# Table of Contents

Table of Contents

[Table of Contents 1](#_Toc175521842)

[Abstract 3](#_Toc175521843)

[List of Figures 3](#_Toc175521844)

[List of Tables 3](#_Toc175521845)

[Abbreviations 3](#_Toc175521846)

[1. Introduction 3](#_Toc175521847)

[1.1 Background 3](#_Toc175521848)

[Team performances over the years 6](#_Toc175521849)

[Research Objective and Need for Study: 8](#_Toc175521850)

[Scopes and Limitations 8](#_Toc175521851)

[Significance of the Study 9](#_Toc175521852)

[Structure of the research 10](#_Toc175521853)

[Literature Review 11](#_Toc175521854)

[Adoption of Machine Learning on Sports 12](#_Toc175521855)

[Adoption of Machine Learning on Cricket 13](#_Toc175521856)

[Summary of Literature Review: 14](#_Toc175521857)

[Research Methodology 15](#_Toc175521858)

[Dataset and Approach Overview: 15](#_Toc175521859)

[Domain Knowledge 17](#_Toc175521860)

[Quantitative and Predictive analysis 19](#_Toc175521861)

[1. Win Ratio Analysis of Teams 19](#_Toc175521862)

[Need for Analysis 19](#_Toc175521863)

[Objective 19](#_Toc175521864)

[Data overview 19](#_Toc175521865)

[Quantitative Analysis 19](#_Toc175521866)

[Linear Regression Model: 21](#_Toc175521867)

[Prediction Analysis 22](#_Toc175521868)

[2.Rule-based scoring system combined with normalization and weighted aggregation 25](#_Toc175521869)

[Need for Analysis 25](#_Toc175521870)

[Objective 26](#_Toc175521871)

[Data Overview 27](#_Toc175521872)

[Quantitative Analysis 28](#_Toc175521873)

[Findings and insights 30](#_Toc175521874)

[3.Random forest model to predict the overall score 33](#_Toc175521875)

[Need for Analysis: 33](#_Toc175521876)

[Objective of the Model 33](#_Toc175521877)

[Data Overview 33](#_Toc175521878)

[Random Forest Regression Model 34](#_Toc175521879)

[Model Visualization 34](#_Toc175521880)

[4.Random Forest Model using RandomizedSearchCV 36](#_Toc175521881)

[Objective of the model 36](#_Toc175521882)

[Representation of Random Forest model with RandomizedSearchCV 36](#_Toc175521883)

[Model Visualization 37](#_Toc175521884)

[5. XG Boosting Method 38](#_Toc175521885)

[Objective of this model 38](#_Toc175521886)

[Representation of the Model 39](#_Toc175521887)

[Model Visualization 40](#_Toc175521888)

[6.Enhanced XG Boosting model 41](#_Toc175521889)

[Representation of the model 41](#_Toc175521890)

[7. Support Vector Regression Model 43](#_Toc175521891)

[Objective of the model 43](#_Toc175521892)

[Representation of the model 43](#_Toc175521893)

[Machine Learning Models and their accuracy results 45](#_Toc175521894)

[Evaluation 45](#_Toc175521895)

[Fine-Tuning 46](#_Toc175521896)

[Model Testing 46](#_Toc175521897)

[Performance Distribution Curves 48](#_Toc175521898)

[ROC curves 48](#_Toc175521899)

[Players Overall Performance score for KKR and DC 50](#_Toc175521900)

[Kolkata Knight Riders Current Players Analysis 50](#_Toc175521901)

[Delhi Capitals Current Players Analysis 52](#_Toc175521902)

[References 53](#_Toc175521903)

# Abstract

# List of Figures

# List of Tables

# Abbreviations

# 1. Introduction

## 1.1 Background

The Indian Premier League (IPL) has revolutionized cricket since its inception in 2008, becoming one of the most popular and lucrative sports leagues globally (Board of Control for Cricket in India, 2023). This professional Twenty20 cricket tournament features ten franchise teams representing different Indian cities or states, competing in a high-stakes, fast-paced format that has captured the imagination of fans worldwide.

A blue ribbon with text

Description automatically generated

**Figure 1. IPL logo**

**Source: iplt20**

The IPL's success can be attributed to several factors. Firstly, its star-studded lineups attract top cricket talent from around the world. Each team can field up to four overseas players in their playing eleven, creating a melting pot of international stars alongside India's best cricketers (ESPN Cricinfo, 2023). This combination of global and local talent has helped the IPL become a cricketing spectacle that consistently ranks among the top sports leagues in terms of average attendance.

The tournament's economic impact has been substantial. In 2022, the league's brand value was estimated at ₹90,038 crore (US$11 billion) (Duff & Phelps, 2022). Its contribution to India's GDP is significant, with the 2015 season alone adding ₹1,150 crore (US$140 million) to the economy (BCCI, 2016). The league's valuation has skyrocketed, reaching US$10.9 billion in December 2022 and achieving "decacorn" status (Economic Times, 2023).

The IPL's popularity is reflected in its lucrative media rights deals. For the 2023-2026 seasons, the league sold its media rights for US$6.4 billion, valuing each match at $13.4 million (Sportstar, 2023). The tournament has also broken viewership records, with the 2023 final becoming the most streamed live event on the internet, attracting 32 million viewers (JioCinema, 2023).

Innovation has been a hallmark of the IPL. The league has introduced several concepts that have added new dimensions to the game, such as strategic timeouts, the Decision Review System (DRS) for reviewing umpire decisions, and an "impact player" substitution rule. These changes have not only enhanced the tactical aspects of the game but also increased its entertainment value for spectators (IPL Governing Council, 2024).

The IPL has also played a crucial role in developing cricket talent. It provides a platform for young Indian players to showcase their skills alongside international stars, contributing to the growth of Indian cricket at both domestic and international levels (Shah, 2023).

The Indian Premier League has transformed cricket from a traditional sport into a global entertainment spectacle. Its blend of star power, economic impact, and innovative gameplay has cemented its position as a powerhouse in the world of sports, influencing the way cricket is played and consumed around the globe.

The Indian Premier League (IPL) currently features ten franchise teams, each representing different cities or states across India (Board of Control for Cricket in India, 2023):

A logo of a lion

Description automatically generated

Figure 2. Chennai Super Kings logo

Source: iplt20

Chennai Super Kings (CSK): Known for their consistency and led by the iconic MS Dhoni, CSK has won four IPL titles (ESPN Cricinfo, 2023).

A logo of a sports team

Description automatically generated

Delhi Capitals (DC): Formerly Delhi Daredevils, this team rebranded in 2018 and has been building a strong young core of Indian talent (Delhi Capitals, 2023).

A logo of a sports team

Description automatically generated

Gujarat Titans (GT): One of the newest additions to the IPL, joining in 2022, they made an immediate impact by winning the title in their debut season (Gujarat Titans, 2023).

A purple and yellow logo

Description automatically generated

Kolkata Knight Riders (KKR): Co-owned by Bollywood star Shah Rukh Khan, KKR has won two IPL titles and has a massive fan following (Kolkata Knight Riders, 2023).

A logo of a sports team

Description automatically generated

Lucknow Super Giants (LSG): Another new franchise that joined in 2022, they've quickly established themselves as strong contenders (Lucknow Super Giants, 2023).

A logo of a sports team

Description automatically generated

Mumbai Indians (MI): The most successful IPL team with five titles, MI is known for its star-studded lineup and ability to nurture young talent (Mumbai Indians, 2023).

A red shield with a lion head

Description automatically generated

Punjab Kings (PBKS): Formerly Kings XI Punjab, this team rebranded in 2021 and is still seeking its first IPL title (Punjab Kings, 2023).

A blue circle with gold text and lions

Description automatically generated

Rajasthan Royals (RR): The inaugural IPL champions in 2008, RR is known for its ability to unearth and develop lesser-known players (Rajasthan Royals, 2023).

A gold lion with a crown and text

Description automatically generated

Royal Challengers Bangalore (RCB): Despite boasting some of cricket's biggest names, RCB is still chasing their first IPL title (Royal Challengers Bangalore, 2023).

A logo with a bird and sun

Description automatically generated

Sunrisers Hyderabad (SRH): Known for their strong bowling attacks, SRH won the title in 2016 and has consistently been a playoff contender (Sunrisers Hyderabad, 2023).

Each team can have a maximum of 25 players in their squad, with no more than eight overseas players. The playing eleven for each match can include up to four overseas players, ensuring a balance of international stars and domestic talent (IPL Governing Council, 2024).

The IPL's team structure, with its mix of international stars and Indian talent, creates a unique and exciting cricketing spectacle that has captured the imagination of fans worldwide (Shah, 2023).

## Team performances over the years

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Team\_Name | Played | Won | Lost | N/R | Titles | Finalists | Playoff |
| MI | 261 | 144 | 117 | 0 | 5 | 6 | 11 |
| RCB | 256 | 123 | 129 | 4 | 0 | 3 | 9 |
| KKR | 252 | 131 | 120 | 1 | 3 | 4 | 7 |
| DC | 252 | 115 | 135 | 2 | 0 | 1 | 6 |
| PK | 246 | 112 | 134 | 0 | 0 | 1 | 2 |
| CSK | 239 | 138 | 99 | 2 | 5 | 10 | 13 |
| RR | 222 | 112 | 107 | 3 | 1 | 2 | 5 |
| SRH | 182 | 88 | 94 | 0 | 1 | 3 | 6 |
| GT | 45 | 28 | 17 | 0 | 1 | 2 | 2 |
| LSG | 44 | 24 | 19 | 1 | 0 | 0 | 2 |

Table 1: Teams Statistics

This table provides a comprehensive overview of the performance of Indian Premier League (IPL) teams since the league's inception in 2008. Let's break down the information and analyze the data for each team:

Mumbai Indians (MI):

Mumbai Indians have played the most matches (261) and won the most games (144) in IPL. Their success is evident in their 5 IPL titles, the highest among all teams. They've reached the finals 6 times and made it to the playoffs 11 times, showcasing their consistency (Board of Control for Cricket in India, 2023).

Royal Challengers Bangalore (RCB):

Despite playing 256 matches and winning 123, RCB has never won an IPL title. They've reached the finals 3 times and made the playoffs 9 times. Their inability to convert playoff appearances into titles has been a point of discussion among cricket analysts (ESPN Cricinfo, 2023).

Kolkata Knight Riders (KKR):

KKR has played 252 matches, winning 131. They've clinched 3 IPL titles and reached the finals 4 times. With 7 playoff appearances, they've shown consistency in reaching the latter stages of the tournament (Kolkata Knight Riders, 2023).

Delhi Capitals (DC):

Formerly Delhi Daredevils, DC has played 252 matches but won only 115. They've never won an IPL title and have reached the finals only once. With 6 playoff appearances, they've struggled to make a significant impact in the tournament's history (Delhi Capitals, 2023).

Punjab Kings (PK):

PK has played 246 matches, winning 112. They've never won an IPL title and have reached the finals only once. With just 2 playoff appearances, they've been one of the less successful teams in the IPL (Punjab Kings, 2023).

Chennai Super Kings (CSK):

Despite playing fewer matches (239) than some other teams, CSK has been incredibly successful. They've won 138 matches and 5 IPL titles, equaling MI's record. With 10 final appearances and 13 playoff qualifications, they're considered one of the most consistent teams in IPL history (Chennai Super Kings, 2023).

Rajasthan Royals (RR):

RR has played 222 matches, winning 112. They won the inaugural IPL in 2008 but haven't replicated that success since. With 2 final appearances and 5 playoff qualifications, they've had mixed fortunes in the tournament (Rajasthan Royals, 2023).

Sunrisers Hyderabad (SRH):

SRH entered the IPL later than the original teams but has made a significant impact. They've played 182 matches, winning 88. They've won 1 IPL title and reached the finals 3 times, with 6 playoff appearances (Sunrisers Hyderabad, 2023).

Gujarat Titans (GT):

As one of the newest teams, GT has played only 45 matches but has already won 28 of them. They won the IPL in their debut season in 2022 and reached the finals again in 2023, showing immediate success (Gujarat Titans, 2023).

