IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Unsupervised Classification Of Individual Batting Performances In The Ipl: A Data-Driven Approach

[1]Vineetha S Das, [2]Jayashankar P

[1]Associate Professor, [2]M.Tech student
[1]Mechanical Engineering Department,
[1]College of Engineering, Trivandrum, India

Abstract: This study investigates the application of unsupervised learning techniques, specifically K-means and hierarchical clustering, to classify individual player batting performance in the Indian Premier League (IPL). Leveraging match-level data from the 2008 to 2024 seasons, player performance was analyzed based on three distinct feature sets: runs scored, a combination of runs and strike rate, and a novel score metric (runs multiplied by strike rate). To ensure robust classification, only innings where a player faced at least 10 balls or was dismissed were considered. A total of six clustering experiments were conducted, and the resulting cluster quality was evaluated using the silhouette score. Comparative analysis revealed that K-means clustering, when applied to the score metric, yielded the most discernible and well-separated clusters of batting performance. This research demonstrates the potential of unsupervised learning in categorizing player batting performance in the high-stakes environment of the IPL, offering valuable insights for team management, player evaluation, and strategic decision-making.

Key words- K-means, Hierarchical clustering, score metric, performance

I. INTRODUCTION

Analysing individual batting performances within a match in the Indian Premier League (IPL) provides critical insights into a player's contribution to that game. Key indicators of a batsman's performance in an innings are the total runs they score and the rate at which they score them (strike rate). These two factors largely define the impact of an innings on the match outcome. A batsman might score a high number of runs at a slow pace, or score fewer runs very quickly, each having different implications . In limited-overs cricket, slow run-scoring, even without a batsman losing their wicket, is generally detrimental and can lead to a team's defeat rather than victory [1]. Therefore, classifying individual batting performances based on these fundamental metrics is crucial for a comprehensive evaluation.

Traditional methods often look at runs and strike rate in isolation or as averages over a season. However, clustering techniques can help identify natural groupings of innings that exhibit similar combinations of runs and strike rate. This can reveal distinct types of batting performances that might not be obvious through simple statistical summaries. By applying unsupervised learning, the study categorizes individual innings based on the runs scored and the efficiency of scoring within that innings.

This study explores the application of K-means and hierarchical clustering to classify individual player batting performances within matches in the IPL. The study analyses match-level data from the 2008 to 2024 seasons, focusing on innings where players faced at least 10 balls or were dismissed. By utilizing three feature sets for the analysis: the total runs scored in the innings, the strike rate achieved in the innings, and a score metric derived from both (runs multiplied by strike rate). The silhouette score is used to evaluate the quality of the clusters obtained from each method and feature set. This research aims to demonstrate the effectiveness of unsupervised learning in categorizing in-match batting performances in the IPL based on their run-scoring and strike rate characteristics, providing a valuable tool for analysing and understanding the variety of batting contributions in the league.

II. LITERATURE SURVEY

Barr and Kantor (2004) proposed a novel criterion for comparing and selecting batsmen in limited-overs cricket, emphasizing that traditional batting averages alone are insufficient due to the time dimension of the game. They introduce a two-dimensional graphical representation, akin to a risk-return framework, plotting strike rate against the probability of getting out to gain direct and comparative insights into batting performance. Within this framework, they develop a selection criterion that combines both batting average and strike rate, illustrating its application with data from the 2003 World Cup.

Hermanus H. Lemmer (2004) significantly contributed to cricket performance analysis by proposing a single, comprehensive measure for batting, moving beyond traditional averages to encompass consistency and strike rate. This work underscored the critical importance of strike rate in limited-overs formats, arguing for its inclusion in evaluating a batsman's overall impact. This foundational research directly informs the present study's development of a combined score metric and its application of unsupervised learning to categorize individual batting performances in the T20 format.

