Predicting Team Success in the Indian Premier League Cricket 2024 Season Using Random Forest Analysis

Swamynathan Sanjaykumar¹ABCDE, Subhashree Natarajan²BCDE, Ponnusamy Yoga Lakshmi¹BCE and Farjana Akter Boby³BD

¹SRM Institute of Science and Technology
²RV University
³Daffodil International University

Authors’ Contribution: A – Study design; B – Data collection; C – Statistical analysis; D – Manuscript Preparation; E – Funds Collection

Corresponding Author: Swamynathan Sanjaykumar, E-mail: sanjayswaminathan007@gmail.com
Accepted for Publication: April 6, 2024
Published: April 30, 2024
DOI: 10.17309/tmfv.2024.2.16

Abstract

Background. Random Forest is a popular machine learning algorithm used for classification and regression tasks. The study purpose is to investigate the use of Random Forest machine learning to predict the winning chances of teams in the 2024 Indian Premier League (IPL) season.

Objectives. By analyzing comprehensive player statistics, including matches played, batting and bowling averages, as well as fielding contributions, the study aims to understand the factors that influence team success in T20 cricket and provide actionable insights for team management, betting markets, and cricket enthusiasts.

Material and methods. The study involved 10 cricket teams competing in the IPL 2024 season. Player statistics and match outcomes data from previous IPL seasons were collected and cleaned, with missing values addressed through imputation. The features were based on player statistics, including both aggregate measures and calculated metrics. A Random Forest is chosen as the machine learning model, trained using preprocessed data, with features derived from player statistics as input and match outcomes as the target variable. The dataset was split into training and validation sets, using methodologies such as cross-validation to ensure a robust model performance.

Results. The Random Forest model demonstrates strong predictive performance, with a low Mean Squared Error (MSE) of 8.2174, Root Mean Squared Error (RMSE) of 2.8666, and a high R-Squared value of 0.9173, indicating approximately 91.73% variance explained in the data. Chennai Super Kings emerge as frontrunners with a predicted performance percentage of 83.4%, while teams like Punjab Kings and Lucknow Super Giants show lower predicted performance percentages, suggesting potential areas for its improvement.

Conclusions. This study reveals the effectiveness of Random Forest machine learning in forecasting IPL match outcomes based on player statistics. It provides valuable insights into team dynamics and offers actionable recommendations for team management and cricket enthusiasts. The findings enrich our understanding of IPL match dynamics, contribute to the evolution of cricket analytics, and promote greater engagement with sport, ultimately enhancing the fan experience in the IPL.

Keywords: Indian Premier League, Random Forest, machine learning, team performance prediction, cricket analytics.

Introduction

Cricket, often described as a gentleman’s game, has transcended its traditional boundaries to become a global phenomenon, captivating audiences with its blend of athleticism, strategy, and sheer spectacle. Within this vibrant tapestry of cricketing culture, the Indian Premier League (IPL) stands as a beacon of innovation, drawing players from across the cricketing world and electrifying audiences with its fast-paced format and star-studded line-ups (Subburaj et al., 2023; Kapadia et al., 2022).
The Indian Premier League (IPL) stands as a testament to the global appeal and excitement surrounding cricket, captivating audiences worldwide with its blend of athleticism, drama, and sheer spectacle. As the cricketing world gears up for the upcoming IPL season in 2024, anticipation is at an all-time high, with fans eagerly awaiting the clash of titans and the quest for supremacy on the field (Sanjaykumar et al., 2023). Central to the IPL’s allure is the element of unpredictability, where every match is a battleground where teams vie for victory, driven by a potent mix of skill, strategy, and determination. In this dynamic and ever-evolving landscape, the ability to forecast winning chances emerges as a tantalizing challenge, offering valuable insights into the intricate dynamics that shape match outcomes (Passi and Pandey, 2018; Sumathi et al., 2023).

The Random Forest machine learning technique is a versatile algorithm renowned for its effectiveness in handling complex datasets. It operates by aggregating the predictions of multiple decision trees, mitigating overfitting through feature selection and bootstrap sampling. With its ability to handle high-dimensional feature spaces and deliver robust predictions, Random Forest stands out as a valuable tool in sports analytics. By leveraging historical data, it can uncover patterns and identify key factors influencing outcomes, making it particularly well-suited for predicting player performance, match results, and tournament dynamics in sports like cricket (Abebe et al., 2020; Passi & Pandey, 2018).

