Unsupervised Machine Learning for Bowler Performance Analysis in Indian Premier League (IPL) and Cross-League Prediction

Vijay Kumar Varadarajan [1], P. S. Metkewar [2], Dhananjay S. Deshpande [3,*], Farhaan Bhola [4], Anuja Bokhare [2], and Deshinta Arrova Dewi [1]

[1] Faculty of Data Science and Information Technology, Department of Computer Science and Engineering, INTI International University, Nilai, Malaysia
[2] School of Computer Science and Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India
[3] MBAESG, School of Management, Ajeenkya D Y Patil University, Pune, India
[4] Faculty of Computer Studies, Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Pune, India
Email: vijayakumar.varadarajan@adypu.edu.in (V.K.V.); pravin.metkewar@gmail.com (P.S.M.); drdeshpande.dhananjay@gmail.com (D.S.D.); farhaanrbhola@gmail.com (F.B.); anuja.bokhare@gmail.com (A.B.); deshinta.ad@newinti.edu.my (D.A.D.)
*Corresponding author

***Abstract*—The selection of players in cricket, particularly in formats like the Indian Premier League (IPL), is crucial for a team's success. This process involves analyzing various player attributes, including their batting and bowling performances, to create a balanced team. In cricket, each player's role—batsman, bowler, or all-rounder—must be clearly defined. This clarity helps in optimizing team dynamics and ensuring that players can perform at their best. For instance, a well-rounded team might include specialists who excel in specific areas, during matches. Recent advancements in Artificial Intelligence (AI) have opened new avenues for player analysis. Unassisted AI models can evaluate bowlers based on their historical performances in the IPL. By leveraging data from past matches, these models can predict future performances, aiding selectors in making informed decisions about player selection. This study employs predictive modelling to analyze bowler performance using unsupervised learning. The developed AI model can extend its applicability beyond the IPL, offering insights into bowlers from other leagues. There are some discriminative features of players' performance as well as team strategy, which can be obtained by applying machine learning in Bowler Performance Analysis. This data-driven approach provides a decision-making and fast response possibility to managers and analysts in playing T20 matches 2018. By combining various techniques, e.g., K-means and mean-shift, teams are also equipped to select players more effectively, establish strategic and success rate, etc.**

***Keywords*—sports analytics, bowler performance, unsupervised Machine Learning (ML), k-means clustering, cross-league prediction, process innovation, silhouette score, clustering algorithms, machine learning in sports**

I. Introduction

The Indian Premier League (IPL) has evolved in just a few years to be one of the most extravagantly viewed and conducted sporting events globally. These interests are borne out in a few delicious numbers: IPL 2023 reached an estimated audience of over 500 million viewers around the world with a significant growth in digital viewership via streaming offerings like JioCinema. At the same time, the final match of IPL 2023 was watched by 120 million viewers breaking all the records. The estimate of IPL's total revenue generation in 2023 was around $1.5 billion USD, with a large share being generated from selling media rights, sponsorships, and merchandise.

Cricket is one of the most expensive sports in the world, and India hosts one of the largest fan clubs. Information Investigation has tracked down its direction into cricket. In addition, it helps teams make more informed decisions and prepare matchups to understand how one team will play against another. With a combination of expertise, technique, and engagement, the Indian Head Association (IPL) is driving this evolution. In the IPL, bowlers are key to a team's success, and study of their performance is crucial while choosing a team and developing an approach. This research paper attempts to bridge this gap using AI algorithms.

In the case of robotizing cricket crew selection, Vetukuri *et al.* [1] have developed a general model that uses recurrent neural networks. This general model helps in providing the team and tactical preparation insight. Very recently, Kapadiya *et al.* [2] have presented an algorithm to predict player performance in cricket. Study completed by Vetukuri *et al.* [1] and Kapadiya *et al.* [2] plays a very major role in our study in order to evaluate the outcomes. This would consider elements like the

Manuscript received November 21, 2024; revised January 15, 2025; accepted February 19, 2025; published December 5, 2025.

playing surface and the weather that can significantly affect players' skills.

For the in-depth study of the performance of bowlers in the Indian Premier League, this research paper will be using unsupervised machine learning techniques, specifically K-means clustering and Mean Shift cluster analysis. Unsupervised learning helps identify the pattern of the bowler's performance. The outcome helps to find distinct groups of bowlers with the same feature.

Thus, our research paper aims to bridge this gap by employing K-means clustering and Mean Shift cluster analysis for bowler performance analysis in IPL and extending the model for cross-league prediction. As it enables the recognition of patterns and data from unlabelled or unclassified data, unsupervised learning techniques—more specifically, clustering techniques—are utilized broadly in bowling analytics. Owing to the quantity of performance data that are being gathered and owing to the fact that most of these data are poorly classified in a way that would expose the analysts to actionable information, this activity is broadly linked with sports.

To meet the growing demands of competitive decision-making in cricket, advanced AI tools are increasingly being utilized.

II. Literature Review

Cricket Data Analysis and Team Selection Research Use of machine learning and data analytics in cricket analysis are the emerging area of study. Many studies have worked with outcome prediction, optimum selection of a team, and even with decision-making processes within the game of cricket. Jaswanth *et al.* [3] explores how different performance metrics impact the prediction of the outcomes of IPL matches. They pointed out that player statistics, batting average, bowling economy rate, and strike rate determine match results. It uses historical performance data, player stats, match conditions, and a myriad of other factors to model and forecast match outcomes. However, predicting the game is a complex process in itself as the game itself is fairly random in nature, but with the right data and techniques, you are bound to be able to predict the scores. The research process involved the use of machine learning algorithms to process historical data to make more accurate predictions of future match outcomes [3]. The focus on IPL data highlights the competitive nature and the value of statistical analysis in professional cricket.