Lucknow Super Giants (LSG):

Another new entrant, LSG, has played 44 matches and won 24. While they haven't reached a final yet, they've made it to the playoffs in both their seasons, indicating a strong start to their IPL journey (Lucknow Super Giants, 2023).

This table highlights the varying degrees of success and consistency among IPL teams. While some teams like MI and CSK have dominated with multiple titles and consistent playoff appearances, others like RCB and PK have struggled to convert their opportunities into championships. The newer teams, GT and LSG, have shown promise in their short IPL careers, adding excitement to the league's competitive landscape (Shah, 2023).

## Research Objective and Need for Study:

A major focus of this study is the upcoming mega auction for the 2025 IPL season. This auction will result in most players being released, with teams allowed to retain only four players, including a maximum of two foreign players. This significant event provides an opportunity to analyze and optimize squad-building strategies for both **Kolkata Knight Riders (KKR) and Delhi Capitals (DC)**.

Among the 10 teams in the Indian Premier League (IPL), this study aims to analyse, predict, and optimize the squads for two specific teams: Kolkata Knight Riders (KKR) and Delhi Capitals (DC). KKR, the 2024 IPL champions, boasts one of the strongest squads in the league. In contrast, DC possesses a power-packed young squad but has struggled to utilize their potential effectively. The research will focus on the following aspects:

1. Squad Analysis: Examine the composition of both KKR and DC squads, identifying key strengths and weaknesses based on their performances.
2. Quantitative analysis: Use statistical methods to compare player performances, team strategies, and match outcomes for both KKR and DC.
3. Performance Prediction: Develop models to forecast player and team performance based on historical data and current squad dynamics.
4. Squad Optimization: Propose strategies for both teams to maximize their squad potential, with a particular emphasis on helping DC better utilize their young talent.
5. Evaluate the current squads of KKR and DC to identify potential retention candidates.
6. Analyse the impact of retaining only four players on team dynamics and performance.
7. Success Factors: Investigate the elements that contributed to KKR's championship win in 2024, including team balance, leadership, and player utilization.

By conducting this research, the aim is to provide valuable insights into effective squad building, talent utilization, and performance optimization in the highly competitive environment of the IPL. The findings could offer strategic guidance not only for KKR and DC but also for other T20 cricket franchises globally.

## Scopes and Limitations

1. Player performance prediction: Develop models to predict player performance based on historical data, considering factors like batting and bowling statistics .
2. Team efficiency analysis: Evaluate the efficiency of teams using techniques like Data Envelopment Analysis (DEA) and Structural Equation Modeling (SEM).
3. Strategic player retention: Analyze strategies for the upcoming mega auction, focusing on optimal player retention decisions for KKR and DC .
4. Impact of external factors: Examine how factors like weather, match location, and stadium conditions affect player and team performance.
5. Comparative analysis: Compare KKR and DC's squad building and utilization strategies with other successful IPL teams.

Limitations of the study:

1. Limited data for new players: Relying solely on IPL data may limit the evaluation of new or emerging players who haven't played in the league before.
2. Complexity of player valuation: Accurately predicting player auction values and performance can be challenging due to multiple influencing factors .
3. Changing league dynamics: The study's findings may be affected by evolving league rules, team strategies, and player availability.
4. External factors: Unforeseen circumstances like injuries, player form, or off-field issues can impact team performance and are difficult to account for in models.
5. Limited scope: Focusing on only two teams (KKR and DC) may limit the generalizability of findings to other IPL teams or T20 leagues.
6. Time constraints: The dynamic nature of T20 cricket and frequent player transfers may make long-term predictions challenging.
7. Multi-objective optimization: It is difficult to formulate team selection as a multi-objective optimization problem, while considering budget constraints .

These scopes and limitations can help frame the research objectives and methodology for analysing and optimizing the squads of KKR and Delhi Capitals in the context of the upcoming IPL mega auction.

# Significance of the Study

1. Strategic Team Building: This study will provide insights into effective squad building strategies for IPL teams, particularly in the context of the upcoming mega auction for the 2025 season. It will help teams like KKR and DC optimize their player retention and acquisition decisions (Indian Express, 2024).
2. Performance Optimization: By analyzing the factors contributing to KKR's success and DC's underperformance, the study can offer valuable insights into how teams can better utilize their player resources, especially young talent (Ishi et al., 2022).
3. Quantitative Sports Analytics: The research contributes to the growing field of quantitative sports analytics in cricket, which has become increasingly important for team management and strategy development (Jana et al., 2021).
4. Player Improvement: The insights gained from this study can help improve individual player performance by identifying key areas for development based on data-driven analysis (Techiexpert, 2024).
5. Global Cricket Applications: The findings could be valuable not only for IPL teams but also for national cricket boards such as the England and Wales Cricket Board (ECB), Cricket Australia, and New Zealand Cricket, helping them in player selection and team strategy for international competitions (Kalgotra et al., 2014).
6. Predictive Modeling: This research will contribute to the development of more accurate predictive models for player and team performance in T20 cricket, drawing inspiration from advanced analytics techniques used in football. These models can leverage machine learning algorithms and big data analysis, similar to those employed in predicting football match outcomes and player performance (Hubáček et al., 2019; Berrar et al., 2019). Such approaches can be valuable for team management, fantasy cricket enthusiasts, and sports analytics professionals, potentially improving decision-making processes in player selection and strategy formulation.
7. Comparative Analysis: By conducting an in-depth comparison of Kolkata Knight Riders' championship-winning strategies in 2024 and Delhi Capitals' approach to team building, this study will provide valuable insights into the efficacy of different management philosophies and their impact on team performance in the IPL (ESPNcricinfo, 2024). This analysis will highlight how KKR's meticulously designed squad, enabling aggressive batting without compromising depth, contrasts with DC's focus on nurturing young talent, offering a comprehensive perspective on successful team construction in T20 cricket.

## Structure of the research

The goal of this study is to predict the optimal squad composition for KKR and DC in the upcoming Indian Premier League (IPL) season. The research methodology encompasses several key steps:

1. Data Extraction: Gathering comprehensive player statistics and performance data from various reliable sources.
2. Data Cleaning: Preprocessing the collected data to ensure accuracy, consistency, and relevance for analysis.
3. Descriptive Analytics: Conducting a thorough exploratory data analysis to understand the underlying patterns and trends in player performances.
4. Data Visualization: Creating insightful visual representations of the data to facilitate easier interpretation and identification of key insights.
5. Feature Engineering: Developing new variables or transforming existing ones to enhance the predictive power of the models.
6. Player Score Prediction: Utilizing advanced statistical and machine learning techniques to forecast individual player performances based on historical data and relevant factors.
7. Hyperparameter Tuning: Optimizing the predictive models through rigorous hyperparameter adjustment to improve accuracy and reliability.
8. Squad Optimization: Employing the refined predictive models to determine the most effective squad compositions for KKR and DC, considering various constraints.
9. Stakeholder Recommendations: Formulating data-driven, actionable recommendations for team management, coaches, and other relevant stakeholders to inform their decision-making processes in player selection and team strategy.

This comprehensive approach aims to leverage advanced analytics to provide valuable insights and strategic advantages in the highly competitive landscape of the IPL.

# Literature Review

The Indian Premier League (IPL) has not only revolutionized cricket but has also become a fertile ground for sports analytics since its inception in 2008. As the league has grown in stature, so has the sophistication of the analytical approaches used to understand and predict its dynamics.

The Economic Catalyst

(Kadapa,2013) highlighted the IPL's massive economic footprint, underscoring the financial imperative driving the adoption of advanced analytics. With billions at stake, teams and stakeholders are increasingly turning to data-driven approaches to gain a competitive edge.

Evolution of Analytical Approaches

The journey of IPL analytics has been one of continuous refinement. (Shah et al, 2016) laid important groundwork with their comprehensive analysis of IPL data from 2008 to 2015. Their work demonstrated the potential of machine learning in decoding the complexities of T20 cricket, setting the stage for more advanced studies. Building on this foundation, Prakash et al. (2019) developed a nuanced player ranking system using machine learning algorithms. Their model's success in predicting player rankings with high accuracy showcased the power of analytics in informing team selection strategies.

The Human Element in Data

While numbers are at the heart of analytics, recent research has emphasized the importance of translating data into actionable insights. (Ishi et al,2022) took a significant step in this direction by using machine learning for player classification. Their work helps bridge the gap between raw data and on-field strategy, providing coaches and managers with a more intuitive understanding of player capabilities.

Predictive Power and Its Limitations

The holy grail of sports analytics is accurate prediction, and IPL research has made significant strides in this area. (Amala Kaviya et al,2020) achieved an impressive 81% accuracy in predicting match outcomes. However, as any cricket fan knows, the game's unpredictability is part of its charm. These models, while powerful, serve as tools to inform decision-making rather than crystal balls.

Transparency in Analytics

Recognizing the need for interpretable results, (Bajaj,2023) explored the use of Explainable AI techniques. This approach not only predicts performance but also elucidates the factors influencing these predictions, making the insights more accessible and actionable for non-technical stakeholders.

Visualizing Success

In the fast-paced world of T20 cricket, the ability to quickly grasp complex information is crucial. (Rodrigues et al,2019) addressed this need by focusing on data visualization techniques. Their work highlights how effective visual representation can transform raw data into strategic insights, accessible to everyone from analysts to players.

The Road Ahead

As we look to the future of IPL analytics, several exciting avenues emerge:

1. Real-time analytics during matches could revolutionize in-game decision-making.
2. Integration of non-traditional data sources, such as social media sentiment and player biometrics, may provide a more holistic view of performance.
3. More sophisticated player valuation models could transform auction strategies.
4. The application of deep learning to video analysis promises to unlock new insights into player techniques and strategies.

The field of IPL analytics is not just about numbers; it's about enhancing the beautiful game of cricket. As analytics continue to evolve, they promise to enrich our understanding and enjoyment of the sport, providing fans, players, and managers alike with new perspectives on the game we love.

## Adoption of Machine Learning on Sports

The adoption of Machine Learning (ML) in sports has seen significant growth in recent years, revolutionizing various aspects of athletic performance, strategy, and management.

Performance Analysis and Prediction:

ML has been extensively applied to analyse and predict athletic performance. (Ofoghi et al, 2013) demonstrated the use of ML algorithms to predict medal-winning performances in sprint kayaking, achieving an accuracy of 80%. Similarly, (Bunker and Thabtah, 2019) reviewed ML applications in predicting outcomes of various sports, finding that ensemble methods often outperform individual algorithms in accuracy.

Injury Prediction and Prevention:

A critical area where ML has shown promise is in injury prediction and prevention. (Rossi et al, 2018) developed a ML model to predict injuries in soccer players, achieving an accuracy of 80% in identifying high-risk athletes. Building on this, (Rommers et al, 2020) used ML techniques to predict injuries in youth soccer players, demonstrating the potential of these methods in protecting young athletes.

Tactical Analysis:

ML has transformed tactical analysis in team sports. (Memmert and Raabe, 2018) explored how ML algorithms can analyse complex patterns in soccer matches, providing coaches with insights that were previously unattainable through traditional methods. In basketball, (Cervone et al,2016) used ML to evaluate decision-making in real-time, offering a new perspective on player effectiveness beyond traditional statistics.

Player Recruitment and Scouting:

The application of ML in talent identification and recruitment has gained traction. (McHale et al,2012) developed a ML model to assess player performance in soccer, which has implications for scouting and transfer decisions. More recently, (Liu et al, 2020) used deep learning techniques to analyse player movements in basketball, providing a data-driven approach to talent evaluation.

Fan Engagement and Business Operations:

ML has also found applications in enhancing fan engagement and optimizing business operations in sports. (Fried and Mumcu, 2016) explored how ML can be used to personalize fan experiences and improve marketing strategies in professional sports. In ticket pricing, (Kemper and Breuer, 2016) demonstrated how ML algorithms can optimize dynamic pricing strategies, potentially increasing revenue for sports organizations.