The IPL, with its fast-paced format and star-studded line-ups, presents a unique challenge for predicting winning chances. Team dynamics, player form, pitch conditions, and match strategy all play pivotal roles in shaping the outcome of a match, demanding a nuanced and data-driven approach to analysis (Wickramasinghe, 2014; Bai & Bai, 2021).

Moreover, the IPL’s evolving ecosystem, marked by player auctions, tactical innovations, and fluctuating fortunes, adds another layer of complexity to the prediction task. While historical data provides a foundation for analysis, it must be complemented by real-time insights and contextual understanding to capture the dynamic nature of the tournament accurately (Aburas et al., 2018; Bunker and Thabtah, 2019). By leveraging a comprehensive dataset encompassing past IPL seasons, player statistics, match conditions, and other pertinent variables, we aim to develop a predictive model capable of estimating the likelihood of each team winning a match in the upcoming season. Through rigorous analysis and validation, we seek to uncover the underlying patterns and trends that drive match outcomes in the IPL (ESPNcricinfo; Indian Premier League official website).

The purpose of the research. To employ Random Forest machine learning to predict winning chances of teams in the 2024 Indian Premier League season. By analyzing comprehensive player statistics, including the number of matches played, batting averages, bowling averages, and fielding contributions (e.g., catches taken), we aim to understand the factors that influence team success in T20 cricket. The implications of the study extend beyond mere prediction, offering actionable insights for team management, betting markets, and cricket enthusiasts alike. By identifying key factors that influence winning chances, teams can optimize their strategies, fine-tune their player selections, and enhance their competitive edge in the tournament.

Material and Methods

Participants

The participants of this study encompass 10 cricket teams slated to compete in the upcoming Indian premier league T20 2024. These teams represent diverse cricketing franchises and comprise professional players selected to represent their teams in the tournament. The dataset utilized for analysis comprises player statistics and match outcomes sourced from previous IPL and other pertinent T20 tournaments leading up to the IPL T20 cricket tournament 2024 (ESPNcricinfo; Indian Premier League official website).

Study Organization

Comprehensive player statistics, including the number of matches played, batting averages, bowling averages, and fielding contributions (e.g., catches taken), are gathered for each player participating in the IPL 2024 in Indian premier league official website (figure 1).

The collected data undergoes meticulous cleaning to rectify inconsistencies or inaccuracies, with missing values addressed through imputation or suitable handling techniques. Relevant features are derived from player statistics, incorporating aggregate measures and calculated metrics (Sanjaykumar et al., 2023; Vetukuri et al., 2019). A random forest is chosen as the machine learning model for predicting match outcomes. The model is trained using the preprocessed data, with features derived from player statistics as input and match outcomes as the target variable. The dataset is split into training and validation sets, employing methodologies like cross-validation to ensure robust model performance (Hudnurkar & Rayavarapu, 2022; Baboota & Kaur, 2019).

The purpose of the research. To employ Random Forest machine learning to predict winning chances of teams in the 2024 Indian Premier League season. By analyzing comprehensive player statistics, including the number of matches played, batting and bowling averages, and fielding contributions, we aim to understand the factors that influence team success in T20 cricket. The implications of the study extend beyond mere prediction, offering actionable insights for team management, betting markets, and cricket enthusiasts alike. By identifying key factors that influence winning chances, teams can optimize their strategies, fine-tune their player selections, and enhance their competitive edge in the tournament.
Statistical Analysis

To evaluate the employ Random Forest machine learning to predict winning chances of teams in the 2024 Indian Premier League cricket season. Random Forest, several statistical metrics are employed, these metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared Metric (R²). The MSE quantifies the average of the squared differences between predicted and actual performance values, while RMSE represents the square root of this average, maintaining the same units as the original data. MAE assesses the average absolute variance between predicted and observed performance values, and R² evaluates the goodness of fit of the model, ranging from 0 to 1 (Wickramasinghe, 2020; Lakshmi et al., 2024).

\[
\text{Mean Squared Error (MSE)} = \frac{1}{n} \sum (X_{\text{predict}} - X_{\text{actual}})^2 \\
\text{Root Mean Squared Error (RMSE)} = \sqrt{\text{MSE}} \\
\text{R-Squared (R²)} = 1 - \frac{\sum (X_{\text{actual}} - X_{\text{predict}})^2}{\sum (X_{\text{actual}} - X_{\text{mean}})^2}
\]

In the provided equation, \(X_{\text{actual}}\) represents the observed performance values, \(X_{\text{predict}}\) denotes the predicted performance, \(X_{\text{mean}}\) stands for the mean of the actual performance values, and "n" symbolizes the number of data points. This equation serves as a crucial tool in calculating the R-squared Metric (R²), which assesses the goodness of fit of predictive models by measuring the proportion of variance in the dependent variable (match outcomes, in this context) that is explained by the independent variables (player statistics).