Murali [4] have introduced a Data Envelopment Analysis (DEA) model to optimize team selection. DEA is a non-parametric approach applied for assessing the efficiency of decision-making units. Here, it measures the performance of cricket players based on multiple metrics such as batting, bowling, and fielding.

The purpose of the method is to find out which player will be most efficient for any particular role within a team, which can then be used practically in selecting domestic and international cricket teams. Dakhani and Maginmani [5] have utilized the machine learning model for the prediction of future performance based on historical data. In their research, they have classified players' accuracy in batting and bowling by using classification algorithms like decision trees and support vector machines. They are integrating the data, such as past performances, conditions of play, and player fitness, to develop predictive models to help coaches make informed decisions on player selection. Punjabi *et al.* [6] surveys various methods used in cricket team selection. Punjabi *et al* [6] explains and emphasizes the use of clustering, regression analysis, and neural networks machine learning techniques in analyzing player performance data and recommending the most effective team composition. Swetha and Saravanan [7] note the added benefits of machine learning by providing not only the selection of the best players but also predicting match outcomes according to the team composition and player's strengths. Swetha and Saravanan [7] discusses the various factors influencing cricket match outcomes. Through analysis of attributes such as team strength, player form, pitch conditions, and weather, the study identifies key factors that contribute to winning outcomes. The research suggests a combination of statistical analysis and expert insights can improve the accuracy of predictions. Finally, Abbas and Haider [8] concentrated on predicting the winner based on past data. Using machine learning models, they integrated the features of team strength, player form, and match conditions into predicting the outcome of cricket matches. This research contribute to the broader field of match prediction and further show the utility of machine learning for predicting outcomes in dynamic, real-time sports environments.

Fu [9] analyzes and predicts the performance of bowlers and batters in the Indian Premier League cricket tournament by using machine learning algorithms. Fu [9] utilizes traditional statistical models alongside modern machine learning algorithms such as logistic regression, decision trees, and gradient boosting. Emphasis is placed on identifying which model offers the best predictive accuracy across diverse match contexts. Fu [9] finds that machine learning models outperform baseline approaches, especially when incorporating complex spatial and temporal data. The study also explores feature importance to determine which variables contribute most to accurate predictions. This research contributes to sports analytics by providing actionable insights into model design and feature selection, ultimately supporting performance evaluation, coaching decisions, and recruitment strategies in professional soccer [9].

Santra *et al.* [10] have focus on IC clustering to analyze IPL bowlers' performance. By segmenting bowlers based on performance metrics, Santra *et al.* [10] identify distinct groups that exhibit similar characteristics. The study reveals that clustering can effectively categorize bowlers into performance tiers, providing valuable insights for team selection and match strategy. This approach not only enhances understanding of individual contributions but also supports tactical decisions during matches. Nipane *et al.* [11] employing unsupervised learning to predict bowler performance in the IPL. They utilize a dataset comprising historical bowling statistics to develop

a model that identifies key factors influencing bowler success. Their findings indicate that factors such as match conditions and opposition strength play critical roles in performance outcomes. This research adds depth to the existing literature by illustrating how unsupervised techniques can yield actionable insights for teams [11]. Gour and Khan [12] have investigated cross-league analytics using various machine learning models to compare player performances between different cricket leagues. The study emphasizes the importance of contextual factors when interpreting performance data across leagues, revealing that certain models outperform others depending on the league-specific characteristics. This comparative analysis enhances the understanding of how players adapt their skills across different competitive environments. Chakraborty *et al*. [13] have delved into the application of unsupervised algorithms for analysing IPL players' performances. Their research highlights how these algorithms can uncover hidden patterns within player data, leading to improved predictions of future performances. Srikantaiah *et al.* [14] have argued that leveraging unsupervised learning can enhance traditional analytics methods, providing teams with a more comprehensive understanding of player capabilities. Srikantaiah *et al.* [14] have employed clustering combined with dimensionality reduction techniques to predict player performance in the IPL. By reducing the complexity of the dataset while maintaining essential information, they achieve more accurate predictions regarding player contributions during matches. This innovative approach showcases the effectiveness of integrating multiple analytical techniques in sports analytics. Santhosh *et al.* [15] present a comparative study focusing on cross-league predictions between the IPL and Caribbean Premier League (CPL). They analyze various machine learning algorithms to determine their effectiveness in predicting match outcomes based on player performances across both leagues. The findings indicate significant differences in predictive accuracy depending on the chosen model, emphasizing the need for tailored approaches when analyzing different leagues. Herath and Wijenayake [16] have evaluate bowler performance using machine learning techniques within an IPL context. Their research highlights key metrics influencing bowler success and offers a framework for assessing individual performances through advanced analytical methods. This study contributes valuable insights into how machine learning can refine traditional evaluation processes in cricket. Miller *et al.* [17] have combined clustering techniques with cross-league prediction methodologies to assess IPL bowlers' performances comprehensively. Their analysis reveals how different clusters of bowlers perform under varying conditions, providing teams with strategic insights for match preparation and player selection.

