

RESEARCH ARTICLE

 OPEN ACCESS**Received:** 08-08-2024**Accepted:** 08-05-2025**Published:** 30-05-2025**Editor:** Special Issue Editors: Dr. N. Pothanna, Dr.T. Jayashree and Dr. V. Ganesh Kumar

Citation: Vadla GK, Bai KJL, Teja YB, Chandra MT, Bharath S (2025) A Methodological Approach Utilising Factor Analysis and Clustering to Evaluate the Strengths and Weaknesses of IPL Players. Indian Journal of Science and Technology 18(SP1): 28-36. <https://doi.org/10.17485/IJST/v18s1.icamada21>

* **Corresponding author.**ganeshkumar_v@vnrvjiat.in**Funding:** None**Competing Interests:** None

Copyright: © 2025 Vadla et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](#))

ISSN

Print: 0974-6846

Electronic: 0974-5645

A Methodological Approach Utilising Factor Analysis and Clustering to Evaluate the Strengths and Weaknesses of IPL Players

Ganesh Kumar Vadla^{1*}, K Jhansi Lakshmi Bai², Yerram Bhanu Teja³, Muvvala Thoran Chandra³, Siripuram Bharath³

¹ Department of Mathematics, VNRVJIET, Hyderabad, Telangana, India² Department of CSE, VNRVJIET, Hyderabad, Telangana, India³ Department of CSE & CSBS, VNRVJIET, Hyderabad, Telangana, India

Abstract

Objectives: To analyze the strengths and weaknesses of IPL players for improvement, and suggest targeted improvements using factor analysis and cluster analysis. Data to keep track of the immense data factors. **Methods:** This study delves into a comprehensive examination of player performance in the Indian Premier League (IPL) using statistical methods. The integrated approaches of data collection, cleaning, and statistical methods like factor and cluster analysis provide a comprehensive framework for IPL player performance data. The partitioning of the data into distinct groups and performing dimensionality reduction techniques are applied. **Findings:** Here, we analyse and visualize the data. The probable weaknesses of batsmen and bowlers are obtained from our analysis. This provides insight into the areas where the player can overcome their weaknesses and offers targeted training programs, thereby enhancing overall team performance in the IPL. Two factors are identified in the variables of data. Players clusters are analysed based on the factors. **Novelty:** This approach can be used to find the performance of individuals and also a comparison between the batsmen and bowlers. The result will be based on many factors and variables retrieved from the dataset.

Keywords: Factor Analysis; Clustering; Data Visualisation; Player weaknesses; Dimensionality reduction

1 Introduction

In recent years, sports analytics has drawn a lot of interest in the scholarly community, offering unprecedented opportunities to gather insights from vast datasets in professional sports such as cricket. This paper presents an exciting opportunity as there is an excess of real-time events that can be analysed to gain better insights. The IPL, founded by the BCCI in 2008, is a highly competitive 20 cricket league. The IPL franchises are competing vigorously during the summer vacation, showcasing the pinnacle of cricketing talent and strategy.

The strength of batters was evaluated using a new methodology that took into account individual performance, the role of the team, and the interactions between the team members⁽¹⁾. Static networks, or the accumulation of game results over time, are the traditional source of suggested ranking systems. The Deep Player Performance Index (DPPI) is recommended for use in the assessment of Twenty20 cricket players based on their performance⁽²⁾. DPPI uses a combination of K-Means clustering and Random Forest algorithm. The purpose of the research was to develop a prediction model that makes use of significant quantitative variables in order to provide a forecast regarding the auction price of a candidate⁽³⁾. Through a comparison of the player's real performance and the auction price paid, this study sheds light on the player's economic worth. A technique of Linear Programming was used for evaluating the efficiency measurement of bowlers, batsmen, and all-rounder in IPL 2019⁽⁴⁾. Agnes and fvizcluster, which distinguish players according to their performances and are used to predict a player's category, have been employed in a hierarchical approach to cluster the players⁽⁵⁾.

