Introduction:**

   The bowling economy rate, a crucial metric in cricket, measures the number of runs a bowler allows per over-delivered. This metric is a vital indicator of a bowler's success,

directly mirroring their capacity to limit the rate at which the other team scores. A lower economy rate indicates a bowler with greater efficiency in minimizing the number of runs given throughout their overs. The visual representation offers a comprehensive examination of the bowling economy rates exhibited by players across different stages of IPL matches from 2016 to 2022. The data used in this study was carefully retrieved and filtered from a complete dataset of the IPL) covering the period from 2008 to 2022. The aggregation process was conducted using PostgreSQL. The graph consists of three unique segments, each containing a line graph that illustrates the average economy rate of bowlers during different phases of a match: PowerPlay, Middle Overs, and Death Overs.

2. **Data Overview:**

   The data presented pertains to the Bowling Economy Rate KPI, a crucial indicator for evaluating player performance in the IPL. The dataset has been divided into three phases of a T20 cricket competition: the Powerplay, Middle Overs, and Death Overs. The economy rate, a metric that measures the number of runs given per over, provides valuable insights into a bowler's proficiency in limiting the opposition's scoring. The present investigation examines the performances of bowlers during each phase, aiming to elucidate their efficacy in sustaining a low economy rate.

3. **Data Filtering Slicer:** A tool for data filtering is known as a slicer. To accurately capture the dynamics of recent seasons of the IPL, a slicer has been utilized to facilitate data filtering starting from 2016. This strategy decision ensures the analysis of the most up-to-date trends and patterns in the economic rates of bowling. The present section will analyze the collected data and present the findings from the analysis.

4. **Powerplay - Early Impact:**

   The early impact of powerplay in the mentioned context reveals notable performers and areas for improvement. BB Sran and KC Cariappa demonstrated exceptional performance during the powerplay, exhibiting commendable economy rates of 8.84 and 8.88, respectively. Conversely, GHS Garton and KJ Abbott encounter difficulties in maintaining economic rates during this initial phase, with rates of 11.78 and 12.72, respectively, suggesting potential opportunities for enhancement.

5. **Middle Overs - Stabilization Phase:**

During the stabilization phase, also known as the middle overs, certain bowlers such as SL Malinga (13.05) and CR Brathwaite (13.12) have displayed exceptional skill in maintaining low economy rates. This indicates their ability to restrict the scoring rate during mid-phase innings effectively. Most bowlers tend to experience an increase in their economy rates compared to the powerplay phase, emphasizing the difficulty of containing the opposition's scoring as the innings progresses.

6. **Death Overs - Crucial Finish:**

    - final overs of a cricket match, commonly called the "Death Overs," play a crucial role in determining the game's outcome. In this context, T Natarajan and AJ Tye have demonstrated excellent skills as experts in the Death Overs.
    - With economy rates of 7.1 and 7.10, respectively, they have effectively contained the opposition's scoring rate during high-pressure situations. Their ability to restrict runs during these critical moments showcases their proficiency in this aspect of the game.
    - Although D Pretorius (9.12) had a somewhat elevated economy rate in this phase, it is crucial to acknowledge the inherent difficulty in sustaining a low economy rate during the Death Overs.

7. **Strategic Decisions:**

    - **Powerplay specialists:** Teams may strategically utilize bowlers with exceptional economy rates during the Powerplay phase to capitalize on early dismissals and pressure the opposing team's batting order.
    - **Middle Overs Stability:** During the Middle Overs phase, it is imperative to prioritize stability and containment. Bowlers like SL Malinga and CR Brathwaite, who excel in this phase, can be critical assets for their respective teams.
    - **Death Overs Finishers:** In the context of cricket, the Death Overs necessitate the presence of bowlers who can effectively curtail the occurrence of boundary hits and impede late-stage accelerations in the scoring process. Players such as T Natarajan and AJ Tye can conclude innings effectively, providing teams with a valuable advantage.
    - **Match Situation-Based Selection:** In the context of team selection, teams should employ a flexible strategy and make decisions regarding the choice of bowlers

based on the particular circumstances of the match. The utilization of a bowler in high-pressure situations can be strongly impacted by their performance in various phases.

8. **Key Takeaways:**

   - Significant Observations: Bowlers frequently experience an increased trend in their economy rates as the cricket match progresses from the Powerplay phase to the Middle Overs stage and ultimately to the Death Overs phase.
   - Identifying the strengths of bowlers during different parts of the game can offer teams significant strategic opportunities.
   - Maintaining a low economy rate during the Death Overs poses a significant difficulty, yet it can greatly impact the outcomes of matches.

4.3 Team Performance

Introduction:

The Mumbai Indians, a prominent franchise within the IPL, have established a renowned and esteemed reputation in T20 cricket. Having achieved an impressive tally of five championship triumphs in the seasons preceding 2022, this team has established itself as the epitome of excellence in collective performance. For the last seven years, Mumbai Indians have won the championship title three times. So, the analysis of 2017 will be comprehensive to understand each KPI of Mumbai Indians and other disqualified teams. Concerning 2019 and 2020, a holistic analysis of all the KPI's has been performed. Examining six commonly used KPI's that provide comprehensive insights into their batting and bowling abilities to comprehend the reasons for their superior performance compared to their competitors over these years.

When considering batting performance, evaluate the Batting Average, Strike Rate, and Boundary Percentage. The Batting Average metric quantifies the average number of runs a team's batters scored. The Strike Rate statistic evaluates the batters' proficiency in scoring runs rapidly. On the other hand, the Boundary Percentage metric indicates the frequency of boundary hits, which suggests the team's aggressive purpose.

When considering bowling, assess their Bowling Average, Economy Rate, and Dot Ball Percentage. The bowling average is a statistical measure representing the average number of runs conceded by the team's bowlers per wicket. The bowling economy rate is a metric used to assess the team's bowlers' effectiveness in restricting the number of runs scored by the opposition. Additionally, the dot ball percentage indicates the team's bowlers' ability to create pressure by delivering dot balls.

The KPI's provide insights into the multifaceted elements of Mumbai Indians' strategic approach, implementation, and overall superiority. Through a comparative analysis of Mumbai Indians' performance concerning the disqualification of other clubs during the specified seasons, Uncover the strategic choices that contributed to their success in the Indian Premier League. The subsequent analysis will provide a comprehensive perspective on the continual elevation of standards and establishing new milestones in T20 cricket by this formidable cricketing entity. The overall architecture of data for team performance is explained in 3.7

4.3.1 Comprehensive Team Performance Analysis for 2017

Batting Average Performance for 2017:

The Mumbai Indians demonstrated their dominance in the IPL by examining their batting performance in 2017. The data will be analyzed to gain insight into the significant impact of their batting skills on their success throughout the season.

1. **Ribbon Chart Analysis:**

   The ribbon chart can visually compare batting performances in both victorious and defeated matches throughout multiple seasons. The data presented provides a vivid depiction of the Mumbai Indians' batting average fluctuations compared to other teams over the 2017 season. The Mumbai Indians' supremacy is further substantiated by the significant discrepancy observed in batting averages between their team and the other teams in victorious matches.

2. **Batting Performance in Matches Won:**

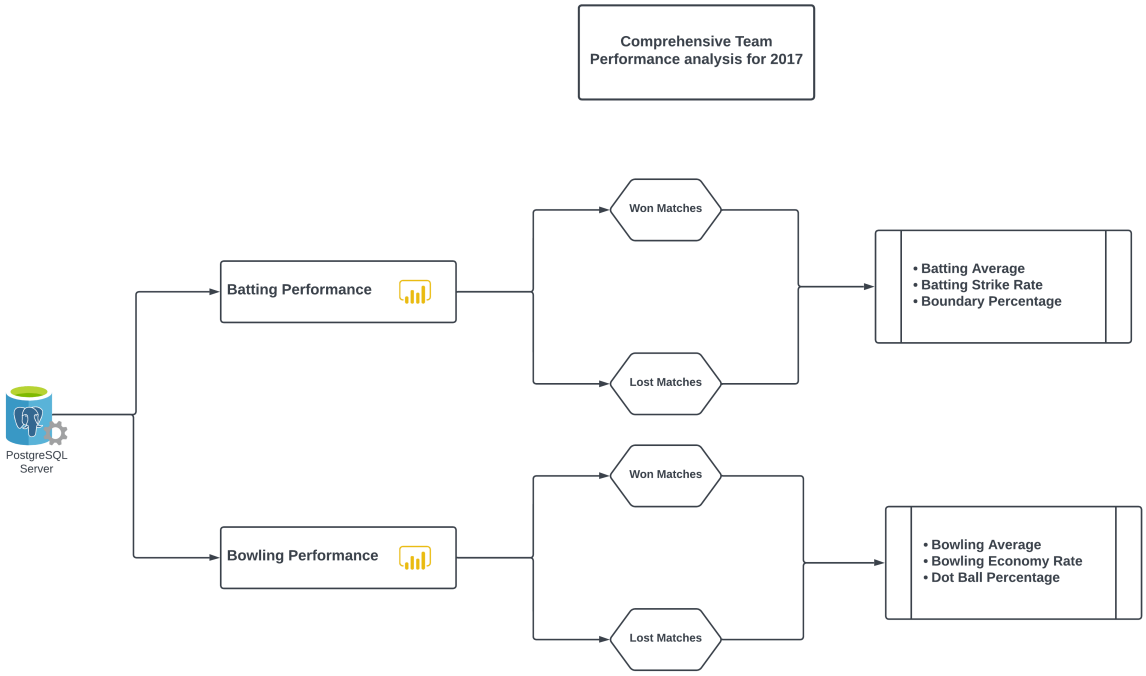

Figure 4.13: Data Overview of 2017

With a batting average of an impressive 212.8, the Mumbai Indians amassed 1,748 runs in victories. This fact highlights their capacity to score runs and consistently build solid bases for success. Their batting lineup performed at its peak during these games.

3. **Strategic Insights for won matches:**

   - **Consistency in Run Scoring:** The Mumbai Indians have shown notable proficiency in frequently achieving high run scores in their victorious matches. The observed stability in the batting lineup can be attributed to a carefully constructed team composition, deliberate lineup arrangement, and successful collaborations among players. Preserving this equilibrium in prosperous matches is vital to guarantee a consistent scoring rate and establish a solid basis for achieving triumph.
   - **Reliable Batsmen:** A critical strategic advantage is the presence of reliable batters who can regularly execute in diverse situations. During the 2017 season, the Mumbai Indians appeared to possess players of notable caliber in their lineup. The consistent performance of these dependable batters played a crucial role in enabling the team to achieve competitive scores, substantially contributing to their overall success.

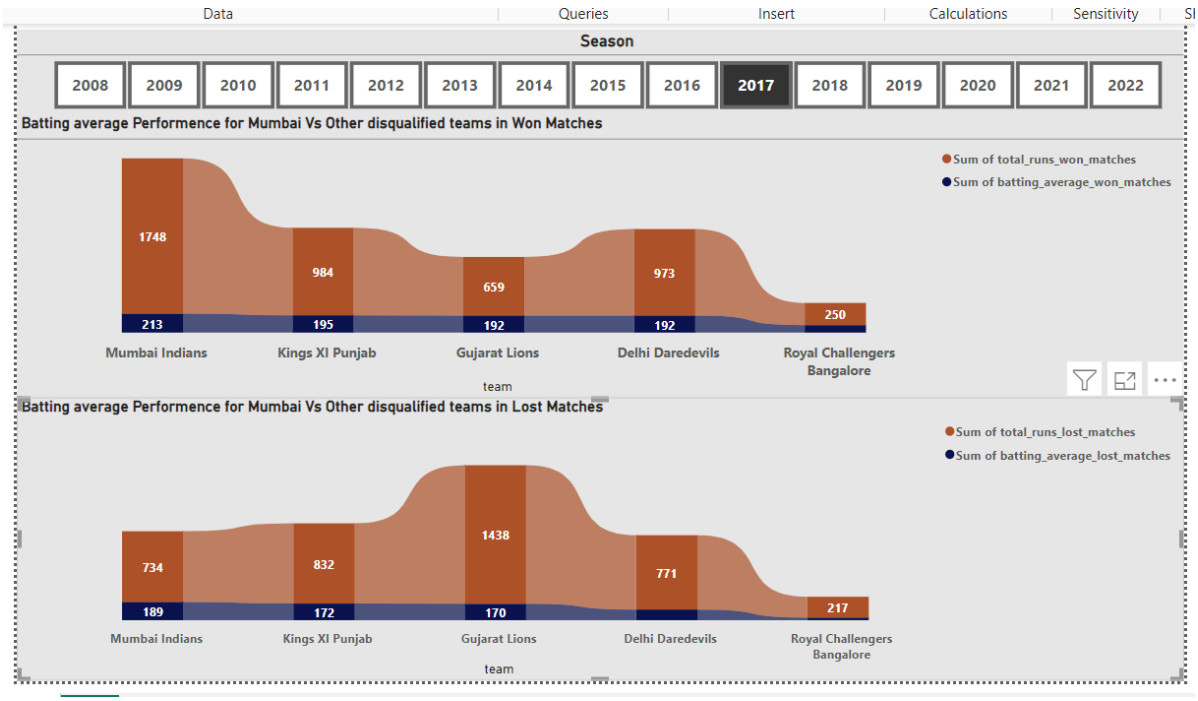

Figure 4.14: Batting Average Analysis in 2017

- **Match Awareness:** The Mumbai Indians demonstrated high match awareness by effectively comprehending match scenarios and adapting their strategy accordingly. The ability to adjust played a crucial role in their strategic success. The ability to correctly interpret the game and make appropriate judgments, such as strategically increasing the run rate or consolidating their position when necessary, was a distinguishing characteristic of their batting performance throughout the matches in which they emerged victorious.