Next most important sport after football is Cricket. The most popular league or the IPL for T20 domestic leagues around the globe. Cricket includes a number of data and statistics involved. Several parameters can predict what a cricket match's outcome can be in the game of Cricket. Machine Learning can come together with these factors involving a cricket match to give an outcome to a game. This research has concentrated on analyzing the features of the cricket matches in IPL. Moving further towards the analysis of the Indian Premier League, this paper has rated the Batsmen and Bowlers in a unique way based on their performance. A few crucial factors like team form and team strength in predicting the match outcome apart from the conventional features like the toss, venue of the games etc., have been added. Further, a novel analysis of Batting and Bowling has been proposed based on Batting Index and Bowling Index. Machine Learning algorithms like SVM, Logistic Regression, Decision Tree, Random Forest and Naive Bayes have been applied, for match predictions. The results have been plotted based upon which algorithm is giving accuracy. The Decision Tree and the Logistic Regression algorithm have also provided an accuracy over 87% and 95%, respectively.

Due to the rising demand for cricket analytics, cricket fans and analysts can research on the performance evaluation of players. Various studies have been performed into the player classification and prediction models using machine learning approaches. Anik *et al.* [18] have presented a methodology for the prediction of a cricket player's performance in One Day International (ODI) matches. The proposed model makes use of statistical data and a feature selection algorithm for the prediction of the number of runs scored by the batters and the number of runs given up by the böwlers. Furthermore, Saikia [19] have focused at creating machine learning techniques for the classification of cricket all-rounders, opening the door for player role identification within a team.

Based on the improvement of performance, a realistic approach for Star Cricketer Prediction (SCP) has been illustrated to identify star players who would perform particularly well going forward by Ahmad *et al.* [20]. They provide a list of the top players that has been verified by the International Cricket Council (ICC) rankings and use supervised machine learning models for binary categorization of great cricket players. Khot *et al.* [21] have used statistical analysis and machine learning to forecast the future stars of cricket. They analyzed basic features for batters and bowlers and built a support vector machine model for prediction. Balasundaram *et al.* [22] have explored data mining-based player classification in cricket, including different formats and classifications with decision trees, support vector machines, and random forests.

Furthermore, a novel approach was introduced by Wickramasinghe [23] to quantify the current form of cricket teams, particularly all-rounders, using strength-based indices and machine learning. This study paved the way for assessing team composition and performance in dynamic cricket matches. In the context of IPL, a study was conducted on machine learning methods for team formation and winner prediction in cricket (Ishi and Patil [24]). They used various supervised learning techniques to predict the outcome of IPL matches based on the performance of players and other factors.

Ahmed *et al.* [25] and Arvindhan *et al.* [26] have developed a multi-objective genetic optimization methodology for team selection in cricket, which is quite vital for batting strength and other factors like bowling, fielding, and even the captaincy. It took the Indian Premier League as a test case to search for an optimal team which outperforms in batting and bowling performances within the constraints of budget.

The genetic optimization-based approach, it provides value to cricket analytics because it allows the simultaneous optimization of several, usually conflicting issues, thus including maximization of a player's performance, minimization of injury risk, and balancing the team's batting and bowling strengths. Therefore, analysts can come up with the best possible combinations, strategies, and decisions based on various scenarios by applying genetic algorithms. There is a new possibility for the data-driven guidance of these tactical decisions like the player selection, match planning, and the decision-making process with which the team can improve its performance and efficiency.

Sumithra *et al.* [27] have focused on Deep learning technique in order to extract meaningful insights from the database. However, despite these significant contributions in cricket analytics, a dearth of search focuses explicitly on bowler performance analysis using unsupervised machine learning techniques, especially in the context of IPL discussed by Abdulla *et al.* [28]. Thus, our research paper aims to bridge this gap by employing K-means clustering and Mean Shift cluster analysis for bowler performance analysis in IPL and extending the model for cross-league prediction.

Bose *et al.* [29] have evaluated player performance in cricket game. The objective of the study is to verify grouping of player based on performance measures without considering predefined labels. The study shows data-driven framework for team selection process. Sarlis and Tjortjis [30] have used data mining techniques to analyze how player position, age, and injury history influence performance and economic value in the NBA. The result of the study focuses on insights into player evaluation by linking on-court metrics with financial outcomes. In summary, these papers collectively advance the field of sports analytics by applying various machine learning techniques to predict cricket players' performances, particularly within the context of the IPL and other leagues. They illustrate the potential for these methodologies to enhance decision-making processes in sports management and strategy development.

Historical cricket analytics have been, and the majority of cricket literature is, more focused on batsmen since they are more visible and performance can be better quantified via metrics such as batting average, strike rate, and boundary percentage. Studies look closely at how conditions, opposition and a range of other factors influence batsmen. Batsmen hog the limelight when it comes to playing a match, but bowlers too play a pivotal role in achieving victory. While most studies on bowlers center on wickets taken, economy rate, and bowling average, they lack consideration for more nuanced metrics, such as fluctuations based on pace, spin, bounce, and overall field placement strategies that would determine bowling effectiveness.

Based on the literature review, the following research gaps can be identified:

Bowler Performance Analysis using Unsupervised Learning: Even with recent developments in cricket analytics, minimal research has explicitly addressed the utilization of unsupervised machine learning methods, such as K-means clustering and Mean Shift, for in-depth bowler performance analysis. The gap is highly noticeable in IPL, where insights for bowlers based on data are scarce and performance analysis becomes more biased toward batsmen.

Incorporating Sophisticated Metrics in Bowler Analysis: Most current research on bowlers is based on metrics such as wickets taken, economy rate, and bowling average. More sophisticated metrics such as pace variations, spin, bounce, and field placement strategies are not considered, which are essential in measuring bowling effectiveness. More advanced models incorporating these factors could provide better insights.