Amin and Sharma (2014) proposed a two-stage method combining Ordered Weighted Averaging (OWA) operators and regression to measure and rank batting parameters in Twenty20 (T20) cricket. Their study, utilizing data from 40 batsmen in the 2011 Indian Premier League (IPL), identified Strike Rate (S/R) as the most important batting parameter, followed by Highest Score (HS) and Average (Avg). A key finding was that the ranking of batting parameters remained insensitive to changes in OWA weights, reinforcing the robustness of their findings. This research provides a scientific basis for prioritizing batting attributes in the T20 format, aligning with the current study's emphasis on strike rate and its combined impact with runs for effective player performance classification.

Rath (2021) conducted a study on the application of the K-means clustering algorithm for classifying cricket players based on their career statistics, specifically runs scored and wickets taken. The research aimed to categorize players into distinct roles such as batsmen, bowlers, and all-rounders, demonstrating the utility of K-means in identifying inherent player capabilities. The findings suggested that K-means effectively grouped players with similar statistical profiles, providing a foundation for team selection processes.

Sumathi and Prabu (2023) presented a system for cricket player performance prediction and evaluation using a combination of machine learning algorithms, including linear regression, K-means clustering, and random forest models. Their work aimed to analyze player performance to aid in team formation and training, with K-means specifically used to classify players into 'n' clusters based on shared characteristics. The study demonstrated that K-means effectively grouped players by their total runs, with the best-performing cluster identified for selecting top players.

Sanjay Raajesh et al. (2024) addressed the crucial need for data-driven precision in cricket team selection and player analysis. Their project utilized the K-means clustering algorithm to categorize players based on historical performance data, with a specific focus on strike rate.

Jhansi Rani et al. (2020) explored the combined application of neural networks with K-means and hierarchical clustering to predict the best batsman or bowler for specific match conditions in the Indian Premier League (IPL). Their work focused on ball-by-ball analysis to inform real-time strategic decisions, aiming to optimize team performance. The study found that a hybrid approach, combining K-means and hierarchical clustering, yielded improved accuracy in player categorization, thereby aiding in the selection of an ideal team composition.

Jayanth et al. (2018) developed a comprehensive system for cricket match outcome prediction, team structure analysis, and player recommendation. Their work utilized K-means clustering as an unsupervised learning method to group similar players based on their past performance statistics, which then informed a K-Nearest Neighbor (KNN) classifier to recommend a preferred role for new or less-experienced players. The study highlighted the importance of player profiling and team composition in predicting match outcomes, and it demonstrated the utility of clustering in identifying player archetypes.

III. METHODOLOGY

The data required for the study has been collected from the website cricsheet [9]. The runs and strike rate for each batsman inning were the 2 features required. The proposed metric, named the score metric, was calculated for each batsman inning. The three metrics on which clustering was done are:

1) Runs
2) Runs and strike rate
3) Score metric = Runs * Strike rate

Both K-means and hierarchical clustering were applied to the 3 metrics. To objectively evaluate the performance of the various clustering approaches and feature sets, the silhouette score was employed. The silhouette score measures how similar an object is to its own cluster compared to other clusters, with values ranging from -1 to +1. A score closer to +1 indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters, suggesting a good clustering structure.

IV. RESULT AND DISCUSSION

Fig. 1 illustrates the K-means clustering results when individual batting innings are grouped solely based on the total runs scored. As observed, the innings are broadly categorized into three distinct clusters, which can be interpreted as "Poor", "Good", and "Excellent" innings based on their run contribution.

While this clustering provides a straightforward categorization based on run volume, it presents a significant drawback, particularly in the context of Twenty20 (T20) cricket. The inherent limitation is that innings with the same number of runs, but vastly different strike rates, are grouped into the same cluster. In the fast-paced and limited-overs format of T20 matches, the strike rate (runs scored per 100 balls faced) is a critically important metric. A batsman scoring 30 runs off 15 balls (strike rate 200) has a fundamentally different impact on a match compared to a batsman scoring 30 runs off 30 balls (strike rate 100), even though their run contribution is identical. The former represents an aggressive, quick-scoring innings vital for accelerating the run rate, while the latter might signify a slower, anchor-type role, which can sometimes be detrimental if the required run rate is high.