4. **Batting Efficiency in Lost Games:**

   Mumbai Indians scored 734 runs at a batting average of 188.8, even in games they lost. Batting Efficiency implies that regardless of the match's outcome, their batting lineup was competitive and capable of posting high totals.

5. **Strategic Insights for Matches Lost:**

   - **Competitive Batting Lineup:** The Mumbai Indians exhibited a competitive batting lineup, consistently achieving a high average run rate even in matches where they faced defeat. This observation exemplifies the very competitive composition of their hitting lineup. The capacity to attain formidable scores in losses confers

a strategic edge. This suggests that, under unfavorable circumstances, the team's batters consistently showed strong performance, potentially attributable to astute player selection and strategic approaches.

- **Optimizing Player Roles:**The optimization of player roles is a strategic consideration that can have a significant impact on team performance. The Mumbai Indians exhibited a strategic deployment of their players, resulting in the noteworthy contribution of their batters, even in instances of defeat. This underscores the need for players who can adjust to various circumstances within the team's strategy framework.
- **Opportunities for Improvement:** Analyzing the batting performance in lost matches presents significant growth opportunities. Comprehending the reasons behind the inability to translate these encounters into triumphs enables the implementation of strategic modifications in domains such as player fitness, team spirit, and match tactics.

6. **Comparison with Other clubs:**

   The Mumbai Indians' supremacy is evident by examining other clubs' hitting statistics from the 2017 season. Kings XI Punjab had a far lower total of 984 runs than their nearest rival in batting average among matches won.

   In conclusion, Mumbai Indians' outstanding batting performance during the 2017 IPL season was essential to their success. Their consistency in run scoring and flexibility in many match scenarios testify to the tactical choices and team dynamics that went into their championship success that year.

Batting Strike Rate Analysis for 2017:

1. **Line Chart Analysis - Matches Won:**

   The line chart illustrates the link between strike rate and total runs in matches won by highlighting the data points with markers. It graphically shows how Mumbai Indians, who had the best strike rate, amassed 1,444 runs. They excelled because they struck a balance between racking up runs and doing so quickly.

2. **Batting Strike Rate Analysis in 2017 - Matches Won:**

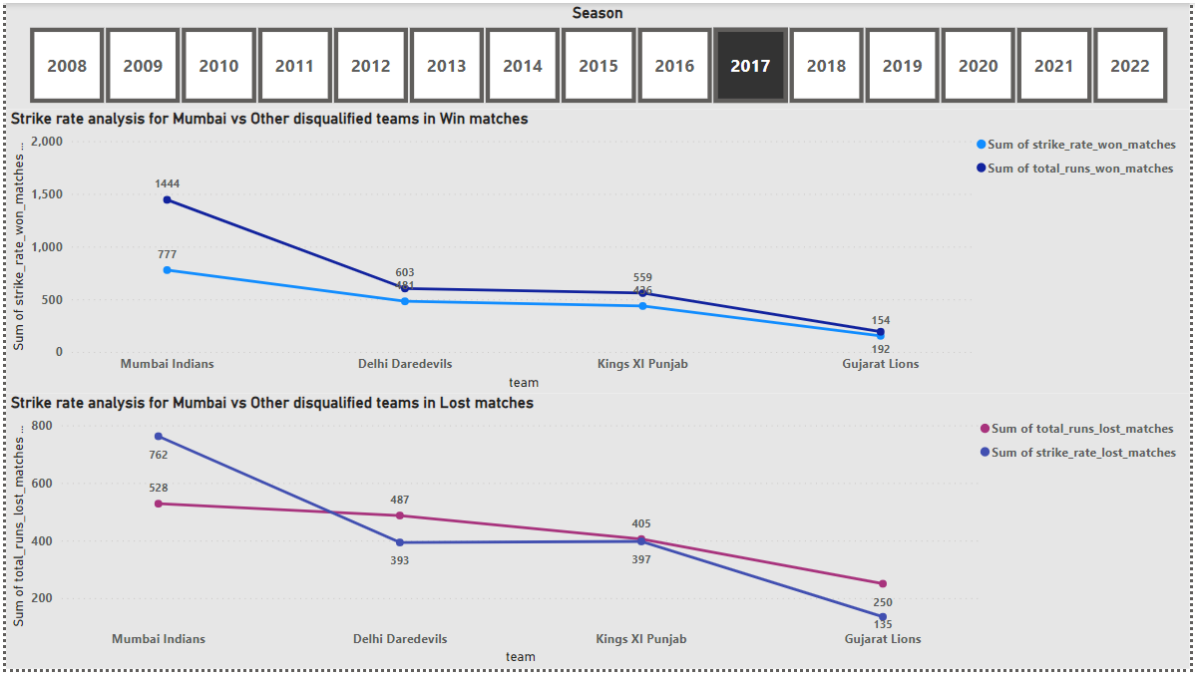

Figure 4.15: Batting Strike Rate Analysis in 2017

The Mumbai Indians displayed a great batting strike rate, totaling 777.29, in victories throughout the 2017 IPL season. This figure demonstrates their capacity to score runs quickly while also accumulating runs.

3. **Strategic Analysis - Won Games:**

   - **Aggressive Start:** Mumbai Indians' propensity for aggressive, high-strike-rate innings openings reveals the efficiency of their opening batters in laying a solid foundation for the squad.
   - **Effective Powerplay:** The tone of the inning can be set by scoring frequently during the Powerplay overs (the first six). The Mumbai Indians' approach in this phase probably included using an aggressive style of stroke play and making the most of fielding limitations.

4. **Analysis of the Line Chart for Lost Matches**

   The line graph for lost matches highlights the association between total runs and strike rate. Mumbai Indians scored 528 runs in these games while keeping a respectable strike rate. The graph shows that while their batting strike rate remained competitive, more was needed to win those games.

5. **2017 Batting Strike Rate Analysis - Lost Games**

   The Mumbai Indians struggled to keep their high-hitting strike rate of 762.27, which they lost in the 2017 season. Despite being respectable, this rate is noticeably lower than their strike rate in victories.

6. **Strategic Insights - Matches Lost:**

   The Mumbai Indians' batting strike rate analysis from the 2017 IPL season gives a comprehensive picture of their performance. Their success largely depended on their capacity to balance aggression and adjust to various match scenarios.

   - **Strategic Adaptability:** Mumbai Indians demonstrated their strategic versatility by consistently having a high batting strike rate in victories. Thanks to their versatility, they flourished in various game phases, from aggressive Powerplay over beginnings to team-building in middle overs.
   - **Setting the Tone:** The team's strategy for the Powerplay played a crucial role in establishing the tenor of their innings. Utilizing the fielding limits, they tried to get as many runs as possible during this phase. Early in the game, this tactic put the opposition under strain.
   - **Strategies of the other Team:** To deal with the Mumbai Indians, the other teams may have modified their approaches. They probably concentrated on getting early wickets and messing with Mumbai's batting rhythm.

7. **Overall Insights:**

   According to the Analysis, Mumbai Indians maintained a high batting strike rate throughout the 2017 IPL season, especially in victories. However, it might have been difficult for them to repeat this success in lost matches. Their ability to adapt to match scenarios and strategically chase targets significantly influenced their performance.

Boundary Percentage Analysis for Batting

1. **Piechart analysis:**

   The Mumbai Indians' aggressive intent is highlighted by their high boundary percentage in victories and losses. They sought to put pressure on their opponents by maintaining a scoreboard with regular limits. Mumbai's plan may depend on significant impact players

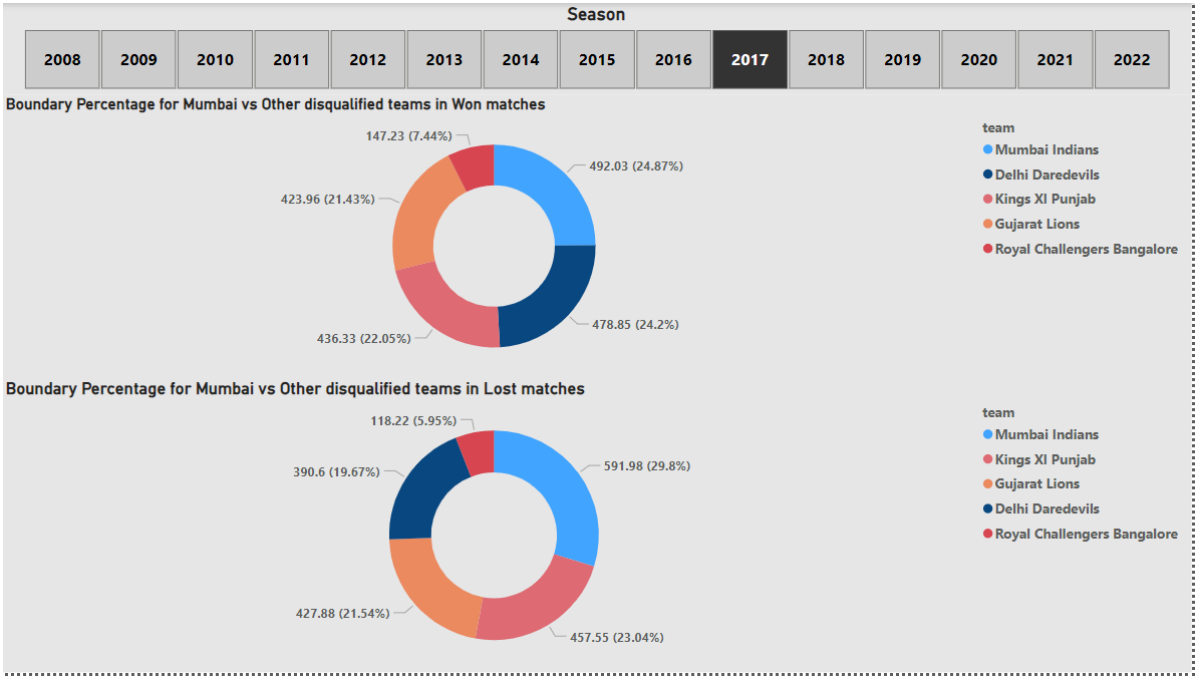

Figure 4.16: Boundary Percentage Analysis for Batting in 2017

who were excellent at hitting boundaries. These players' aggressive stroke play had the power to change the course of a game.

2. **Boundary Percentage Analysis in 2017 - Matches Won:**

   Mumbai Indians have shown their skill in hitting boundaries in victories throughout the 2017 IPL season. In these matches, their boundary percentage was 492.03, demonstrating their ability to constantly find the ropes and maintain a commanding presence.

3. **Strategic Insights for won matches:**

   - **Consistency in Boundary Hitting:** The Mumbai Indians demonstrated a noteworthy level of consistency in their capacity to achieve boundary hits throughout the matches they were victorious in during the 2017 Indian Premier League season. The observed consistency in performance can be seen as a valuable strategic advantage, indicating the presence of a formidable batting order of reliable power hitters who routinely manage to hit the ball towards the boundary ropes. Maintaining such consistency imposes significant pressure on the other team and guarantees consistent scoring in victorious matches.
   - **Aggressive Intent:** Their significant boundary percentage in victories highlights

the Mumbai Indians' aggressive scoring mentality. As mentioned above, consistently amassing boundaries is intended to maintain a steady increase in points and construct substantial innings. The manifestation of an aggressive intent can be attributed to the deliberate formulation of strategies and the allocation of player roles that promote the adoption of aggressive shot-making techniques.

- **Impact Players:** The Mumbai Indians team is expected to have heavily depended on prominent impact players renowned for their exceptional ability to strike boundaries. These players can significantly impact a match's outcome through their assertive and decisive batting style. Identifying and cultivating such players is a vital component of their strategic methodology for victory in games.

4. **Boundary Percentage Analysis in 2017 - Matches Lost:**

   Mumbai Indians maintained a respectable boundary percentage of 591.98, even in the lost games. Boundary percentage demonstrates their determined pursuit of goals, regardless of the game's outcome.

5. **Strategic Insights for lost matches:**

   - **Maintaining Aggression:** Despite experiencing defeat in specific matches, the Mumbai Indians consistently exhibited a noteworthy boundary percentage, demonstrating their unwavering commitment to maintaining an aggressive approach. The observed strategic consistency implies that the individuals in question persist in employing an offensive strategy, irrespective of the result of the match. The individuals consistently tend to apply force upon the bowlers from the other team.
   - **Pressure on Opponents:** The aggressive tactics employed by Mumbai in their unsuccessful encounters might present formidable challenges for their adversaries. Adversaries may have encountered difficulties restraining the bold batting display and facing challenges in disrupting collaborative efforts between the batters. The strategic factor employed by Mumbai demonstrates their ability to keep their rivals in a state of alertness consistently.
   - **Impactful Hitters:** The utilization of players with exceptional boundary-hitting skills is a critical strategic element, regardless of the outcome of matches. The team's strategic game strategy involves relying on these individuals, even in instances of defeat, to optimize scoring possibilities and sustain an advantage in boundary-hitting.