Cross-League Performance Analysis: Although there is work on comparing performance between cricket leagues, e.g., the IPL and Big Bash League, a more specific method of cross-league prediction is necessary. Leagues have different features, and more specialized models might be necessary to manage the contextual variations in player performance and match results between leagues.

Integration of Dynamic Environmental Factors: Current models usually measure performance against historical statistics and conditions of matches, but do not always account for the dynamic interaction of real-time environmental factors (like strength of opposition, pitch conditions, and weather). Future development could involve incorporating these factors more systematically into prediction models to enhance accuracy.

III. Holistic Model Construction For Bowler Success

Most research does not comprehensively combine several variables affecting bowler success in actual match situations. A potential avenue for future research may be the construction of holistic, multidimensional models that measure bowler performance from a blend of past performance, player conditioning, tactical team choice, and environmental factors.

A. Dataset Description

This dataset for research work was retrieved from the trusted source: https://cricsheet.org/downloads, which delivers cricket data in JavaScript Object Notation (JSON) format. This dataset comprises complete information for matches for all seasons of Indian Premier League (up to 2022), with relevant player statistics and match results and more features.

A comprehensive dataset is one that covers significant variables, thereby ensuring that the data remain diverse and representative, which in turn maximizes the reliability

of analysis. With minimal bias induced, it allows for more accurate and generalized insights.

B. Data Content

The dataset consists of data from different IPL seasons, taking into account an all-encompassing analysis of player performances over the long term. Every match passage in the dataset includes the following key features:

- Match Data: Date of the match, venue, and participating teams.
- Innings Details: For each inning, information about the batting team, bowling team, total runs scored, and wickets taken.
- Batsman Execution: The nuances of each batsman's display, comprising runs ored, balls faced, strike rate, and excusal type.
- Bowler Execution: The execution figures of each bowler, including the number of overs bowled, runs leaked, wickets taken, and economy rate.
- Other Highlights: The dataset could contain extra variables, including more extras, fall of wickets, and partnerships.

Utilizing cricsheet.org as the information source will ensure the reliability and authenticity of the dataset since it follows tight information collection and approval processes.

C. Data Collection and Pre-processing

As mentioned above, the dataset for this research was downloaded from https://cricsheet.org/downloads/, which is a reliable source for JSON cricket data. Data of bowlers of the Caribbean Premier League was also taken randomly to be used for cross-league validation in later stages. In preparation of data for analysis, the following pre-processing steps were followed: This included dealing with missing values, transforming a player's name so they were standardized and transduced categorical variables to facilitate applying machine learning algorithms in any way possible. This was also associated with finding the averages and runs per match along with wickets per match and no ball deliveries per match for moving towards the study. And, feature selection was considered.

D. Unsupervised Machine Learning—Clustering

To dissect bowler execution, solo AI methods were utilized. In particular, bunching calculations, like K-implies and Progressive Bunching (Mean Bunch Examination), were applied to bunch bowlers with comparative performance qualities. The elbow strategy was utilized to decide the quantity of groups (K).

E. Feature Selection

For the grouping examination, vital elements, for example, number of no balls, complete runs given, wickets per match, runs per match, and no balls per match were chosen in light of their significance in portraying bowler execution. The picked highlights filled in as the reason for bunching bowlers into unmistakable gatherings.

F. Bowler Performance Analysis

The bunching results were then deconstructed to gain insights into the bowler execution designs. Representation methods, such as dissipate plots and group profiles, were utilized to understand the attributes of each bunch. The exhibition of bowlers inside each bunch was considered, factoring in the recognition of top-performing bowlers and those in need of improvement.

G. Cross-League Prediction

Models were trained on the data of one cricket league, that is, IPL in the current scenario, and evaluated on the data of a different league, for instance, another T20 league. The above unsupervised models then came into play to predict the performance of some of the bowlers for different leagues so that they can determine how they would go in the IPL.

Cross-league validation was implemented by splitting the dataset into distinct training and testing sets, ensuring that data from different leagues or groups didn't overlap between training and testing phases. This method helps to evaluate the model's generalizability across diverse groups. Feature normalization was performed to standardize the scale of input features, typically using techniques like Min-Max scaling for normalization, to ensure that no single feature disproportionately influenced the model. This process improves the model's performance and stability by making sure each feature contributes equally.

IV. Bowler Performance Analysis

Data for 473 bowlers who have bowled at least one over in the Indian Premier League (IPL) was collected and processed for the Bowler Performance Analysis. The data was first merged into a single Pandas DataFrame for further analysis. Relevant performance metrics were calculated with the following steps:

A. Data Processing and Feature Calculation

Data Pre-processing could be happened based on following parameters:

Data Cleaning: This includes the processes of cleaning data by correcting or deleting wrong data. Key techniques include: Handling Missing Values: Filling in the gaps in missing values through statistical techniques including mean or median. Outlier Detection: The process of locating outliers and either removing or modifying them. Noise Reduction: The processes of smoothing or filtering out noise from the data set. Feature Selection: The strategy of selecting important features in dataset to improve the performance of the model as well as control overfitting. In general terms, feature selection is very important in machine learning because it increases the prediction accuracy of the model, diminishes overfitting, and makes the results easier to understand. Filter Methods assess subsets of features for their statistical merits independent of any machine learning algorithm and the features are obtained by Correlation Coefficients technique which measures the association between the features and the focal target.

- Matches Played: The number of appearances a bowler had made as is computed based on data source.
- Performance Metrics Calculation: Key performance. The metrics, which were "runs per match", "no balls per match", and "wickets per match", came out. This was done by grouping the columns "noballs", "runs", and "wickets" according to the name of the bowler and then calculating the corresponding averages for each one of them.
- Feature Selection: Based on the derived performance metrics, the last five features used for training the model were determined: "number of no balls", "total runs given", "wickets per match", "runs per match", and "no balls per match".