The analysis of player performance was conducted utilising the IPL matches and ball-by-ball dataset sourced from Kaggle⁽⁶⁾. Batting performance evaluation utilised adjusted batting scores alongside the stochastic dominance method for comparative analysis, whereas bowling performance assessment employed a metric known as Combined Bowling Rate. Principal Component Analysis was used to uncover cricketing performance indicators that significantly affect IPL 2023 players⁽⁷⁾. Hamed Taherdoost focused his attention on defining the notion of data preparation as well as the process of data analysis. Next, a comprehensive investigation is carried out into the statistical tools that are utilised in the most often employed methodologies, which include descriptive, explanatory, and inferential analysis⁽⁸⁾. The efficacy of a team is commonly assessed through its win-loss ratio, which is impacted by a multitude of factors. The factors influencing the outcome of the game encompass the individual performances of players from both teams, as well as the uncertainties associated with the current situation. Modeling the behavior of a cricket team is a non-trivial task, given that a cricket team can be regarded as a complex social system⁽⁹⁾.

Priyanka et. al. discussed the broader application of machine learning techniques in sports, including cricket, demonstrating how predictive modeling can forecast player performances in IPL matches⁽¹⁰⁾. By applying statistical methods like cluster and factor analysis, researchers can probe into complex player metrics such as strike rate, batting average, bowling rate to analyse the performance of players in past matches⁽¹¹⁾. These methods not only give the performance measures of the players but also highlight the areas of improvement for enhancing the capabilities of player strategies⁽¹²⁾. Rory B and Fadi T presented a critical examination of the academic literature in the field of machine learning, with a particular emphasis on the use of artificial neural networks (ANN) for the prediction of sporting results⁽¹³⁾. For the purpose of predicting a player's performance, Chetan Kapadiya et al. presented a method that would make use of a weather dataset in conjunction with statistics on cricket matches⁽¹⁴⁾.

An approach to batting position prediction based on historical player profiles was suggested⁽¹⁵⁾ by Santra et al. The association between the goal variables and the batsman-associated attributes was determined by exploratory data analysis.

A significant number of studies have been undertaken regarding predictions for cricket matches and assessments of player performance. The conventional ranking methodology relies exclusively on win/loss ratios for teams and players, which are determined by batting averages or strike rates. Numerous studies have been performed on the ranking of cricket players utilising social network analysis; however, these studies have focused solely on individual performances.

This paper contributes to the evolving landscape of sports analytics, where empirical evidence drives strategic decisions and fosters continuous improvement in player performance and team management. This paper aims to conduct a comprehensive examination of IPL player data from the year 2008 to 2023 through data-driven analysis. By leveraging the insights from historical match data, the paper aims to focus on the strengths and weaknesses of players by offering recommendations for training programs and strategic refinements such as player selection, team management, coaching staff, and enhancing overall team performance.

2 Methodology

2.1 Data Set

Data preparation involves the conversion of data into a machine-readable numerical format, enabling its use in specific analysis programs like SAS or SPSS. The data preparation process involves several steps, including data coding, data entry, handling missing values, and data transformation. Data URL: https://www.kaggle.com/datasets/jamiewelsh2/ball-by-ball-ipl?select=ball_by_ball_ipl.csv

The IPL is an annual men's Twenty20 (T20) cricket league held in India. The league consists of ten franchise teams, each representing a different city. The Board of Control for Cricket in India (BCCI) established the league in the year 2008. The competition is typically scheduled to take place during the summer months, specifically between the months of March and May, on an annual basis. It gains very high viewership every year. The data set consists of data from 2008 to 2023 with the size of

239692. We loaded the dataset into a Pandas Data Frame and filtered it to include relevant columns ('batsman', 'batsman_runs', 'dismissal_kind', 'bowler', etc.). We created new features based on the type of runs scored by batsmen ('1s', '2s', '3s', '4s', '6s') and also based on dismissal types ('caught', 'bowled', 'lbw', 'runout') for bowlers.

Table 1. Batsman Runs

batsman	1s	2s	3s	4s	6s	batsman_runs
A Ashish Reddy	83	20	1	16	15	280
A Chandila	4	0	0	0	0	4
A Chopra	21	2	0	7	0	53
A Choudhary	13	1	0	1	1	25
A Dananjaya	2	1	0	0	0	4
A Flintoff	23	2	1	5	2	62
A Kumble	21	1	0	3	0	35
A Mishra	172	22	0	29	5	362
A Mithun	8	2	0	4	1	34
A Mukund	11	2	0	1	0	19

The pre-processed data is then subjected to find insights including characteristics and patterns within the dataset. In EDA descriptive statistics, data visualization techniques for exploring the distribution of responses, identifying outliers, and uncovering potential correlations between variables. Heat maps can show the correlation between different variables in a dataset.