Bowling Performance for 2017:

Figure 4.17: Bowling Average Performance in 2017

1. **Bowling Average - Matches Won:**

   The Mumbai Indians displayed impressive bowling in games they won during the 2017 IPL season, accumulating 65.06 in bowling average while taking 86 wickets. The Bowling Average shows how good they are at lowering the opposition's run total and getting big wickets.

2. **Strategic Insights - Matches Won:**

   - **Wicket-Taking Ability:** Mumbai Indians' bowlers' ability to regularly take wickets is a significant factor in their success in winning matches. Wicket-taking ability pressures the opposition while limiting their opportunities to score runs. Strategies to target key opposition batters might have played a role.
   - **Balanced Attack:** The total number of wickets indicates a balanced attack with contributions from various bowlers. It implies that Mumbai Indians made a team effort to secure wickets rather than relying mainly on one or two bowlers.

3. **Bowling Average - Matches lost:**

   In matches lost during the 2017 season, the Mumbai Indians' bowling performance remained competitive, with a bowling average of 63 and 18 wickets taken. Although they did not secure as many victories, their bowling unit maintained its effectiveness.

4. **Strategic Insights - Matches Lost:**

   - **Competitive Bowling:** Mumbai Indians' bowling unit maintained a respectable average even in games where they lost. This implies that the bowlers kept up their fight, and the defeats might have been due to other elements like batting performance or team tactics.
   - **Addressing shortcomings:** The team may address any bowling shortcomings they may have discovered or concentrate on areas where they might improve after losing a game. Continual assessment and adaptation likely played a part in maintaining a competitive edge.

5. **Overall Findings:**

   According to the data, Mumbai Indians had a strong bowling attack in the 2017 IPL season that could secure victory and stay competitive even in defeat. The team's tactics focused on wicket-taking prowess, balanced bowler contributions, and match-situation adaptation.

   In conclusion, Mumbai Indians' bowling performance data shows their effectiveness at gaining wickets to secure victory and remaining competitive in games they lose. Their bowling unit's flexibility and balanced approach helped them play well overall during the 2017 campaign.

Bowling Economy Rate:

1. **Insights from Bar Chart Analysis:**

   The bar chart analysis provides an excellent visual representation of the team's economy rate in both won and lost matches throughout the 2017 IPL season. It demonstrates how Mumbai Indians' bowling unit remained reliable in all scenarios while maintaining a competitive economy rate.

Figure 4.18: Bowling Economy Rate Performance in 2017

2. **Bowling Economy Rate Analysis in 2017 - Matches Won:**

   The Mumbai Indians had a fantastic economy rate in games they won during the 2017 IPL season, totaling 70.09. This statistic demonstrates their bowlers' ability to limit runs while efficiently claiming wickets.

3. **Strategic Insights - Matches Won:**

   - **Economical Bowling:** The Mumbai Indians' bowlers routinely produced economical spells by maintaining a strict line and length. By increasing pressure on the opponent, this tactic enabled breakthroughs and stopped the flow of runs.
   - **Wickets in Hand:** The Mumbai Indians frequently succeeded in taking early wickets in games they won. Pressure being applied to the opponent's middle and lower order stopped the flow of runs and made it difficult for the opposition to pick up speed.

4. **Bowling Economy Rate Analysis in 2017 - Matches Lost:**

   The Mumbai Indians maintained a competitive economy rate in the games they lost in 2017, totaling 68.51. However, this rate is slightly lower than in victories, indicating that their bowlers were still productive even in defeat.

5. **Strategic Insights - Matches Lost:**

   - **Pressure in the Death Overs:** In lost games, Mumbai Indians bowlers may have struggled to stop the opposition from scoring in the final overs. Tactical advancements in their death bowling might have decided these matches.
   - **Chasing Defendable Totals:** The team occasionally struggled to defend lesser totals despite having a solid economy rate. Strategic changes like adopting alternative bowling techniques or establishing more significant goals might have produced better outcomes.

6. **Overall Findings:**

   According to the analysis, the Mumbai Indians' bowlers performed well in both victories and defeats throughout the 2017 IPL season. Their ability to keep their economy rate competitive in both wins and losses emphasizes the reliable performance of their bowling team.

Dot Ball Percentage:

The Mumbai Indians bowled a significant portion of dot balls, totaling 385.59, in victories during the 2017 IPL season. This statistic shows how effectively their bowlers can increase pressure by persistently denying the opposition scoring chances.

1. **Insights from the Pie Chart Analysis:**

   The pie chart analysis shows the percentages of dot balls in both won and lost matches throughout the 2017 IPL season. It demonstrates the influence of this KPI on match outcomes by demonstrating that Mumbai Indians had a much more significant dot ball percentage in their victories compared to their defeats.

2. **Dot Ball Percentage Analysis in 2017 - Matches Won:**

   The Mumbai Indians bowled a significant portion of dot balls, totaling 385.59, in victories during the 2017 IPL season. This statistic shows how effectively their bowlers can increase pressure by persistently denying the opposition scoring chances.

3. **Strategic Insights - Matches Won:**

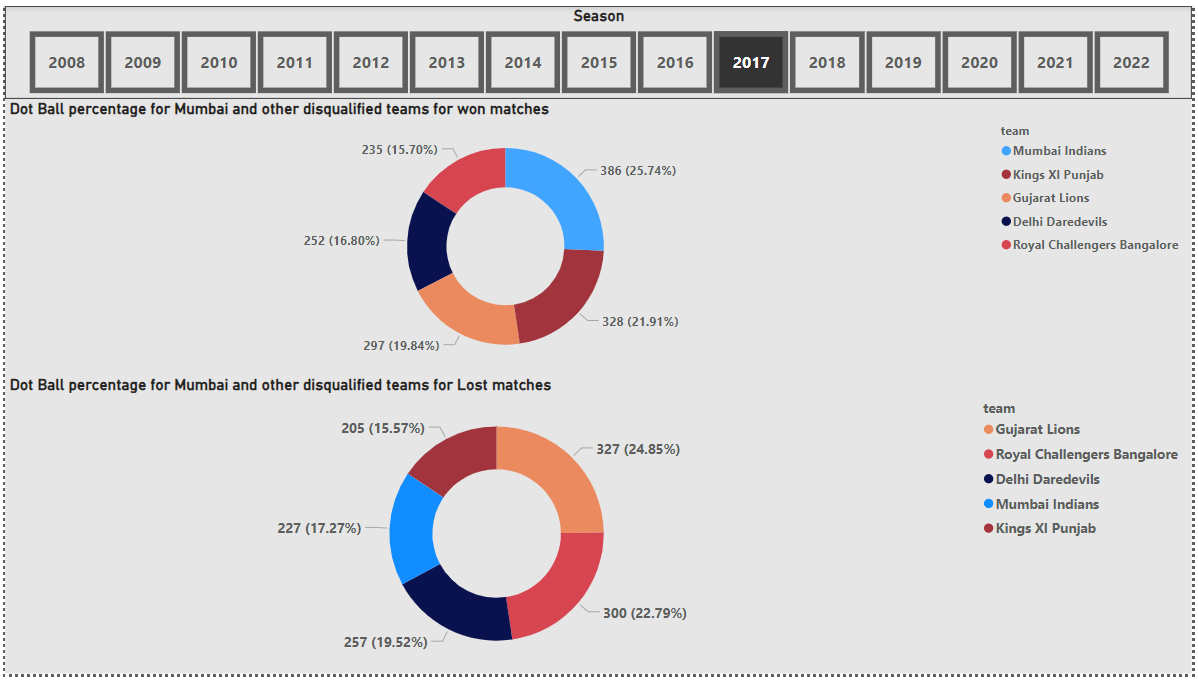

Figure 4.19: Bowling Economy Rate Performance in 2017

- **Building Pressure:** The bowlers for Mumbai Indians concentrated on keeping a precise line and length, which produced many dot balls. By using this tactical strategy, pressure was applied to the batters, causing them to make errors.
- **Wickets and Dot Balls:** Dot balls and regular wickets frequently threw off the flow of the opposition, making it difficult for them to form partnerships. The team's tactical goal was to get wickets while applying pressure.

4. **Dot Ball Percentage Analysis in 2017 - Matches Lost:**

   The Mumbai Indians kept a respectable dot ball percentage in losses during the 2017 season, at 227.48. While slightly lower than in their victories, this indicates that their bowlers still managed to create opportunities by restricting scoring.

5. **Strategic Analysis - Lost Games:**

   Consistency in Dot Balls: Mumbai Indians' bowlers consistently applied dot ball pressure, even when they lost games. These games might have been decided by tactical changes in other areas of the game, such as hitting and fielding.

6. **Overall Takeaways:**

The data shows that Mumbai Indians' success in 2017 IPL was greatly influenced by their dot ball percentage. Their bowlers consistently applied pressure by preventing scoring opportunities for the opponent, which resulted in both successful outcomes and valiant efforts in defeats.

4.3.2 Holistic Team Performance Analysis for 2019

Overall Performance 2019:

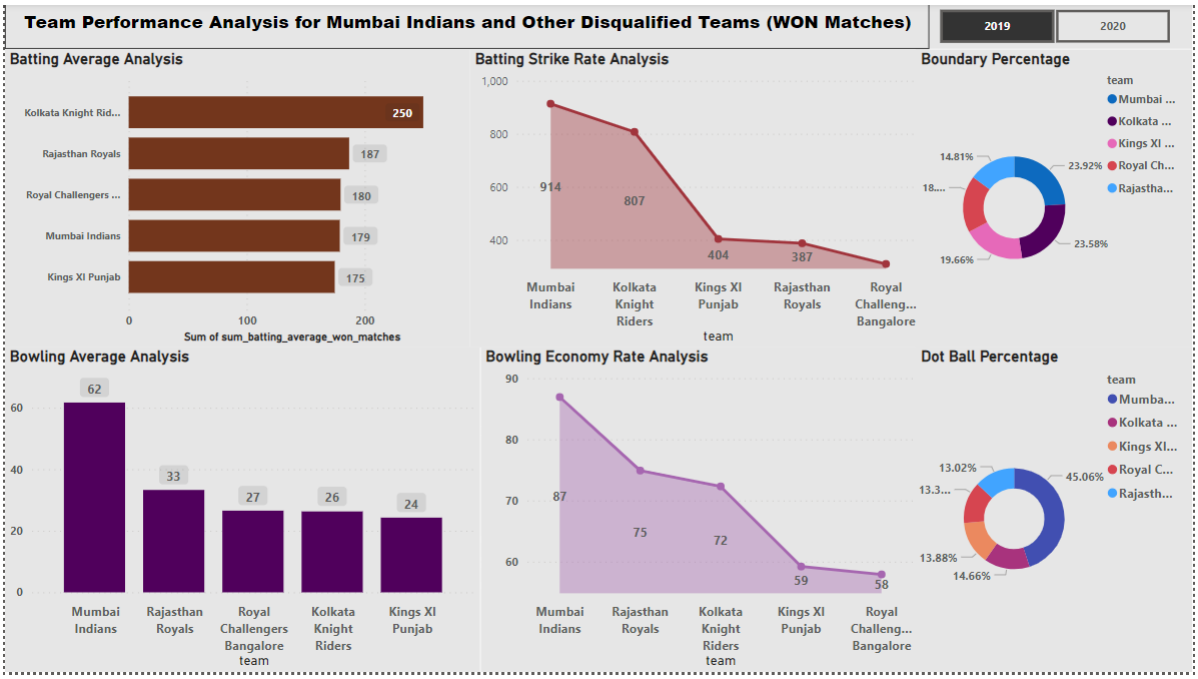

Figure 4.20: Holistic Dashboard for 2019

Mumbai Indians vs. Other Disqualified Teams - 2019 IPL Season

1. **Batting Average:**

   Mumbai Indians' batting average in victories was 178.88, demonstrating their ability to score runs consistently to form partnerships and anchor innings efficiently.

2. **Bowling Average:**

   The Mumbai Indians demonstrated their skill at capturing wickets, averaging 61.69 in games they won. Their bowlers routinely broke opposition partnerships, putting pressure on the batting side.

3. **Batting Strike Rate:**

   Mumbai Indians had a fantastic batting strike rate of 913.55 in games they won, demonstrating their capacity to score runs quickly and emphasizing their aggressive attitude when batting.

4. **Bowling Economy Rate:**

   - The team's bowling economy rate in victories was 86.93, indicating that they efficiently managed the flow of runs.
   - They pressured the opposition batters by bowling with precise lines and lengths.

5. **Boundary Percentage:**

   Mumbai Indians won matches with a boundary percentage of 504.91, demonstrating their ability to find the gaps and strike boundaries consistently. Aggressive yet thoughtful shot selection was a vital element.

6. **Dot Ball Percentage:**

   The squad achieved a dot ball percentage of 385.59 in games they won, demonstrating their capacity to increase pressure by pressuring batters into playing dot balls.

7. **Strategic Insights - Matches Won:**

   - **Balanced Batting:** The Mumbai Indians' batting order was balanced, with numerous players making contributions and less reliance on individual performances.
   - **Wicket-Taking Bowlers:**The side had bowlers who could end partnerships, take big wickets, and apply pressure.
   - **Explosive Start:** By taking advantage of fielding limitations, aggressive openers establish the tone for the Powerplay.
   - **Calculated Aggression:** Batsmen demonstrated calculated aggression by emphasizing boundary hitting while reducing risks.
   - **Data-Driven Decisions:** Identifying the adversary's weaknesses and developing plans required data analytics.

