Automated Selection of Optimal Cricket Team Using Machine Learning

by

Mohammad Shariful Alam Mollah
20201146
Dipjyoty Biswas Niloy
20301279
Hasnat Khalid Ruhani
20301283
Azmain Azam Chowdhury
20301272
Risha Tasnin
20301025

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
2024

Declaration

It is hereby declared that

1. This thesis is my/our own original work conducted while completing my/our degree at Brac University.
2. The thesis does not include material previously published or written by others, except where it has been appropriately cited with full and accurate references.
3. The thesis does not contain material that has been accepted or submitted for any other degree or diploma at any university or institution.
4. All significant sources of assistance have been acknowledged.

Student's Full Name & Signature:

Mohammad Shariful Alam Mollah
20201146

Dipjyoty Biswas Niloy
20301279

Hasnat Khalid Ruhani
20301283

Azmain Azam Chowdhury
20301272

Risha Tasnin
20301025

Approval

The thesis/project titled "Automated Selection of Optimal Cricket Team Using Machine Learning" submitted by

1. Mohammad Shariful Alam Mollah (20201146)
2. Dipjyoty Biswas Niloy (20301279)
3. Hasnat Khalid Ruhani (20301283)
4. Azmain Azam Chowdhury (20301272)
5. Risha Tasnin (20301025)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October, 2024.

Examining Committee:

Supervisor:
(Member)

Md. Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Dibyo Fabian Dofadar
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The process of selecting the best 11 players for a cricket team is a complex and critical task that requires considering various factors such as individual player performance, team dynamics, and match conditions. Traditional methods of the team selection system depend on manual analysis, experts' opinions, which can be time- consuming and can be biased. This thesis aims to develop an automated approach using Machine Learning (ML) techniques to assist in the selection of the optimal cricket team. ML algorithms are employed to analyze and extract meaningful patterns and insights from the dataset. Here we will consider a range of performance indicators, such as batting and bowling average, batting strike rate and bowling economy rate, etc hepls us to determine the key attributes that creates a major role of success for a cricket team. These algorithms learn from historical data and identify patterns to create a predictive model for player selection. By including indicators like player endurance, injury history, and recovery time frames, the model provides a more complete picture of a player's total contribution to the team. This technique assures that players are selected not just based on their present form and talents, but also on their physical preparedness and endurance throughout a tournament. This automated system provides objective and data-driven insights, reducing biases and human errors in the selection process.This selection method will draw the explanation for choosing this team over other selections. It will assist cricket team management, coaches, and selectors in making informed decisions, maximizing team performance, and optimizing player utilization. Moreover, the model adapts to different formats of the game like Test, One-Day International (ODI), and Twenty20 (T20) formats and each requiring unique strategies and player attributes. For instance, while a Test match may emphasize endurance and technique, a T20 match prioritizes aggression and quick decision-making. The system uses tailored algorithms for each format, ensuring the selection is optimized for the specific demands of the match at hand. The integration of that technology with cricket team selection has the potential to reshape the sport and elevate team strategies to new levels. The potential of this system extends beyond selection, potentially influencing training methods and in-game tactics, marking a new era in the technological evolution of cricket.

Keywords: Machine Learning; Performance; Cricket; Prediction; Objective; Player utilization; Integration of technology.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cricket is a royal sport that has an illustrious history that originated in the 16th century in England. It gained popularity in the 17th century. Today it is a world famous sport over 20 countries are now playing this game and now it has become a people's emotion.This is a complex sport with many variables and player statistics, and careful consideration is often required when choosing the best combination of players for a match. People always try to find out the best player for their national team. Today it is the most important part in cricket to figure out the best player for the team. Selection of the best 11 is one of the most crucial decisions in cricket sports. As cricket has evolved into a global sensation, encompassing the emotions of millions, the need for an impartial and data-driven approach to team selection becomes paramount. In the past, those who watched a lot of cricket and offered their comments performed this task, But occasionally, this could take a while and might not always be just.

Now, a study titled "Automated Selection of Optimal Cricket Team Using Machine Learning" aims to rectify that. Moreover, It is comparable to selecting the top cricket squad using really sophisticated computer tools. Also, these software applications employ a technique known as machine learning, which is excellent at identifying patterns in large amounts of data.

This study tested a variety of factors, including a player's bowling or batting performance. To determine who is performing exceptionally well, they use statistics like batting and bowling average, batsman strike rate, and bowling economy rate. Moreover, it will also check player performance against a specific team or opponent, and the player's performance and records at that specific venue, Also it will determine these with weather conditions. Finally, a model is created using all of this data that can forecast which players belong on the team. The use of machine learning algorithms adds impartiality and efficiency to the team selection process, minimizing biases inherent in human perspectives. The model's goal is to help cricket team management, coaches, and selectors make better decisions by utilizing historical data on player performance, field conditions, and opponent information. The resultant ideal lineups are not only based on previous results, but also adaptive to unique playing scenarios, guaranteeing that the squad is prepared to tackle a variety of obstacles.

The greatest benefit is that this approach is impartial and free of bias, unlike human opinions and also it can process vast amounts of real-time data. This study aims to automate and streamline the team selection process by leveraging historical data on

player performance, field conditions, opposing team performance analysis, and various other factors. It can also provide an important justification for why it selected these particular players. Managers, coaches, and selectors of cricket teams can utilize this technique to make informed choices that will improve the team's performance and make the most of each player. By using machine learning techniques, this study seeks to develop a model that can recommend optimal lineups based on specific playing situations and desired players' team. Furthermore, this system's adaptability to changing match dynamics offers a significant advantage, particularly in tournaments where teams face different opponents in quick succession.

Finally, we can say that this is significant because it may alter how cricket teams are selected, improving and enhancing them. This paper explains how they did it and why it's a fantastic idea for the world of sports. It will contribute to the growth of cricket and unbiased sports. In essence, the use of machine learning in cricket ushers in a new era in which technology, strategy, and skill combine to propel the sport to unparalleled heights.

1.1 Motivation

The motivation for this study came from the need to improve the cricket team selection process, which has become increasingly complex as the sport has evolved. The normal selection process relies entirely on people's own judgment, and is often criticized for biases and biases. To select the best player one has to analyze several factors like the performance of the player, his form, his past match statistics and the statistics of the opposition team players. As this process is time-consuming, it is also difficult for humans to do these things properly. Which can be faulty at any time. It is possible to select a sound team by analyzing the statistics of the players' past matches and the statistics of the players of the opposing team. Machine learning has clear potential to streamline this entire process. Machine learning empowers all players by reviewing data and processing that data to provide a clear recommendation. The main aim of this research is to automate the player selection process to reduce bias and improve decision making so that modern cricketers can better prepare for the challenges ahead. This study seeks to bridge the gap between tradition and technology, making cricket team selection more efficient, fair, and aligned with the growing demand for analytical precision in sports.

1.2 Research Objective

This research work is done with the intention of developing a keen, machine-learning-based system that can recommend, with speed and accuracy, the perfect team composition in any particular situation. Our squad selection system automatically tries to leverage the power of data-driven insights and predictive modeling to make wise player choices. The main objectives and key deliverables of the study are presented below:

1. To increase a cricket team's overall performance and competitiveness through player selection using data-driven criteria that optimize each player's skills and match-specific flexibility. .

2. To create a system that enables for quick player statistics and game scenario analysis, making quick and accurate team suggestions and decreasing the time and labor needed for manual selection.

3. To increase objectivity and transparency while reducing prejudice and subjective evaluations involved in the selection of human teams.

4. To make sure that team resources, such as players' abilities and skills, are used as effectively as possible in order to improve match outcomes and team success.

5. A solution that can adapt to different cricket formats including Tests, One-Days, and T20s along with keeping other variables such as pitch type, weather and opponents into consideration.

6. To generate insightful robust data on player performance patterns highlighting their strengths and weaknesses along with long-term player development and tactical planning.

7. To reduce the possibility of making poor judgments by using statistical and historical data to reduce mistakes in team selection.

8. To enhance the efficiency of the team selection process by integrating real-time data feeds during matches, allowing for dynamic adjustments based on evolving player performances and changing match dynamics.

9. To set a standard for ethical considerations in automated team selection, addressing issues about data privacy, algorithmic bias, and fair representation, and assuring the responsible and ethical use of machine learning in sports.