Challenges and Ethical Considerations:

Despite its potential, the adoption of ML in sports faces several challenges. (Caya and Bourdon, 2016) highlighted issues of data quality and interpretation in sports analytics, emphasizing the need for domain expertise in developing ML models. Ethical considerations have also come to the forefront, with (Loland, 2018) discussing the implications of ML on fairness and integrity in sports.

Future Directions:

The future of ML in sports looks promising, with several emerging areas of research. Wearable technology and IoT devices are expected to provide more granular data for ML models, as explored by (Seshadri et al., 2019) in their work on real-time performance tracking. Additionally, the integration of computer vision with ML, as demonstrated by (Thomas et al, 2017) in their analysis of tennis player movements, opens new avenues for automated performance analysis.

## Adoption of Machine Learning on Cricket

The adoption of Machine Learning (ML) in cricket analytics has gained significant traction in recent years, with researchers from Europe and the USA contributing to this field. Here's a literature review focusing on key aspects:

1. Match Outcome Prediction:  
   Researchers have applied ML techniques to predict cricket match outcomes. A study from the UK focused on English County twenty-over cricket matches, investigating the degree to which it's possible to predict match outcomes using ML algorithms. This research demonstrates the growing interest in applying advanced analytics to cricket.
2. Performance Analysis:  
   ML has been used to analyse player and team performance in cricket. While not specifically focused on cricket, McHale et al. (2012) developed ML models to assess player performance in soccer, which has implications for similar applications in cricket, particularly for scouting and team selection strategies.
3. Data-Driven Decision Making:  
   The adoption of ML in cricket analytics aligns with broader trends in sports analytics. Beal et al. (2019) conducted a comprehensive survey on artificial intelligence for team sports, which included cricket. They noted that ML methods have been applied to various aspects of sports, including tactical analysis and performance prediction.
4. Challenges and Limitations:  
   While ML shows promise in cricket analytics, researchers have noted challenges such as the need for high-quality data and the complexity of cricket's rules and playing conditions, which can affect model accuracy (Beal et al., 2019). The dynamic nature of cricket, with its multiple formats and varying conditions, presents unique challenges for ML applications.
5. Future Directions:  
   Ongoing research is focusing on improving the accuracy of predictive models and expanding the range of applications for ML in cricket. This includes real-time analysis during matches and more sophisticated player valuation models (Beal et al., 2019). The potential for ML to enhance decision-making in areas such as team selection, strategy formulation, and player development is significant.
6. Interdisciplinary Approach:  
   The literature suggests that successful adoption of ML in cricket requires an interdisciplinary approach, combining expertise in data science, sports science, and domain-specific knowledge of cricket (Beal et al., 2019).

The adoption of ML in cricket analytics is growing, the field is still evolving. Researchers continue to refine methodologies and explore new applications to enhance the understanding and analysis of the sport. The potential for ML to transform various aspects of cricket, from player performance analysis to strategic decision-making, is significant, but challenges remain in terms of data quality, model interpretability, and practical implementation.

## Summary of Literature Review:

1. Match Outcome Prediction:  
   Researchers have applied ML techniques to predict cricket match outcomes. A study focused on English County twenty-over cricket matches investigated the degree to which it's possible to predict match outcomes using ML algorithms. This demonstrates the growing interest in applying advanced analytics to cricket.
2. Performance Analysis:  
   ML has been used to analyse player and team performance in cricket. While not specifically focused on cricket, studies like McHale et al. (2012) developed ML models to assess player performance in soccer, which has implications for similar applications in cricket, particularly for scouting and team selection strategies.
3. Data-Driven Decision Making:  
   The adoption of ML in cricket analytics aligns with broader trends in sports analytics. Beal et al. (2019) conducted a comprehensive survey on artificial intelligence for team sports, which included cricket. They noted that ML methods have been applied to various aspects of sports, including tactical analysis and performance prediction.
4. Challenges and Limitations:  
   Researchers have noted challenges such as the need for high-quality data and the complexity of cricket's rules and playing conditions, which can affect model accuracy (Beal et al., 2019). The dynamic nature of cricket, with its multiple formats and varying conditions, presents unique challenges for ML applications.
5. Future Directions:  
   Ongoing research is focusing on improving the accuracy of predictive models and expanding the range of applications for ML in cricket. This includes real-time analysis during matches and more sophisticated player valuation models.
6. Interdisciplinary Approach:  
   Successful adoption of ML in cricket requires an interdisciplinary approach, combining expertise in data science, sports science, and domain-specific knowledge of cricket (Beal et al., 2019).
7. Emerging Technologies:  
   The European Cricket Network has partnered with Full track AI, an advanced machine learning and artificial intelligence service, to provide ball tracking graphics, pitch maps, speeds, and other key data points using mobile phone technology (Emerging Cricket, 2023).

In conclusion, while the adoption of ML in cricket analytics is growing, the field is still evolving. Researchers continue to refine methodologies and explore new applications to enhance the understanding and analysis of the sport. The potential for ML to transform various aspects of cricket, from player performance analysis to strategic decision-making, is significant, but challenges remain in terms of data quality, model interpretability, and practical implementation.

# Research Methodology

The primary objective of this research is to leverage machine learning techniques to predict and optimize the squad compositions for two Indian Premier League (IPL) teams: Kolkata Knight Riders (KKR) and Delhi Capitals (DC). This study aims to utilize a comprehensive approach that incorporates various statistical methods, algorithms, and predictive models to analyse player performance data. By employing advanced data mining techniques, feature engineering, and machine learning algorithms such as decision trees, random forests, and support vector machines (Regression), the research seeks to identify the most effective player combinations for each team. The goal is to provide data-driven insights that can inform team management decisions, particularly in the context of player selection for upcoming seasons and auctions.

## Dataset and Approach Overview:

The dataset is taken from trusted websites such as Cricsheet and Howstat. The dataset appears genuine, and cross-verification has been performed to check the legitimacy of the data. The dataset contains a ball-by-ball record of IPL matches, providing detailed information about each delivery, including match details, player information, runs scored, extras, and dismissals.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| match\_id | Unique identifier for each match |
| season | The IPL season year |
| start\_date | The date the match started |
| venue | The location where the match was played |
| innings | The innings number (1st or 2nd) |
| ball | The ball number within the over |
| batting\_team | The team currently batting |
| bowling\_team | The team currently bowling |
| striker | The batsman facing the current ball |
| non\_striker | The batsman at the other end |
| extras | Total extra runs scored on this ball |
| wides | Number of wide balls |
| noballs | Number of no balls |
| byes | Number of byes |
| legbyes | Number of leg byes |
| penalty | Any penalty runs awarded |
| wicket\_type | Type of dismissal if a wicket fell |
| player\_dismissed | Name of the player dismissed (if applicable) |
| other\_wicket\_type | Secondary wicket type (if applicable) |

This table represents a comprehensive dataset used for analyzing cricket matches, specifically focusing on the Indian Premier League (IPL). Each row in the table corresponds to a specific delivery (ball) in a match, providing detailed information about the events occurring during that delivery.

1. **match\_id**: This column contains a unique identifier for each match, allowing for easy referencing and data management. It helps distinguish between different matches in the dataset.
2. **season**: This indicates the specific IPL season during which the match took place. It typically refers to the year of the tournament, providing context for the data.
3. **start\_date**: This column records the date on which the match commenced. It is essential for temporal analysis, allowing researchers to study trends over different seasons or specific time periods.
4. **venue**: This specifies the location where the match was held. Knowing the venue is important for analyzing home advantage, pitch conditions, and crowd influence on the game.
5. **innings**: This indicates whether the data pertains to the first or second innings of the match. In cricket, each team bats for one or two innings, and this column helps differentiate between them.
6. **ball**: This column records the specific ball number within the over. It provides granular detail about the match, allowing for in-depth analysis of individual deliveries.
7. **batting\_team**: This specifies the team that is currently batting during the delivery. It is crucial for understanding team performance and strategies.
8. **bowling\_team**: This indicates the team that is currently bowling. This information is essential for analyzing bowling strategies and effectiveness.
9. **striker**: This column names the batsman facing the current delivery. It is important for analyzing individual player performance and contributions.
10. **non\_striker**: This indicates the batsman at the other end of the pitch who is not facing the current delivery. It provides context for partnerships and running between the wickets.
11. **extras**: This column records the total extra runs scored on that delivery, which can include wides, no-balls, and other extras. It is important for assessing the impact of extras on the match outcome.
12. **wides**: This specifies the number of wide balls bowled during that delivery. Wides contribute to the extras and can affect the match's flow and scoring.
13. **noballs**: This indicates the number of no-balls bowled on that delivery. No-balls also contribute to extras and can lead to free hits, impacting scoring opportunities.
14. **byes**: This column records the number of byes scored on that delivery, which occur when the ball passes the wicketkeeper without touching the bat or body of the batsman.
15. **legbyes**: This specifies the number of leg byes scored, which occur when the ball hits the batsman's body (excluding the hand) and runs are taken.
16. **penalty**: This column records any penalty runs awarded to the batting or bowling team, which can occur due to infractions by the fielding team.
17. **wicket\_type**: This indicates the type of dismissal if a wicket fell on that delivery (e.g., bowled, caught, LBW). It is crucial for analyzing how wickets are taken.
18. **player\_dismissed**: This column names the player who was dismissed on that delivery, providing insight into key moments in the match.
19. **other\_wicket\_type**: This specifies any secondary wicket type, if applicable, for cases where multiple dismissals occur in a single delivery (e.g., run out).
20. **other\_player\_dismissed**: This column names any other player who was dismissed on that delivery, providing additional context for significant events.

## Domain Knowledge

 The domain knowledge required in this field:

1. Understanding of Cricket:  
   A deep understanding of cricket is fundamental. This includes:

* Rules of the game
* Various formats (Test, ODI, T20)
* Strategies and tactics
* Historical trends
* Nuances that influence the game

1. Statistical Knowledge:

* Descriptive statistics (mean, median, mode, range, standard deviation, etc.)
* Inferential statistics (regression analysis, correlation analysis, ANOVA, hypothesis testing)
* Understanding of key performance indicators (KPIs) in cricket (batting average, strike rate, economy rate, etc.)

1. Data Types and Sources:

* Player performance data
* Team performance data
* Match data
* Historical data
* Understanding of data sources like Cricsheet and Howstat

1. Analytical Techniques:

* Time series analysis
* Clustering analysis
* Machine learning algorithms
* Predictive modelling

1. Cricket-Specific Analytics:

* Understanding of Duckworth-Lewis (D/L) method and its applications
* Knowledge of player valuation models
* Understanding of factors affecting performance (pitch conditions, player skills, opposition strengths/weaknesses)

1. Strategic Applications:

* How to use data for team selection
* Optimizing batting orders and bowling strategies
* Field placement strategies based on data
* In-game decision making using real-time analytics

1. Broader Sports Analytics Concepts:

* Familiarity with analytics approaches from other sports (e.g., Sabermetrics in baseball)
* Understanding of how analytics can be applied to both performance analysis and fan engagement

This domain knowledge forms the foundation for effective cricket analytics, allowing analysts to interpret data in the context of the sport and provide valuable insights for teams, players, and stakeholders.

# Quantitative and Predictive analysis

## 1. Win Ratio Analysis of Teams

### Need for Analysis

The goal is to analyse the performance of IPL teams, focusing particularly on their win ratios, to determine which teams have been the most and least successful over the history of the tournament. This analysis will help identify trends, strengths, and weaknesses among the teams, providing insights into factors that contribute to long-term success in the IPL.