8. **Strategic Insights for Other Teams:**

- **Balanced Lineup:** To avoid relying too heavily on a few players, put together a balanced batting lineup.
- **Wicket-Taking Bowlers:** Create or acquire bowlers who can capture significant wickets.
- **Explosive Start:** By choosing aggressive openers, emphasize the value of a dynamic start during the Powerplay.
- **Calculated Aggression:** Encourage batters to play aggressively but strategically, concentrating on hitting boundaries.
- **Dot Ball Pressure:** Apply pressure to the opposition by accurately bowling dot balls.

9. **Overall Insights:**

   The Mumbai Indians' success in the 2019 IPL season can be due to their strong team effort. Their success was mainly due to their capacity to keep a balanced team, take wickets when they mattered, and adjust to various match circumstances. To better their performance, other IPL clubs can take inspiration from Mumbai Indians' tactics.

   By implementing these tactical suggestions, clubs can improve their chances of winning the IPL by batting and bowling with a balanced and dynamic approach, aiming for reliable and competitive performances.

Holistic Dashboard 2019 - Lost Matches

The Mumbai Indians encountered difficult circumstances in the matches they lost during the 2019 IPL season. In terms of KPI, let us compare their performance to that of other teams:

1. **Batting Average:**

   Mumbai Indians retained a good average while losing games, as evidenced by their batting average 156.8 in those games. It was also clear that they could form partnerships and anchor innings.

2. **Bowling Average:**

   The Mumbai Indians have shown their ability to take wickets even under challenging situations, with a bowling average of 37.21 in losing games. Their bowlers were still capable of dismantling rival partnerships.

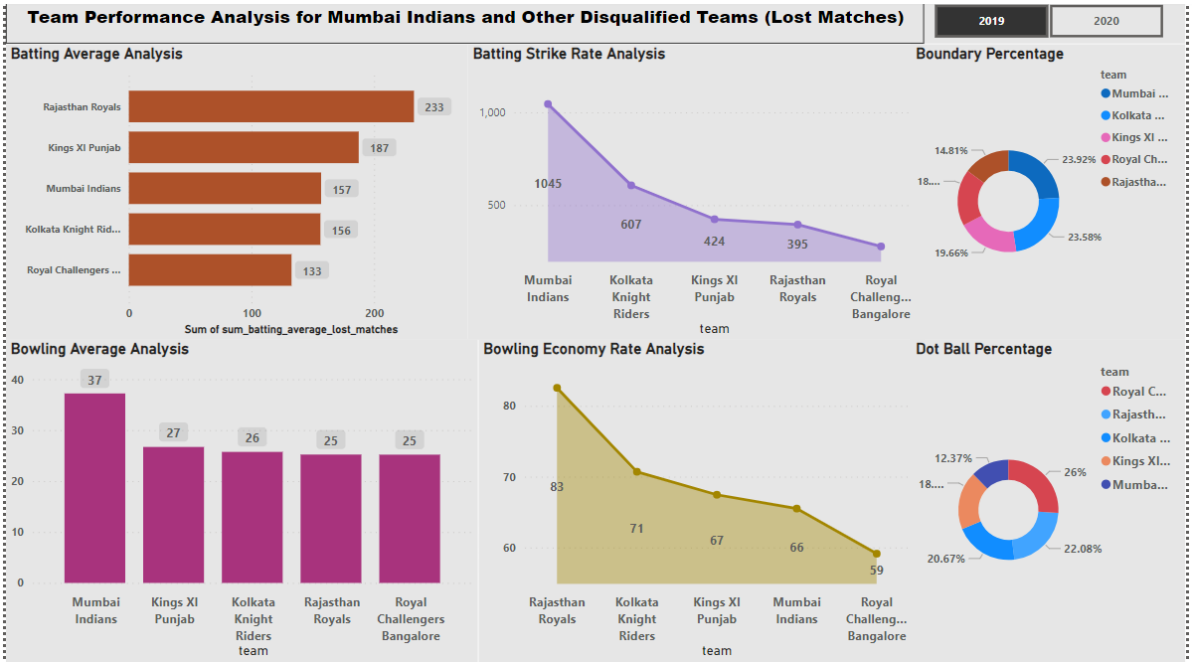

Figure 4.21: Holistic Dashboard for Lost Matches in 2019

3. **Batting Strike Rate:**

   Mumbai Indians displayed calculated aggression even under poor conditions, maintaining a high batting strike rate of 1044.97 in losing matches, demonstrating their aggressive attitude to scoring runs.

4. **Bowling Economy Rate:**

   Despite the loss, the team's bowlers effectively contained the opposition, recording a bowling economy rate of 65.5 in losing games, which is competitive and demonstrates their ability to limit the flow of runs.

5. **Boundary Percentage:**

   Mumbai Indians' ability to find openings and hit boundaries even under challenging circumstances was shown by their boundary percentage 156.8 in lost games. Aggressive yet thoughtful shot selection continued to be a crucial component.

6. **Dot ball percentage;**

   Even in defeat, the club secured a dot ball percentage of 445.35, demonstrating their capacity to apply pressure by making batters play dot balls. Using dot balls to provide pressure was a recurring element of their plan.

7. **Strategic Analysis - Lost Games:**
    - **Resilient Batting:** Mumbai Indians batting showed perseverance, maintaining a respectable average even in defeat.
    - **Bowlers who could take key wickets:** The squad had bowlers who could end partnerships even when they were in danger of losing.
    - **Aggressive Mentality:** The team's high strike rate and boundary percentage in losing games indicate they were not afraid to be aggressive.
    - **Economy Control:** Despite the defeats, the bowlers of the squad were nevertheless successful in restricting the flow of runs, a crucial component of their plan of attack.
8. **Strategic insights for Other teams**
    - **Maintain Resilience:** Motivate players to maintain a respectable batting average despite dire circumstances.
    - **Wicket-Taking Bowlers:** Create or obtain bowlers who can disrupt critical alliances even in defeat.
    - **Consistent Aggression:** No matter the game outcome, instill an aggressive mindset in the team by emphasizing maintaining a high strike rate and boundary-hitting.
    - **Economy Control**: Make sure people understand how crucial it is to keep their spending under control, even in trying circumstances.
9. **Overall Insights:**

    Based on the Mumbai Indians' play in games they lost throughout the 2019 IPL season, they kept their aggressive approach to batting and bowling. Even in challenging circumstances, they can maintain essential KPIs like batting strike rate and dot ball percentage, demonstrating their adaptation and tenacity. This information emphasizes consistency and unwavering tactics while dealing with setbacks. It underlines that maintaining important KPIs can make a difference in games that do not go as planned, an essential lesson for other IPL teams.

4.3.3 Holistic Team Performance Analysis for 2020

Holistic Dashboard 2020 - Won Matches

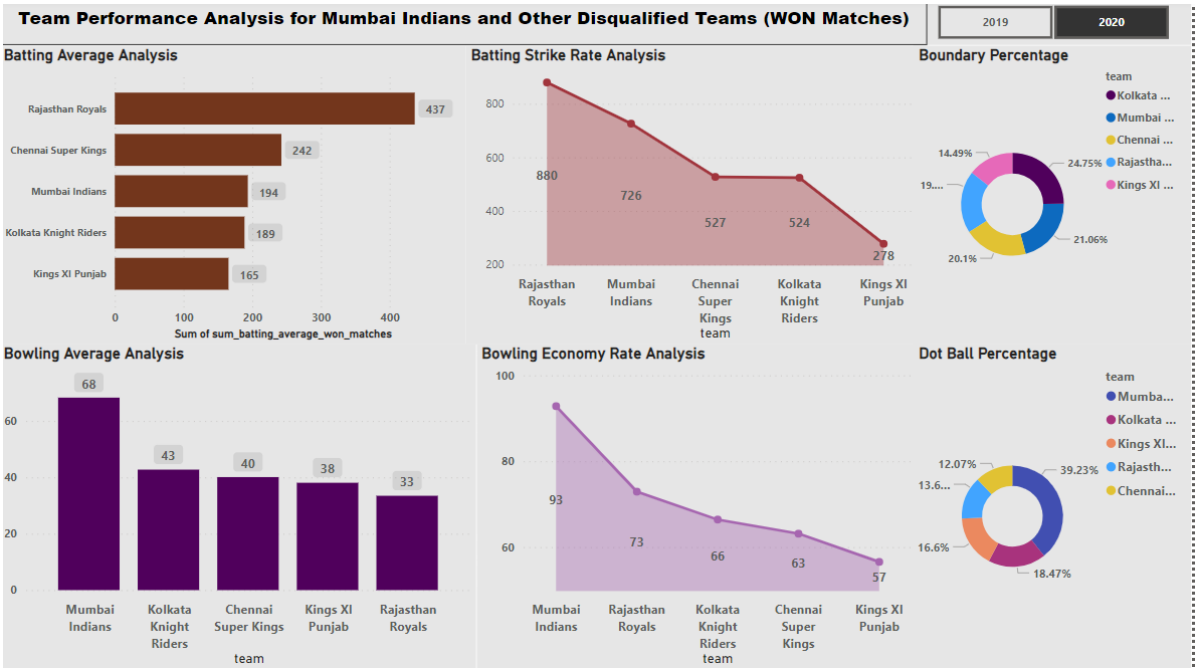

Figure 4.22: Holistic Dashboard for won Matches in 2020

The Mumbai Indians won multiple games during the 2020 IPL season, demonstrating their skill in both bowling and batting. Using KPI's, let us examine their performance and assess how it stacks up to that of other teams:

1. **Batting Average:**

   Mumbai Indians' strong 193.55 batting average in victories highlights their capacity to construct innings and make a sustained impact with the bat.

2. **Bowling Average:**

   The side consistently bowled at a competitive average of 68.3 in victories, demonstrating their potency in getting vital wickets during the opposition's innings.

3. **Batting Strike Rate:**

   In games they won, Mumbai Indians displayed an exceptional batting strike rate of 726.13, demonstrating their aggressive approach and capacity for scoring rapidly.

4. **Bowling Economy Rate:**

   With a 92.83 bowling economy rate in victories, the squad successfully managed the flow of runs and put pressure on the opposition.

5. **Boundary percentage:**

   The Mumbai Indians' successful matches had a boundary percentage of 465.62, demonstrating their ability to find gaps and strike boundaries consistently.

6. **Dot Ball Percentage:**

   Their 193.55 dot ball percentage in wins shows how effectively they can apply pressure and get batters to play dot balls.

7. **Strategic Insights - Matches Won:**

   - **Consistency in Batting:** Mumbai Indians kept their batting average high, showing they were reliable in starting and concluding innings.
   - **Wicket-Taking Bowlers:** The team had bowlers who could end critical partnerships, take significant wickets, and keep a respectable bowling average.
   - **Aggressive Batting:** They could score rapidly and set lofty goals because of their strong batting strike rate and aggressive attitude.
   - **Economy Control:** Despite the aggression, their bowlers managed the economy rate to stop the opposition from scoring quickly.
   - **Boundary Hitting:** Their high boundary percentage shows that they can regularly find the boundary, which increases their total number of rapid runs.
   - **Pressure Building:** Their ability to increase pressure on the opposition by reducing scoring opportunities is indicated by their dot-ball percentage

8. **Strategic Insights for other teams**

   - **Consistent Batting:** Encourage players to maintain a high batting average to make contributions consistently.
   - **Wicket-Taking Bowlers:** Create or obtain bowlers who can destroy partnerships and take significant wickets.
   - **Balanced Aggression:** Focus on preserving a high strikeout rate while assuring restrained aggression.
   - **Economy Control:** Teach bowlers to successfully manage the flow of runs without sacrificing their ability to take wickets.

- **Boundary Hitting:** As part of the team's batting plan, encourage boundary-hitting.
- **Pressure Building:** Concentrate on getting the opposition to play more dot balls to increase pressure.

9. **Overall Analysis:**

   Mumbai Indians' success during the 2020 IPL season shows their well-rounded and practical strategy. They kept their batting and bowling averages high, assuring consistency throughout their victories. They were strong opponents because they could exert control over the game through aggressive batting, wicket-taking bowlers, and economy control.

   This information emphasizes how crucial it is to strike the ideal balance between aggression and restraint in batting and bowling to win the IPL.