10. To promote diversity and justice in team selection by reducing the impact of external variables such as regional prejudices, senior player influence, and political ties on merit-based player picks.

In this research objective portion, we have presented a way for an innovative process of cricket squad selection. By leveraging machine learning techniques, we aim to utilize traditional methods to present a robust competitive and transparent approach for squad selection. Our method is adaptable to altering the cricket environment, and our data-driven approach optimizes cricket team selection, which ensures fairness and prioritize data driven solution.

1.3 Research Methodology

This study focuses on developing a machine learning-based system for selecting an optimal cricket team. It is designed to handle a large volume of player performance data according to their previous matches in different countries. The methodology begins with the collection and preprocessing of extensive cricket datasets, including player statistics, their previous match records, their recent forms and opposition team performance. Preprocessing steps include normalizing the data, handling missing values, and engineering features like batting averages, bowling economy, strike rates, and player consistency to enhance model training.

The research focuses on selecting and training the data by using unsupervised clustering machine learning models such as K-means, Hierarchical, DBSCAN. These models were chosen for their proven ability to cluster the data, particularly based on the performance of the player; this model divides the players into two clusters, good and bad. And based on these two clusters the model created an optimal 11 player for specific matches. The study also employs dimensionality reduction techniques like Principal Component Analysis (PCA) to simplify the feature space, making the clustering process more efficient and interpretable.
Techniques like the elbow method and also silhouette analysis are used to find the ideal number of clusters for models like K-Means. The elbow approach assists in determining the point at which within-cluster variation is not much improved by adding more clusters and silhouette analysis evaluates the consistency of data points within clusters. Additionally, this study using Exponential K-Means Clustering (ExKMC) is implemented to refine cluster boundaries, offering a more adaptive approach that adjusts cluster sizes based on player performance dynamics, ensuring a group of similar players.
The Silhouette score method works to evaluate the model and measures the clusters how they are formed and separated. These measures help us to understand the position of those groups of clusters. That also identifies the similar performance player and places them in a group. After forming the clusters, these groups can take a specific role in forming a cricket team. This makes it easier for coaches and selectors to find the best players for different matches.
Moreover the research also focuses on computational efficiency of the models and creating a system that can process large datasets quickly and in the meantime maintain the cluster formation. For that purpose PCA and ExKMC models are used because these models are particularly important for reducing computational complexity.
By using unsupervised learning models and clustering methods, we aim to improve the way of picking up the best players team using a more data driven approach that can adjust different match situations. With careful testing and evaluation of the models, the study aims to build a reliable system for finding the best team lineup, helping selectors make more informed and fair choices.

1.4 Research Problem

Cricket has been a worldwide famous sport since its birth. It has popularity in every continent. Like Football and, the Olympics, cricket has created its own fanbase. In 1975, the first Cricket World Cup was organised and since then cricket kept evolving in a different way. There was a time when finding a good player was challenging but now as cricket is at its peak time, there are a lot of talented players who have been producing in every country. As it is a game, in each match two teams are playing and each of the team consists of 11 players, in this era for every country it has been a challenge to every team to select the best 11 players combination to take away success from the opponent.

So to choose the best 11 combinations, every team has a team selection panel for those who do this selection part. It is a traditional way of selecting the 11 players. That selection panel closely monitors players' performance at the domestic and international levels. They try to evaluate players' skills and form from the stats. On the other hand, this panel also has to analyze the opponent team's performance, skills, and weaknesses. Then they work with the team coach and captain to make a game strategy based on

weather and pitch be the team squad given more priority to batting weighted team or bowling. If the team is batting weighted then which batsmen should play or if it is bowling a weighted team will it be given more priority to spin bowlers or seamers, these are the strategies they take for the team's success. Following all of these, the selection panel is sometimes criticized by fans or the media. There are plenty of reasons for that kind of accusation. Selection panels often consist of former cricketers or individuals with deep bonds to the cricket community. This can sometimes lead to biases based on personal preferences or regional affiliations, impacting the fairness of player selections. Eventually, it often happens that the selection panel chooses certain players even after a consequent poor performance in matches. Sometimes, it happens that senior players may have a strong influence on the selection process, leading to decisions that prioritize individual interests over the team's best interests. Sometimes, for political affiliations with the board may happen that some players may be included or excluded from the squad without any specific reasons. For these reasons, in this era, the traditional way of selecting squads are not fruitful anymore. Also this conventional approach of team selection has limitations in terms of subjectivity and time-consuming manual examination. Human biases and differing viewpoints within the selection panel might result in unsatisfactory judgments. Furthermore, the dynamic nature of cricket, as impacted by elements such as changing weather conditions and shifting player forms, necessitates a more responsive and data-driven strategy. The research challenge is to provide a simplified, impartial, and efficient approach for analyzing large datasets using machine learning. An automated approach might ease the stress on selection panels, offering objective insights and streamlining the process of choosing the optimum 11 players.

In this era of advanced technology, machine learning has been a promising system that is helpful to human's complex works. Eventually, machine learning can find patterns or insights of given data or stats. In a cricket match, a player's form plays a vital role. This is something unpredictable. So consistent performance is an attribute that is looked over before selecting a player in the main squad. On the other hand, there are more attributes like weather conditions, venue, pitch conditions, stats against the opponent team, experience, key player, etc. Using all of these attributes, and feeding the data into ML models, will give fair and unbiased output of best 11 players combination. Furthermore, the dynamic nature of the game necessitates real-time flexibility. Machine learning algorithms can monitor and evaluate live match data, providing a constant stream of information for in-game modifications and guaranteeing that the selected 11 players are optimal for the ever-changing conditions of a cricket match. This combination of modern technology and cricket dynamics has the potential to revolutionize how teams are created, strategies are developed, and, ultimately, success is gained in this respected sport

Predicting the best 11 players for a cricket team using machine learning involves a complex decision-making process that considers various player attributes, recent performances, team requirements, and match-specific conditions. Several machine learning models and techniques can be applied to address this task. There are many types of machine learning models to perform this task. Using Logistic Regression we can find out players performance predictions using batting average, bowling economy, strike rate, consistency, opposition analysis, etc. Random Forest Classifier on the other hand another model to that can give insights of any specific feature that is giving more scores

to any other models. Using that selectors can understand that specific feature carefully. Gradient Boosting Algorithm can capture the complex relation between a player's attribute and team selection criteria or specifically that strategy the team management wants to adopt. This model can produce a prediction that considers both individual player's performance and team dynamics.Additionally, Support Vector Machines can be used to classify the players by given attributes and stats of the player. Then management can choose the player for his specific skill to specific role to keep the team dynamic and balanced.

So it is clear that our aim is to give that hectic and time-consuming task to a computer to analyze the given data and stats to give the output of the best squad for the team which will be transparent as all the data can be visualized and also fair and unbiased. In essence, by employing machine learning models such as Logistic Regression, Random Forest Classifier, Gradient Boosting Algorithm, and Support Vector Machines, our objective is to revolutionize the conventional approach to cricket team selection. This transition from manual, subjective decision-making to data-driven, algorithmic forecasts promotes openness and justice. It streamlines the difficult work of selecting the top 11 players while also encouraging a more dynamic and balanced squad. This disruptive strategy has the ability to reshape the cricket environment by making judgments based on data, expertise, and unbiased analysis, therefore setting new standards for excellence.

1.5 Workplan

The proposed model focuses on the automated selection of an optimal cricket team. We began by gathering the dataset through raw data extraction from the official ESPN Cricinfo website. Following data collection, we explored the dataset by performing initial analyses to understand its characteristics and structure.
This flowchart is a machine learning model flowchart for predicting anomalies in cricket data. To execute this model perfectly there are major three steps:

1.Problem Defining:The main goal is to create an automated method for choosing the best cricket team. This entails specifying the precise objectives and results that the model is anticipated to produce. To comprehend the current approaches and best practices in team selection, clustering techniques, and cricket data analytics, a comprehensive literature research is carried out. Player statistics, performance indicators, and other pertinent data are gathered from the official ESPN Cricinfo website. The performance of the model is then verified by choosing appropriate testing sets.