2.2 Approaches

Heat map is used to find the correlation between the features in the data. Figure 1 shows the highest correlation between total_runs and batsman_runs and total_runs and extra_runs. It indicates that total_runs of a particular player depends on the batsman_runs as well as extra_runs. There is a high correlation among total_runs, batsman_runs and extra_runs, these three features must be considered for the prediction.

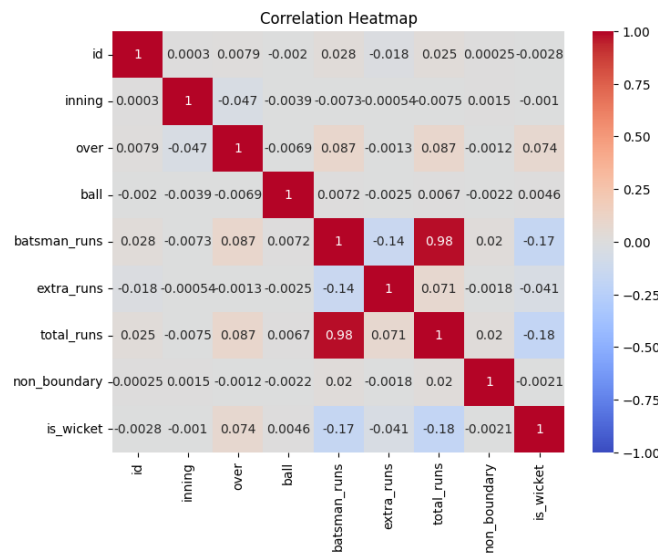


Fig 1. Heat Map for IPL dataset

2.2.1 Clustering

In the context of IPL data analysis, cluster analysis can be applied to identify patterns and groupings within the dataset related to player performance, match dynamics, and team strategies. The goal of cluster analysis in IPL data analysis is to divide the dataset into clusters or groups such that the data points within each cluster are more similar to data points in other clusters.

In our paper, we used Silhouette score to find the number of clusters in the dataset. Objects are ranked according to their degree of similarity to their own cluster (cohesion) and to other clusters (separation). The silhouette score is a metric used to assess the quality of clustering. A higher silhouette score indicates that the object is well-matched to its own cluster and poorly matched to neighbouring clusters. This suggests a clear distinction between clusters, indicating a good separation of data points. Figure 2 represents the silhouette score for 20 clusters, here is highest silhouette score at two clusters.

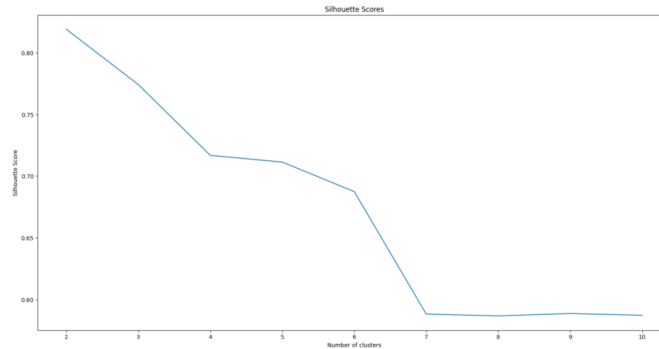


Fig 2. Silhouette methods to find the no. of clusters

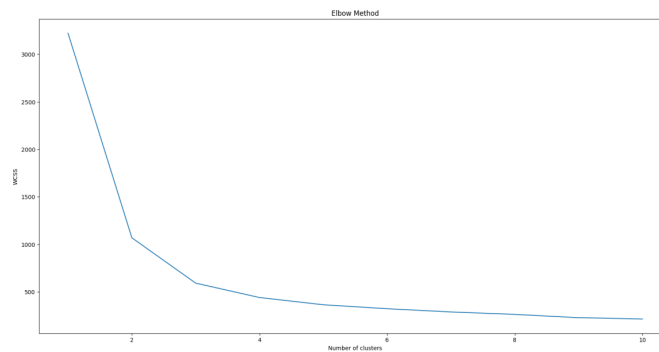


Fig 3. Elbow methods to find the no.of clusters

Figure 3 shows the possible clusters as 2 in the data using elbow method. Elbow method uses within-cluster sum of squares (WCSS) and finds a point where the curve or WCSS having stable variation. By comparing the two methods both are giving the number of clusters as 2. So in this research no.of clusters used in K-means algorithm is 2.