Holistic Dashboard 2020 - Lost Matches

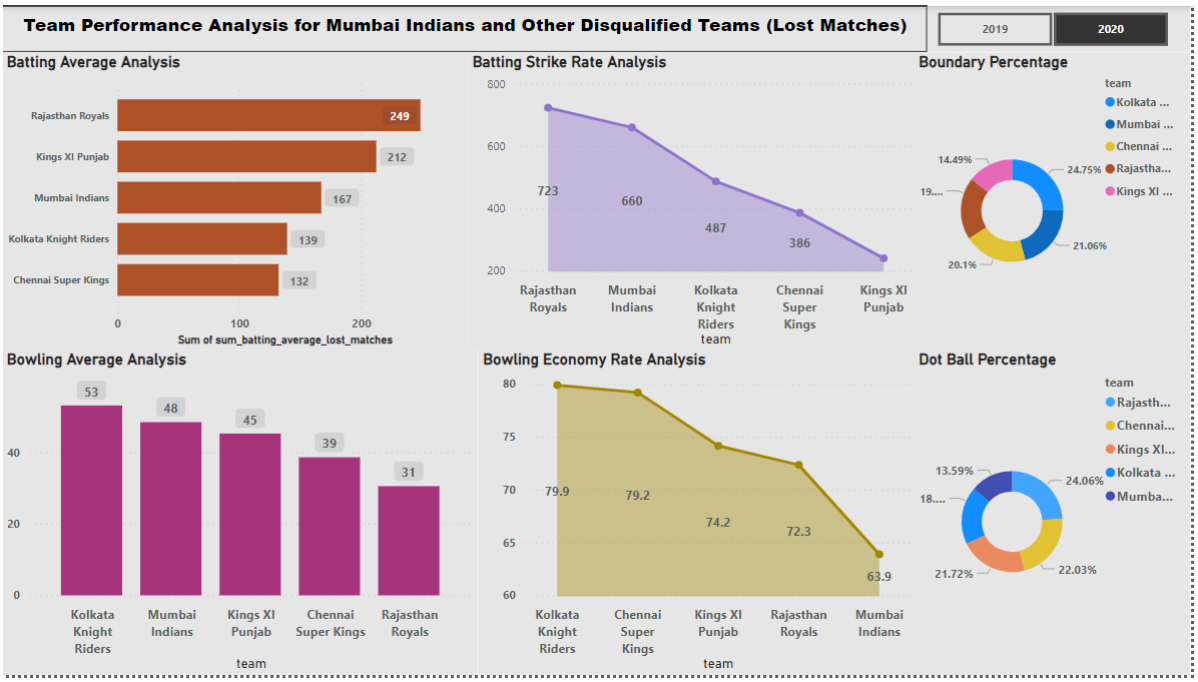

Figure 4.23: Holistic Dashboard for Lost Matches in 2020

The Mumbai Indians experienced difficulties in a few games during the 2020 IPL season, resulting in defeats. The following KPI's will be used to examine their performance in lost

games and compare it to that of other teams:

1. **Batting Average:**

   The Mumbai Indians had a batting average of 167.3 in the games they lost, which shows that they could score runs but had trouble turning their early starts into significant totals.

2. **Bowling Average:**

   Their bowlers appeared to find it difficult to regularly take wickets, as evidenced by their 48.46 bowling average in games they lost.

3. **Batting Strike Rate:**

   The losing matches had a batting strike rate of 660.13, indicating that they continued to score at a commendable pace but could not earn victories.

4. **Bowling Economy Rate:**

   The Mumbai Indians could efficiently stem the flow of runs in lost games, with a bowling economy rate of 63.88, but it did not always result in triumphs.

5. **Boundary Percentage:**

   They had a boundary percentage 438.42 in lost games, demonstrating their capacity to identify gaps and hit boundaries.

6. **Dot Ball Percentage:**

   These games had a dot ball percentage of 167.3, demonstrating their ability to pressure opposing batters by requiring them to play dot balls.

7. **Strategic Insights - Matches Lost:**

   - **Competitive Batting:** The club maintained a decent batting average despite losing, demonstrating a desire to compete.
   - **Difficulties Taking Wickets:** Their bowling average revealed they had trouble regularly taking wickets.
   - **High Batting Strike Rate:** This team plays aggressively, looking to score as soon as possible.
   - **Effective Economy Control:** Bowlers consistently controlled their economy rate, but victories did not always accompany it.

- **Pressure Creation:** According to the dot ball percentage, they successfully created pressure but could not use it to their advantage.

8. **Strategic Insights for Other Teams:**
   - **Competitive Batting:** Focus on batting at a competitive level, even in defeat.
   - **Wicket-Taking Bowlers:** Create plans for routinely taking wickets, particularly in crucial situations.
   - **Effective Economy Control:** Teach bowlers to manage the flow of runs properly.
   - **Boundary Conversion:** Consider turning boundary opportunities into long runs.
   - **Pressure Creation:** Use the increasing pressure to the team's advantage by pressing additional dot balls.
9. **Overall Insights:**

   Mumbai Indians' performance in unsuccessful matches during the 2020 IPL season shows a competitive team despite losing. They had trouble capturing wickets consistently and turning boundaries into game-winning performances, but they upheld high standards in their batting and ran rate control. This information highlights the value of all-around performance and the necessity of capitalizing on decisive circumstances to secure victories in the IPL.

5 Conclusion and Future work

5.1 Conclusion

The IPL player and team performance analysis has shed important light on the inner workings of one of the world's most well-liked T20 cricket leagues. A deeper understanding of the tactics and performance of teams and players over several seasons by carefully analyzing KPI's like batting average, strike rate, boundary percentage, bowling average, economy rate, and dot ball Percentage.

5.1.1 Player Performance Conclusion:

Several important insights and conclusions have been drawn from examining player performance in the IPL. Rishabh Pant, David Warner, KL Rahul, and Andre Russell are just a few players who have excelled in all facets of batting. Andre Russell's spectacular hitting in the Death Overs, David Warner's dependable play, KL Rahul's ability to start solid innings, and Rishabh Pant's explosive Powerplay batting have all made a big difference in the success of their teams. BB Sran, Kagiso Rabada, and T Natarajan, among others, have demonstrated excellent bowling averages and economy rates. Three players in particular have stood out: BB Sran's efficiency in the Powerplay, Kagiso Rabada's skill in getting wickets, and T Natarajan's dependability in the Death Overs.

1. **Machine Learning Models: Random Forest and SVM**

   Machine learning models, notably Random Forest and Support Vector Machine SVM, were used to forecast player performance based on previous data. These models have offered insightful information regarding the forecasting of batting and bowling averages. The accuracy of these models was assessed using the RMSE statistic.

   Random Forest and SVM models fared well regarding batting average prediction, with low RMSE values. Both the Model suggests that these models successfully predict a

player's average batting performance.

In comparison to predictions of batting average, the RMSE values for both models for predicting bowling average were relatively higher. RMSE score implies that while these models can shed light on a player's anticipated batting success, precisely forecasting the bowling average may be more difficult due to the different elements affecting bowling performance.

2. **Strategic Insights:** Strategic choices greatly influenced the performance of the players. Based on each player's strengths and the circumstances of the match, teams tactically used them. For instance, the choice of openers, middle-order batters, and finishers was made with the team's success in mind. Bowlers, like Powerplay specialists or Death Overs gurus, were given specialized duties to maximize their influence on the game.

3. **Key Players:** Players like Rishabh Pant and David Warner consistently put in strong batting performances throughout several seasons. Their contributions to their respective teams' batting averages, strike rates, and boundary percentages have made them priceless assets.

   Kagiso Rabada and T Natarajan have excelled in the bowling department. Natarajan's economy rate during pivotal moments of the game and Rabada's ability to regularly take wickets have made them indispensable to their teams' success.

4. **Analysis of Batting Strike Rate:**

   The relevance of players performing well during particular parts of IPL matches was shown by examining batting strike rates during various phases of the games. Some players excel at giving the Powerplay off to explosive starts, while others play steadily over the Middle Overs. Additionally, a select few can compete well in the Death Overs. Teams must know these subtleties to optimize their batting order and exploit player strengths.

5. **Analysis of Bowling Economy Rates:**

   In the IPL, controlling the game and limiting the opposition's batting depends on bowling economy rates. Particularly during Powerplays and Death Overs, players like BB Sran and Kagiso Rabada have demonstrated their ability to manage the flow of runs. Bowlers who can get wickets while maintaining a low economy rate are vital to their teams.

5.1.2 Team Performance Conclusion:

1. **2017 Comparison:**

   The five-time acrshortipl winner Mumbai Indians outperformed other disqualified teams in both batting and bowling, according to our analysis of their 2017 performance. They consistently had a high batting average and strike rate, notably when they won games. In matches they won, they made strategic decisions based on aggressive starts and efficient Powerplay goals. Furthermore, they stand out for their fielding and capacity to make critical catches under duress.

2. **2019 Comparison:**

   The Mumbai Indians showed their supremacy in several areas, including batting, bowling, and strategic decision-making, in the team performance analysis for 2019. They consistently had strong batting averages and strike rates, notably when they won games. Assertive Powerplay starts and efficient Powerplay scoring were crucial strategic choices that aided their success. Additionally, their bowlers exhibited superb control in the final overs, limiting the opportunity for the opposition to score runs.

3. **2020 Comparison:**

   Mumbai Indians had a respectable batting strike rate even in defeat in 2020, continuing their strong play. They demonstrated outstanding bowling control and an efficient economy rate but needed help getting wickets regularly. It was impressive how they could exert pressure by pressing dot balls. Additionally, as shown by their tactical choices in different game phases, they were able to adjust to shifting match circumstances, which made them a formidable squad.

5.2 Future Work:

1. **Prediction Modeling:** The creation of prediction models that might foretell player and team performance in IPL matches is one intriguing topic for future research. To produce accurate forecasts, machine learning techniques can be used to examine historical data, player statistics, match circumstances, and other variables.

2. **Advanced Analytics:** Advanced statistical and analytical tools can be used to understand player performance better. Advanced analytics involves investigating player-versus-

player matches, assessing player behavior in various situations, and spotting performance trends across several seasons.

3. **Real-time analytics:** Real-time analytics can offer timely insights during IPL games. Live data streams can help teams make split-second choices like field placements, bowling changes, and batting order alterations.

4. **Performance Enhancement:** Through individualized coaching and training programs, teams can use data-driven insights to improve player performance. Better on-field performance can arise from pinpointing each player's improvement areas, whether in technique, physical fitness, or mental toughness.

5. **Fan Engagement:** Data analytics can improve fan engagement. The viewing experience can be enhanced by giving spectators interactive data, player comparisons, and predictive analytics.

6. **Strategies for IPL player auctions:** Teams can use prior player performance statistics to guide their tactics. Choosing players with worth based on their prior performance might help teams be more economically arranged.

7. **Injury Prevention:** Data analytics can aid in managing and preventing injuries. Teams can make data-driven judgments about player rotations and rest times by keeping track of players' workloads and injury histories.

8. **Evolution of Performance Metrics:** Performance metrics' choice and assessment could change over time. Cricket may see the emergence of new KPIs, and measures and teams may need to adjust their plans accordingly.

   In conclusion, performance metrics and data analytics have become crucial components of the IPL, giving players and spectators a deeper understanding of the game. The abundance of data presents fascinating opportunities for boosting player effectiveness, team tactics, and the spectator experience. Data-driven insights will become increasingly important in T20 cricket as the game develops.

Bibliography

Agrawal Shilpi, Suraj Pal Singh, J. K. S. (2018). Predicting results of indian premier league t-20 matches using machine learning. *8th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 67–71.

Alamar, B. and Mehrotra, V. (2011). The rapidly evolving world of sports analytics. *Part I. Analytics Magazine*, pages 33–37.

Allana, S. (2018). Ipl has no negative impact on test cricket, a study proves. *crictracker.com*. accessed on August 28, 2018, 22:14 IST, `https://www.crictracker.com/ipl-has-no-negative-impact-on-test-cricket-a-study-proves/`.

Bai, Z. and Bai, X. (2021). Sports big data: management, analysis, applications, and challenges. *Complexity*, pages 1–11.

Barman, H. (2020). A web-based support system to measure fielding performance in cricket. *Management and Labour Studies*, 45(2):159–167.

Bhattacharjee, D. and Pahinkar, D. G. (2012). Analysis of performance of bowlers using combined bowling rate. *International Journal of Sports Science and Engineering*, 6(3):184–190.

Bhoyar, P. and Agrawal, P. (2020). Exploratory data analysis of indian premier league: An empirical study. *IJFANS International Journal of Food and Nutritional Sciences*, 11(3):4125–4130.

Borooah, V. K. and Mangan, J. E. (2010). The ”bradman class”: An exploration of some issues in the evaluation of batsmen for test matches. *Journal of Quantitative Analysis in Sports*, 6(3):1877–2006.

Bousdekis, A., Mentzas, G. D., and Lepenioti, K. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 60.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–25.

Burns, N. and Grove, S. K. (2001). *Intervention Research*. WB Saunders Company, Philadelphia, 4th edition.

Castellano, J., Perea, A., Alday, L., and Mendo, A. H. (2008). The measuring and observation tool in sports. *Behavior Research Methods*, 40:898–905.

Chattopadhyay, A. (2015). Spot fixing in ipl: Exploring the archaeology of spot. *International Journal of Research (IISRR)*, 1(3):17–40.

Chellapilla, D. P., C, P., and Singh, S. (2016). A new category-based deep performance index using machine learning for ranking ipl cricketers. *International Journal of Electronics, Electrical, and Computational System (IJEECS) ISSN*, 5(2):37–47.

Christian, L. S. and Yanyan, S. (2015). Measuring expert performance for serious games analytics: From data to insights. In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 101–134. Springer International Publishing Switzerland.

Davide Chicco, M. J. W. and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse, and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623.

De Silva, V., Caine, M., Tilson, P., and Smith, B. (2018). Player tracking data analytics as a tool for physical performance management in football. *Sports*, 6(4):130.

Dinesh, A. P. (2014). Outcome of the extra delivery in cricket. *International Journal of Engineering Research & Technology (IJERT)*, pages 990–999.

Dodge, Y. (2003). The oxford dictionary of statistical terms.

Elia, M., Ofer, A. H., and Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5(4):213–222.

Enderwick, P. and Nagar, S. (2010). The indian premier league and indian cricket: Innovation in the face of tradition. *Journal of Sponsorship*, 3(2):130–143.

Ferran Vidal-Codina, Evans Nicolas, B. E. F. and Billingham, J. (2022). Automatic event detection in football using tracking data. *Sports Engineering*, 25(1):1–15.