Results

The results presented here serve to enrich our understanding of IPL match dynamics, providing valuable insights that can inform decision-making processes and enhance the fan experience. Through a combination of data-driven analysis, we aim to contribute to the ongoing evolution of cricket analytics and promote greater engagement with the sport.

Table 1. Random Forest Model Performance Metrics analysis

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Square Error</td>
<td>8.2174</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>2.8666</td>
</tr>
<tr>
<td>R-Squared</td>
<td>0.9173</td>
</tr>
</tbody>
</table>

The Random Forest (RF) model exhibits strong predictive performance, as evidenced by the low Mean Square Error (MSE) of 8.2174 and Root Mean Square Error (RMSE) of 2.8666. The R-Squared value of 0.9173 indicates that the model explains approximately 91.73% of the variance in the data, suggesting a high level of accuracy in predicting team winning chances (Table 1). These metrics collectively demonstrate the effectiveness of the RF model in forecasting IPL match outcomes, providing valuable insights for stakeholders and decision-makers (Fig. 3).
providing valuable insights for stakeholders and decision-makers (Kaur et al., 2021).

The performance prediction graph of IPL teams based on the Random Forest model highlights Chennai Super Kings as the frontrunners with a predicted performance percentage of 83.4%. This aligns with the study’s aim of forecasting team success and suggests that player statistics, such as matches played, batting and bowling averages, and fielding contributions, contribute to Chennai Super Kings’ favorable outlook for the upcoming IPL season. Conversely, teams like Punjab Kings (71.2%) and Lucknow Super Giants (73.8%) display lower predicted performance percentages, indicating areas for improvement in their player statistics and team dynamics (Fig. 4) (Bhattacharjee & Talukdar, 2020; Van Eetvelde et al., 2021).

The study enriches our understanding of IPL match dynamics and provides actionable insights for team management, betting markets, and cricket enthusiasts. By identifying key factors that influence winning chances, teams can optimize strategies, fine-tune player selections, and enhance their competitive edge (Men, 2022; Turhan & Canpolat, 2023). The study’s combination of data-driven analysis and advanced modeling techniques contributes to the ongoing evolution of cricket analytics and promotes greater engagement with the sport, ultimately enhancing the fan experience and decision-making processes in the IPL (Šuštaršić et al., 2022; Gu et al., 2023).

Conclusions

This study employs Random Forest machine learning to forecast team winning probabilities for the 2024 Indian Premier League (IPL) season, utilizing a wide array of player statistics such as matches played, batting and bowling averages, and fielding contributions. The model demonstrates robust predictive capability, reflected in the low Mean Square Error (MSE) of 8.2174, Root Mean Square Error (RMSE) of 2.8666, and high R-Squared value of 0.9173, signifying an explanatory power of approximately 91.73% in the data variance. Chennai Super Kings emerge as the favored contenders with a predicted performance percentage of 83.4%, in accordance with the study’s objective of gauging team success. Conversely, Punjab Kings (71.2%) and Lucknow Super Giants (73.8%) exhibit lower predicted performance percentages, implying potential areas for enhancement in player statistics and team dynamics.

Acknowledgments

We extend our heartfelt gratitude to the Indian Premier League (IPL) and ESPN for providing access to valuable datasets and resources essential for the completion of this study. Additionally, we acknowledge the Board of Control for Cricket in India (BCCI) for their continued support and dedication to promoting cricketing excellence.

Conflict of Interest

We declare no conflicts of interest regarding the publication of this study.