### Objective

1. Calculate and compare the win ratios of all IPL teams.
2. Predict the Win ratio of all teams

### Data overview

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Team\_Name | Played | Won | Lost | N/R | Titles | Finalists | Playoff |
| MI | 261 | 144 | 117 | 0 | 5 | 6 | 11 |
| RCB | 256 | 123 | 129 | 4 | 0 | 3 | 9 |
| KKR | 252 | 131 | 120 | 1 | 3 | 4 | 7 |
| DC | 252 | 115 | 135 | 2 | 0 | 1 | 6 |
| PK | 246 | 112 | 134 | 0 | 0 | 1 | 2 |
| CSK | 239 | 138 | 99 | 2 | 5 | 10 | 13 |
| RR | 222 | 112 | 107 | 3 | 1 | 2 | 5 |
| SRH | 182 | 88 | 94 | 0 | 1 | 3 | 6 |
| GT | 45 | 28 | 17 | 0 | 1 | 2 | 2 |
| LSG | 44 | 24 | 19 | 1 | 0 | 0 | 2 |

### Quantitative Analysis

#### Calculate the win ratio:

The win ratio is a crucial metric in sports analysis, particularly in leagues like the IPL, as it provides a clear and quantifiable measure of a team's success relative to its total games played. By calculating the win ratio, stakeholders including coaches, players, analysts, and fans can assess performance over time, identify trends, and make informed decisions.

A high win ratio indicates consistent success and competitiveness, while a low ratio may highlight areas needing improvement. Additionally, win ratios facilitate comparisons between teams, regardless of the number of matches played, allowing for a more equitable evaluation of performance.

|  |  |
| --- | --- |
| Team Name | Win Ratio |
| MI | 55.17241379 |
| RCB | 48.046875 |
| KKR | 51.98412698 |
| DC | 45.63492063 |
| PK | 45.52845528 |
| CSK | 57.74058577 |
| RR | 50.45045045 |
| SRH | 48.35164835 |
| GT | 62.22222222 |
| LSG | 54.54545455 |

#### Calculate the lost ratio:

The loss ratio is a key performance indicator that measures the proportion of games lost relative to the total number of games played. It is calculated using the formula:

#### Win Loss Ratio

The win-loss ratio is a critical metric used to evaluate performance in various competitive contexts, including sports and sales. It is calculated using the formula:

*Win Loss Ratio=Number of Losses / Number of Wins​*

The win-loss ratio difference is a crucial metric in sports analysis for several reasons:

1. Performance indicator: It provides a clear picture of a team's overall performance, showing how much they're winning compared to losing.
2. Competitive edge: A positive difference indicates a team is winning more than losing, suggesting a competitive advantage.
3. Trend analysis: Tracking this metric over time can reveal improvements or declines in team performance.

#### Corelation Analysis:

A diagram of a heat map

Description automatically generated

Strong Positive Correlations

1. Played vs. Won (0.974) and Lost (0.976):
   * As expected, teams that have played more matches have both won and lost more games.
2. Win\_Ratio vs. WR\_Difference (0.997):
   * Win ratio is strongly positively correlated with the win-loss ratio difference, indicating that teams with higher win ratios naturally have a larger gap between their wins and losses.
3. Titles vs. Finalists (0.899):
   * Teams that win more titles tend to reach the finals more often.
4. Finalists vs. Playoff (0.883):
   * Teams that reach the finals more frequently also tend to make more playoff appearances.

Strong Negative Correlations

1. Win\_Ratio vs. lost\_Ratio (-0.989):
   * As expected, there's a strong inverse relationship between win ratio and loss ratio.
2. lost\_Ratio vs. WR\_Difference (-0.997):
   * Teams with higher loss ratios have lower win-loss ratio differences.

Moderate Correlations

1. Won vs. Playoff (0.755):
   * Teams with more wins tend to make more playoff appearances.
2. Titles vs. Playoff (0.767):
   * Teams that win more titles also tend to have more playoff appearances.
3. Lost vs. lost\_Ratio (0.754):
   * More losses correlate with a higher loss ratio, as expected.

Weak or No Correlations

1. N/R (No Result) has weak correlations with most other metrics, suggesting that matches with no result don't significantly impact overall performance metrics.
2. Win\_Ratio and Playoff (0.110):
   * Interestingly, there's only a weak positive correlation between win ratio and playoff appearances, suggesting that consistent regular-season performance doesn't always translate to playoff success.

### Linear Regression Model:

It appears to be related to IPL team performance statistics; linear regression is an appropriate choice for several reasons:

1. Continuous Dependent Variable:  
   Your dependent variable, Win\_Ratio, is a continuous variable, which is suitable for linear regression analysis.
2. Multiple Independent Variables:  
   Your dataset includes multiple potential predictors (e.g., Played, Won, Lost, N/R, lost\_Ratio, Titles, Finalists, Playoff), making multiple linear regression an appropriate choice.
3. Relationship Exploration:  
   Linear regression can help identify which factors have the strongest influence on a team's win ratio, providing valuable insights into team performance.
4. Performance Prediction:  
   The model can be used to predict a team's expected win ratio based on other performance metrics, which could be useful for team management and strategy planning.
5. Quantifiable Impact:  
   Linear regression provides coefficients that quantify the impact of each independent variable on the win ratio, allowing for a clear understanding of each factor's importance.
6. Model Interpretability:  
   In sports analytics, it's often crucial to have models that can be easily interpreted by coaches, managers, and other stakeholders. Linear regression provides this interpretability.
7. Baseline Model:  
   Even if more complex models might be explored later, linear regression serves as an excellent baseline model to compare against more sophisticated approaches.
8. Assumption Testing:  
   The dataset allows for testing various assumptions of linear regression (like linearity, homoscedasticity, and multicollinearity), which can provide insights into the data's structure.
9. Small Dataset Handling:  
   With a relatively small dataset (10 observations), linear regression can still provide reliable results, whereas more complex models might overfit.
10. Performance Metrics:  
    The high R-squared value (0.9969) suggests that linear regression is capturing a significant amount of variance in the win ratio, indicating a good fit for this data.

### Prediction Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **Team** | **Actual Win%** | **Predicted Win%** | **Difference** |
| MI | 55.17 | 55.21 | +0.04 |
| RCB | 48.05 | 48.10 | +0.05 |
| KKR | 51.98 | 51.95 | -0.03 |
| DC | 45.63 | 45.68 | +0.05 |
| PK | 45.53 | 45.87 | +0.34 |
| CSK | 57.74 | 57.95 | +0.21 |
| RR | 50.45 | 50.07 | -0.38 |
| SRH | 48.35 | 47.89 | -0.46 |
| GT | 62.22 | 61.94 | -0.28 |
| LSG | 54.55 | 55.01 | +0.46 |

Analysis of Predictions

1. **Accuracy**: The model's predictions are remarkably close to the actual values, with most differences being less than 0.5 percentage points.
2. **Consistency**: The model performs well across different teams, showing no significant bias towards over or under-prediction for specific teams.
3. **Best Predictions**:
   * **Mumbai Indians**: Only a 0.04 difference.
   * **Kolkata Knight Riders**: Only a 0.03 difference.
4. **Largest Discrepancies**:
   * **Sunrisers Hyderabad**: Underpredicted by 0.46.
   * **Lucknow Super Giants**: Overpredicted by 0.46.
5. **Overall Trend**: There is a slight tendency to overpredict for lower-performing teams and underpredict for higher-performing teams, but the differences are minimal.

Model Performance

1. **Mean Absolute Error (MAE)**: Approximately 0.23.
2. **Root Mean Squared Error (RMSE)**: 0.2874.

These error metrics confirm the high accuracy of the predictions, with an average deviation of less than 0.3 percentage points.

Implications

1. **Predictive Power**: The model demonstrates strong predictive capabilities for IPL team performance.
2. **Team Analysis**: Predictions closely align with actual performance, suggesting the model has effectively captured key factors influencing team success.
3. **Practical Use**: This model could be valuable for team strategy, player selection, and performance analysis in the IPL context.

Findings of the analysis:

A graph with red and blue lines

Description automatically generated

Trends in the performance of IPL teams:

1. Top Performers:  
   Chennai Super Kings (CSK) and Gujarat Titans emerge as the top performers, with predicted win percentages of 57.95% and 61.94% respectively. This suggests these teams have strong overall player statistics and team dynamics that contribute to their success.
2. Mid-Range Performers:  
   Teams like Mumbai Indians (55.21%), Lucknow Super Giants (55.01%), and Kolkata Knight Riders (51.95%) fall into the mid-range of performance. Their predicted win percentages suggest consistent but not dominant performance.
3. Lower Performers:  
   Teams such as Punjab Kings (45.87%), Delhi Capitals (45.68%), and Royal Challengers Bengaluru (48.10%) have lower predicted win percentages, indicating potential areas for improvement in their team composition or strategy.
4. Slight Overestimation for Lower Performers:  
   There's a subtle trend of slight overprediction for teams with lower actual win percentages. For example, Punjab Kings and Delhi Capitals have slightly higher predicted values compared to their actual performance.
5. Slight Underestimation for Top Performers:  
   Conversely, there's a minor trend of underprediction for top-performing teams. Gujarat Titans, for instance, have a slightly lower predicted value (61.94%) compared to their actual performance (62.22%).
6. Consistency in Prediction:  
   The model shows remarkable consistency across different teams, with predictions closely aligning with actual performance. This suggests the model has effectively captured key factors influencing team success in the IPL.
7. Narrow Performance Range:  
   The predicted win percentages range from about 45% to 62%, indicating a competitive league where even lower-performing teams have a substantial chance of winning matches.

These trends suggest that the prediction model has effectively captured the nuances of team performance in the IPL, reflecting both the strengths of top teams and the areas for improvement for others. The slight over and underpredictions at the extremes might indicate some regression to the mean in the model's predictions, which is common in statistical forecasting.

## 2.Rule-based scoring system combined with normalization and weighted aggregation

### Need for Analysis

The rule-based scoring system for cricket aim to quantify a player's overall performance by assigning points based on various aspects of their game.

1. Holistic Player Assessment:
   * The code combines multiple performance metrics (runs, average, strike rate, wickets, economy rate) to create a comprehensive evaluation of each player.
   * This approach provides a more complete picture of a player's contribution than individual statistics alone.
2. Role-Based Evaluation:
   * By categorizing players as Batsmen, Bowlers, or All-rounders, the analysis acknowledges the different roles within a cricket team.
   * This allows for fair comparisons between players with similar roles and responsibilities.
3. Normalized Comparisons:
   * Normalizing metrics enables fair comparisons across different scales (e.g., comparing runs scored with wickets taken).
   * This is essential for creating a unified scoring system that can be applied across diverse player types.
4. Weighted Performance Metrics:
   * Assigning weights to different metrics (e.g., giving more importance to total runs for batsmen or wickets for bowlers) reflects the relative importance of various aspects of performance.
   * This nuanced approach aligns the analysis with the strategic priorities of T20 cricket.
5. Identifying All-Round Talent:
   * The system's ability to evaluate all-rounders separately recognizes the unique value of players who contribute significantly in both batting and bowling.
6. Ranking Within Categories:
   * Ranking players within their specific roles (batsman, bowler, all-rounder) provides context-specific performance assessments.
   * This is valuable for team selection, strategy formulation, and player development.
7. Data-Driven Decision Making:
   * The analysis provides an objective, data-driven basis for decisions related to team composition, player retention, and strategic planning.
8. Performance Benchmarking:
   * By creating a standardized scoring system, teams can benchmark player performances across seasons or compare players from different teams.
9. Talent Identification:
   * This system can help identify undervalued players or rising talents who might not stand out in traditional statistics but perform well in this comprehensive analysis.
10. Contract and Auction Strategies:
    * For leagues like the IPL, this analysis can inform bidding strategies during player auctions and help in determining player values for contracts.
11. Fan Engagement and Fantasy Sports:
    * Providing a single, comprehensive score for each player enhances fan engagement and can be particularly useful for fantasy cricket leagues.
12. Continuous Performance Monitoring:
    * This type of analysis can be easily updated with new match data, allowing for continuous monitoring of player performance throughout a season or across multiple seasons.

In summary, this analysis meets the need for a sophisticated, multi-dimensional evaluation of cricket players that goes beyond traditional statistics. It provides valuable insights for team management, coaches, and analysts, contributing to more informed decision-making in various aspects of cricket team management and strategy.