Gaur, K. P. and Bhattacharjee, D. (2016). On finding the most compatible batting average. *Journal of Applied Quantitative Methods*, pages 50–57.

Ghai, K. and Zipp, S. (2020). Governance in indian cricket: Examining the board of control for cricket in india through the good governance framework.

Goyal, C. (2017). Ipl-2017 cross country cluster analysis. *International Journal of Computer Science Trends and Technology (IJCST)*, 5(4):117–122.

Gupta, A. (2003). The globalization of sports, the rise of non-western nations, and the impact on international sporting events. *The International Journal of the History of Sport*, 26(12):1779–1790.

Kanungo, V. and Bomatpalli, T. (2019). Data visualization and toss-related analysis of ipl teams and batsmen performances. *International Journal of Electrical and Computer Engineering*, 9(5):4423–4431.

Kapadiya, C., Shah, A., Adhvaryu, K., and Barot, P. (2020). Intelligent cricket team selection by predicting individual players' performance using efficient machine learning technique. *International Journal of Engineering and Advanced Technology*, 9(3):3406–3409.

Karetnikov, A. (2019). Application of data-driven analytics on sport data from a professional bicycle racing team.

Karthik, K., Gokul, K. S., Manjunath, V. K., and Shashank, S. (2020). Analysis and prediction of fantasy cricket contest winners using machine learning techniques. In *Evolution in Computational Intelligence, Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1*.

Kaviya, A., Mishra, A. S., and Valarmathi, B. (2020). Comprehensive data analysis and prediction on ipl using machine learning algorithms. *International Journal on Emerging Technologies*, 11(3):218–228.

Kumar, P. D., Ghosh, D. N., and Mondal, A. C. (2011). A mcdm approach for evaluating bowlers' performance in ipl. *Journal of Emerging Trends in Computing and Information Sciences*, 2(11).

Lin, T., Chen, Z., Johanna, B., and Hanspeter, P. (2022). Wearable sensors for monitoring the internal and external workload of the athlete. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):962–971.

Martínez, E. C., Cristaldi, M., and Grau, R. J. (2009). Design of dynamic experiments in modeling for optimization of batch processes. *Industrial and Engineering Chemistry Research*, 48(7).

Martínez, E. C., Cristaldi, M., and Grau, R. J. (2013). Dynamic optimization of bioreactors using probabilistic tendency models and bayesian active learning. *Computers and Chemical Engineering*, 49.

Mirjalili, F. (2019). A brief history of cricket in india. "https://www.timesknowledge.in/know-india/a-brief-history-of-cricket-in-india-1421-1.html. Last accessed on January 25, 2019, 5:00 PM IST.

Moore, A., Turner, D., and Johnstone, J. A. (2012). A preliminary analysis of team performance in english first-class twenty-twenty (t20) cricket. *International Journal of Performance Analysis in Sport*, 12(1):188–207.

Nasim, F., Yousaf, M. A., and Masood, S. (2020). Data-driven probabilistic system for batsman performance prediction in a cricket match. *Intelligent Automation & Soft Computing*, 36(3):2866–2876.

Nasrabadi, M. N. (2007). Pattern recognition and machine learning. *Journal of Electronic Imaging, 16(4)*.

Neeraj, P. and Hardik, W. (2016). Applications of modern classification techniques to predict the outcome of odi cricket. "https://www.sciencedirect.com/science/article/pii/S1877050916304653.

Nimmagadda, A., Teja, N. N. S., and Raju, C. G. (2018). Cricket score and winning prediction using data mining. *International Journal of Advanced Research and Development*, pages 300–301.

Panda, D. (2018). How india's improbable victory changed indian cricket. *Sportskeeda*. Last accessed on October 18, 2018, 15:00 IST, https://www.sportskeeda.com/cricket/t20-world-cup-2007-india-s-improbable-victory-and-its-significance.

Peniel, B. (2015). Research design.

Pers, J., Kovacic, S., and Vuckovic, G. (2005). Analysis and pattern detection on large amounts of annotated sport motion data using standard sql. *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*.

Petersen, C. J., Pyne, D. B., and Portus, M. (2011). Comparison of player movement patterns between 1-day and test cricket. *Journal of Strength and Conditioning Research*, 25(5):1368–1373.

Prakash, J., Khandelwal, M., and Pradhan, T. (2015). Evaluation of ipl teams and players using association, correlation, and classification rules. In *2015 International Conference on Computer, Communication and Control (IC4)*, pages 1–6. IEEE.

Ramachandra, G. (2003). A corner of a foreign field: The indian history of a british sport.

Roger, M. E. and Alberto, M.-O. (2009). Sage handbook of quantitative methods in psychology,.

Rory, B. P. and Fadi, T. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33.

Saikia, H., Bhattacharjee, D., and Bhattacharjee, A. (2012). Is ipl responsible for cricketers' performance in twenty20 world cup. *International Journal of Sports Science and Engineering*, 6(2):96–110.

Schumaker, R., Solieman, K. O., and Chen, H. (2010). Predictive modeling for sports and gaming. In *Sports Data Mining*, pages 55–63.

Sen, A. (2018). Five indian players who made it big after shining in the ipl. Last accessed on April 18, 2018, 12:46 IST, `https://www.news18.com/cricketnext/news/five-indian-players-who-made-it-big-after-shining-in-the-ipl-1725423.html`.

Seshadri, D., Li, R., and Drummond, C. (2022). The quest for: Embedded visualization for augmenting basketball game viewing experiences. *NPJ Digital Medicine*, 2(1):71.

Shah, P. (2022). Predicting outcome of live cricket match using duckworth-lewis par score. *International Journal of Systems Science and Applied Mathematics*, 6(7S):83–85.

Shah, S., Hazarika, P. J., and Hazarika, J. (2017). A study on performance of cricket players using factor analysis approach. *International Journal of Advanced Research in Computer Science*, 8(6):656–660.

Sharma, S. K. (2013). A factor analysis approach in performance analysis of t-20 cricket. *Journal of Reliability and Statistical Studies*, 6(1):69–76.

Shashi, K. (2013). How sustainable is the strategy of the indian premier league-ipl? a critical review of 10 key issues that impact the ipl strategy. *International Journal of Scientific and Research Publications (IJSRP)*, 3(12):1–11.

Sherif, A. (2016). *Practical Business Intelligence*. Packt Publishing Ltd.

Singh, A. and Bagchi, A. (2020). Impact of indian premier league on test cricket in india. *Annals of Tropical Medicine and Public Health*, 23(17).

Singh, G., Bhatia, N., and Singh, S. (2011). Fuzzy logic-based cricket player performance evaluator. *IJCA Special Issue on Artificial Intelligence Techniques - Novel Approaches and Practical Applications*, 1(3):11–16.

Singh, R., Saikia, H., and Bhattacharjee, D. (2015). Predictive modeling for sports and gaming. *A Journal of Management Research*, 14(1):14–24.

Singh, S. (2011). Measuring the performance of teams in the indian premier league. *American Journal of Operations Research*, 1(3):180–183.

Swartz Tim, Gill S Paramjit, B. D. and deSilva Basil M (2006). Optimal batting orders in one-day cricket. *Computers & Operations Research*, 33(7):1939–1950.

szymon. Kozak Jan, G. and przemyslaw, J. (2023). Knowledge discovery in databases for a football match result.

Thomas, J., Silverman, J., and Nelson, J. (2015). *Research Methods in Physical Activity (7th Edition)*. Human Kinetics.

Thomas, O. O. and Lawal, O. R. (2020). Exploratory research design in management sciences: An x-ray of literature. *Annals of the University Dunarea de Jos of Galati: Fascicle I, Economics and Applied Informatics*.

Vidit, K. and Tulasi, B. (2019). Data visualization and toss related analysis of ipl teams and batsmen performances. *International Journal of Electrical and Computer Engineering (IJECE)*, 9(5):4423–4432.

Yang, L., Javed, M. F., and Salem (2022). Comparative analysis of the optimized knn, svm, and ensemble dt models using bayesian optimization for predicting pedestrian fatalities: An advance towards realizing the sustainable safety of pedestrians. pages 1–18.

6 Appendix

The data set contains two CSV files called **matches.csv** and **ball by ball.csv**

The database and tables were created in Postgres

The code for creating tables is as follows:

```
CREATE TABLE cricket_stats (
    ID INTEGER,
    innings INTEGER,
    overs NUMERIC,
    ballnumber INTEGER,
    batter VARCHAR(100),
    bowler VARCHAR(100),
    non-striker VARCHAR(100),
    extra_type VARCHAR(100),
    batsman_run INTEGER,
    extras_run INTEGER,
    total_run INTEGER,
    non_boundary BOOLEAN,
    isWicketDelivery BOOLEAN,
    player_out VARCHAR(100),
    kind VARCHAR(100),
    fielders_involved VARCHAR(100),
    BattingTeam VARCHAR(100)
);

CREATE TABLE cricket_matches_3 (
    ID serial PRIMARY KEY,
    City varchar(100),
    Date date,
    Season varchar(20),
    MatchNumber varchar(20), -- Using varchar for mixed-format match numbers
    Team1 varchar(100),
    Team2 varchar(100),
    Venue varchar(200),
    TossWinner varchar(100),
    TossDecision varchar(20),
    SuperOver boolean,
    WinningTeam varchar(100),
    WonBy varchar(20),
    Margin varchar(50),
    Method varchar(50),
```

```
    Player_of_Match varchar(100),
    Team1Players text[], -- Using an array for player names
    Team2Players text[], -- Using an array for player names
    Umpire1 varchar(100),
    Umpire2 varchar(100)
);
```