Therefore, while clustering by runs provides a basic performance segregation, it fails to capture the efficiency and pace of scoring, which is paramount in T20 cricket. This highlights the necessity of incorporating strike rate, or a combined metric, to achieve a more nuanced and contextually relevant classification of batting performances.

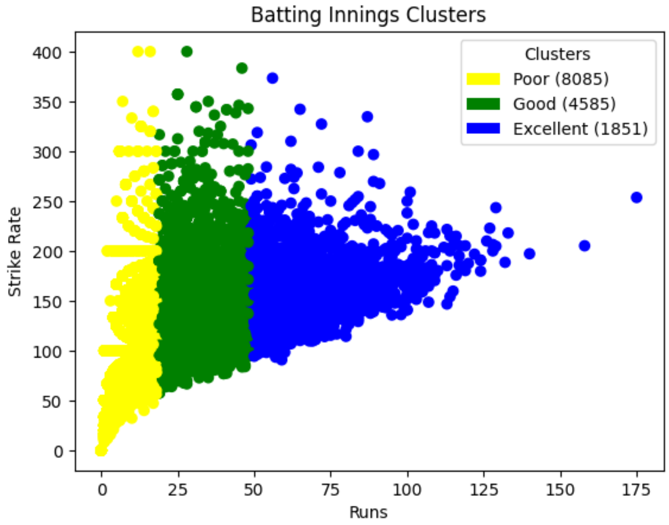

Fig.1 K-means clustering result for runs

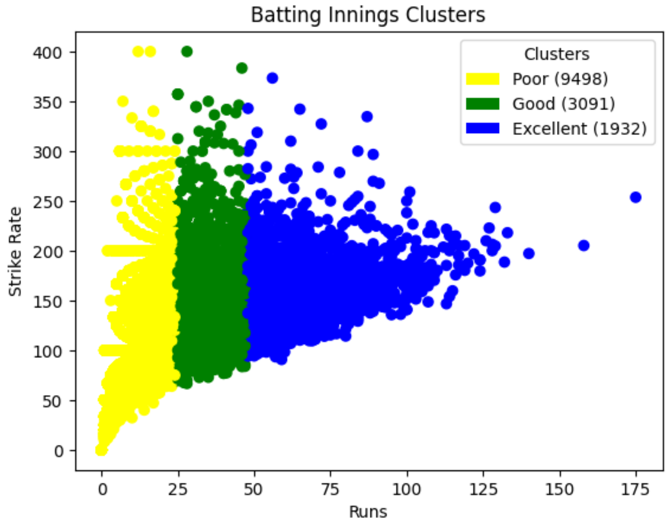

Fig.2 Hierarchical clustering result for runs

Fig. 3 illustrates the K-means clustering results when individual batting innings are classified using both runs scored and strike rate as input features. This approach was adopted to address the limitations observed in the runs-only clustering, which failed to account for the crucial element of scoring pace in T20 cricket. The K-means algorithm effectively groups innings where higher runs are consistently combined with a higher strike rate, forming distinct and intuitively interpretable clusters.

In contrast to the K-means approach, the hierarchical clustering (Fig.4) results for the same 'Runs and Strike Rate' features presented less desirable and occasionally counter-intuitive groupings. A notable observation is that within the range of 25 to 50 runs, some innings characterized by demonstrably higher strike rates were categorized into a 'good innings' cluster, while other innings exhibiting lower strike rates were unexpectedly assigned to an 'Excellent' innings cluster. This inconsistency suggests that the hierarchical algorithm, in this specific application, did not consistently align its clustering with an intuitive performance hierarchy where higher efficiency (strike rate) combined with a reasonable run count should elevate an innings' perceived quality. Such anomalies indicate that the hierarchical clustering method struggled to robustly differentiate and rank batting performances based on the combined impact of runs and strike rate in a manner that fully captures their significance in the T20 format.