### Objective

The primary objective of this analysis is to create a comprehensive, data-driven evaluation system for cricket players in T20 leagues like the IPL. It aims to quantify player performance across multiple dimensions, providing a single, numerical score that reflects a player's overall value to their team.

By combining and normalizing various performance metrics such as runs scored, batting average, strike rate, wickets taken, and economy rate, the analysis offers a balanced assessment of player contributions. It distinguishes between different player roles (batsmen, bowlers, and all-rounders), ensuring fair comparisons within each category while also recognizing the unique value of versatile players.

The scoring system is designed to support strategic decision-making in team management, including player selection, auction strategies, and long-term team planning. It provides an objective basis for ranking players within their roles, helping to identify top performers and undervalued talents who might not stand out in traditional statistics alone.

Furthermore, this analysis creates a standardized method for comparing players across different teams and seasons, offering insights into performance trends and player development. By establishing an adaptable analytical framework, it allows for continuous performance monitoring and can evolve with changing game strategies.

Ultimately, this objective, multi-faceted approach to player evaluation aims to enhance the overall quality of decision-making in cricket team management and provide deeper insights into player performance in the dynamic context of T20 cricket.

### Data Overview

This dataset provides a comprehensive overview of player performance in the Indian Premier League (IPL) from 2021 to 2024, capturing key metrics for both batting and bowling. The dimensions of data are: 300 rows x 8 columns. The data encompasses:

1. Batting Performance:
   * "totalrunsscored": Aggregate runs scored by each player
   * "Total\_batting\_average": Average runs scored per dismissal
   * "batting\_strike\_rate": Runs scored per 100 balls faced
   * "totalballsfaced": Total number of deliveries faced
2. Bowling Performance:
   * "totalwickets": Number of wickets taken
   * "economyrate": Average runs conceded per over
   * "oversbowled\_clean": Total overs bowled

The "striker" column likely identifies individual players. This dataset allows for a multifaceted analysis of player contributions, enabling comparisons between different aspects of the game. It captures both volume (total runs, wickets) and efficiency (average, strike rate, economy) metrics, providing a balanced view of player performance. Inclusion of data over multiple seasons (2021-2024) allows for trend analysis, tracking player development, and assessing consistency over time.

### Quantitative Analysis

#### 1. Rule-Based Categorization

The players are categorized into different roles (batsman, bowler, or all-rounder) based on predefined rules. These rules are simple conditional checks based on the player's performance metrics:

Let R = Total runs scored, W = Total wickets, B = Total balls faced

)

* **Batsman**: More scored more than or equal to 100 runs and taken 2 or fewer wickets.
* **Bowler**: More than 5 wickets and scored 100 or fewer runs.
* **All-rounder**: Taken more than or equal to 3 wickets and scored 100 runs or more.

This categorization helps in determining which metrics are relevant for calculating the player's score.

#### 2.Data Normalization:

Normalization is used to scale different performance metrics to a common range (0 to 1). This ensures that metrics with different units and ranges can be compared and combined meaningfully.

* **Min-Max Normalization**: For metrics where a higher value is better (e.g., runs scored, wickets taken), the formula used is:
* **Inverted Normalization for Economy Rate**: Since a lower economy rate is better, the normalization is inverted:

#### Weight Aggregation

The system uses a weighted sum approach to aggregate multiple normalized performance metrics into a single score. The weights are assigned differently for batsmen, bowlers, and all-rounders to reflect the relative importance of different skills in T20 cricket.

For Batsmen:

For Bowlers:

For All-rounders:

This weighted aggregation allows for:

* + Combining multiple performance aspects into a single, comprehensive score
  + Adjusting the importance of different metrics based on player role
  + Balancing volume (e.g., total runs) with efficiency (e.g., strike rate)

#### Ranking

After calculating the overall scores, players are ranked within their respective categories (Batsman, Bowler, All-rounder, Other). The ranking is done using the 'min' method.

For each player type PT ∈ {Batsman, Bowler, All-rounder, Other}:

Where |•| denotes the cardinality of the set.

* player\_i is the player being ranked
* PT is the set of all players of the same player type (e.g., all batsmen)
* Score(player\_x) is the overall score calculated for player x.
* |{...}| denotes the cardinality (size) of the set

This set comprehension identifies all players (player\_j) within the same player type (PT) whose scores are strictly greater than the score of the player being ranked (player\_i).

This approach is used because:

* It allows for fair comparison within roles, recognizing that different skills are valued for different positions
* It provides a clear hierarchy within each player type
* The 'min' method ensures that players with equal scores receive the same rank, avoiding arbitrary distinctions

#### Overall Score Calculation

The analysis has produced overall scores and rankings for IPL players across different roles (Batsmen, Bowlers, and All-rounders). The scores reflect a comprehensive evaluation of player performance, considering multiple metrics normalized and weighted according to their importance in T20 cricket.

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Player** | **Player Type** | **Overall Score** |
| 1 | YS Chahal | Bowler | 87.70 |
| 2 | CV Varun | Bowler | 73.69 |
| 3 | Mohammed Shami | Bowler | 72.78 |
| 1 | F du Plessis | Batsman | 71.37 |
| 2 | Shubman Gill | Batsman | 70.19 |
| 3 | RD Gaikwad | Batsman | 69.86 |
| 1 | Rashid Khan | All-rounder | 54.49 |
| 2 | AD Russell | All-rounder | 51.74 |
| 3 | RA Jadeja | All-rounder | 51.51 |

This table highlights the top-ranked players in each category (Bowlers, Batsmen, and All-rounders) along with their overall scores. It provides a clear overview of the leading performers in the IPL based on the analysis conducted from 2021 to 2024. This ranking system provides valuable insights for team selection, strategy formulation, and player evaluation in the IPL context, offering a data-driven approach to assessing player contributions across different roles.

### Findings and insights

A graph of different colored rectangles

Description automatically generated

This bar chart displays the average overall score for Batsmen and Bowlers, excluding any players categorized as "Other Players." It provides a clear comparison of performance between these two key player types, highlighting which group tends to perform better on average.

Distribution of Overall Scores by Player Type

A diagram of different colored boxes

Description automatically generated

This box plot shows the distribution of overall scores for batsmen, bowlers and all-rounders. It highlights the median, quartiles, and any outliers within each category. This visualization is useful for understanding the spread of scores and identifying top performers in each group.

Overall Score vs. Total Runs Scored (for All-rounder):

A graph with blue dots

Description automatically generated

This scatter plot illustrates the relationship between total runs scored and overall score for Batsmen. Each point represents a player, and the player's name is annotated next to their respective point. This visualization helps identify how total runs correlate with overall scores among all-rounders.

Composition of Overall score:

A graph of different colored bars

Description automatically generated

This stacked bar chart visualizes the composition of the overall score for the top 10 players in the dataset. Each bar represents a player, and the segments within each bar show the normalized contributions of different performance metrics to their overall score. The metrics included are total runs scored, batting average, batting strike rate, total wickets, and economy rate.

The chart allows for easy comparison of how different players achieve their high overall scores. Some players may excel in run-scoring, while others might contribute more through wicket-taking or economic bowling. This visualization provides a comprehensive view of what makes these top players stand out, highlighting their strengths across various aspects of the game.

Density distribution of overall scores:

A diagram of different colors

Description automatically generated

This violin plot will show the density distribution of overall scores for Batsmen, Bowlers, and All-rounders, allowing for easy comparison of score distributions across these player types. The width of each "violin" represents the density of data points at that score level, providing insight into the most common score ranges for each player type.

Performance metrics of All-rounders

A diagram of a diamond with colored lines

Description automatically generated with medium confidence

Each player is represented by a unique coloured line connecting their scores across all metrics. The area enclosed by a player's line visually represents their overall performance - a larger area generally indicates better all-round performance.

This chart effectively highlights the strengths and weaknesses of each all-rounder. For example, a player might excel in wicket-taking (high totalwickets\_norm) but have a lower economy rate (lower economyrate\_norm). Similarly, it shows how players balance their bowling and batting performances.

The radar chart provides a quick, intuitive way to compare these top all-rounders and understand their performance profiles across multiple aspects of the game.

## 3.Random forest model to predict the overall score

### Need for Analysis:

The random forest regression model is an excellent choice for predicting a player's Overall\_score based on the provided dataset. This model can effectively process the diverse set of features, including batting statistics (e.g., totalrunsscored, batting\_strike\_rate), bowling metrics (e.g., totalwickets, economyrate), and the crucial Player\_type category. By leveraging these varied inputs, the model can discern complex patterns that contribute to a player's overall performance rating. The random forest's ability to handle non-linear relationships is particularly valuable, as the impact of certain stats on overall score may not be strictly linear. Feature importance analysis will reveal which attributes most significantly influence the Overall\_score prediction, potentially offering insights into the weighting of different skills. This model can adapt to the distinct scoring criteria for batsmen and bowlers, thanks to the Player\_type feature. The inclusion of normalized features allows for fair comparison across different statistical scales. By employing this model, overall\_score predictions can be generate accurate that holistically represent a player's performance, accounting for their specific role and the nuanced interplay of their various cricket statistics.

### Objective of the Model

To develop and implement a random forest regression model that accurately predicts the Overall\_score for cricket players based on their comprehensive performance statistics, including batting and bowling metrics. The model aims to provide a data-driven, unbiased evaluation of player performance that can be used for team selection, player ranking, and strategic decision-making in cricket management and analysis.

### Data Overview

Based on the dataset you provided, here's an overview of the data you're working with for your random forest regression model to predict player Overall\_score: The dataset contains various cricket player statistics, including both batting and bowling metrics. Key features include:

1. Batting statistics: totalrunsscored, Total\_batting\_average, batting\_strike\_rate, totalballsfaced
2. Bowling statistics: totalwickets, economyrate, overs bowled
3. Normalized versions of features (totalrunsscored\_norm, Total\_batting\_average\_norm, batting\_strike\_rate\_norm, totalwickets\_norm, economyrate\_norm,)
4. Player\_type: Categorizes players as Batsman or Bowler or All-rounder
5. Overall\_score: The target variable you're predicting
6. Rank: Player ranking based on Overall\_score

The dataset includes players with diverse roles (batsmen and bowlers), allowing the model to learn patterns specific to each player type. The presence of both raw and normalized features provides flexibility in how the model interprets the data.

A red and blue squares with white text

Description automatically generated

This matrix suggests that batting performance has a stronger influence on Overall\_score. The model will likely give more weight to batting and bowling statistics, especially totalrunsscored and Total\_batting\_average, and overall economy when predicting Overall\_score.

### Random Forest Regression Model

#### Evaluation

1. Mean Squared Error (MSE): 3.365822488403902
   * This is a relatively low MSE, suggesting that on average, the model's predictions deviate from the actual Overall\_score by about √3.37 ≈ 1.84 points.
   * Given that the Overall\_score likely spans a wider range, this level of error is quite small.
2. R-squared Score: 0.9921912055446288 (99.22%)
   * This is an extremely high R-squared value, indicating that your model explains about **99.22%** of the variance in the Overall\_score.
   * It suggests a very strong fit between your model's predictions and the actual Overall\_scores.

The model demonstrates excellent predictive power, capturing almost all of the variability in the Overall\_score based on the provided features.

With an R-squared of 99.22%, the model's predictions are very closely aligned with the actual scores, leaving only about 0.78% of the variance unexplained.

The low MSE further confirms the high accuracy of the predictions.

### Model Visualization

Actual vs Predicted Values Scatter Plot shows how well the predicted values align with the actual values. Points closer to the red dashed line indicate better predictions.

A graph with a red line and blue dots

Description automatically generated

Residuals Plot:

This plot helps identify any patterns in the residuals (prediction errors). Ideally, the residuals should be randomly scattered around the horizontal line at y=0.

A diagram of a plot

Description automatically generated

Prediction Error Distribution

This histogram shows the distribution of prediction errors. A distribution cantered around zero and symmetric indicates good model performance.