2.Model Selection and training:This stage begins with exploratory data analysis (EDA) to understand the characteristics and structure of the dataset.The dataset is narrowed down by focusing on present and upcoming under-19 players after the data has been processed to remove errors and inconsistencies. High-dimensional problems are addressed by using dimensionality reduction techniques like Principal Component Analysis (PCA), and the data is arranged in a structured style appropriate for clustering approaches. A technique for handling null values is mean imputation. To find the best team, a variety of clustering methods are used, such as K-Means, Hierarchical Clustering, and DBSCAN. Using the preprocessed dataset, the model is trained, its parameters changed, and cross-validation is carried out to guarantee robustness.

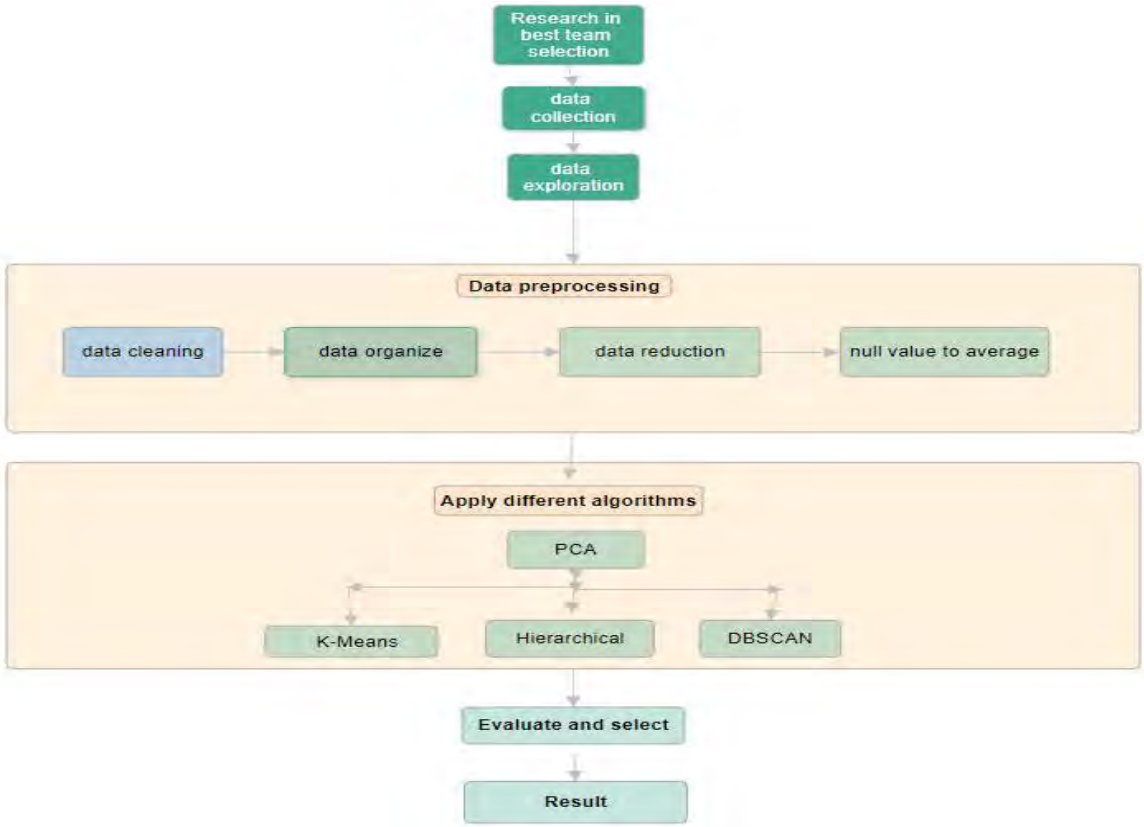

Figure 1.1: Work Plan

3. Decision: To find the best algorithm for team selection, the clustering results are assessed using metrics such as cluster consistency, Davies-Bouldin index, and silhouette score. These findings are used to determine the ideal team arrangement.

Chapter 2

Literature Review

From 1950 to 2023, machine learning has been evolving so has it's usefulness in many sectors of our lives. From our health-related issues to identifying bank frauds in every part of our life it has been a crucial part. It has also been used in cricket to take out insights of cricketers' performance, predicting scores to many other things that are related to cricket. In this research, we can have some help from some of these papers.

Like Patil et al. (2020) have introduced a system that can generate the best 11-player combinations for a cricket match [12]. So that the team management's workload can be minimized. This system values the player's past performances like strike rate, recent form of a batsman, bowling average, wickets, and economic rate for bowlers. Moreover, it values the opponent team's player's performance and venue as well. The authors used the data that was retrieved from websites, especially from ESPN. Technically authors used web scraping to create their own data set. The authors have applied two algorithms and these are the 'Random Forest Algorithm' and the 'Decision Tree Classifier'. According to the authors, the Random Forest Algorithm gave them the best result. The system they created to create the best 11-player combinations gave them a list of 11 players. However, how the team is effective or the accuracy rate is not mentioned.

Amin and Sharma (2012) worked on all kinds of players' stats so that one could pick the best batsmen, bowlers, all-rounders, and wicket-keepers for the T20 format matches [2]. Basically, here the authors had chosen the DEA method to differentiate the batsmen, bowlers, all-rounders, and wicket keepers using multiple outputs. Specifically, for the batsmen the authors had chosen here the highest scores, average run, strike rate, and how often 4's and 6's had been hit by a batsman as a factor of that DEA method for the batsmen so that the method can rank the batsmen with its linear and aggregation calculation that gives a score using those factors. Same for the bowlers, the factors are average total runs over total wickets, strike rate is total balls over total wickets, economy is total Runs over total overs, and total wickets taken by that specific bowler. For the All-rounders, the authors had chosen all the attributes of the batsmen as well as the bowlers for the scores and rank. For the wicket keeper, the authors had chosen the attributes of the batsmen. So applying these according to the authors either a club or a country can select their best team for the upcoming T20 match. Here in this research, the authors had chosen the 2007 T20 World Cup stats of the player for the test purpose. Then the authors ran a full analysis of the stats of the IPL T20 season 4. However, the accuracy rate was not mentioned in the paper.

The authors of this research [9] worked on all kinds of players' stats so that the authorities could pick the best batsmen, bowlers, all-rounders, and wicket-keepers for the ODI format matches. In this paper, the authors picked the statistics from January 2015 to September 2017 and these were retrieved from Howstat.com. But before selecting the best 15-player squad, the authors first predicted the outcome by another research paper (Croucher J S 2000 Proceedings of the fifth Australian Conference on Mathematics and Computers in Sport (Sydney University of Technology Sydney, NSW) pp 95106). Then the authors used integer optimization programming for the selection of the best team. Here the players were categorized into batsmen, batting all-rounder, wicketkeeper, bowler all-rounder, and bowler. As the attributes, the authors had chosen batting average, strike rate, consistency for the batsmen and bowling average, bowling economy, and consistency as well. But to calculate consistency, for both bowlers and batsmen, the authors used standard deviation and bowling or batting average. According to the authors, the limitation of this paper is the lack of information about wicketkeepers. Like except batting stats for the wicketkeepers what else attributes can be used to determine the best wicketkeeper. However, this method is also useful not only for ODI formats but for the T20 and Test Matches. On the other hand, the authors stated that this same method can be used in other sports such as football.

Sumathi et al. (2023) proposed a system that predicts the performance of a cricket player [23]. Here authors performed machine learning methods in a serial way. At first, the authors had gathered the pre-processed dataset from Kaggle. Then performed linear regression model, then K-means model, and lastly Random forest Analysis to predict the performance. After completing the process the authors found 14 clusters with the value of '1' which means the best players. According to the authors, it is a successful model that can be useful for any other datasets related to any other sports or games to rank and identify the best players.

In this research [19] the authors proposed the prediction of the efficient players that can be suggested by a system. Here AI and ML are used for this certain problem. At first, the authors collected the dataset from Kaggle with over 2000 data points and processed the dataset eliminating some attributes like country, date of birth, serial ID, etc. as the authors are predicting for only IPL. Here a random forest classifier is used for this system. After applying this model, the model predicted the outcome with 92% of accuracy. The authors stated that with more decision trees the accuracy can be more accurate. According to the authors, this system can be used in other games if the system is organized.They also check on other attributes that are required in each sport the feature selection can be changed and further moved on.