2.2.2 Factor Analysis (FA)

In complicated settings with many variables, the approach described in this paper is a useful tool for analysing and finding patterns and correlations. The intended result is achieved by reducing the variables to either a new variable or a reduced collection of derived factors. The purpose of the approach is not to provide predictions regarding the outcome variables. The term "latent factor analysis" refers to a set of statistical techniques that are employed to determine the underlying factors that influence observable variables⁽¹⁶⁾. The aggregated data was standardized using StandardScaler to ensure that all features have a consistent scale. This is crucial for clustering algorithms such as K-means. The K-means clustering algorithm was utilized to group batsmen or bowlers based on their performance metrics after standardizing the data.

By loading the dataset into a Pandas Data Frame, filtered it to include only the relevant columns ('batsman', 'bowler', 'batsman_runs') and also ('batsman', 'bowler', 'dismissal_count'). These factor scores represent the contribution of each factor to the observed variation in batsman runs against bowlers and dismissal counts variation against batsman.

3 Results and Discussion

In the framework of the Indian Premier League (IPL), the interactions between individual batters and bowlers usually fascinate cricket fans. An exemplary matchup that has garnered attention is the one between Virat Kohli and Jasprit Bumrah. Often facing up against Bumrah, a bowler noted for his great accuracy and deadly yorkers, Kohli is well-known for his aggressive batting approach and constant performance. Kohli has gathered a significant number of runs against Bumrah over several seasons, proving his ability in managing even the most difficult deliveries. The contest between Kohli and Bumrah has witnessed Kohli's ability to adjust his technique, frequently employing innovative shot selection to effectively counter Bumrah's diverse range of pace and swing variations. The interactions between teams in the IPL highlight the extraordinary talent of the players, which generates thrilling events and vital runs that usually have a major influence on the flow of the game.

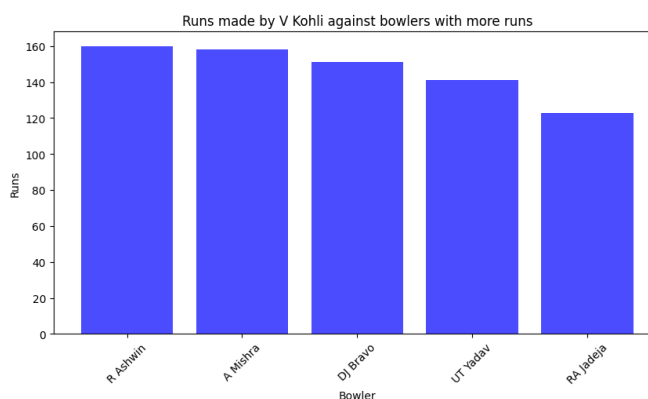


Fig 4. R uns against bowlers for V Kohli

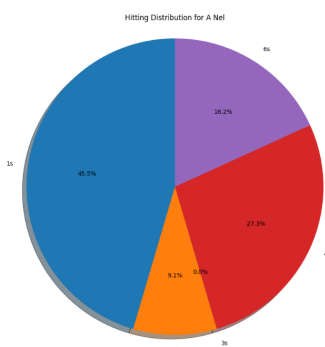


Fig 5. Hitting distribution of user defined input player

As per Figure 5, V Kohli made more of his IPL runs in spin bowling comparatively than fast bowling. Hitting distribution of Virat Kohli at IPL is shown in Figure 4. It shows that he scores more singles than boundaries. Also, it tells that he scores more fours than sixes in IPL.

The dismissal rate of a batsman in the IPL is an important metric that indicates their level of consistency and reliability when facing pressure. The dismissal distribution in the IPL offers a comprehensive breakdown of how a bowler obtains wickets using different methods of dismissal. The distribution of dismissals in cricket typically includes categories such as caught, bowled, stumped, LBW, and run-out. The analysis of these patterns aids in the comprehension of a bowler's strategies and strengths.