B. *Unsupervised Machine Learning—K-means Clustering*

- Elbow Method: The elbow method (see Fig. 1) was applied to determine the optimal number of clusters (K) for the K-means clustering algorithm. The elbow method plots visually identified where the within-cluster sum of squares levels off, indicating the optimal number of clusters.
- In the elbow method for K-means clustering, selecting $k = 4$ is determined by observing a plot of the number of clusters k against the Within-Cluster Sum of Squares (WCSS). The optimal k is identified at the elbow point, where the rate of decrease in WCSS sharply changes, indicating that adding more clusters yields diminishing returns in terms of improved clustering quality.

This point suggests a balance between model complexity and variance, making $k = 4$ a suitable choice for capturing the underlying data structure effectively without overfitting.

- K-means Clustering: The K-means clustering algorithm was then trained with $k = 4$ (as determined by the elbow method). This algorithm divided the bowlers into four clusters based on their selected performance metrics.
- Cluster Visualization: The clustering results were visualized using a scatter plot (see Fig. 2), where each data point represented a bowler, and the color represented the assigned cluster.

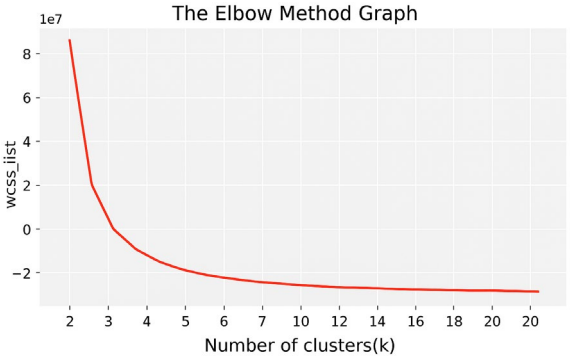

Fig. 1. Visualization of clusters using k means clustering.

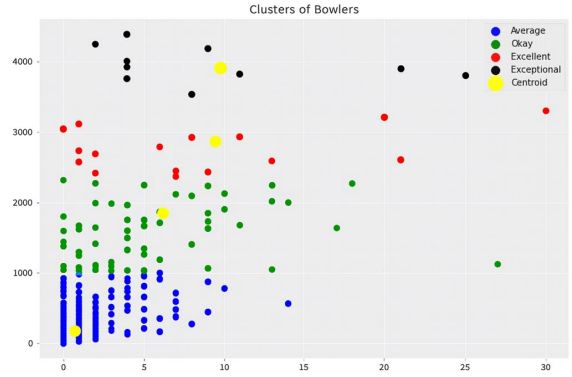

Fig. 2. Visualization of clusters using k-means cluster analysis. (The X-axis represents "no balls" and the y-axis represents "balls bowled").

TABLE I. K-MEANS VS. MEAN SHIFT

Feature	K-Means Clustering	Mean Shift Clustering
Predefined Number of Clusters	Requires the number of clusters (k) to be specified.	Does not require the number of clusters to be specified.
Cluster Shape	Assumes clusters are spherical (circular in high dimensions).	Can find clusters of arbitrary shapes.
Computational Complexity	Generally faster, $O(n \times k \times d)$, where n is the number of data points, k is the number of clusters, and d is the number of features.	Computationally more expensive, $O(n^2)$ or higher, due to the need to compute distances for all points.
Sensitivity to Initial Conditions	Sensitive to initial centroid placement, which can lead to suboptimal solutions.	Less sensitive to initial conditions, but can still be affected by bandwidth selection.
Cluster Size Balance	Tends to favor creating clusters of roughly equal size, which may not reflect real-world data.	Can form clusters of varying sizes based on the density of data points.
Handling Outliers	Sensitive to outliers, as they can pull centroids away from true clusters.	More robust to outliers, as it focuses on the density of points rather than individual centroids.
Scalability	Scales well to large datasets and is computationally efficient.	May not scale well to large datasets, especially without optimization techniques.
Interpretability	Simple to understand and interpret, with clearly defined centroids.	Harder to interpret due to density-based clusters and no clear centroid.
Convergence	Converges quickly, but may only find local optima depending on the initial centroids.	Can take longer to converge, but provides more flexibility in terms of cluster discovery.
Use Case Suitability	Ideal for situations where the number of clusters is known and clusters are expected to be relatively equal and spherical.	Ideal for more complex datasets with arbitrary-shaped clusters or when the number of clusters is unknown.

Mean Shift clustering has some benefits in comparison to K-means clustering (discussed in Table I) that makes it favorable for use in different scenarios. Some of these include: No Need for Predefined Number of Clusters: In contrast to K-means which needs the number of clusters to be chosen beforehand, Mean Shift determines the number of clusters based on the data provided. This feature becomes useful especially when one does not know how many clusters to break the data into prior to analysis. Robustness to Noise: In general, the Mean Shift exhibits more robustness to the noise and outliers that are present in the data context. Its density map based strategy helps it in minimizing the noise and concentrating on the important data items thereby improving clustering results on real world datasets. No Initial Centroid Sensitivity: In K-means, the final results of the clusters are greatly dependent on how the centroids in the beginning. Mean Shift starts with every data sample as a cluster center candidate which decreases initialization sensitivity and can improve the repeatability of the algorithm over different runs.