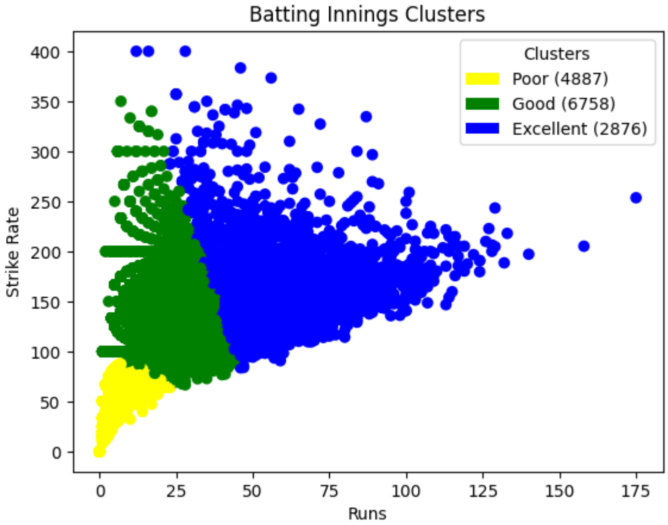

Fig.3 K-means clustering result for runs and strike rate

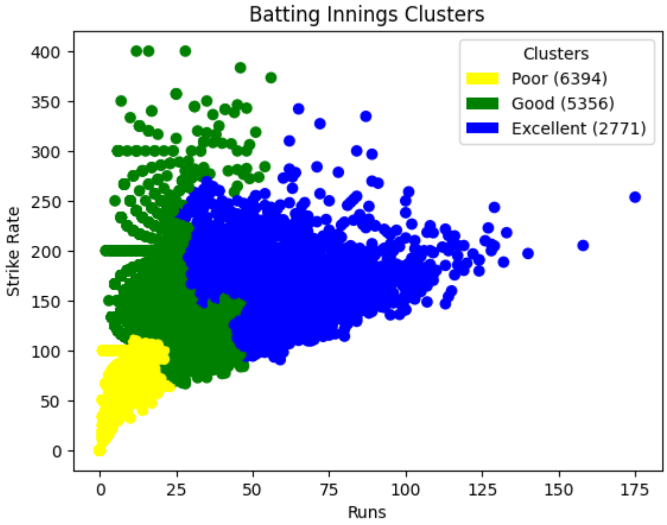

Fig.4 Hierarchical clustering result for runs and strike rate

Clustering based on the proposed score metric (Runs * Strike Rate) yielded the most insightful and intuitively correct classifications of individual batting innings, as depicted in the Fig.5 and Fig.6. A key advantage of this metric is its ability to inherently address the limitation observed with the 'runs-only' approach: innings with the same number of runs but a higher strike rate are appropriately grouped into a higher performance class. This characteristic is crucial in T20 cricket, where the efficiency of scoring holds as much significance as the volume of runs. The score metric effectively synthesizes both aspects, ensuring that an innings' classification reflects its true impact, accounting for both accumulation and acceleration.

Both K-means clustering and hierarchical clustering exhibited a similar positive trend when applied to the score metric, demonstrating that this feature space inherently promotes better separation of performance levels. The clusters formed across both algorithms generally show a logical progression from lower to higher impact innings. However, a subtle but notable difference was observed: hierarchical clustering grouped a comparatively smaller number of innings as 'excellent' compared to K-means. This suggests a potentially stricter or more conservative boundary definition by hierarchical clustering for the highest performance tier, possibly due to its agglomerative nature, which can lead to larger, more encompassing clusters at higher levels of the hierarchy, thus consolidating some truly excellent innings into the 'Good' category. Nevertheless, the fundamental principle of assigning higher-impact innings (defined by a higher score metric) to superior performance classes remained consistent across both algorithms, validating the strength of the proposed score metric as a powerful discriminator for T20 batting performance.