Jha et al. (2022) presented a hybrid approach for team selection in fantasy cricket using 'Recursive feature elimination', 'Random forest', and 'Genetic algorithm' [21]. The author utilized an ICC dataset that has a numerical analysis of 332 cricket players and is extensively used by other researchers. They tested the model on 8 features and utilized it with Random Forest Recursive Feature Elimination (RF-RFE). The accuracy using this approach is 84.2% for bowlers and 82% for batsmen. The accuracy may be enhanced if they use more datasets.

Another research paper[11] presented an efficient machine-learning technique that can select the best player by predicting individual players' performance. The author uti-

lized a holistic dataset and described 8 factors. They tested only two factors: weather statistics and players' past records. Which have a total of 24 features. The authors tested the features by utilizing SVM, Naive Bias, Decision Tree, and Random Forest. They discovered that the random forest algorithm gives 93.73% accuracy and it gives the highest accuracy among the algorithms. They discovered the right features, if they use total factors then the accuracy will be more accurate.

Mukherjee et al. (2023) presented a spectral clustering framework for selecting the best substitute player[20]. The author utilized the ESPNCricinfo dataset by scraping their website. This paper had 20 features and they used Pearson Correlation Coefficient. They get 77.50% accuracy from this. The authors use two algorithms DBSCAN and Spectral. They conclude that this spectral algorithm gives much more accuracy and also takes less time compared to DBSCAN from the same dataset.

In another study[3], titled "Automated Player Selection for Sports Teams using Competitive Neural Networks,"a new method was used to optimize football team selection. While the dataset size is not specified, neither the source of their competi- tive neural network model returned very encouraging performances, reaching up to 60 percent accuracy rate in team outcome predictions. This current study fills a critical gap in sports analytics field

The research paper,[27] titled "Enhancing Cricket Team Selection Through a Priority-Based Optimisation Model," offers a novel method for selecting cricket teams that combines the Analytical Network Process (ANP) and Multiple Criteria Decision Making (MCDM). The research is based on actual player attributes from Pakistan's cricket squad in October 2019. The ANP model evaluates players within this framework using standards like batting, bowling, all-around skills, and wicket keeping, with a careful analysis of consistency ratios to assure correctness. In the end, this technique is a useful tool for building successful cricket teams.

S. Banerjee et al. (August 2019) authored a paper presenting a system that recommends players for selection in a team based on heuristic player rankings and a greedy algorithm. [7]. Their algorithm assists decision makers in identifying the team and suggests alternate players if the desired player is unavailable. The paper considers derived features quantifying them to assign scores to each player in the pool. These scores are used to rank the players aligning with known IPL player rankings. Additionally each player is assigned a score on a scale of 1-10. The findings of this study highlight the inclusion of two higher level clusters; batting all rounders and bowling all rounders. However it should be noted that the greedy algorithm for team selection may result in combining ranked players with those ranked lower. To address this limitation further investigation into programming approaches could be worthwhile to ensure a balanced inclusion of high quality players, within the team.

In another research paper [14] a method is suggested for classifying cricket shots using a forest algorithm. They base their approach on the idea that the posture of a batsman's the factor in determining the type of cricket shot they play.Therefore, they use MediaPipe, a framework for creating multimodal, cross-platform applied machine learning Pipelines, to extract the human postures from an image as a set of keypoints.Results of the experiment reveal that the suggested model exceeds the current answer by 5% and

obtains an F1-score of 87%.Additionally, they suggest a similarity assessment method to find the cricket shot image that is the most similar to the user's cricket image from those of well-known international cricket players. Cricket players will be able to examine, enhance, and track their batting performances without the need for a coach thanks to the mobile application they built based on their solution.

Surendran et al. (2023) the author of this paper has introduced a productive way to work on 'Indian Premier League' (IPL) Data Exploration duration (2008-2020) using Python [22].With the help of Python library like 'Numpy' for Scientific Computing, 'Pandas' for Data Analysis, and finally 'Matplotlib' and 'Seaborn' for Data Visualization. By using these application modules, preprocessing, data analysis, and visualization are used to develop a model that forecasts the chances of team winning or not.This paper focuses on player performance consistency, particularly batsman performance, and it tackles the study that is done for the most men of the match,It also consider, the top batsmen, and the top 10 performers on the most runs. In this study, 816 games were used along with toss-related breakdowns including total toss victories and decisions taken by each squad after winning the toss during the course of the tournament.

Mahbub et al.(2021), the authors of this paper, represent the idea of identifying the squad of eleven players for the Bangladesh cricket team with the help of Machine learning algorithms [15]. The dataset was obtained by the authors via the Espncricinfo website, and they also gathered historical data from highlights of a few games. They choose nine features for bowlers and eleven for batters. Next, develop a rating based on various characteristics that may be used to gauge a player's performance. Without taking any players' recent performance into account, all ratings are determined based on the general profile. After creating ratings for each batsman and bowler, they created three models using the SVM, Naive Bayes, and Random Forest machine learning techniques. then used a dataset to train each model. With an accuracy of 94 percent for the batter dataset and 93% for the bowler dataset, they achieved their maximum accuracy on the support vector machine.

Another paper [16] developed a effective genetic algorithm (GA) and recurrent neural network (RNN)-based model for selecting a cricket squad. The proposed approach is a hybrid one, which includes a genetic algorithm with the concept of RNN in order to select efficient players. Herein, this has utilized historical statistics of players for creating initial feature matrices, which were then refined using the GA so as to minimize the loss factor. Further assessment was done by RNN after refinement of matrices for assigning final scores to individual players. This resulted in a parallel rank table that would help a team selector to select players rapidly for any upcoming match or tournament. Experimental validation, referencing three real datasets, showed the impressive results of the model, with frequent outperforming compared to manual team selection. On an overall basis, the accuracy of 98.5 percent was excellent in predicting player lists of matches compared to manual selections.

Ishi et al. (2022) present an enhanced model for team selection using unbiased techniques and consider various factors like batting and bowling averages, opponent strengths, and weaknesses. Nature-inspired algorithms, specifically CS-PSO (Cuckoo Search and Particle Swarm Optimization), are used for significantly feature optimization to improve machine learning models prediction [18]. The authors used data from p ublicly available

sources. They measure the accuracy in five factors- batsman strength, Bowler Strength, batting all-rounder, bowling all-rounder, and wicket-keeper performance. Batsman strength is calculated based on 25 characteristics, 23 for bowlers, 45 features for batting/bowling all-rounders, and 23 for wicketkeepers. In all respects, SVM with CS-PSO offers the maximum accuracy in every way. CS-PSO's accuracy is superior than that of the individual CS and PSO methods. Combining feature optimization techniques with machine learning models results in the highest prediction accuracy. This approach achieves high accuracy in selecting players for different roles, ranging from 92.63% to 97.29% for different player categories.

Chapter 3

Description of the Data

3.1 Data Collection

Our thesis, "Automated Selection of Optimal Cricket Team Using Machine Learning" focuses on selecting the top-performing cricket players for the Bangladesh national team using machine learning. Firstly, the dataset was meticulously collected from the ESPN Cricket official website which is a publicly available online open source, specifically utilizing the StatsGuru section, which serves as a comprehensive database for player statistics. Additionally, the procedure of gathering data was extensive and had to be done manually by humans to be precise. We started by looking for each player on our own, covering both established players of the national team and bright prospects for the Under-19 squad. For our, work we collected performance data for every player in the Test, ODI, and T20 formats. In order to do this, several datasets with various match formats and circumstances had to be gathered. On top of that, in StatsGuru platform, we searched each player's name and meticulously recorded their performance data. This included batting and bowling averages, strike rates, and other key performance indicators. Overall, We focused on collecting the average performance of each player against different countries, ensuring we captured their effectiveness in various competitive scenarios. In addition to performance metrics, we also considered factors like batting position and weather conditions, which were crucial for contextual analysis, this will help our system to be adaptable to various cricket environments. The collection of data from both the established players and the Under19 squad ensured that our system could produce fair suggestions and a wider range of data was collected and used to make sure our system could be adaptable to altering the cricket environment. This comprehensive data collection approach allowed us to create a robust dataset, laying the groundwork for our machine learning models aimed at optimizing team selection for the Bangladesh national cricket team.