Python program for player performance

Code for Most Valuable Batter

```
import pandas as pd
import plotly.graph_objects as go
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error

# Load the data into a pandas DataFrame (replace 'data.csv' with your data file)
data = pd.read_csv(r"C:\Users\pja06\Downloads\batting_average_by_season_202309071622.csv")

# Data Preprocessing (e.g., handling missing values, converting columns)

# Calculate average runs per inning
data['average_runs_per_inning'] = data['total_runs'] / data['total_innings']

# Convert 'season' column to integers
data['season'] = data['season'].str.extract('(\d+)').astype(int)

# Filter the dataset for the year 2016 and onwards
filtered_data = data[data['season'] >= 2016]

# Filter the dataset to include only players with more than 10 innings and batting average > 30
filtered_data = filtered_data[(filtered_data['total_innings'] > 10) & (filtered_data['batting_average'] > 30)]

# Split data into features (X) and target (y)
X = filtered_data[['total_innings', 'total_runs', 'average_runs_per_inning']]
y = filtered_data['batting_average']

# Split 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)

# Random Forest regression model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Predictions on the testing data
y_pred = model.predict(X_test)
```

```
38
39 # Evaluate the model (calculate RMSE)
40 rmse = mean_squared_error(y_test, y_pred, squared=False)
41 print("Root Mean Squared Error (RMSE):", rmse)
42
43 # Rank players based on predicted batting averages
44 filtered_data['predicted_batting_average'] = model.predict(X)
45 ranked_data = filtered_data.sort_values(by='predicted_batting_average', ascending=False)
46
47 # New column for labels, set to None for all data points
48 filtered_data['labels'] = None
49
50 # Assign labels to the top 10 batters
51 top_10_batters = ranked_data.head(10)
52 filtered_data.loc[top_10_batters.index, 'labels'] = top_10_batters['batsman']
53
54 # Define a condition for color and text position assignment
55 def color_and_text_position_condition(row):
56     batting_average = row['predicted_batting_average']
57     if batting_average < 40:
58         color = 'darkred' # Dark red for batting average below 40
59         marker_symbol = 'triangle-up'
60         text_position = 'bottom right' # Text position for dark red
61     elif 40 <= batting_average < 50:
62         color = 'blue' # Blue for batting average between 40 and 50
63         marker_symbol = 'circle'
64         text_position = 'bottom right' # Text position for blue
65     else:
66         color = 'green' # Green for batting average 50 and above
67         marker_symbol = 'square'
68         text_position = 'top left' # Text position for green
69     return color, marker_symbol, text_position
70
71 # New columns for color, marker symbol, and text position
72 filtered_data['color'], filtered_data['marker_symbol'], filtered_data['text_position'] =
       zip(*filtered_data.apply(color_and_text_position_condition, axis=1))
73
74 # Mapping of color to legend label
75 color_legend_mapping = {
76     'darkred': 'Below 40',
77     'blue': 'Between 40 and 50',
78     'green': '50 and Above'
79 }
80
81 # Create a figure
82 fig = go.Figure()
83
84 # Traces for each color threshold and display data labels
85 for color in color_legend_mapping:
86     color_data = filtered_data[filtered_data['color'] == color]
87     fig.add_trace(go.Scatter(
```

```
        x=color_data['predicted_batting_average'],
        y=color_data['batting_average'],
        mode='markers+text', # Include text labels
        marker=dict(
            size=10,
            opacity=0.7,
            symbol=color_data['marker_symbol'],
            line=dict(width=2),
            color=color
        ),
        text=color_data['labels'], # Display data labels
        textposition=color_data['text_position'],
        name=color_legend_mapping[color] #legend label
    ))

#legends for different categories of batting averages
fig.add_trace(go.Scatter(x=[], y=[], mode='markers', name='Below 40', marker=dict(size=10, opacity=0.7,
      symbol='triangle-up', line=dict(width=2), color='darkred')))
fig.add_trace(go.Scatter(x=[], y=[], mode='markers', name='Between 40 and 50', marker=dict(size=10, opacity=0.7,
      symbol='circle', line=dict(width=2), color='blue')))
fig.add_trace(go.Scatter(x=[], y=[], mode='markers', name='50 and Above', marker=dict(size=10, opacity=0.7,
      symbol='square', line=dict(width=2), color='green')))

# Update layout options
fig.update_layout(
    title='Most valuable Batters (Random Forest)',
    xaxis_title='Predicted Batting Average',
    yaxis_title='Batting Average',
    legend=dict(title='Batting Average Thresholds')
)

# Show the plot
fig.show()

#For SVM MODEL
 Load the data into a pandas DataFrame (replace 'data.csv' with your data file)
data = pd.read_csv(r"C:\Users\pja06\Downloads\batting_average_by_season_202309071622.csv")

# Data Preprocessing (e.g., handling missing values, converting columns)

# Calculate average runs per inning
data['average_runs_per_inning'] = data['total_runs'] / data['total_innings']

# Convert 'season' column to integers
data['season'] = data['season'].str.extract('(\d+)').astype(int)

# Filter the dataset for the year 2016 and onwards
filtered_data = data[data['season'] >= 2016]

# Filter the dataset to include only players with more than 10 innings and batting average > 30
filtered_data = filtered_data[(filtered_data['total_innings'] > 10) & (filtered_data['batting_average'] > 30)]
```

```
136
137 # Split data into features (X) and target (y)
138 X = filtered_data[['total_innings', 'total_runs', 'average_runs_per_inning']]
139 y = filtered_data['batting_average']
140
141 # Split data into training and testing sets
142 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
143
144 # Train an SVM regression model
145 model = SVR(kernel='linear')
146 model.fit(X_train, y_train)
147
148 # Predictions on the testing data
149 y_pred = model.predict(X_test)
150
151 # Evaluate the model (calculate RMSE)
152 rmse = mean_squared_error(y_test, y_pred, squared=False)
153 print("Root Mean Squared Error (RMSE):", rmse)
154
155 # Rank players based on predicted batting averages
156 filtered_data['predicted_batting_average'] = model.predict(X)
157 ranked_data = filtered_data.sort_values(by='predicted_batting_average', ascending=False)
158
159 #new column for labels, set to None for all data points
160 filtered_data['labels'] = None
161
162 # labels to the top 10 batters
163 top_10_batters = ranked_data.head(10)
164 filtered_data.loc[top_10_batters.index, 'labels'] = top_10_batters['batsman']
165
166 # Define a condition for color and text position assignment
167 def color_and_text_position_condition(row):
168     batting_average = row['predicted_batting_average']
169     if batting_average < 40:
170         color = 'darkred' # Dark red for batting average below 40
171         marker_symbol = 'triangle-up'
172         text_position = 'bottom right' # Text position for dark red
173     elif 40 <= batting_average < 50:
174         color = 'blue' # Blue for batting average between 40 and 50
175         marker_symbol = 'circle'
176         text_position = 'bottom right' # Text position for blue
177     else:
178         color = 'green' # Green for batting average 50 and above
179         marker_symbol = 'square'
180         text_position = 'top left' # Text position for green
181     return color, marker_symbol, text_position
182
183 # condition to create new columns for color, marker symbol, and text position
184 filtered_data['color'], filtered_data['marker_symbol'], filtered_data['text_position'] =
        zip(*filtered_data.apply(color_and_text_position_condition, axis=1))
185
```

```
# Mapping of color to legend label
color_legend_mapping = {
    'darkred': 'Below 40',
    'blue': 'Between 40 and 50',
    'green': '50 and Above'
}

# Create a figure
fig = go.Figure()

# Add traces for each color threshold and display data labels
for color in color_legend_mapping:
    color_data = filtered_data[filtered_data['color'] == color]
    fig.add_trace(go.Scatter(
        x=color_data['predicted_batting_average'],
        y=color_data['batting_average'],
        mode='markers+text', # Include text labels
        marker=dict(
            size=10,
            opacity=0.7,
            symbol=color_data['marker_symbol'],
            line=dict(width=2),
            color=color
        ),
        text=color_data['labels'], # Display data labels
        textposition=color_data['text_position'],
        name=color_legend_mapping[color] # Use the legend label
    ))

# legends for different categories of batting averages
fig.add_trace(go.Scatter(x=[], y=[], mode='markers', name='Below 40', marker=dict(size=10, opacity=0.7,
      symbol='triangle-up', line=dict(width=2), color='darkred')))
fig.add_trace(go.Scatter(x=[], y=[], mode='markers', name='Between 40 and 50', marker=dict(size=10, opacity=0.7,
      symbol='circle', line=dict(width=2), color='blue')))
fig.add_trace(go.Scatter(x=[], y=[], mode='markers', name='50 and Above', marker=dict(size=10, opacity=0.7,
      symbol='square', line=dict(width=2), color='green')))

# Update layout options
fig.update_layout(
    title='Most valuable Batters (SVM)',
    xaxis_title='Predicted Batting Average',
    yaxis_title='Batting Average',
    legend=dict(title='Batting Average Thresholds')
)

# Show the plot
fig.show()
```

Code for Most valuable bowler

```
import pandas as pd
import plotly.graph_objects as go
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error

# Load the data into a pandas DataFrame (replace 'bowling_data.csv' with your data file)
data = pd.read_csv(r"C:\Users\pja06\Downloads\bowling_averages_202309081302.csv")

# Data Preprocessing (e.g., handling missing values, converting columns)
# Convert 'season' column to integers
data['season'] = data['season'].str.extract('(\d+)').astype(int)

# Filter the dataset for seasons from 2016 onwards
filtered_data = data[data['season'] >= 2016]

# Split the dataset to include only bowlers with more than 10 wickets and a lower bowling average
filtered_data = filtered_data[(filtered_data['wickets'] > 15) & (filtered_data['bowling_average'] < 25)]

# Split data into features (X) and target (y)
X = filtered_data[['total_runs', 'wickets', 'bowling_average']]
y = filtered_data['bowling_average']

# Split 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)

# Train a Random Forest Regressor model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the testing data
y_pred = model.predict(X_test)

# Evaluate the model (calculate RMSE)
rmse = mean_squared_error(y_test, y_pred, squared=False)
print("Root Mean Squared Error (RMSE):", rmse)

# Rank bowlers based on predicted bowling averages
filtered_data['predicted_bowling_average'] = model.predict(X)
ranked_data = filtered_data.sort_values(by='predicted_bowling_average')

# scatter plot to visualize the actual vs. predicted bowling averages for all bowlers
fig = go.Figure()

# Define color and marker based on specified ranges
def get_color_marker(bowling_average):
    if bowling_average >= 20:
        return 'darkred', 'triangle-up'
    elif 16 <= bowling_average < 20:
        return 'darkblue', 'circle'
    else:
```

```
        return 'green', 'square'

# The color and marker based on the bowling average
filtered_data['color'], filtered_data['marker'] = zip(*filtered_data['bowling_average'].apply(get_color_marker))

# Add data points to the plot with reversed axes and bowler labels for averages below 16
for color in ['darkblue', 'green', 'darkred']:
    subset = filtered_data[filtered_data['color'] == color]

    for index, row in subset.iterrows():
        if color == 'green':
            # Alternate text positions for green data points
            if index % 4 == 0:
                text_position = 'top right'
            elif index % 4 == 1:
                text_position = 'top left'
            elif index % 4 == 2:
                text_position = 'bottom left'
            else:
                text_position = 'bottom right'
        else:
            text_position = 'top left' if color == 'green' else 'bottom right'

        if row['bowling_average'] < 16:
            fig.add_trace(go.Scatter(
                x=[row['bowling_average']],
                y=[row['predicted_bowling_average']],
                text=[row['bowler']], # Include bowler's name as text label
                mode='markers+text', # Display text on the graph
                marker=dict(size=10, color=row['color'], symbol=row['marker']),
                showlegend=False,
                textposition=text_position,
                textfont=dict(size=10),
            ))
        else:
            fig.add_trace(go.Scatter(
                x=[row['bowling_average']],
                y=[row['predicted_bowling_average']],
                mode='markers',
                marker=dict(size=10, color=row['color'], symbol=row['marker']),
                showlegend=False
            ))

# The reversed X-axis and Y-axis labels
fig.update_xaxes(title_text='Predicted Bowling Average', autorange='reversed')
fig.update_yaxes(title_text='Actual Bowling Average', autorange='reversed')

# Title to the plot
fig.update_layout(title='Actual vs. Predicted Bowling Averages (All Bowlers - Random Forest)')

#single legend for "Bowling Average Threshold" with corresponding measures and colors
```

```
legend_colors = {
    'darkblue': '16-20',
    'green': '<16',
    'darkred': '>=20',
}

for color, measure in legend_colors.items():
    fig.add_trace(go.Scatter(
        x=[None],
        y=[None],
        mode='markers',
        marker=dict(size=10, color=color, symbol='square'),
        showlegend=True,
        name=f'Bowling Avg: {measure}',
    ))

# Show the plot
fig.show()

# Capture the data used to create the scatter plot
scatter_data = []

for color in ['darkblue', 'green', 'darkred']:
    subset = filtered_data[filtered_data['color'] == color]

    for index, row in subset.iterrows():
        data_point = {
            'Bowler': row['bowler'],
            'Bowling Average': row['bowling_average'],
            'Predicted Bowling Average': row['predicted_bowling_average']
        }
        scatter_data.append(data_point)

# Create a data frame from the captured data
scatter_df = pd.DataFrame(scatter_data)

# the path where you want to save the CSV file
csv_path = r'C:\Users\pja06\Downloads\scatter_data.csv'

# the DataFrame to a CSV file
scatter_df.to_csv(csv_path, index=False)

# Print a message to confirm that the data has been exported
print(f"Scatter data has been exported to {csv_path}")


#For SVM MODEL
# Load the data into a pandas DataFrame (replace 'bowling_data.csv' with your data file)
data = pd.read_csv(r"C:\Users\pja06\Downloads\bowling_averages_202309081302.csv")

# Data Preprocessing (e.g., handling missing values, converting columns)
```

```
# Convert 'season' column to integers
data['season'] = data['season'].str.extract('(\d+)').astype(int)

# Filter the dataset for seasons from 2016 onwards
filtered_data = data[data['season'] >= 2016]

# Split the dataset to include only bowlers with more than 10 wickets and a lower bowling average
filtered_data = filtered_data[(filtered_data['wickets'] > 15) & (filtered_data['bowling_average'] < 25)]

# Split data into features (X) and target (y)
X = filtered_data[['total_runs', 'wickets', 'bowling_average']]
y = filtered_data['bowling_average']

# Split 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)

# Train a Support Vector Regression (SVR) model
model = SVR(kernel='linear')
model.fit(X_train, y_train)

# predictions on the testing data
y_pred = model.predict(X_test)

# Evaluate the model (calculate RMSE)
rmse = mean_squared_error(y_test, y_pred, squared=False)
print("Root Mean Squared Error (RMSE):", rmse)

# Rank bowlers based on predicted bowling averages using SVR
filtered_data['predicted_bowling_average'] = model.predict(X)
ranked_data = filtered_data.sort_values(by='predicted_bowling_average')

# scatter plot to visualize the actual vs. predicted bowling averages for all bowlers
fig = go.Figure()

# Define color and marker based on specified ranges
def get_color_marker(bowling_average):
    if bowling_average >= 20:
        return 'darkred', 'triangle-up'
    elif 16 <= bowling_average < 20:
        return 'darkblue', 'circle'
    else:
        return 'green', 'square'

# color and marker based on bowling average
filtered_data['color'], filtered_data['marker'] = zip(*filtered_data['bowling_average'].apply(get_color_marker))

# Data points to the plot with reversed axes and bowler labels for averages below 16
for color in ['darkblue', 'green', 'darkred']:
    subset = filtered_data[filtered_data['color'] == color]

    for index, row in subset.iterrows():
```

```
        if color == 'green':
            # Alternate text positions for green data points
            if index % 4 == 0:
                text_position = 'top right'
            elif index % 4 == 1:
                text_position = 'top left'
            elif index % 4 == 2:
                text_position = 'bottom left'
            else:
                text_position = 'bottom right'
        else:
            text_position = 'top left' if color == 'green' else 'bottom right'

        if row['bowling_average'] < 16:
            fig.add_trace(go.Scatter(
                x=[row['bowling_average']],
                y=[row['predicted_bowling_average']],
                text=[row['bowler']], # Include bowler's name as text label
                mode='markers+text', # Display text on the graph
                marker=dict(size=10, color=row['color'], symbol=row['marker']),
                showlegend=False,
                textposition=text_position,
                textfont=dict(size=10),
            ))
        else:
            fig.add_trace(go.Scatter(
                x=[row['bowling_average']],
                y=[row['predicted_bowling_average']],
                mode='markers',
                marker=dict(size=10, color=row['color'], symbol=row['marker']),
                showlegend=False
            ))

# Customize the reversed X-axis and Y-axis labels
fig.update_xaxes(title_text='Predicted Bowling Average', autorange='reversed')
fig.update_yaxes(title_text='Actual Bowling Average', autorange='reversed')

# Add a title to the plot
fig.update_layout(title='Actual vs. Predicted Bowling Averages (All Bowlers - SVR)')

# Create a single legend for "Bowling Average Threshold" with corresponding measures and colors for SVR
legend_colors_svr = {
    'darkblue': '16-20',
    'green': '<16',
    'darkred': '>=20',
}

for color, measure in legend_colors_svr.items():
    fig.add_trace(go.Scatter(
        x=[None],
        y=[None],
```

```
        mode='markers',
        marker=dict(size=10, color=color, symbol='square'),
        showlegend=True,
        name=f'Bowling Avg: {measure}',
    ))

# Show the plot
fig.show()
```

```
CREATE MATERIALIZED VIEW batting_average_by_season AS
SELECT
    cs.batter AS batsman,
    cm3.Season,
    COUNT(DISTINCT cm3.ID) AS total_innings,
    SUM(cs.batsman_run) AS total_runs,
    CAST(ROUND(SUM(cs.batsman_run) * 1.0 / COUNT(DISTINCT cm3.ID)) AS INT) AS batting_average
FROM
    cricket_stats AS cs
JOIN
    cricket_matches_3 AS cm3 ON cs.ID = cm3.ID
GROUP BY
    cs.batter, cm3.Season;


create materialized view bowling_averages as
SELECT cricket_scores.bowler,
       matches.season,
       SUM(cricket_scores.total_run - cricket_scores.extras_run) AS total_runs,
       COUNT(CASE WHEN cricket_scores.is_wicket_delivery THEN 1 END) AS wickets,
       CASE WHEN COUNT(CASE WHEN cricket_scores.is_wicket_delivery THEN 1 END) > 5
       THEN ROUND(SUM(cricket_scores.total_run - cricket_scores.extras_run) / COUNT(CASE WHEN
            cricket_scores.is_wicket_delivery THEN 1 END), 2)
       ELSE 0 END AS bowling_average
FROM cricket_scores
INNER JOIN matches ON cricket_scores.id = matches.id
GROUP BY cricket_scores.bowler, matches.season
HAVING COUNT(DISTINCT matches.id) >= 10 and
    COUNT(cricket_scores.is_wicket_delivery) > 10
order by bowling_average;
```