Silhouette Score: The Clustering Quality Assessment Index (CQAI) also referred to as Silhouette Coefficient Scales easily answers questions pertaining to the quality of a given clustering. Likewise: High Degree of Separation: Contents prone to score over 0.7 are likely to place z values greater than 0.7, indicating that the variety amongst clusters is clear and easy to pinpoint. Hence there is a minute chance that members of different clusters have close similarity to one another. Fusion and divergence: This suggests that the value is considerably large due to the presence of strong bonds and diverging elements making a cluster. That being said, ensuring that the clusters are similar to each other and dissimilar to others. Demonstrative of how well data has been grouped: In a way, being partitioned into such scores should be of help in proving how effective the algorithm used is. This implies that the number of required sub-clusters has been attained at this point and further expansion is pointless. In this case, there is no need to explain why fewer numbers—Or still fewer numbers will give better results as accuracy increases with more crystallization. To sum up, a coefficient ranging between 0.7763 and 1 suggests that strong levels of determination and cohesion were present bringing forth the conclusion that the algorithm was able to efficiently cluster the data into strong and salient groups.

C. *Unsupervised Machine Learning—Mean Shift Clustering*

Mean Shift Clustering: The same set of the bowler performance metrics applied the Mean Shift clustering algorithm.

This algorithm automatically selects the number of clusters by itself, without the setting of K beforehand, as shown in Fig. 3.

Cluster Visualization: The Mean Shift clustering results were visualized using a scatter plot, like the K-means clustering visualization. The elbow point at $k = 4$ indicates that adding more clusters beyond this point does not significantly improve the model's performance, suggesting that four clusters effectively capture the data's underlying structure while avoiding overfitting. This balance allows for a meaningful representation of the data without unnecessary complexity.

Cricket dataset clusters entail different implications for team strategy and player utilization. Top Performers are the mainstay of the team, habitually providing match-winning contribution and should be relied upon to consolidate in times of adversity. All-rounders, being versatile, help balance batting and bowling, thereby adding utility to team combination and assisting in providing flexibility in the team combination and rotations. Specialist Batsmen are important for building innings and partnerships, especially in difficult batting conditions and during chase situations; thus, their placement in the order should be optimized. Specialist Bowlers hold the key to breaking partnerships and controlling the game's pace; their roles in the respective phases (i.e. death overs, powerplays) of matches are critical for winning. Finally, Match Changers thrive under pressure: Their propensity to alter the course of events at critical junctures makes them apt for the high-stakes phases of the game, in many cases tipping the scales in favor of the winning team in tight contests.

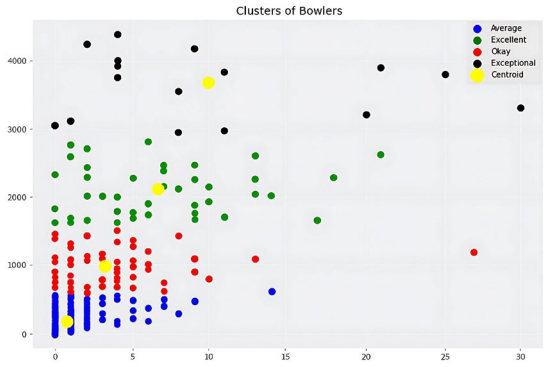

Fig. 3. Visualization of clusters using mean shift cluster analysis. (The X-axis represents "no balls" and the y-axis represents "balls bowled").

- Silhouette Score

The quality of the Mean Shift clustering was assessed using the silhouette score, which measured the cohesion and separation of the clusters. A higher silhouette score (0.7078 in this case) indicated well-defined clusters.

- Analysis and Interpretation

The results of the clustering based on K-means and Mean Shift algorithms were insightful in terms of bowler performance patterns. Clustering the data helped to identify top bowlers with similar attributes for those in other clusters and their areas of improvement.

- Cross-League Prediction

In the Cross-League Prediction phase, data from bowling performance in the Caribbean Premier League (CPL) was also gathered and pre-processed with the same steps on the Indian Premier League dataset for uniformity. The goal was to check the cross-league prediction ability of previously trained K-means and Mean Shift models in predicting the categories for the performance of the CPL dataset. For this purpose, five sample records have been

provided to facilitate the prediction of outcomes. Table II illustrates five sample records from the data set.

- Prediction Results

Surprisingly, all five sample records in the CPL dataset were identically predicted by K-means and Mean Shift clustering models. This consistency in predictions across both models reinforced the clustering algorithms' robustness and applicability to cross-league predictions. 2 players can be considered potential talents that can be very useful in the leagues, getting a wicket almost every time they bowl. In such leagues of T20, including such players in the team could be essential.

- Use of K-means Clustering to predict the cross-league game

The model trained on IPL bowler CPL data is applied to predict bowling performance categories for the given set of bowlers in the dataset. The results are presented below: array [0, 1, 0, 0, 1].

- Use of the mean shift cluster to predict the cross league

Similarly, the CPL dataset was used to predict performance categories using a mean shift clustering model which is also trained on IPL data. The results are as follows: array ([0, 1, 0, 0, 1], dtype = int64).

Cross-league prediction validity is an evolving field that benefits from advancements in machine learning techniques and innovative methodologies like cross-prediction. While significant progress has been made, ongoing research is essential to address the inherent challenges and improve the robustness of predictive models across diverse sporting contexts.