3.2 Data Cleaning and Preparing

In the initial stages of data cleaning and preparation, we focused on selecting only the selective players for our analysis so that the earlier stage of our dataset could be more precise. Basically, this included both current national team players and promising upcoming players from the Under-19 in our dataset. Finally, Our main aim was to generate the best team composition by combining the already established talents from the national team and emerging talents from the Under-19 players. One significant

Figure 3.1: The Dataset and Collection Process

challenge we encountered was the issue of high dimensionality in our dataset since cricket data is vast containing a large number of variables. To address this, we undertook several key steps:

3.2.1 Dataset Collection:

We meticulously collected data from publicly available online open sources for each player both from the Bangladesh National team and the Under-19 team, focusing solely on performance metrics which were critical for our analysis against individual teams of other countries.

3.2.2 Dataset Organization:

We carefully organized the dataset into a well-structured format to ensure optimal performance and accuracy of our system for selecting players for the team. To create this well-structured format we included player names as well as their respective relevant features such as performance metrics and other relevant attributes. The features were carefully selected in an organized way so that there was consistency and easy access to data points. It was also ensured that the dataset had the correct dimensions required for our processing so that our machine learning models could work effectively on our dataset.

3.2.3 Dataset Reduction:

We streamlined our data by focusing solely on each player's average[17] performance against each team, discarding other columns that were not essential for our primary objective. Because the average consist of player overall performance of each match. This reduction in dimensionality was necessary to improve overall performance of our model.

3.2.4 Handling Null Values:

We addressed any missing values in our dataset by filling them through a systematic process of imputation to ensure the integrity and completeness of our dataset. Specifically, we filled in the missing values using the calculated average values along the horizontal axis, this axis represents each player‘s average performance against other countries to ensure no gaps in the datasets that could affect the model's accuracy or performance.

3.2.5 Calculating Players Form

In our study, we applied the approach developed by Passi and Pandey[5],in their paper they proposed a formula for calculating a player‘s form based on their previous performances. The authors decided to separate the concept of "form" into two separate equations, Batting Form and Bowling Form. We adopted these equations and calculated them by using the player's previous match records and applied them to our dataset. These forms are

3.2.6 Batting Performance

$$\text{Batting Form} = (0.4262 \times \text{Average} + 0.2566\ times\text{No.of innings} + 0.1510 \times \text{SR} + 0.0787 \times\text{Centuries} + 0.0556 \times \text{Fifties} - 0.0328 \times \text{Zeros}) \quad [6]$$

By combining important performance metrics like average, innings played, strike rate (SR), hundreds, fifties, and zeros, the Batting Performance formula calculates a player's batting form [6]. Every element is given a weight that corresponds to its significance in assessing overall performance. While the number of innings signifies experience, which frequently results in increased confidence and performance, the average gauges a player's scoring effectiveness. In order to gain momentum in modern cricket, a player's strike rate measures how quickly they score. While zeros indicate moments of failure that can affect confidence and team dynamics, milestones like hundreds and fifty demonstrate a player's capacity to make a substantial contribution to match outcomes.

This all-inclusive algorithm produces a score that accurately depicts a player's present batting form, giving coaches and analysts a trustworthy way to forecast future results. The formula provides a comprehensive picture of a player's abilities and patterns by combining a number of key performance indicators. A high batting form score indicates confidence and competence in the player's abilities and indicates that they are likely to continue or even improve their performance in future matches. On the other hand, a low score can portend difficulties down the road, leading coaches to think about changing their approach or player rotation.

In order to determine how well a player is expected to perform in the near future, teams can also compute recent form by looking at the player's performance during the previous five games. With the use of this recent form analysis, one may ascertain whether a player is ready to perform well based on their most recent performances. In the end, team managers can make well-informed decisions on player selection and strategy by using the batting performance formula. Teams may increase the likelihood that they will win games and maximize their lineups to take advantage of players who are performing well right now by utilizing both recent performance measures and overall form.

3.2.7 Bowling Performance

$$\text{Bowling Form} = 0.3269 \times \text{No. of overs} + 0.2846 \times \text{No. of innings} + 0.1877 \times \text{SR}$$
$$+0.1210 \times \text{Average} + 0.0798 \times \text{FF} \quad [6]$$

By processing the data in this systematic manner, we were able to simplify the dataset and effectively resolve the dimensionality issues that could have hindered our model's accuracy. Moreover, the reduction of dimensionality was essential to ensure that only the most relevant and meaningful performance metrics were used for selecting players. This targeted approach allowed us to ignore extraneous variables thereby making our dataset more compact and easy to work with in terms of both computational efficiency and memory requirements. This approach allowed us to focus on the most relevant performance metrics, such as averages, strikes, and rates to build a more robust model that can make decisions based on the most relevant attributes. All these steps helped us facilitate more accurate clustering and identification of the best players for the Bangladesh national team.

By combining key performance metrics such the number of overs bowled, innings, strike rate (SR), bowling average, and a performance factor (FF), the Bowling Performance formula determines a player's bowling form [6]. Each element is given a weight that corresponds to its importance in assessing bowling effectiveness as a whole. A player that has a high bowling form score is likely to do well in subsequent games, demonstrating a consistent ability to take wickets and keep the run rate low. On the other hand, a lower score can indicate possible problems in the road, warning analysts and coaches about areas that need work.

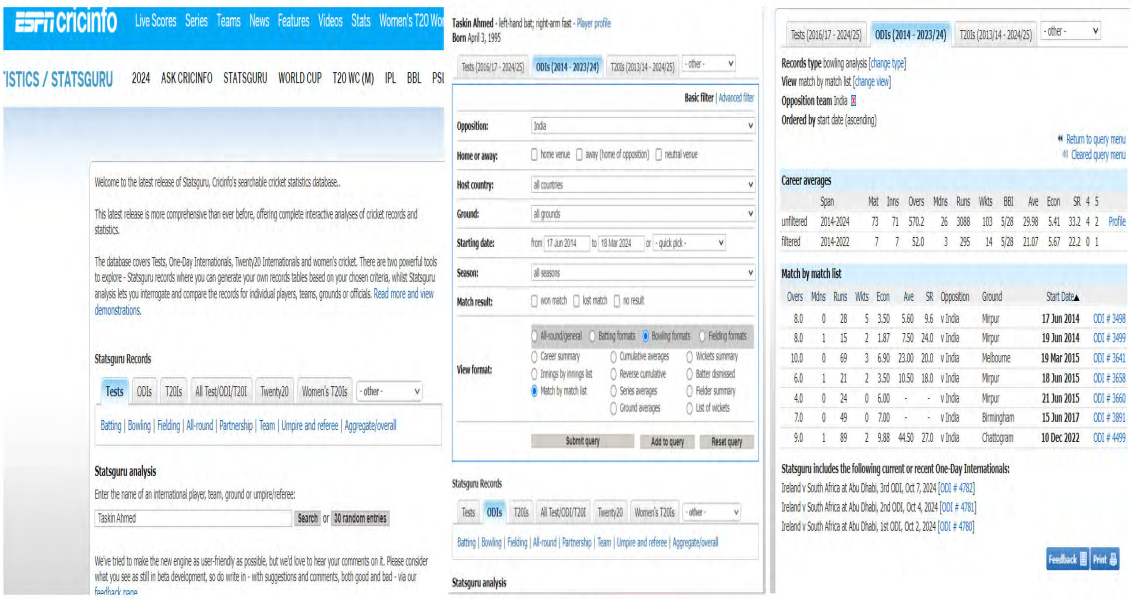

Figure 3.2: Recent form dataset collection Process

Teams can determine recent form, which is essential for comprehending a player's present level of play, by examining a player's performance over the previous five games. This helps them make even more accurate forecasts. This recent form analysis helps determine

whether a bowler is likely to put on excellent performances in future games by offering insights into their current rhythm and effectiveness. Teams are able to make well-informed decisions about player selection and strategy by integrating recent performance information with overall bowling form. By making sure that players who are performing at their best are given the chance to improve their team's performance, this strategy increases the chances of success.