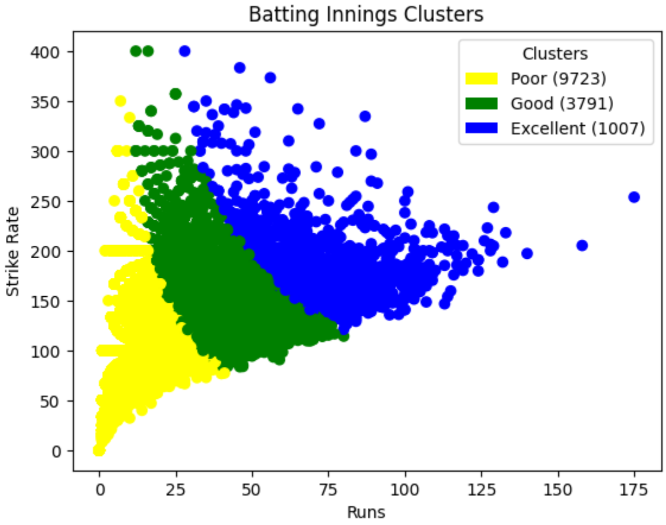

Fig.5 K-means clustering result for score metric

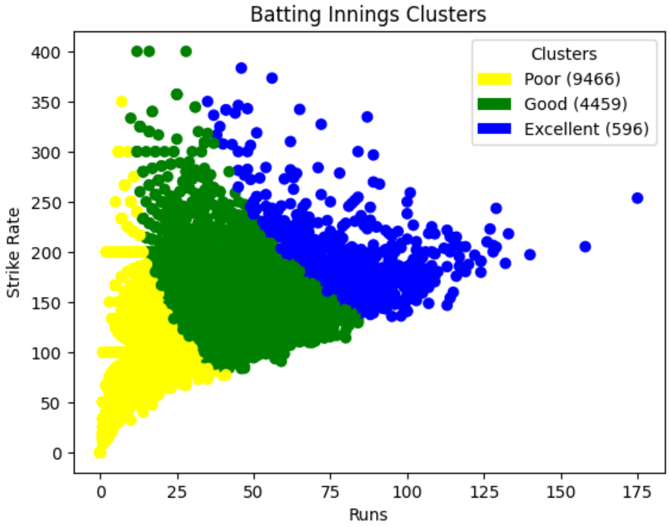

Fig.6 Hierarchical clustering result for score metric

As evident from Table.1, the clustering based on the 'Score Metric' consistently yielded the highest silhouette scores for both K-means (0.65) and Hierarchical Clustering (0.64). Furthermore, K-means clustering generally outperformed Hierarchical Clustering across all feature sets, albeit by a small margin in most cases. The lowest silhouette scores were observed when clustering was performed using 'Runs and Strike Rate' as independent features, indicating less distinct or well-separated clusters in those configurations.

Table .1 Silhouette score for each clustering

	K-means clustering	Hierarchical clustering
Runs	0.60	0.58
Runs and strike rate	0.44	0.40
Score metric	0.65	0.64

The evaluation of clustering performance using the silhouette score provided clear insights into the efficacy of different unsupervised learning configurations for classifying IPL individual batting performances. The consistently higher silhouette scores for the 'Score Metric' (Runs * Strike Rate) across both K-means (0.65) and Hierarchical Clustering (0.64) indicate that this combined feature provides a more robust and discriminative basis for grouping batting innings. This suggests that considering both the volume of runs and the pace at which they are scored, as encapsulated by the score metric, leads to more coherent and meaningful clusters of batting performance. An innings with a high score metric is indicative of both significant run accumulation and efficient scoring, allowing for a better differentiation of performance types.