TABLE II. SAMPLE RECORDS FROM CARIBBEAN PREMIER LEAGUE USED FOR TESTING THE MODELS

bowler name	no_balls	runs_given	total_wickets	total_matches	wickets_per_match	runs_per_match	no_balls_per_match
JP Duminy	0	91	4	6	0.67	15.17	0.00
Sohail Tanvir	7	1356	62	55	1.13	24.65	0.13
DJG Sammy	2	701	21	32	0.66	21.91	0.06
Pas Mcsween	2	48	1	2	0.5	24	1
Kok Williams	26	1744	83	63	1.32	27.68	0.41

V. DISCUSSION

A valuable insight into bowler performance patterns in the Indian Premier League IPL was provided by an analysis of Bowler Performance using unsupervised machine learning techniques, namely K-means and Mean Shift Clustering. Top performers were identified with the help of distinct characteristics. These analytics help in planning strategy and choosing players for team management.

The cross league prediction achievement showed that the trained clustering models were well adapted to the Caribbean Premier League dataset. The consistency of performance categories predictions in both leagues has confirmed the model's applicability and prospects for identifying skill across different T20 cricket leagues. This analysis can be used by teams to target bowlers with a track record of success in their individual IPL categories, which will improve team dynamics and results across different leagues.

Actual implications of the outcome of prediction results across leagues are important for management in teams and strategies in the recruitment of players. In this way, data-generated insights can help teams decide and identify potential talents within the CPL based on the performance profiles that will suit the team. Additionally, teams can modify their bowling strategies and tactics according to the bowlers with a cricket background by observing the common performance characteristics of each league. Nevertheless, it is important to consider that the study has certain limitations such as data representativeness and feature selection, which may affect the prediction accuracy. Further analysis can also be done to find how the selected bowlers fare against different types of batters to enhance the selection further and find a talent that can help the team in various situations.

VI. CONCLUSION

The Bowler Performance Analysis is an innovative analysis tool that dives into the cricket bowling performance to understand the hidden trends. By implementing unsupervised machine learning methodology, it correctly identifies and categorizes talents present in each of the various leagues, laying open patterns invisible to conventional analysis approaches. This analysis is key in maximizing the value of cricket analysis, as it giving the cricket teams enough insight to be able to make their decisions rationally.

Managers and analysts in the IPL are confronted with decision-making based on data, handling players, and strategy formulation. Managers use data to choose players and coordinate team relationships so that players are at their optimal fitness and morale levels, while analysts assist through offering live data, tracking opponent tactics, and connecting with the fans through sentiment analysis. Managers and analysts must work together to formulate strategies and sustain success throughout the season.

In T20 cricket, detailed bowler performance statistics are crucial for strategic planning, player choice, and match strategy. These measures identify up-and-coming talent and evaluate existing player performance, providing teams with a competitive advantage. Ultimately, Bowler Performance Analysis enables teams to make data-driven decisions, improving flexibility and boosting success in the high-speed environment of T20 cricket.

K-means and Mean Shift are good choices when speed, scalability, and simplicity are considerations. Whereas

K-means is efficient and speedy on large data sets but has difficulty with non-spherical clusters and needs the number of clusters to be known beforehand, Mean Shift is more flexible and can deal with arbitrary-shaped clusters but is computationally costly for big data sets.

Our results show that unsupervised clustering can help IPL teams identify potential talent and optimize player selection strategies.

The silhouette score measures how similar each point is to its own cluster compared to other clusters, with a score closer to +1 indicating well-separated clusters. A score near 0 suggests overlapping clusters, while a negative score indicates poor clustering or wrong assignments.

Model Bias: K-means and Mean Shift both work under the principle that the clusters inherently contain patterns in the data. But when the data is biased (for instance, when there are underperformances owing to external issues such as injuries or team strategy), the models might yield biased results.

Small Dataset for CPL Validation: The comparatively small dataset for the Caribbean Premier League (CPL) may not cover the entire range of performances of players, which can be the cause of overfitting or wrong generalization by the model. A more extensive and diversified dataset would make the model more dependable.

League-Specific Conditions: Even though clustering algorithms yield comparable outcomes across leagues, they fail to consider league-specific conditions such as pitch conditions, weather, or varying playing styles, which might affect bowler performance and confine the applicability of the model in cross-league comparisons.

To deal with the foundation for future work on this topic, one must think about how using advanced methods like deep learning may improve the examination of the bowlers' performance. The recurrent neural network- and LSTM-network-based deep learning algorithms might be used to uncover the time-trend patterns in the performance data that would establish more precise and long-lasting forecasts.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

P. S. Metkewar has prepared abstract for the research, Vijay Kumar Varadarajan's has contributed in literature review, methodology part, Dhananjay S. Deshpande prepare the layout of the paper, Farhaan Bhola analyse the data, Anuja Bokhare worked on literature review part, and Deshinta Arrova Dewi conducted the research; P. S. Metkewar wrote the paper; all authors had approved the final version.

Funding

This research was funded by INTI International University, Malaysia, Grant Number: INTI-FDSIT-02-11-2023.

Acknowledgment

The authors wish to thank MIT WPU for providing the research environment and required infrastructure.

References

[1] V. S. Vetukuri, N. Sethi, and R. Rajender, "Generic model for automated player selection for cricket teams using recurrent neural networks," *Evolutionary Intelligence*, vol. 14, pp. 971–978, Sep. 2020.

[2] C. Kapadiya, A. Shah, K. Adhvaryu *et al.*, "Intelligent cricket team selection by predicting individual players' performance using efficient machine learning technique," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 3, pp. 3406–3409, Feb. 2020.

[3] K. Jaswanth, K. Arun, B. D. Deekshith *et al*., "Analyzing and predicting outcomes of IPL cricket data," *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 3, pp. 1125–1128, Mar. 2022.

[4] A. Murali, "A DEA model for selection of cricket team players," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 9, pp. 749–752, Sep. 2020.