Figure 7 and Figure 8 deals with the bowlers' statistics in IPL. Particularly, most of the dismissal of batsmen by Nehra are caught which is shown in Figure 7. Figure 8 explains the extras given by bowler in IPL. It confirms that A Zampa had given extras in the form of wides than remaining.

When applied to data from the IPL, factor analysis is a statistical tool for discovering hidden connections between different performance indicators. It is most helpful when attempting to identify a limited set of latent factors that account for the vast majority of the variation in a dataset containing many observed variables (such as batting average, strike rate, amount of sixes,

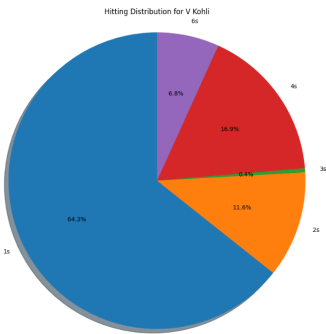


Fig 6. Bowler runs distribution

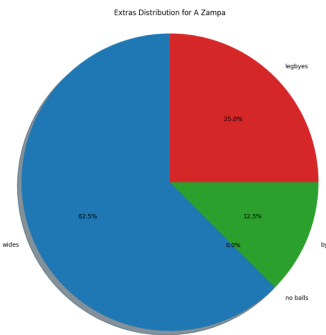


Fig 7. Dismissal distribution for bowler

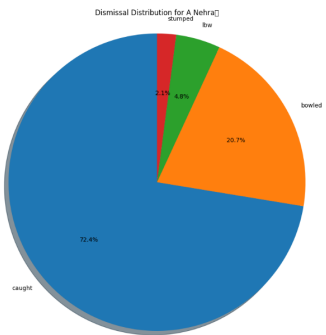


Fig 8. Extras distribution for bowler

Table 2. Two Factors in the considered data		
Variable	Factor 1 (Aggressive batting)	Factor 2 (Bowling Efficiency)
Runs	0.943	0.028
Strike Rate	0.380	-0.360
Sixes	0.987	0.002
Wickets	-0.069	0.692
Economy	0.083	-0.434

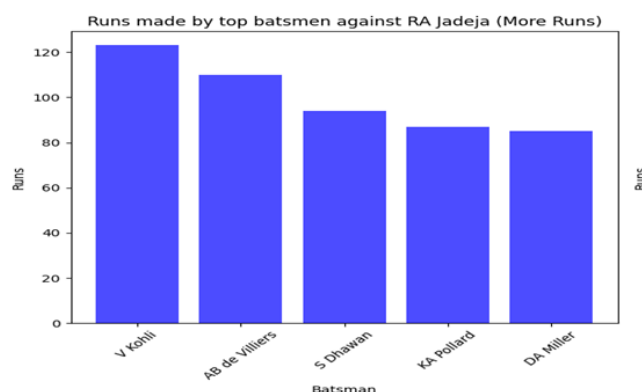


Fig 9. R uns made by batsman against RA Jadeja

wickets, economy rate, etc.). Using Factor analysis, we identified the hidden patterns, player's strengths and roles. Factor 1 heavily loads on Runs and Sixes, suggesting it represents Aggressive Batting Performance. Factor 2 has high loadings on Wickets (positive) and Economy (negative), suggesting it's tied to Bowling Efficiency, high wicket-taking with low runs conceded.

The ML experiments are conducted using the Weka tool, which is a widely used ML suite. The Simple K Means cluster analysis algorithm is being utilized with a specified number of clusters set to 2. The suggested method helps the coach to choose several playing techniques based on their fit for a particular game. This feature distinguishes our approach from the several player statistics available on specialised websites. Often with the aim of estimating betting chances, generic assessments of teams are produced using player statistics.