PostgresSQL Query for Most Impactful Batter

```
    create materialized view strike_rate as (
SELECT
    cricket_stats.batter,
    cricket_matches_3.Season,
    COUNT(cricket_stats.ballnumber) AS BallsFaced,
    SUM(cricket_stats.batsman_run) AS RunsScored,
    CASE
```

```
        WHEN cricket_stats.overs <= 6 THEN 'Powerplay'
        WHEN cricket_stats.overs > 6 AND cricket_stats.overs <= 15 THEN 'Middle Overs'
        ELSE 'Death Overs'
    END AS OverType,
    ROUND((SUM(cricket_stats.batsman_run) * 100.0) / COUNT(cricket_stats.ballnumber), 2) AS StrikeRate
FROM
    cricket_stats
    INNER JOIN cricket_matches_3 ON cricket_stats.ID = cricket_matches_3.ID
GROUP BY
    cricket_stats.batter,
    cricket_matches_3.Season,
    OverType
HAVING
    COUNT(cricket_stats.ballnumber) > 100
```

PostgresSQL Query for Most Impactful Bowler

```
SELECT
    cm.season AS Season,
    cs.bowler AS Bowler,
    count(cs.overs) AS TotalBalls,
    SUM(cs.total_run) AS TotalRunsConceded,
    CASE WHEN count(cs.overs) = 0 THEN 0
         ELSE ROUND(SUM(cs.total_run) * 6 / count(cs.overs), 2)
    END AS EconomyRate
FROM
    cricket_stats cs
JOIN
    cricket_matches_3 cm ON cs.ID = cm.ID
WHERE
    cs.overs IS NOT NULL
    AND cs.overs <= 6 -- Consider only powerplay overs (first 6 overs)
    cs.overs > 6 AND cs.overs <= 15 -- Consider middle overs (overs 7 to 15)
    and cs.overs > 15 -- Consider death overs (overs 16 onwards)
GROUP BY
    cm.season, cs.bowler
HAVING
    COUNT(DISTINCT cm.id) > 3
ORDER BY
    Season ASC, EconomyRate desc;
```

PostgreSQL Query For Team Performance

Batting Performance Query for 2017,2019 and 2020

```sql
    CREATE MATERIALIZED view batting_average_for_team AS
SELECT
    cs.batter AS batsman,
    cm3.Season,
    teams.team_name AS team,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) AS total_runs_won_matches,
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) AS won_matches,
    CASE
        WHEN COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) = 0 THEN 0
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) * 1.0 /
             COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END), 2)
    END AS batting_average_won_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run ELSE 0 END) AS total_runs_lost_matches,
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) AS lost_matches,
    CASE
        WHEN COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) = 0 THEN 0
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run ELSE 0 END) * 1.0 /
             COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END), 2)
    END AS batting_average_lost_matches
FROM
    cricket_stats cs
JOIN
    cricket_matches_3 cm3 ON cs.ID = cm3.ID
JOIN
    teams ON (cs.BattingTeam = teams.team_name)
WHERE
    cm3.WinningTeam IS NOT NULL
GROUP BY
    cs.batter, cm3.Season, teams.team_name
HAVING
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) > 2
ORDER BY
    batsman, Season, team;



CREATE MATERIALIZED view Batting_strikerate_for_team AS
   SELECT
    cm3.Season,
    teams.team_name AS team,
    cs.batter AS batsman,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) AS total_runs_won_matches,
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) AS won_matches,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) AS balls_faced_won_matches,
    CASE
        WHEN COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) = 0 THEN 0
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) * 100.0 / SUM(CASE
             WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END), 2)
    END AS strike_rate_won_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run ELSE 0 END) AS total_runs_lost_matches,
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) AS lost_matches,
```

```sql
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) AS balls_faced_lost_matches,
    CASE
        WHEN COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) = 0 THEN 0
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run ELSE 0 END) * 100.0 / SUM(CASE
             WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END), 2)
    END AS strike_rate_lost_matches
FROM
    cricket_stats cs
JOIN
    cricket_matches_3 cm3 ON cs.ID = cm3.ID
JOIN
    teams ON (CASE WHEN cs.BattingTeam = teams.team_name THEN cs.BattingTeam ELSE cs.battingteam END = teams.team_name)
WHERE
    cm3.WinningTeam IS NOT NULL
GROUP BY
    cs.batter, cm3.Season, teams.team_name
HAVING
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) > 2
    AND SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) > 100
ORDER BY
    batsman, Season, team;

 CREATE MATERIALIZED view Boundary_percentage_for_team AS
  SELECT
    cm3.Season,
    teams.team_name AS team,
    cs.batter AS batsman,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) AS total_runs_won_matches,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND (cs.batsman_run = 4 OR cs.batsman_run = 6) THEN cs.batsman_run
         ELSE 0 END) AS boundary_runs_won_matches,
    CASE
        WHEN SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) = 0 THEN NULL
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND (cs.batsman_run = 4 OR cs.batsman_run = 6) THEN
             cs.batsman_run ELSE 0 END) * 100.0 / SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run
             ELSE 0 END), 2)
    END AS boundary_percentage_won_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run ELSE 0 END) AS total_runs_lost_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name AND (cs.batsman_run = 4 OR cs.batsman_run = 6) THEN cs.batsman_run
         ELSE 0 END) AS boundary_runs_lost_matches,
    CASE
        WHEN SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run ELSE 0 END) = 0 THEN NULL
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam != teams.team_name AND (cs.batsman_run = 4 OR cs.batsman_run = 6) THEN
             cs.batsman_run ELSE 0 END) * 100.0 / SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.batsman_run
             ELSE 0 END), 2)
    END AS boundary_percentage_lost_matches
FROM
    cricket_stats cs
JOIN
    cricket_matches_3 cm3 ON cs.ID = cm3.ID
JOIN
    teams ON (cs.BattingTeam = teams.team_name)
```

```
WHERE
    cm3.WinningTeam IS NOT NULL
GROUP BY
    cs.batter, cm3.Season, teams.team_name
HAVING
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) > 2
    OR COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) > 2
    AND SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.batsman_run ELSE 0 END) > 200
ORDER BY
    batsman, Season, team;
```

Bowling Performance Query for 2017,2019 and 2020

```
    CREATE MATERIALIZED view Bowling_average_for_team AS
SELECT
    cm3.Season,
    teams.team_name AS team,
    cs.bowler AS bowler,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.total_run ELSE 0 END) AS runs_conceded_won_matches,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND cs.isWicketDelivery = TRUE THEN 1 ELSE 0 END) AS
        wickets_won_matches,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) AS balls_bowled_won_matches,
    CASE
        WHEN SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND cs.isWicketDelivery = TRUE THEN 1 ELSE 0 END) = 0 THEN
            NULL
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.total_run ELSE 0 END) * 1.0 / (SUM(CASE WHEN
            cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) / 6), 2)
    END AS bowling_average_won_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.total_run ELSE 0 END) AS runs_conceded_lost_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name AND cs.isWicketDelivery = TRUE THEN 1 ELSE 0 END) AS
        wickets_lost_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) AS balls_bowled_lost_matches,
    CASE
        WHEN SUM(CASE WHEN cm3.WinningTeam != teams.team_name AND cs.isWicketDelivery = TRUE THEN 1 ELSE 0 END) = 0 THEN
            NULL
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.total_run ELSE 0 END) * 1.0 / (SUM(CASE WHEN
            cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) / 6), 2)
    END AS bowling_average_lost_matches
FROM
    cricket_stats cs
INNER JOIN
    cricket_matches_3 cm3 ON cs.ID = cm3.ID
INNER JOIN
    teams ON (cm3.team1_id = teams.team_id OR cm3.team2_id = teams.team_id)
WHERE
    cm3.WinningTeam IS NOT NULL
GROUP BY
    cm3.Season, teams.team_name, cs.bowler
HAVING
```

```
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND cs.isWicketDelivery = TRUE THEN 1 ELSE 0 END) > 3
    AND (COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) > 2
    OR COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) > 2)
ORDER BY
    bowler, Season, team;



CREATE MATERIALIZED view Bowling_economy_rate_for_team AS
SELECT
    cm3.Season,
    teams.team_name AS team,
    cs.bowler AS bowler,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.total_run ELSE 0 END) AS runs_conceded_won_matches,
    SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) AS balls_bowled_won_matches,
    CASE
        WHEN SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) = 0 THEN NULL
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN cs.total_run ELSE 0 END) * 6.0 / SUM(CASE WHEN
             cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END), 2)
    END AS economy_rate_won_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.total_run ELSE 0 END) AS runs_conceded_lost_matches,
    SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) AS balls_bowled_lost_matches,
    CASE
        WHEN SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) = 0 THEN NULL
        ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN cs.total_run ELSE 0 END) * 6.0 / SUM(CASE WHEN
             cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END), 2)
    END AS economy_rate_lost_matches
FROM
    cricket_stats cs
JOIN
    cricket_matches_3 cm3 ON cs.ID = cm3.ID
JOIN
    (
        SELECT
            team_id,
            team_name
        FROM
            teams
    ) teams ON (cm3.team1_id = teams.team_id OR cm3.team2_id = teams.team_id)
WHERE
    cm3.WinningTeam IS NOT NULL
GROUP BY
    cs.bowler, cm3.Season, teams.team_name
HAVING
    COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) > 2
    OR COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) > 2
ORDER BY
    bowler, Season, team;

CREATE MATERIALIZED view Bowling_dotball_percentage_for_team AS
SELECT
```

```
80      cm3.Season,
81      teams.team_name AS team,
82      cs.bowler AS bowler,
83      SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND cs.total_run = 0 THEN 1 ELSE 0 END) AS dot_balls_won_matches,
84      SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) AS balls_bowled_won_matches,
85      CASE
86          WHEN SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END) = 0 THEN NULL
87          ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam = teams.team_name AND cs.total_run = 0 THEN 1 ELSE 0 END) * 100.0 /
                SUM(CASE WHEN cm3.WinningTeam = teams.team_name THEN 1 ELSE 0 END), 2)
88      END AS dot_ball_percentage_won_matches,
89      SUM(CASE WHEN cm3.WinningTeam != teams.team_name AND cs.total_run = 0 THEN 1 ELSE 0 END) AS dot_balls_lost_matches,
90      SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) AS balls_bowled_lost_matches,
91      CASE
92          WHEN SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END) = 0 THEN NULL
93          ELSE ROUND(SUM(CASE WHEN cm3.WinningTeam != teams.team_name AND cs.total_run = 0 THEN 1 ELSE 0 END) * 100.0 /
                SUM(CASE WHEN cm3.WinningTeam != teams.team_name THEN 1 ELSE 0 END), 2)
94      END AS dot_ball_percentage_lost_matches
95  FROM
96      cricket_stats cs
97  JOIN
98      cricket_matches_3 cm3 ON cs.ID = cm3.ID
99  JOIN
100     teams ON (cm3.team1_id = teams.team_id OR cm3.team2_id = teams.team_id)
101 WHERE
102     cm3.WinningTeam IS NOT NULL
103 GROUP BY
104     cs.bowler, cm3.Season, teams.team_name
105 HAVING
106     COUNT(DISTINCT CASE WHEN cm3.WinningTeam = teams.team_name THEN cm3.ID END) > 2
107     OR COUNT(DISTINCT CASE WHEN cm3.WinningTeam != teams.team_name THEN cm3.ID END) > 2
108 ORDER BY
109     bowler, Season, team;
```