A graph with blue bars and a blue line

Description automatically generated

## 4.Random Forest Model using RandomizedSearchCV

### Objective of the model

Hyperparameter Optimization:

The objective includes finding the optimal combination of hyperparameters for the random forest model using RandomizedSearchCV. This aims to improve model performance beyond what's achievable with default settings.

Model Robustness:

By using cross-validation within RandomizedSearchCV, the new objective includes finding a model that performs well across different subsets of the data, enhancing its generalizability.

### Representation of Random Forest model with RandomizedSearchCV

Let f̂(x) be the Random Forest prediction for input x. The Random Forest model is an ensemble of decision trees, and its prediction is the average of the predictions of all trees:

Where:

* M is the number of trees (n\_estimators in your grid)
* T\_m(x) is the prediction of the m-th tree

Each tree T\_m is constructed as follows:

1. Bootstrap sampling (if bootstrap=True):  
   Draw n samples with replacement from the training data, where n is the number of training samples.
2. At each node of the tree:  
   a. Select k features randomly, where k is determined by max\_features:
   * If max\_features='sqrt', k = √p, where p is the total number of features
   * If max\_features='auto', it's the same as 'sqrt' for regression

b. Find the best split among the k features based on mean squared error reduction:  
where I is the impurity measure (variance for regression), and n is the number of samples.c. Split the node if:

* + The number of samples is ≥ min\_samples\_split
  + The depth of the node is < max\_depth (if specified)

1. Stop growing the tree when:
   * A node has ≤ min\_samples\_leaf samples
   * No further splits can improve the model

The hyperparameters you're tuning affect this process as follows:

* n\_estimators: Determines M
* max\_features: Affects k in step 2a
* max\_depth: Limits the depth in step 2c
* min\_samples\_split: Used in step 2c
* min\_samples\_leaf: Used in step 3
* bootstrap: Determines whether step 1 is performed

The final prediction for a new input x is:

RandomizedSearchCV will try different combinations of these hyperparameters to minimize the cross-validation error, typically mean squared error for regression:

Where y\_i are the true values and ŷ\_i are the predicted values.

#### Evaluation

1. Mean Squared Error (MSE): 5.564417967827715
   * This is slightly higher than your previous model (which had an MSE of 3.37).
   * It indicates that, on average, your predictions deviate from the actual Overall\_score by about √5.56 ≈ 2.36 points.
2. R-squared Score: 0.9879148111371704
   * This is still an excellent R-squared value, indicating that your model explains about 98.79% of the variance in the Overall\_score.
   * It's slightly lower than your previous model (which had an R-squared of 0.9922).

### Model Visualization

#### Actual vs Predicted Analysis

A graph with blue dots

Description automatically generated

1. Strong Correlation: The scatter plot should show a very strong linear relationship between actual and predicted values, with points clustering tightly around the diagonal line (y=x).
2. Minimal Scatter: Given the high R-squared value of 0.9879, there is a little scatter or deviation from the diagonal line.
3. Consistent Accuracy: The model's predictions should be consistently accurate across the range of Overall\_scores, without significant bias towards over- or under-prediction.
4. Small Deviations: The MSE of 5.564 suggests that, on average, predictions deviate from actual values by about √5.564 ≈ 2.36 points. This small deviation might be barely noticeable in the plot.
5. Range Coverage: The plot should show that the model performs well across the entire range of Overall\_scores, from low to high values.

#### Distribution of Prediction Errors Analysis

A graph of a distribution of prediction error

Description automatically generated

1. Cantered around Zero: The histogram should be cantered very close to zero, indicating that the model's predictions are unbiased. This means the model is equally likely to slightly overpredict or underpredict.
2. Narrow Distribution: Given the low MSE and high R-squared, you should see a narrow distribution of errors. Most errors will be clustered tightly around zero.
3. Symmetry: The distribution should appear roughly symmetrical, resembling a normal distribution. This suggests that positive and negative errors are equally likely and of similar magnitudes.
4. Smooth KDE Line: The Kernel Density Estimation (KDE) line should show a smooth, bell-shaped curve overlaying the histogram, further emphasizing the normal-like distribution of errors.

## 5. XG Boosting Method

### Objective of this model

1. Handling Complex Relationships

XGBoost is particularly effective in capturing complex, non-linear relationships between features. In the context of cricket, the relationship between various player statistics such as batting average, strike rate, total runs scored, and wickets taken and the Overall\_score is likely to be intricate. XGBoost's ability to model these complex interactions enhances predictive accuracy.

2. Feature Importance

The model provides built-in feature importance scores, which are valuable for identifying the most significant cricket statistics that contribute to a player's Overall\_score. This insight can guide coaching and selection decisions by highlighting which metrics are most predictive of performance.

3. Mixed Data Types

The dataset includes both continuous variables (e.g., batting average, economy rate) and categorical variables (e.g., Player\_type). XGBoost effectively handles both types of data, allowing for a comprehensive analysis of player performance without extensive preprocessing.

4. Flexibility in Loss Functions

XGBoost allows for the customization of loss functions, which can be beneficial for tailoring the model to specific nuances in the calculation of Overall\_score. This flexibility enhances the model's applicability to various performance metrics. This structured explanation provides a comprehensive rationale for the use of the XGBoost model in predicting player Overall\_score, suitable for publication or formal reporting.

### Representation of the Model

1. Data Preparation:
2. Feature Transformation:
3. Model Formulation:

Where:

* + K is the number of trees (n\_estimators in param\_grid)
  + f\_k is the k-th tree in the ensemble

1. Objective Function:
2. Tree Building Process:  
   For each tree f\_k:
3. Hyperparameter Optimization:  
   Using RandomizedSearchCV to optimize over:
4. Final Prediction:  
   For a new scaled input x\_new:
5. Model Evaluation:

### Model Visualization

#### Actual vs predicted analysis

A graph with blue dots and red line

Description automatically generated

The Actual vs Predicted values scatter plot provides a visual representation of the XGBoost model's performance in predicting player Overall\_scores. The plot displays a strong correlation between actual and predicted values, with data points closely aligning along the diagonal red line. This alignment indicates high prediction accuracy, corroborating the model's impressive R-squared value of 0.9930. The tight clustering of points around the diagonal suggests consistent performance across the range of Overall\_scores. The absence of significant deviations or patterns in the scatter reinforces the model's robustness. This visualization, combined with the low Mean Squared Error of 3.2368, demonstrates the XGBoost model's exceptional ability to capture the underlying patterns in the cricket performance data and accurately predict player Overall\_scores.

#### Distribution of Prediction Errors Analysis

A graph of a distribution of prediction error

Description automatically generated

The histogram of prediction errors provides crucial insights into the XGBoost model's performance. The distribution is cantered near zero, indicating unbiased predictions. The narrow, symmetrical shape suggests that errors are small and evenly distributed between over- and under-predictions. The high peak around zero confirms the model's high accuracy, aligning with the low Mean Squared Error of 3.2368. The short tails indicate few large errors, while the overall shape approximates a normal distribution. This visualization corroborates the model's strong R-squared value of 0.9930, demonstrating consistent and accurate predictions across the range of Overall\_scores. The error distribution reinforces the XGBoost model's effectiveness in capturing the nuances of cricket performance metrics for accurate player evaluation.

## 6.Enhanced XG Boosting model

The enhanced XGBoost model, incorporating feature engineering and extensive hyperparameter tuning, demonstrates a robust performance in predicting player Overall\_scores.

### Representation of the model

Feature Engineering:

Where:

Model Structure:

Where K is the number of trees (n\_estimators)

Tree Structure:

T is the number of leaves in the tree

Objective Function:

Where:

Update Rule:

Where η is the learning rate and h\_m is the weak learner

Hyperparameter Space:

θ ∈ {n\_estimators, max\_depth, learning\_rate, subsample, colsample\_bytree, min\_child\_weight, gamma, reg\_alpha, reg\_lambda}

Feature Selection:

where S is the selection function based on feature importance

Final Prediction:

Model Evaluation:

The enhanced XGBoost model for predicting cricket player Overall\_scores incorporates feature engineering and advanced optimization techniques. The model begins with feature engineering, creating new variables like runs\_per\_ball and wickets\_per\_over to capture performance efficiency. The core XGBoost algorithm utilizes an ensemble of decision trees, each contributing to the final prediction. The objective function balances the model's predictive accuracy with its complexity through regularization terms. Hyperparameter optimization is performed using RandomizedSearchCV, exploring a wide range of parameters including tree depth, learning rate, and regularization factors. Feature selection is implemented to focus on the most impactful variables, potentially improving model efficiency and reducing overfitting. The model's performance, with an R-squared of 0.8142 and MSE of 85.5695, indicates a strong predictive capability, explaining 81.42% of the variance in Overall\_scores. This framework allows for a nuanced analysis of cricket performance, considering various aspects of a player's statistics and their complex interactions, providing a robust tool for player evaluation and team strategy formulation.

This model performance is not great when compared with other models, hence less priority is given to the model.

## 7. Support Vector Regression Model

### Objective of the model

The application of Support Vector Regression (SVR) for predicting cricket player Overall\_scores aims to develop a robust and accurate model capable of handling complex, non-linear relationships within performance data (Smola and Schölkopf, 2004). This approach offers several key advantages:

1. Accurate prediction of Overall\_scores using a subset of critical performance metrics.
2. Identification of non-linear patterns in cricket performance data that may be overlooked by simpler models.
3. Development of a generalizable model with strong performance on unseen data, as demonstrated by the high R-squared score of 0.9632 on the test set.
4. Optimization of the balance between model complexity and prediction accuracy through hyperparameter tuning (Cherkassky and Ma, 2004).

SVR is particularly well-suited for this dataset and prediction task due to its:

1. Ability to capture non-linear relationships using the RBF kernel, which is crucial for modelling the complex interactions between player statistics and overall\_score (Drucker et al., 1997).
2. Robustness to outliers, a valuable feature when analyzing sports performance data that may include exceptional individual performances.
3. Effectiveness with datasets containing a moderate number of features, aligning well with the selected cricket statistics.
4. Regularization capabilities through the 'C' parameter, allowing for control over model complexity and mitigation of overfitting (James et al., 2013).
5. Precision control via the 'epsilon' parameter, enabling fine-tuning of the model's predictive accuracy.
6. Versatility in kernel selection, with the RBF kernel proving optimal for this application.

### Representation of the model

Feature Selection and Scaling:

Where X is the feature matrix.

This step standardizes the selected features, ensuring they're on the same scale.

SVR Objective Function:

RBF Kernel:

The RBF kernel was selected as the best performing kernel in your grid search.

Hyperparameter Optimization:

The grid search found the optimal parameters: C = 100, ε = 0.01, kernel = 'rbf'.

Prediction Function:

Model Evaluation:

The SVR model demonstrates strong performance in predicting cricket player Overall\_scores, as evidenced by its high R-squared value (0.9632) and low MSE (16.9326). Compared to other models, SVR's ability to capture non-linear relationships through its RBF kernel and its robustness to outliers make it particularly well-suited for this complex sports data. The model's optimized hyperparameters further enhance its predictive accuracy.

#### Model Visualization

A graph with a red line

Description automatically generated

The Actual vs Predicted plot for the SVR model illustrates its high predictive accuracy. The scatter points closely align with the red diagonal line, indicating strong agreement between actual and predicted Overall\_scores. This visual representation corroborates the model's high R-squared value (0.9632) and low MSE (16.9326), demonstrating the SVR's effectiveness in capturing the underlying patterns in cricket performance data for accurate player evaluation.

#### Distribution of Prediction errors

A graph with numbers and lines

Description automatically generated

The histogram of prediction errors shows a symmetric distribution cantered near zero, indicating unbiased predictions. The narrow spread suggests small errors, confirming the model's high accuracy. This visualization aligns with the low MSE (16.9326) and high R-squared (0.9632) values.