3.2.8 Bowling Consistency

$$\text{Bowling Consistency} = (0.4174 \times \text{No. of overs} + 0.2634 \times \text{No. of innings} + 0.1602 \times \text{SR} + 0.0975 \times \text{Average} + 0.0615 \times \text{FF} \quad [6])$$

An example of a bowler's consistency and performance over time is bowling consistency.Here, The bowler's strike rate (SR), bowling average, number of overs bowled, number of innings played, and a performance-related factor (FF) are all taken into consideration. A thorough understanding of the bowler's consistency across various characteristics is made possible by the weighted formula.

Bowling consistency is one of the crucial for predicting a player's upcoming performances as it predict their future performance and adaptability over time. Actually, it is essential for forecasting their future performances.Moreover, A reliable bowler is more likely to maintain low runs per over and take wickets, which makes them useful under pressure to against team and it helps to take wickets. Analyzing consistency scores helps coaches make informed decisions about team selection and strategy, while also revealing performance trends that can guide training adjustments.Finally, Consistent bowlers can also be compared to their counterparts, which helps with team development. All things considered, a bowler's capacity to produce high-quality outcomes is encompassed by consistency, which is crucial for sustained success in cricket.

Element	Description	Notes (Divide by Zero)
Average	Batting/Bowling average (runs per wicket)	If no wickets, set average to Null
No. of overs	Number of overs bowled	N/A
No. of innings	Number of innings played	N/A
SR (Strike Rate)	Runs per 100 balls for batsmen, balls per wicket for bowlers	N/A
Centuries	Number of centuries scored by the batsman	N/A
Fifties	Number of fifties scored by the batsman	N/A
Zeros (Ducks)	Number of innings where the batsman scored zero runs	N/A
FF (Performance Factor)	Special performance factor for bowlers, indicating economy, key wickets, etc.	N/A

Table 3.1: Explanation of elements in the Batting and Bowling Formulas

Chapter 4

Description of Models

Our research is completely dependent on the performance of the Bangladeshi cricket players. The performance of the players is clearly understood based on their average of each match. Using that average we will separate good players and poor players with the help of clustering. K-Means, Hierarchical, DBSCAN. These models are used to separate players from average values according to their playing conditions. Silhouette Analysis to evaluate the result of the clusters. For these research we know about those models analysis and their process of application.

4.1 CLUSTERING MODELS

4.1.1 Standardization

Standardization transforms numerical features so that they have a mean of 0 and a standard deviation of 1. This process is essential before performing K-means clustering to ensure that no single feature disproportionately affects the clustering due to its scale. The equation for standardization of a feature X_j is:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \quad [1]$$

Where:

- Z_{ij} = Standardized value for the j^{th} feature of player i
- X_{ij} = Original value for the j^{th} feature of player i
- μ_j = Mean of the j^{th} feature across all players
- σ_j = Standard deviation of the j^{th} feature across all players

The formula for calculating the sample standard deviation of the j^{th} feature is:

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2} \quad [1]$$

Where:

- s_j = Sample standard deviation of the j^{th} feature
- n = Number of data points in the sample

4.1.2 K-Means

K-Means algorithm is popularly used to partition a dataset[24]. It is a clustering algorithm which works for unsupervised machine learning. Initially this algorithm selects centroids randomly and then assigns it to the nearest centroid of the datapoint by forming clusters. After that it calculates the mean of each cluster.

$$\mathrm{J} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad [13]$$

Where:

- J is the total within-cluster sum of squares.
- k is the number of clusters.
- C_i is the set of data points assigned to cluster i.
- μ_i is the centroid of cluster i.
- $\|x-\mu_i\|^2$ is the squared Euclidean distance between a data point x and the centroid μ_i.

4.1.3 How K-means Clustering Works in Our Code

Centroid Calculation: During the K-means clustering process, centroids are calculated based on the means of the features for the data points assigned to each cluster. The algorithm iteratively refines these centroids until convergence.

Feature Mean Values: The *player_mean_value* calculated in the code provides a secondary measure to evaluate player performance. It is not used in the clustering process but offers insight into how players rank within their clusters. This only works to rank the players in their respective cluster.

Cluster Formation: Players are grouped into clusters based on their numerical feature values, and the mean values of these features determine the location of the centroids, leading to the formation of distinct clusters. In essence, the K-means algorithm groups

players by considering the distributions of their numerical features, and the mean values of these features help identify and sort players based on their performance within the defined clusters.

4.1.4 Hierarchical:

Hierarchical clustering model is also a part of unsupervised machine learning algorithm[26]. In this model clusters build hierarchy. This model calculates the distance metrics and linkage criterion. With this calculation this model merges smaller clusters to a larger one and similarly breaks a larger cluster to smaller clusters.
In the above diagram, show how Hierarchical clustering model merge a single cluster from many smaller clusters. In the left side shown every single data point to its nearest data point makes a cluster, then after that make a bigger cluster with the nearest data point. This continuing process ultimately makes one big cluster. In the right side the diagram shows exactly the same process with datasets.

4.1.5 How we are performing Hierarchical Clustering in our Dataset

In Hierarchical Clustering there are 4 distinct methods to perform this clustering. These are Single Linkage, Average Linkage, Complete Linkage, Ward's Linkage. Below in the figures we can see the Complete Linkage and the Ward's Linkage can cluster perfectly for the dataset and also give us insights about how many cluster should we choose. For both we have chosen n = 3 which is above the threshold value.

Ward's Linkage

The distance between two clusters A and B is calculated using the formula:

$$d(A, B) = \frac{n_A n_B}{n_A + n_B} \cdot \|\mathbf{c}_A - \mathbf{c}_B\|^2 \quad [4]$$

Where:

- $d(A, B)$: Distance between clusters A and B.
- n_A: Number of points in cluster A.
- n_B: Number of points in cluster B.
- $\mathbf{c}_A$: Centroid of cluster A.
- $\mathbf{c}_B$: Centroid of cluster B.
- $\|\mathbf{c}_A - \mathbf{c}_B\|$: Euclidean distance between the centroids of A and B.

Example: Let's say clustering has 4 players based on their batting averages:

- Player A: 40
- Player B: 42
- Player C: 85
- Player D: 90

Start with each player as their own cluster:

Cluster 1: {A} Cluster 2: {B} Cluster 3: {C} Cluster 4: {D}

Calculate the increase in variance for each potential merge:

- If we merge {A} and {B}, the average is 41.
  Variance = $(40-41)^2 + (42-41)^2 = 2$ units.
- If we merge {C} and {D}, the average is 87.5.
  Variance = $(85-87.5)^2 + (90-87.5)^2 = 12.5$ units.

Merge the clusters with the least increase in variance. Merge {A} and {B}, as the increase in variance is smallest (2 units).
Now we have 3 clusters:

Cluster 1: {A, B} Cluster 2: {C} Cluster 3: {D}

Continue merging until you reach the desired number of clusters.

Complete Linkage

The distance between two clusters A and B is defined as:

$$d(A, B) = \max_{\mathbf{x} \in A, \mathbf{y} \in B} \|\mathbf{x} - \mathbf{y}\| \quad [8]$$

Where:

- $d(A, B)$: Distance between clusters A and B.
- $\mathbf{x}$: A point in cluster A.
- $\mathbf{y}$: A point in cluster B.
- $\|\mathbf{x} - \mathbf{y}\|$: Euclidean distance between points $\mathbf{x}$ and $\mathbf{y}$.

Example: Suppose players' values against 2 features are given:

- A: vsAUS = 160, vsIND = 60
- B: vsAUS = 165, vsIND = 62
- C: vsAUS = 190, vsIND = 90

Start with each player as their own cluster:

Cluster 1: {A} Cluster 2: {B} Cluster 3: {C}

Find the maximum distance between each pair of clusters:

- Max distance between {A} and {B}: 5 units.
- Max distance between {A} and {C}: 30 units.
- Max distance between {B} and {C}: 27 units.

Merge the clusters with the smallest max distance. We merge {A} and {B} since their max distance is the smallest (5 units).
Now we have two clusters:

Cluster 1: {A, B} Cluster 2: {C}

Stop when you reach the desired number of clusters (2 in this case).