[5] M. M. Dakhani and U. H. Maginmani, "Predicting accuracy of players in the cricket using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 5, pp. 6720–6726, May 2020.

[6] V. Punjabi, R. Chaudhari, D. Pal *et al*., "A survey on team selection in game of cricket using machine learning,'' *International Research Journal of Engineering and Technology (IRJET)*, vol. 6, no. 11, pp. 2442–2446, Nov. 2019.

[7] Swetha and K. Saravanan, ''Analysis on attributes deciding cricket winning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 3, pp. 1105–1107, Mar. 2017.

[8] K. Abbas and S. Haider, "3 winner prediction in cricket using machine learning," *British Journal of Sports Medicine*, vol. 51, A3, Dec. 2017.

[9] S. Fu, "Comparative analysis of expected goals models: Evaluating predictive accuracy and feature importance in European soccer," *Applied and Computational Engineering*, vol. 113, pp. 36–45, 2024.

[10] A. Santra, A. Mitra, A. Sinha *et al.*, "Prediction of most valuable bowlers of Indian Premier League (IPL)," in *Proc. International Conference on Data Management, Analytics and Innovation*, 2021, pp. 211–223.

[11] D. Nipane, S. Joseph, R. Karthik *et al.*, "Prediction of IPL player selection from T20 formats using a novel data balancing approach," in *Proc. 2025 International Conf. on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, 2025, pp. 1–8.

[12] P. N. Gour and M. F. Khan, "Utilizing machine learning for comprehensive analysis and predictive modelling of IPL-T20 cricket matches," *Indian Journal of Science and Technology*, vol. 17, no. 7, pp. 592–597, 2024.

[13] S. Chakraborty, L. Dey, S. Maity *et al.*, "Prediction of winning team in soccer game: A supervised machine learning-based approach," in *Advances on Mathematical Modeling and Optimization with Its Applications*, Boca Raton: CRC Press, 2024, ch. 1, pp. 170–186.

[14] K. C. Srikantaiah, A. Khetan, B. Kumar *et al*., "Prediction of IPL match outcome using machine learning techniques," in *Proc. 3rd International Conf. on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, 2021, pp. 399–406.

[15] S. Santhosh, S. Kumar, and C. G. Nagorao, "A statistical study of IPL team performances," *J. Mod. Math. Stat.*, vol. 17, pp. 1–8, 2023.

[16] E. K. Herath and U. Wijenayake, "Statistical and exploratory data analysis on Indian premier league," in *Proc. Conference on Transdisciplinary Research in Engineering*, 2024. https://doi.org/10.31357/contre.v1i1.7386

[17] T. Miller, I. Durlik, A. Łobodzińska, *et al.*, "AI in context: Harnessing domain knowledge for smarter machine learning," *Applied Sciences*, vol. 14, no. 24, 11612, Dec. 2024.

[18] A. I. Anik, S. Yeaser, A. G. M. I. Hossain *et al.*, "Player's performance prediction in ODI cricket using machine learning algorithms," in *Proc. 2018 4th International Conf. on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, 2018, pp. 500–505.

[19] H. Saikia, "Quantifying the current form of cricket teams and

predicting the match winner," *Management and Labour Studies*, vol. 45, no. 2, pp. 151–158, April 2020.
[20] H. Ahmad, S. Ahmad, M. Asif *et al.*, "Evolution-based performance prediction of star cricketers," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1215–1232, 2021.
[21] A. Khot, A. Shinde, and A. Magdum, "Rising star evaluation using statistical analysis in cricket," in *Proc. 2020 International Conf. on Advances in Distributed Computing and Machine Learning*, 2021, pp. 317–326.
[22] A. Balasundaram, S. Ashokkumar, D. Jayashree *et al.*, "Data mining based classification of players in the game of cricket," in *Proc. 2020 International Conf. on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 271–275.
[23] I. Wickramasinghe, "Classification of all-rounders in the game of ODI cricket: Machine learning approach," *Athens Journal of Sports,* vol. 7, no. 1, pp. 21–34, 2020.
[24] M. S. Ishi and J. B. Patil, "A study on machine learning methods used for team formation and winner prediction in cricket," in *Proc. International Conf. on Inventive Computation and Information Technologies*, 2021, pp. 143–156.
[25] F. Ahmed, A. Jindal, and K. Deb, "Cricket team selection using evolutionary multi-objective optimization," in *Proc. International Conf. on Swarm, Evolutionary, and Memetic Computing*, 2011, pp. 71–78.
[26] M. Arvindhan and D. R. Kumar, "Adaptive resource allocation in cloud data centers using actor-critical deep reinforcement learning for optimized loadbalancing," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 5s, pp. 310–318, 2023.
[27] M. G. Sumithra, R. K. Dhanaraj, C. Iwendi, and A. M. Manoharan, *Deep Learning for Cognitive Computing Systems: Technological Advancements and Applications*, Berlin: De Gruyter, 2022.
[28] Z. H. Abdulla, W. Ismail, L. Q. Zakaria *et al.*, "Classification of exercise game data for rehabilitation using machine learning algorithms," in *Proc. International Conf. on Data Science and Emerging Technologies*, 2022, pp. 293–304.
[29] A. Bose, S. Mitra, S. Ghosh *et al.*, "Unsupervised learning based evaluation of player performances," *Innovations in Systems and Software Engineering*, vol. 17, no. 2, pp. 121–130, 2021.
[30] V. Sarlis and C. Tjortjis, "Sports analytics: Data mining to uncover NBA player position, age, and injury impact on performance and economics," *Information*, vol. 15, no. 4, 242, 2024.