References


Прогнозирование успешности команд в Индийской прем’єр-лізі з крикету сезону 2024 року через аналіз застосування методу випадкового лісу

Сваміннатан Санджайкумар лабсд, Субхасрі Натараджан лбсдс, Поннусамі Йога Лакшмілбсд, Фарджана Актер Бобі лбд

1SRM Інститут науки та технологій
2RV Університет
3Міжнародний університет Даффоділ

Авторський вклад: A – дизайн дослідження; B – збір даних; C – статаналіз; D – підготовка рукопису; E – збір коштів

Реферат. Стаття: 6 с., 1 табл., 4 рис., 26 джерел.

Історія питання. Випадковий ліс (англ. "Random Forest") — популярний алгоритм машинного навчання, який використовується для задач класифікації та регресії. Метою роботи є дослідити застосування методу машинного навчання випадкового лісу для прогнозування шансів команд на перемогу в сезоні Індійської прем’єр-ліги (ІПЛ).

Мета дослідження. Шляхом аналізу всебічної статистики гравців, включаючи зіграні матчі, середні показники бігів і боулінгу, а також активність філдингу, дослідження має на меті з’ясувати фактори, що впливають на успіх команди в серії T20 (“Twenty20”), і надати практичну інформацію для менеджменту команд, ринків прогнозів та поціновувачів крикету.

Матеріал та методи. У дослідженні взяли участь 10 команд з крикету, які змагалися в сезоні ІПЛ 2024 року. Статистичні дані про гравців та результати матчів за попередні сезони ПЛ було зібрано та опрацьовано, а відсутні значення отримано за допомогою методу інтерполяції. Характеристики базувалися на статистиці гравців, включаючи як агреговані показники, так і обчислювані метрики. В якості моделі машинного навчання обрано метод випадкового лісу, який використовується для навчання на основі попередньо оброблених даних, з характеристиками, отриманими на основі статистики гравців – в якості вхідних даних та результатами матчів в якості цільової змінної. Для забезпечення надійного функціонування моделі набір даних було розділено на тренувальний та затверджувальний набори з використанням таких методик, як перехресне затвердження.


Результати. Модель випадкового лісу демонструє суттєву прогностичну ефективність з низькою середньоквадратичною похибкою (СКП) 8,2174, коренем із середньоквадратичної похибки (КСКП) 2,8666 і високим значенням R-квадрату 0,9173, вказуючи на пояснення приблизно 91,73% дисперсії в даних. Команда “Chennai Super Kings” лідирує з прогнозованим відсотком результативності 83,4%, в той час як команди “Punjab Kings” та “Lucknow Super Giants” показують нижчий прогнозований відсоток результативності, що свідчить про наявність потенційних можливостей для її покращення.

Висновки. Це дослідження демонструє ефективність застосування методу машинного навчання випадкового лісу для прогнозування результатів матчів ППЛ на основі статистики гравців. Дослідження містить цінну інформацію щодо командної динаміки та практичні рекомендації для менеджменту команд і шанувальників крикету. Отримані результати збагачують наше розуміння щодо динаміки матчів ППЛ, сприяють розвитку аналітики крикету та більшому залученню до спорту, що в кінцевому підсумку підвищує рівень досвіду вболівальників в ППЛ.

Ключові слова: Індійська прем’єр-ліга, випадковий ліс, машинне навчання, прогнозування результативності команд, аналітика крикету.

Information about the authors
Sanjaykumar, Swamynathan: sanjayswaminathan007@gmail.com; https://orcid.org/0000-0001-9945-2223; Department of Physical Education and Sports Sciences, Faculty of Science and Humanities, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Tamil Nadu 603203, India.
Natarajan, Subhashree: subhashreen@rvu.edu.in; https://orcid.org/0000-0002-2329-5935; School of Business, RV University, RV Vidyanikethan Post, 8th Mile, Mysuru Road, Bengaluru 560059, KA, India.
Lakshmi, Ponnusamy Yoga: yogalakp@srmist.edu.in; https://orcid.org/0000-0002-3461-8226; Department of Computer Science, Faculty of Science and Humanities, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Tamil Nadu 603203, India.
Boby, Farjana Akter: boby.pess@diu.edu.bd; https://orcid.org/0000-0003-3168-6888; Department of Physical Education and Sports Science, Faculty of Health and Life Science, Daffodil International University, Daffodil Smart City, Birulia, Savar, Dhaka-1216, Bangladesh.


Received: 16.02.2024. Accepted: 06.04.2024. Published: 30.04.2024

This work is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0)