4.1.6 DBSCAN:

Density-Based Spatial Clustering of Applications with Noise or DBSCAN is a newly popular clustering model where data points are divided into two points. The data points density is considered as noise. The DBSCAN model works by treating some of the data points as key data points. That key-datapoint defines a circle around it of a size equal to a specified radius. And all the data points located inside this circle belong to the same cluster.
DBSCAN is a nested cluster model that can identify high dimensions data points[25]. In this model we considered two major values that divide the data points into many clusters. The maximum radius of the key data points neighborhood, which is Eps(Epsilon). And the other is MinPts, which refers to the minimum number of data points of the radius. Those data points that cannot be settled in any key data points radius that data points are disclosed. This is how the DBSCAN model works.

4.1.7 How we are performing DBSCAN Clustering in our Dataset

As of we know this DBSCAN is using number of samples per iteration as well as the radius from the iteration point, we tuned it further for our dataset. We tuned it to eps= 5 and min-samples = 5. After this tuning we got two clusters on from out data points. Cluster 0 and Cluster -1.

4.1.8 Silhouette Score:

To evaluate the result of the clusters we use silhouette method. Using this we can identify grouped cluster data points of our dataset. This method is used to understand the performance of the clustering of the dataset. Using the silhouette model over the K-means model the result shows us the number of grouped clusters of our dataset is 2.

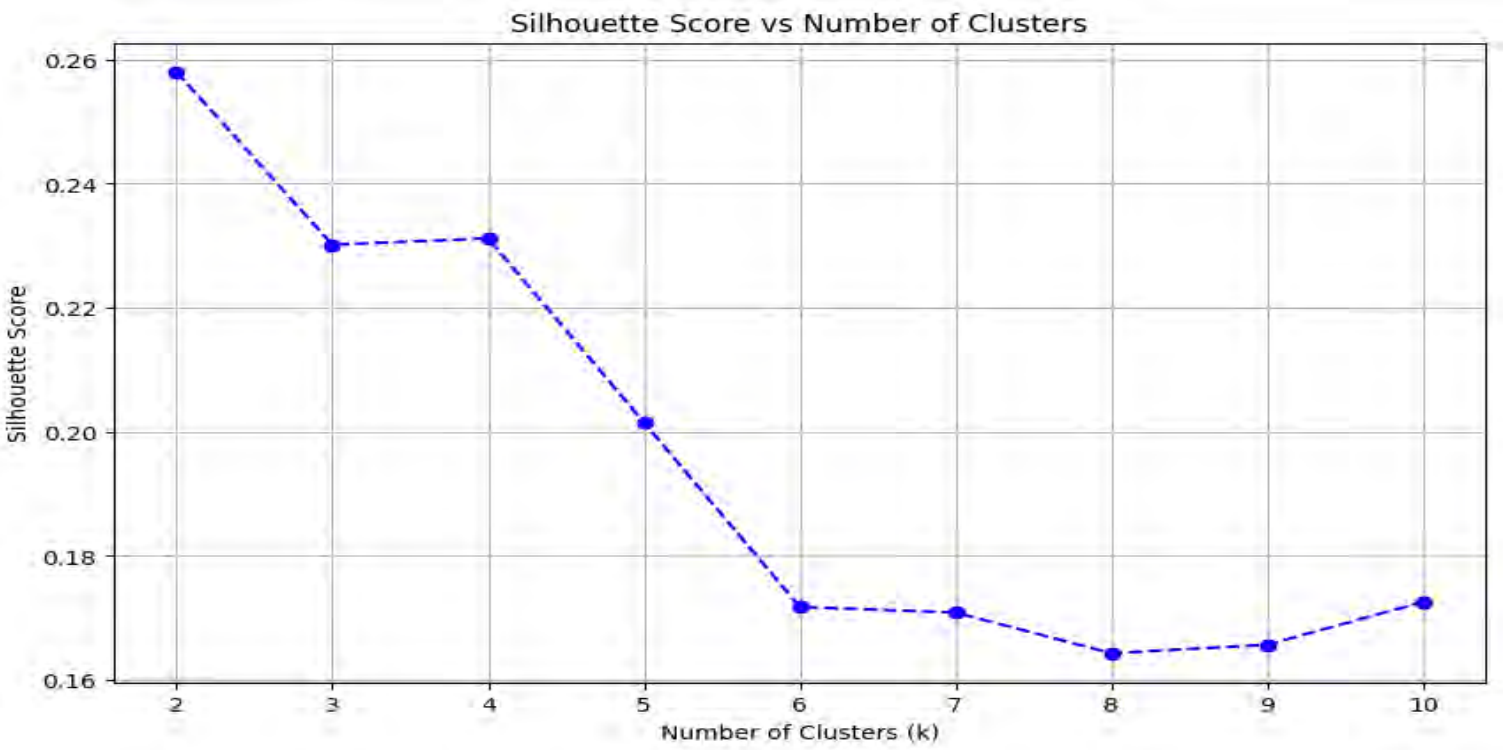

Figure 4.1: Silhouette Score to find the Cluster value for Batting

Silhouette Score for 2 clusters: 0.2580. By using this algorithm we find the result of clusters. When we use the K-Means algorithm, the silhouette model shows how many clusters we should use in this model.This is for our batting dataset. Silhouette Score for 2 clusters: 0.4330 . By using this algorithm we find the result of clusters. When we use the K-Means algorithm, the silhouette model shows how many clusters we should use in this model.This is for our bowling dataset.

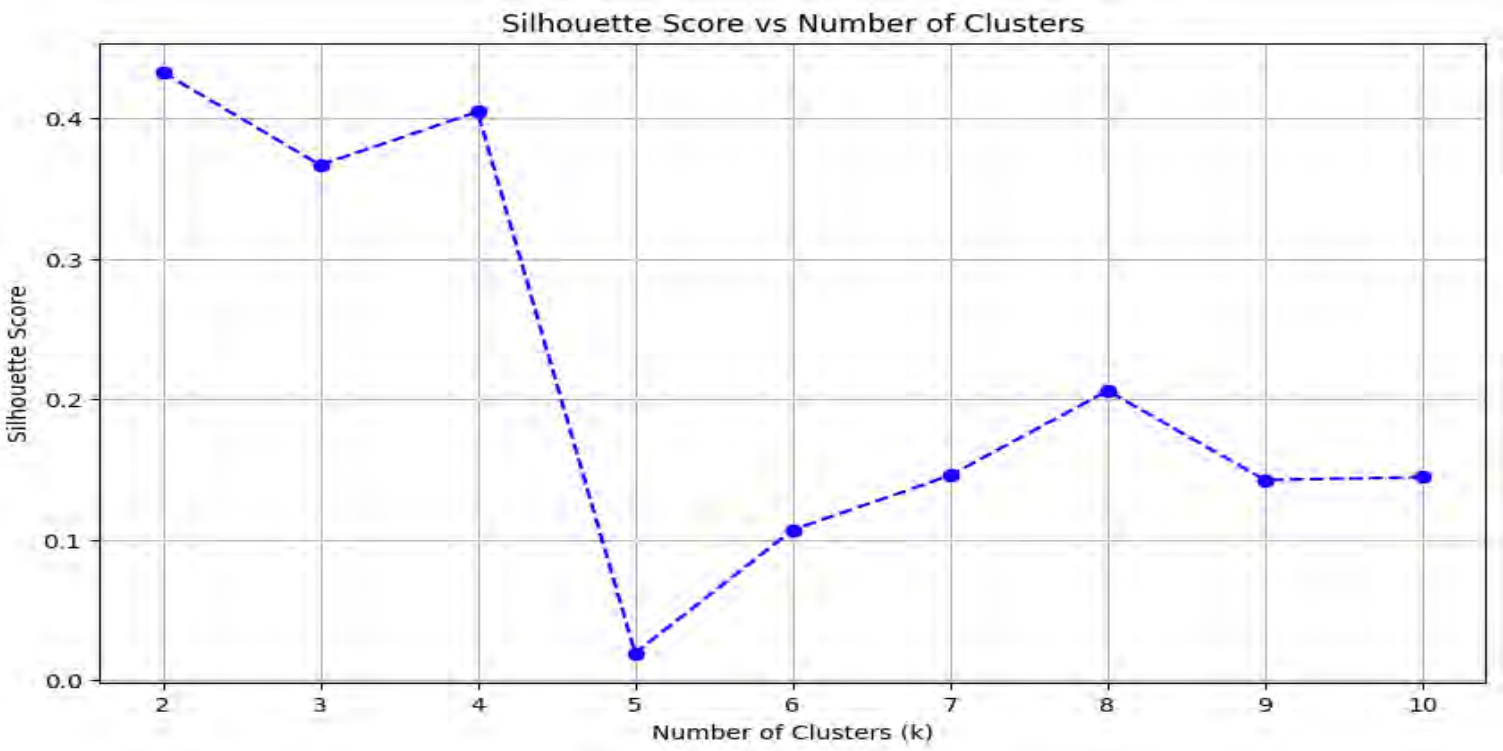

Figure 4.2: Silhouette Score to find the Cluster value for Bowling

4.1.9 Elbow Method:

This machine learning technique is used to find out the optimal number of clusters[24]. This model shows the number of clusters by showing elbow points. This model can visualize perfectly if it uses some algorithm. In the K-Means algorithm, the elbow method works perfectly by determining the best K values. The graph shows a bent elbow when the model finds the optimal K values. We use the elbow model to find the optimal cluster value. That's why we use the K-Means algorithm and after that by using the elbow model we find the cluster values for our research.