## Machine Learning Models and their accuracy results

### Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | MSE | R2 | Accuracy in % |
| Linear Regression Model (used only for team analysis) | 0.08258 | 0.9969 | 99.6% |
| Random Forest Model 1 | 3.3658 | 0.9922 | 99.22% |
| Random Forest Model 2 | 5.5644 | 0.9879 | 98.79% |
| XG Booting Model 1 | 3.2368 | 0.9930 | 99.30% |
| XG Booting Model 2 | 85.570 | 0.8142 | 81.42 |
| Support Vector Regression Model | 16.933 | 0.9632 | 96.32% |

The XG Boosting Model 1 appears to be the most suitable for prediction. It demonstrates the best overall performance with:

1. Lowest Mean Squared Error (MSE) of 3.2368
2. Highest R-squared value of 0.9930
3. Highest accuracy of 99.30%

This model outperforms the others in all three metrics, indicating superior predictive power and fit to the data. The low MSE suggests minimal prediction errors, while the high R-squared value indicates that the model explains 99.30% of the variance in the target variable. The Random Forest Models (1 and 2) also show strong performance, closely following XG Boosting Model 1. The Support Vector Regression Model, while still performing well, has a notably higher MSE and lower R-squared compared to the top performers. XG Boosting Model 2 shows significantly poorer performance compared to the others and would not be recommended for this prediction task. In conclusion, XG Boosting Model 1 would be the most suitable choice for predicting cricket player performance based on these metrics.

### Fine-Tuning

The fine-tuning process led to noticeable improvements in error metrics (lower MSE) and explanatory power (higher R-squared) for both top models. This indicates that the fine-tuning successfully optimized the models to better fit the specific patterns in player performance data. The marginal gains in these already high-performing models suggest that fine-tuning helped capture subtle nuances in the data, potentially leading to more precise player analysis and predictions.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | MSE | R2 | Accuracy in % |
| Random Forest Model 1 | 2.636 | 0.9942 | 99.42% |
| Random Forest Model 2 | 5.5644 | 0.9879 | 98.79% |
| XG Booting Model 1 | 2.49 | 0.9946 | 99.46% |
| XG Booting Model 2 | 85.570 | 0.8142 | 81.42 |
| Support Vector Regression Model | 16.933 | 0.9632 | 96.32% |

 Fine-tuning had a positive impact on the overall performance metrics of top models:

1. Random Forest Model 1:
   * R-squared increased from 0.9922 to 0.9942
   * Accuracy improved from 99.22% to 99.42%
2. XG Boosting Model 1:
   * MSE improved from 3.2368 to 2.49
   * R-squared increased from 0.9930 to 0.9946
   * Accuracy improved to 99.46%

### Model Testing

#### Sample data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Total Runs** | **Batting Average** | **Batting Strike Rate** | **Total Wickets** | **Economy Rate** | **Balls Batted** | **Balls Bowled** |
| Jos Buttler | 391 | 43.44 | 158.62 | 0 | - | 246 | 0 |
| Tymal Mills | 15 | 7.5 | 125 | 16 | 8.2 | 20 | 120 |
| Will Jacks | 230 | 32.86 | 145.57 | 3 | 7.8 | 158 | 30 |
| Liam Livingstone | 185 | 26.43 | 152.89 | 5 | 8.5 | 140 | 40 |
| Reece Topley | 20 | 10 | 111.11 | 11 | 7.9 | 25 | 90 |
| Dawid Malan | 278 | 39.71 | 140.4 | 0 | - | 180 | 0 |
| Sam Curran | 160 | 22.86 | 133.33 | 8 | 8.7 | 130 | 70 |
| Tom Abell | 145 | 24.17 | 128.32 | 2 | 9.2 | 120 | 20 |
| Adil Rashid | 35 | 11.67 | 106.06 | 10 | 7.5 | 30 | 80 |
| Harry Brook | 238 | 47.6 | 172.46 | 0 | - | 150 | 0 |

The data presented in the table comes from the exciting 2023 Hundred tournament held in England (ECB, 2023). 10 players are selected at random to highlight their performance metrics, including total runs scored, batting averages, strike rates, total wickets taken, economy rates, balls batted, and balls bowled (ESPNcricinfo, 2023).

#### Random Forest Model 1Prediction

|  |  |
| --- | --- |
| **Player** | **Predicted Overall score** |
| Jos Buttler | 68.40700315 |
| Tymal Mills | 74.50774429 |
| Will Jacks | 51.35719062 |
| Liam Livingstone | 48.25745156 |
| Reece Topley | 70.82660776 |
| Dawid Malan | 57.29453392 |
| Sam Curran | 44.74423243 |
| Tom Abell | 39.63716106 |
| Adil Rashid | 63.22290778 |
| Harry Brook | 51.47323937 |

Random Forest predictions show a range of scores from 39.64 to 74.51, with an average of around 57. The model seems to predict higher scores for bowlers like Tymal Mills (74.51) and Reece Topley (70.83), while predicting lower scores for some batsmen like Tom Abell (39.64).

#### XG Boost Model 1 Prediction

|  |  |
| --- | --- |
| **Player** | **Predicted Overall score** |
| Jos Buttler | 70.39937 |
| Tymal Mills | 72.09699 |
| Will Jacks | 45.95546 |
| Liam Livingstone | 44.136646 |
| Reece Topley | 67.81441 |
| Dawid Malan | 61.69961 |
| Sam Curran | 46.41379 |
| Tom Abell | 33.50635 |
| Adil Rashid | 48.7163 |
| Harry Brook | 57.242996 |

XGBoost predictions range from 33.51 to 72.10, averaging around 54. This model also predicts high scores for bowlers, with Tymal Mills at 72.10 and Reece Topley at 67.81. However, it predicts lower scores for some players like Tom Abell (33.51) and Will Jacks (45.96).

### Performance Distribution Curves

A graph of a performance curve

Description automatically generated with medium confidence

The performance distribution curves show the spread and frequency of predicted and actual scores for the cricket players.

Random Forest Model Distribution:  
The curve for the Random Forest model predictions likely shows a relatively widespread, with scores ranging from about 39 to 75. The peak of the curve might be around the mid-50s, indicating that the model frequently predicts scores in this range. There may be a slight right skew, suggesting the model tends to predict higher scores more often than lower ones.

XGBoost Model Distribution:  
The XGBoost model's distribution curve probably shows a similar range to the Random Forest model, from about 33 to 72. However, the shape of the curve might be different, possibly with a sharper peak or multiple smaller peaks, reflecting the model's tendency to make more extreme predictions in some cases.

Actual Scores Distribution:  
The curve for actual scores likely shows the widest spread, ranging from about 35 to 90. This curve might have a flatter shape compared to the model predictions, indicating more variability in real-world performance. The peak of this curve could be lower and broader than the prediction curves, suggesting that actual performances are more diverse than what the models predict.

### ROC curves

ROC curves visually represent the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) as the classification threshold changes. Using ROC curves, comprehensively assess and compare models' abilities to distinguish between different levels of cricket player performance (Brownlee, 2018). The AUC provides a single scalar value summarizing the model's performance, making it easy to quickly compare model.

The AUC can be interpreted as the probability that the model ranks a random positive example higher than a random negative example, which is particularly relevant for ranking player performance (Hajian-Tilaki, 2013).

A graph with a line graph

Description automatically generated

1. XGBoost Performance:
   * With an AUC of 0.92, the XGBoost model is showing excellent performance.
   * This score indicates that the model has a very high ability to distinguish between the classes (in this case, likely between high-performing and low-performing players).
   * An AUC of 0.92 suggests that if randomly selected a positive instance and a negative instance, there's a 92% chance that the model will rank the positive instance higher than the negative one.
2. Random Forest Performance:
   * The Random Forest model, with an AUC of 0.79, is showing good performance, though not as strong as the XGBoost model.
   * This score still indicates a decent ability to discriminate between classes, but there's more room for misclassifications compared to the XGBoost model.
3. Comparison:
   * The XGBoost model (0.92) is outperforming the Random Forest model (0.79) by a significant margin.
   * This suggests that XGBoost is better at capturing the underlying patterns in your data for this problem.
4. Interpretation:
   * XGBoost seems to be making more accurate predictions about player performance, with fewer false positives and false negatives compared to the Random Forest model.
   * The Random Forest model, while still good, may be missing some of the nuances that XGBoost is capturing.

Based on the AUC scores, XGBoost (AUC = 0.92) outperforms Random Forest (AUC = 0.79) in predicting cricket player performance. XGBoost's higher AUC indicates superior ability to distinguish between high and low performers. This model demonstrates a 92% probability of correctly ranking players, making it more reliable for performance predictions and team selection decisions in cricket analytics.

# Players Overall Performance score for KKR and DC

## Kolkata Knight Riders Current Players Analysis

Kolkata Knight Riders are the 2024 champions. With Shreyas Iyer’s captaincy they dominated the whole season with their powerful line up. Phil Salt and Suni Narine’s opened the lineup followed by Shreyas Iyer, Venkatesh Iyer, Rinku Singh, Russell, Ramandeep Singh. They had one of the best destructive batting line up. On the other side Starc, Harshit Rana, Varun, Arora are the bowlers. Rahmanullah Gurbaz, Nitish Rana, Manish Pandey, Sakariya, Ankul Roy, Angkrish Raghuvanshi are the substitutes. They add more value to the team.

Using the “rule-based scoring system”, the overall scores for these players are calculated.

This chart shows the performance score for the players who played in the 2024 season. The data is calculated using rule-based scoring system and data taken from 2021 to 2024. Top performers for KKR are Rinku Singh, Varun Chakaravarthy, Venkatesh Iyer, Shreyas Iyer, Andre Russell, Phil Salt, Sunil Narine.

**Batting Dominance**

KKR's batting lineup has been formidable, with several players making substantial contributions:

1. Sunil Narine has emerged as the team's top run-getter, accumulating 488 runs at a strike rate of 180.74.
2. Phil Salt has been a revelation at the top of the order, amassing 435 runs with a blistering strike rate of 182.00.
3. Venkatesh Iyer has shown remarkable consistency, scoring 370 runs at an average of 46.25.
4. Shreyas Iyer has also been a key player, scoring 351 runs at an average of 39.00, further solidifying the middle order. Ramandeep Singh has shown promise with 125 runs in 10 matches, including a highest score of 35 and a strike rate of 205.88.

**All-Round Excellence**

The team's all-round capabilities have been a key factor in their success:

* Sunil Narine has excelled as an all-rounder, complementing his batting prowess with 17 wickets at an economical rate of 6.69 runs per over.
* Andre Russell continues to be a vital asset, contributing 222 runs at a strike rate of 185 while also claiming 19 wickets.

**Bowling strength**

KKR's bowling attack has been equally impressive:

1. Varun Chakaravarthy leads the wicket-taking charts with 21 scalps.
2. Andre Russell has provided crucial breakthroughs, securing 19 wickets.
3. Mitchell Starc, despite a higher economy rate, has taken 17 wickets, including a 4-wicket haul.
4. Harshit Rana has added depth to the bowling lineup with 19 wickets. Vaibhav Arora has made a significant impact, taking 11 wickets in 10 matches at an average of 25.09 and an economy of 8.24

**Emerging Talent**

Angkrish Raghuvanshi has shown promise as a future prospect, scoring 163 runs in 10 matches at a strike rate of 155.23.

Team Balance

KKR's performance this season highlights a well-balanced side with strong contributions from both batsmen and bowlers. The team's ability to score quickly and take regular wickets has been instrumental in their success. This analysis underscores KKR's formidable batting lineup and their balanced bowling attack, with multiple players making significant contributions across various aspects of the game. The team's balanced approach has positioned them as winners in the 2024 IPL season.

The upcoming IPL 2025 retention process poses a significant challenge for Kolkata Knight Riders' management. With a limit of four retention slots and the Right to Match (RTM) option for two more players, KKR faces difficult choices. The team must carefully weigh retaining star performers against nurturing emerging talents, while also considering team chemistry and long-term strategy. This intricate decision-making process is critical for KKR's future success and competitiveness in the league.

## Delhi Capitals Current Players Analysis

Delhi Capitals got 6th position in the 2024 IPL season. despite having a roster filled with talented players capable of high performance, the Delhi Capitals' 2024 season was marred by inconsistencies in both batting and bowling, lack of depth, and challenges in team dynamics. Addressing these issues will be crucial for the management as they look to rebuild and strategize for the upcoming seasons.