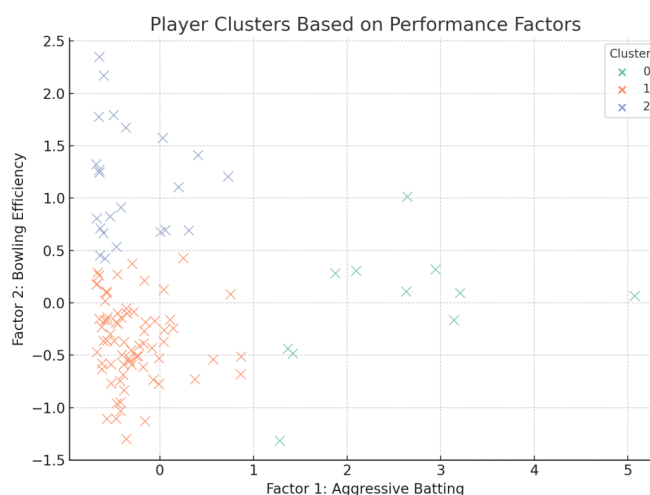


Fig 10. Cluster Analysis of IPL players based on their batting and bowling factors

Figure 10 shows three distinct clusters based on player styles.

- **Cluster 0:** Likely heavy-hitting batters (high Factor 1).
- **Cluster 1:** Likely effective bowlers (high Factor 2).
- **Cluster 2:** Balanced or low-performing players across both dimensions.

ICC ranking system is too complex to be used in a subcontinental competition. It has been noted in the literature that Arijith Ghosh et al. ⁽¹⁶⁾ used SEM and DEA to gauge the effectiveness of bowlers, batsmen, and all-rounders competing in the Indian Premier League 2019. They have studied about the ranking of players by measuring the efficiency of players. Falak Bharadwaj et al. ⁽¹⁷⁾ done comparative study of IPL player performance using different methodologies like decision tree, SVM. In the present study, a complete dataset including in-depth, ball-by-ball details from Indian Premier League (IPL) cricket matches is regarded as data. This dataset records numerous important characteristics for each ball used in the contests and offers a detailed

perspective of every delivery bowled throughout several IPL seasons(2008 to 2023). Weakness and strength of players has been analysed using factor analysis and clustering.

4 Conclusion

This research revealed player weakness, mostly related to bowlers who had experienced more frequency of boundaries and sixes. To increase total team performance in high-pressure moments during IPL matches, clubs and coaches must have this analysis in order to diversify ways to strengthen their batting and bowling lineup. The dismissal rate of a batsman helps us know their level of consistency and reliability when facing pressure. The dismissal distribution in the IPL offers a comprehensive breakdown of how a bowler obtains wickets using different methods of dismissal.

Through the use of latent factors such as batting impact and bowling efficiency, factor analysis simplifies complicated player data by reducing several variables (e.g., runs, strike rate, wickets) to a manageable number. It helps interpret what really drives performance. Factor 1 indicates aggressive batting, irrespective of its documentation in individual statistics. Minimises multicollinearity, defined as the correlation among predictors, thereby enhancing model accuracy for predictive tasks such as forecasting performance or calculating fantasy points.

Cluster analysis categorises players into distinct groups such as Power Hitters, Strike Bowlers, and All-rounders. Applicable for recruitment or document preparation. Teams can identify and focus on particular clusters according to their requirements (for instance, requiring a Cluster 1-type bowler to enhance performance in death overs).

Improving existing models and exploring new parameters to boost forecast accuracy could be the focus of future research. Sports analytics in cricket and beyond can benefit from new insights into player performance dynamics that could emerge from investigating deep learning techniques and using real-time data streams.

5 Acknowledgement

This paper has been presented in “International Conference on Applied Mathematics and Advanced Data Analytics for Industry 5.0 (ICAMADA-2024)” held from 25th to 27th April 2024 at VNR VJIEET, in collaboration with APTSMS and CCOE. The authors express their sincere gratitude to the guest editors for their tireless efforts and dedication by means of comments, editing and overseeing the manuscripts to bring in this final form.

The APC is deferred partially by the Indian Society for Education and Environment.

The authors sincerely thank the Principal, Director and management of VNRVJIEET, Hyderabad, India for providing continuous support and encouragement towards the research work.