Between the two algorithms, K-means clustering generally demonstrated slightly superior silhouette scores compared to hierarchical clustering. While the differences were marginal for the 'Runs' and 'Score Metric' feature sets, the slight edge for K-means, combined with its computational efficiency for larger datasets, supports its selection. The highest overall silhouette score of 0.65 achieved by K-means clustering when applied to the 'Score Metric' strongly supports its effectiveness in identifying well-defined and distinct groups of batting performances. A silhouette score of 0.65 is generally considered to represent a fairly good separation between clusters, suggesting that the identified groups are internally cohesive and externally well-separated.

Conversely, the relatively lower silhouette scores for the 'Runs and Strike Rate' feature set (0.44 for K-means, 0.40 for Hierarchical) suggest that treating runs and strike rate as separate, uncombined features might lead to less distinct or overlapping clusters. This could be because innings with similar runs but vastly different strike rates, or vice-versa, might be grouped when they conceptually belong to different performance categories. The 'Score Metric' effectively consolidates these two dimensions into a single, more informative proxy for overall batting impact within an innings, leading to superior clustering outcomes.

V. CONCLUSION

This study successfully demonstrated the efficacy of unsupervised learning techniques in classifying individual batting performances within the Indian Premier League. By applying K-means and hierarchical clustering to a comprehensive dataset of IPL innings from 2008 to 2024, distinct performance categories were identified across various feature sets. The comparative analysis, rigorously evaluated using the silhouette score, unequivocally established that K-means clustering, when applied to a novel 'score metric' (runs multiplied by strike rate), yielded the most robust and interpretable clusters. This finding underscores the critical importance of integrating both run accumulation and scoring efficiency for a comprehensive assessment of batting impact in the fast-paced Twenty20 format, overcoming the limitations of single-dimensional metrics.

Future research endeavours could extend this methodology by exploring additional unsupervised learning algorithms, such as Gaussian Mixture Models, to potentially uncover more granular performance distinctions. Incorporating a broader array of contextual factors, including pitch conditions, opposition strength, and specific match scenarios (e.g., powerplay, death overs), would further enrich the classification model. Moreover, the identified performance clusters could serve as foundational input for supervised learning models aimed at predicting future player performance or optimizing team selection strategies. Expanding this analytical framework to other limited-overs leagues or international cricket formats would also provide valuable comparative insights into global batting trends.

REFERENCES

[1] Barr, G. D. I., & Kantor, B. S. (2004). A criterion for comparing and selecting batsmen in limited overs cricket. Journal of the Operational Research Society, 55(12), 1266-1274.

[2] Lemmer, H. H. (2004). A measure for the batting performance of cricket players. South African Journal for Research in Sport, Physical Education and Recreation, 26(1), 55-64.

[3] Amin, G. R., & Sharma, S. K. (2014). Measuring batting parameters in cricket: A two-stage regression-OWA method. Measurement, 53, 56-61.

[4] Rath, N. K. K-means clustering algorithm analysis.

[5] Sumathi, M., Prabu, S., & Rajkamal, M. (2023, April). Cricket players performance prediction and evaluation using machine learning algorithms. In 2023 international conference on networking and communications (ICNWC) (pp. 1-6). IEEE.

[6] Raajesh, S., Martin, N., Jiji, J., & Nair, A. (2024, May). Cricket Team Selection and Player Analysis using Data Analytics. In 2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 1-6). IEEE.

[7] Rani, P. J., Kamath, A. V., Menon, A., Dhatwalia, P., Rishabh, D., & Kulkarni, A. (2020, July). Selection of players and team for an indian premier league cricket match using ensembles of classifiers. In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-6). IEEE.

[8] Jayanth, S. B., Anthony, A., Abhilasha, G., Shaik, N., & Srinivasa, G. (2018). A team recommendation system and outcome prediction for the game of cricket. Journal of Sports Analytics, 4(4), 263-273.

[9] https://cricsheet.org/downloads/#experimental