This bar graph data shows overall scores for Delhi Capitals players based on a rule-based scoring system for the 2024 IPL season. Rishabh Pant leads with the highest score of 82.45, followed closely by Jake Fraser-McGurk at 79.97. Key players like Khaleel Ahmed, Kuldeep Yadav, and Mukesh Kumar also scored well, indicating their significant contributions. The scores reflect a combination of batting, bowling, and all-round performances throughout the season. Lower scores for some players suggest either limited opportunities or underperformance, while a few players received no score, due to lack of playing time or poor performance.

**Batting Strength**

1. Rishabh Pant led the batting charts with 446 runs at an impressive average of 40.55 and a strike rate of 155.4.
2. Tristan Stubbs showed excellent form, scoring 378 runs at a high average of 54 and a strike rate of 190.9.
3. Jake Fraser-McGurk emerged as an explosive batsman, scoring 330 runs at a strike rate of 234.04.
4. Abishek Porel contributed significantly with 327 runs at a strike rate of 159.51.

Bowling Strength:

1. Kuldeep Yadav was the standout bowler, taking 16 wickets at an average of 23.37 and an economy of 8.69.
2. Mukesh Kumar impressed with 17 wickets at an average of 21.64.
3. Axar Patel contributed with 11 wickets with a good economy of 7.65.
4. Khaleel Ahmed took 17 wickets with a decent economy of 9.58.

**All-Round Performance:**

1. Axar Patel showcased his all-round abilities, scoring 235 runs and took 11 wickets.
2. Tristan Stubbs, primarily a batsman, also took 3 wickets.

**Emerging Talent:**

1. Jake Fraser-McGurk stood out as a promising talent with his explosive batting.
2. Abishek Porel showed potential as a consistent run-scorer.
3. Rasikh Salam took 9 wickets in 8 matches.

The team's strength clearly lies in its batting, with multiple players capable of scoring quickly. The bowling unit, led by Kuldeep Yadav and supported by Mukesh Kumar and Axar Patel, also performed well. The emergence of young talents like Fraser-McGurk and Porel adds depth to the squad.

The upcoming IPL 2025 retention process presents a significant challenge for Delhi Capitals' management. With a limit of four retention slots and the Right to Match (RTM) option for two additional players, DC faces crucial decisions. The team must carefully balance retaining their star performers against nurturing emerging talents, while also considering team chemistry and long-term strategy.

Key factors in DC's decision-making process include:

1. Rishabh Pant's leadership and batting prowess
2. The all-round abilities of Axar Patel
3. Kuldeep Yadav's consistent spin bowling performances
4. The explosive batting potential of Jake Fraser-McGurk
5. Tristan Stubbs' impressive batting in the previous season

# Squad Optimization

## KKR Squad Optimization and picking best squad

### Current players Prediction

|  |  |
| --- | --- |
| **Player** | **Predicted Overall score** |
| Andre Russell | 62.681908 |
| Angkrish Raghuvanshi | 43.662876 |
| Anukul Roy | 0.01888789 |
| Harshit Rana | 74.76428 |
| Manish Pandey | 18.542007 |
| Mitchell Starc | 66.87246 |
| Nitish Rana | 15.08972 |
| Phil Salt | 68.24854 |
| Rahmanullah Gurbaz | 18.410284 |
| Ramandeep Singh | 47.748146 |
| Rinku Singh | 40.41537 |
| Shreyas Iyer | 62.031254 |
| Sunil Narine | 65.13953 |
| Vaibhav Arora | 54.331146 |
| Varun Chakaravarthy | 74.76428 |
| Venkatesh Iyer | 62.110737 |

# References

* Board of Control for Cricket in India (2023) Indian Premier League. Available at: https://www.iplt20.com/ (Accessed: 15 August 2024).
* BCCI (2016) 'IPL 2015 contributed Rs. 11.5 billion to India's GDP: BCCI', Press Release, 31 October.
* Duff & Phelps (2022) IPL Brand Valuation Report 2022. Mumbai: Duff & Phelps.
* Economic Times (2023) 'IPL becomes decacorn, valuation soars 75% since 2020', 27 December.
* ESPN Cricinfo (2023) Indian Premier League. Available at: https://www.espncricinfo.com/series/indian-premier-league-2023-1345038 (Accessed: 15 August 2024).
* IPL Governing Council (2024) IPL 2024: Playing Conditions. Mumbai: BCCI.
* JioCinema (2023) 'IPL 2023 Final Sets Global Streaming Record', Press Release, 30 May.
* Shah, J. (2023) The IPL Story: Cricket, Commerce and Glamour. New Delhi: Rupa Publications.
* Sportstar (2023) 'IPL media rights sold for Rs 48,390 crore: Disney Star retains TV rights, Viacom18 bags digital package', The Hindu, 14 June.
* Board of Control for Cricket in India (2023) Indian Premier League. Available at: https://www.iplt20.com/ (Accessed: 15 August 2024).
* Delhi Capitals (2023) Official Website. Available at: https://www.delhicapitals.in/ (Accessed: 15 August 2024).
* ESPN Cricinfo (2023) Indian Premier League. Available at: https://www.espncricinfo.com/series/indian-premier-league-2023-1345038 (Accessed: 15 August 2024).
* Gujarat Titans (2023) Official Website. Available at: https://www.gujarattitansipl.com/ (Accessed: 15 August 2024).
* IPL Governing Council (2024) IPL 2024: Playing Conditions. Mumbai: BCCI.
* Kolkata Knight Riders (2023) Official Website. Available at: https://www.kkr.in/ (Accessed: 15 August 2024).
* Lucknow Super Giants (2023) Official Website. Available at: https://www.lucknowsupergiants.in/ (Accessed: 15 August 2024).
* Mumbai Indians (2023) Official Website. Available at: https://www.mumbaiindians.com/ (Accessed: 15 August 2024).
* Punjab Kings (2023) Official Website. Available at: https://www.punjabkingsipl.in/ (Accessed: 15 August 2024).
* Rajasthan Royals (2023) Official Website. Available at: https://www.rajasthanroyals.com/ (Accessed: 15 August 2024).
* Royal Challengers Bangalore (2023) Official Website. Available at: https://www.royalchallengers.com/ (Accessed: 15 August 2024).
* Shah, J. (2023) The IPL Story: Cricket, Commerce and Glamour. New Delhi: Rupa Publications.
* Sunrisers Hyderabad (2023) Official Website. Available at: https://www.sunrisershyderabad.in/ (Accessed: 15 August 2024).
* Rodrigues, M., Vinay, S., Naik, N., Deshpande, S. and Samant, S. (2019). Data visualization and toss related analysis of IPL teams and batsmen performances. [online] ResearchGate.
* ESPNcricinfo (2024). How KKR shaped themselves into the awesome class of 2024. [online] Available at: https://www.espncricinfo.com/story/ipl-2024-final-kkr-vs-srh-how-kkr-shaped-themselves-into-the-awesome-class-of-2024-1435320 [Accessed 15 Aug. 2024].
* Kalgotra, P., Sharda, R. and Chakraborty, G. (2014). Predictive modeling in sports leagues: an application in Indian Premier League. Journal of Sports Analytics, 1(1), pp.19-36.
* Hubáček, O., Šourek, G. and Železný, F. (2019). Exploiting sports-betting market using machine learning. International Journal of Forecasting, 35(2), pp.783-796.
* Berrar, D., Lopes, P. and Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning, 108(1), pp.97-126.
* Amala Kaviya, V.S., 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), pp. 218-228. (Accessed: 15 August 2024).
* Bajaj, A. (2023) 'Prediction of Player Performance for IPL and Analyzing the Attributes Involved, Using Explainable AI', MSc Research Project, National College of Ireland. Available at: https://norma.ncirl.ie/6564/1/ayushibajaj.pdf (Accessed: 15 August 2024).
* Ishi, M., Patil, D.J., Patil, D.N. and Patil, D.V. (2022) 'Winner Prediction in One Day International Cricket Matches Using Machine Learning Framework: An Ensemble Approach', Indian Journal of Computer Science and Engineering, 13, pp. 628–641.
* Kadapa, S. (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, 3.
* Prakash, A., Ghosh, A. and Guha, B. (2019) 'Player Ranking System for IPL Using Machine Learning', International Journal of Sports Analytics, 5(1), pp. 1-12.
* Rodrigues, M., Vinay, S., Naik, N., Deshpande, S. and Samant, S. (2019) 'Data Visualization and Toss Related Analysis of IPL Teams and Batsmen Performances', ResearchGate. Available at: https://www.researchgate.net/publication/336886516\_Data\_visualization\_and\_toss\_related\_analysis\_of\_IPL\_teams\_and\_batsmen\_performances (Accessed: 15 August 2024).
* Shah, R., Ghosh, A. and Guha, B. (2016) 'IPL 2016: A Comprehensive Analysis of the Performance of Teams', International Journal of Sports Analytics, 2(1), pp. 1-15.
* Bunker, R.P. and Thabtah, F., 2019. A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), pp.27-33.
* Caya, O. and Bourdon, A., 2016. A framework of value creation from business intelligence and analytics in competitive sports. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 1061-1071). IEEE.
* Cervone, D., D'Amour, A., Bornn, L. and Goldsberry, K., 2016. A multiresolution stochastic process model for predicting basketball possession outcomes. Journal of the American Statistical Association, 111(514), pp.585-599.
* Fried, G. and Mumcu, C. eds., 2016. Sport analytics: A data-driven approach to sport business and management. Taylor & Francis.
* Kemper, C. and Breuer, C., 2016. How efficient is dynamic pricing for sport events? Designing a dynamic pricing model for Bayern Munich. International Journal of Sport Finance, 11(1), pp.4-25.
* Liu, G., Luo, Y., Schulte, O. and Kharrat, T., 2020. Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Mining and Knowledge Discovery, 34(5), pp.1531-1559.
* Loland, S., 2018. Performance-enhancing drugs, sport, and the ideal of natural athletic performance. The American Journal of Bioethics, 18(6), pp.8-15.
* McHale, I.G., Scarf, P.A. and Folker, D.E., 2012. On the development of a soccer player performance rating system for the English Premier League. Interfaces, 42(4), pp.339-351.
* Memmert, D. and Raabe, D., 2018. Data analytics in football: Positional data collection, modelling and analysis. Routledge.
* Ofoghi, B., Zeleznikow, J., MacMahon, C. and Raab, M., 2013. Data mining in elite sports: a review and a framework. Measurement in Physical Education and Exercise Science, 17(3), pp.171-186.
* Rommers, N., Rössler, R., Goossens, L., Vaeyens, R., Lenoir, M., Witvrouw, E. and D'Hondt, E., 2020. Risk of acute and overuse injuries in youth elite soccer players: Body size and growth matter. Journal of Science and Medicine in Sport, 23(3), pp.246-251.
* Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernàndez, J. and Medina, D., 2018. Effective injury forecasting in soccer with GPS training data and machine learning. PloS one, 13(7), p.e0201264.
* Seshadri, D.R., Drummond, C., Craker, J., Rowbottom, J.R. and Voos, J.E., 2019. Wearable devices for sports: New integrated technologies allow coaches, physicians, and trainers to better understand the physical demands of athletes in real time. IEEE pulse, 10(1), pp.38-43.
* Thomas, G., Gade, R., Moeslund, T.B., Carr, P. and Hilton, A., 2017. Computer vision for sports: Current applications and research topics. Computer Vision and Image Understanding, 159, pp.3-18.
* Cherkassky, V. and Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1), pp.113-126.
* Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J. and Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systems, 9, pp.155-161.
* James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning. New York: Springer.
* Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing, 14(3), pp.199-222.
* Brownlee, J., 2018. How to Use ROC Curves and Precision-Recall Curves for Classification in Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/ [Accessed 25 August 2024].
* Hajian-Tilaki, K., 2013. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine, 4(2), pp.627-635.