References

- 1) Manju MK, Philip AO. Novel method for ranking batsmen in Indian Premier League. *Data Science and Management*. 2023;6(3):158–173. Available from: <https://dx.doi.org/10.1016/j.dsm.2023.06.004>.
- 2) Prakash CD, Verma S. A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket. *Decision Analytics Journal*. 2022;2. Available from: <https://www.sciencedirect.com/science/article/pii/S2772662222000029>.
- 3) Malhotra G. A comprehensive approach to predict auction prices and economic value creation of cricketers in the Indian Premier League (IPL). *Journal of Sports Analytics*. 2022;8(3):149–170. Available from: <https://dx.doi.org/10.3233/jsa-200580>.
- 4) Ghosh A, Dey M, Guha B, Jana S, Sarkar A. Performance evaluation & rankings of players in IPL 2019 by DEA & SEM. *Malaya Journal of Matematik*. 2021;1:46–56. Available from: https://www.researchgate.net/publication/350252203_Performance_evaluation_rankings_of_players_in_IPL_2019_by_DEA_SEM.
- 5) Mansurali A, Harish V, Hussain S, Choudhury T. Profiling the IPL Players—Sports Analytics Through Clustering Algorithms. In: *Machine Intelligence and Data Science Applications*. 2023;p. 53–64. Available from: https://doi.org/10.1007/978-981-99-1620-7_5.
- 6) Patel NJ. Statistical Analysis of IPL Player Performance using Advanced Computational Methods. *International Journal of Computer Applications*. 2024;186(30):0975–8887. Available from: <https://www.ijcaonline.org/archives/volume186/number30/patel-2024-ijca-923882.pdf>.
- 7) Dhakre DDS, Bhattacharya. Cricketing Performance Analysis in IPL 2023: Unravelling Key Influencing Factors through Principal Component Analysis and Biplot Clustering. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*. 2023;9(9):18–23. Available from: <https://ijseas.com/volume9/v9i9/IJSEAS202309103.pdf>.
- 8) Taherdoost H. Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects. *International Journal of Academic Research in Management (IJARM)*. 2022;9(1):1–9. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4178680.
- 9) Neuman Y, Israeli N, Vilenchik D, Cohen Y. The Adaptive Behavior of a Soccer Team: An Entropy-Based Analysis. *Entropy*. 2018;20(10). Available from: <https://dx.doi.org/10.3390/e20100758>.
- 10) Priyanka S, Vysali K, Priyaiyer KB. Prediction of Indian Premier League-IPL 2020 using Data Mining Algorithms. 2020. Available from: <https://dx.doi.org/10.22214/ijraset.2020.2121>.
- 11) Raju N, Ashok R. Measuring the Performance of the Indian Premier League Teams through an Integrated Optimality Analytics by Data Envelopment Analysis Approach. *Mathematical Statistician and Engineering Applications*. 2022;71(4):8918–8940. Available from: https://www.researchgate.net/publication/381486493_Measuring_the_Performance_of_the_Indian_Premier_League_Teams_through_an_Integrated_

[Optimality_Analytics_by_Data_Envelopment_Analysis_Approach.](#)

- 12) Bunker R, Susnjak T. The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review. *Journal of Artificial Intelligence Research*. 2022;73:1285–1322. Available from: <https://dx.doi.org/10.1613/jair.1.13509>. doi:10.1613/jair.1.13509.
- 13) Bunker RP, Thabtah F. A machine learning framework for sport result prediction. *Applied Computing and Informatics*. 2019;15(1):27–33. Available from: <https://dx.doi.org/10.1016/j.aci.2017.09.005>.
- 14) Kapadiya C, Shah A, Adhvaryu K, Barot P, and. Intelligent Cricket Team Selection by Predicting Individual Players' Performance using Efficient Machine Learning Technique. *International Journal of Engineering and Advanced Technology*. 2020;9(3):3406–3409. Available from: <https://dx.doi.org/10.35940/ijeat.c6339.029320>.
- 15) Santra A, Sinha A, Saha P, Das AK. A Novel Regression based Technique for Batsman Evaluation in the Indian Premier League. In: and others, editor. 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), 05-06 September 2020, Kolkata, India. IEEE. 2020;p. 379–384. Available from: <https://ieeexplore.ieee.org/document/9290569>.
- 16) Taherdoost H, Samsul S, Neda J. Exploratory Factor Analysis; Concepts and Theory. *Advances in Applied and Pure Mathematics*. 2022;27:375–382. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4178683.
- 17) Bharadwaj F, Saxena A, Kumar R, Kumar R, Kumar S, Željko Stević. Player Performance Predictive Analysis in Cricket Using Machine Learning. *Revue d'Intelligence Artificielle*. 2024;38(2):449–457. Available from: <https://dx.doi.org/10.18280/ria.380208>.