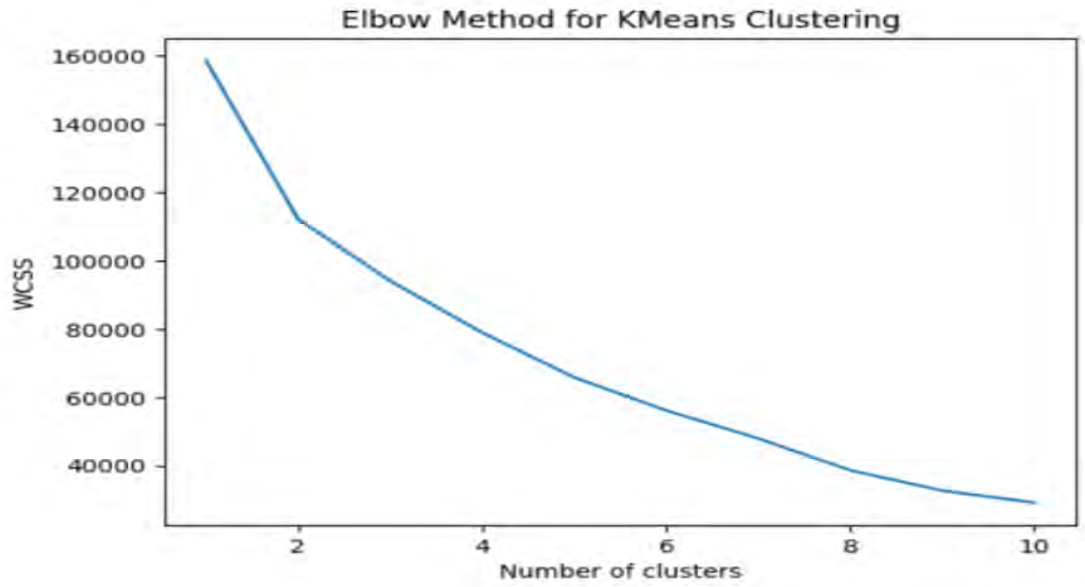

Figure 4.3: Elbow Method to find the Optimal Cluster value for batting

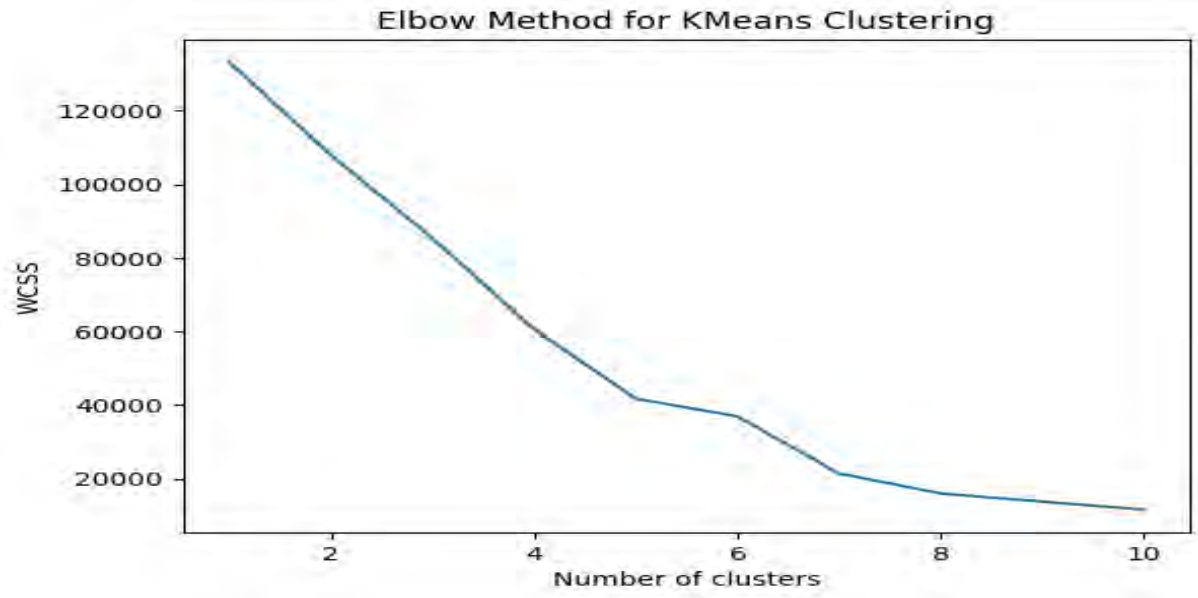

Figure 4.4: Elbow Method to find the Optimal Cluster value for bowling

4.1.10 Principal Component Analysis:

Principal Component Analysis(PCA) is an unsupervised machine learning technique that is used to choose the absolute variable with necessary information[28]. For better analysis this model can transform a high amount of datasets into smaller ones. Because smaller datasets are easy to process and also simpler. To understand the hidden patterns of the dataset and also reduce the unnecessary variables that are not effective on the dataset we use this algorithm. By using this we can clearly understand the data points and also visualize them. We use this algorithm to find and understand the data plots and visualize them. This algorithm clearly shows the values of the clusters with plots of the data points.

4.1.11 ExKMC Model:

The Extended K-Means Clustering (ExKMC) model is the updated version of the K-means Clustering algorithm. The base KMC model has some disadvantages when handling outlines, difficulty shows in varying cluster density of the datasets and has a limitation of non-spherical clusters[10]. But in the standard version which is an ExKMC model that can handle the outliers, the initialization has been improved and also adaptive distance metrics.
The ExKMC model helps to build a decision tree. This tree is based on the feature that describes how the machine learning process is distinction those data points briefly the players are separating based on features shown on this process called ExKMC method which is shortly a decision tree.
Furthermore, the ExKMC model helps to construction of a decision tree which is an essential step in many machine learning procedures. Because this decision tree is built using the characteristics that set the data points apart and moreover it makes it evident how the clustering process makes distinctions between different groups. The ExKMC approach offers information on how players might be grouped according to performance indicators, playing styles, or other pertinent characteristics when it comes to player analysis. In the end, this procedure not only improves comprehension of player dynamics but also facilitates better decision-making in areas like strategy creation and recruitment.

Chapter 5

Result and Analysis

5.1 Model Implementation and Result

5.1.1 K-Means:

We have run the k-means model with the values of k =2 and random state=42 where we get two different clusters with two different mean values and which players fall under those clusters. With these clusters we got the Silhouette Score for 2 clusters: 0.2938

Cluster 1: Aminul Islam, Hasan Mahmud, Arafat Sunny, Mustafizur Rahman, Nasum Ahmed, Rubel Hossain, Shamim Hossain, Shoriful Islam, Tanzid Hasan, Yasir Ali, Sohag Gazi, 'Taijul Islam, Taskin Ahmed, Mohammad Naim, Mehidy Hasan Miraz, Mohammad Mithun, Mominul Haque, Sabbir Rahman—— with mean value of 19.015

Cluster 0: Mohammad Saifuddin, Mosaddek Hossain, 'Mushfiqur Rahim, Towhid Hridoy, Afif Hossain, Anamul Haque, Imrul Kayes, Liton Das, Mahmudullah, Najmul Hossain Shanto, Nasir Hossain, Shakib Al Hasan, Soumya Sarkar, Tamim Iqbal — with mean value of 26.463

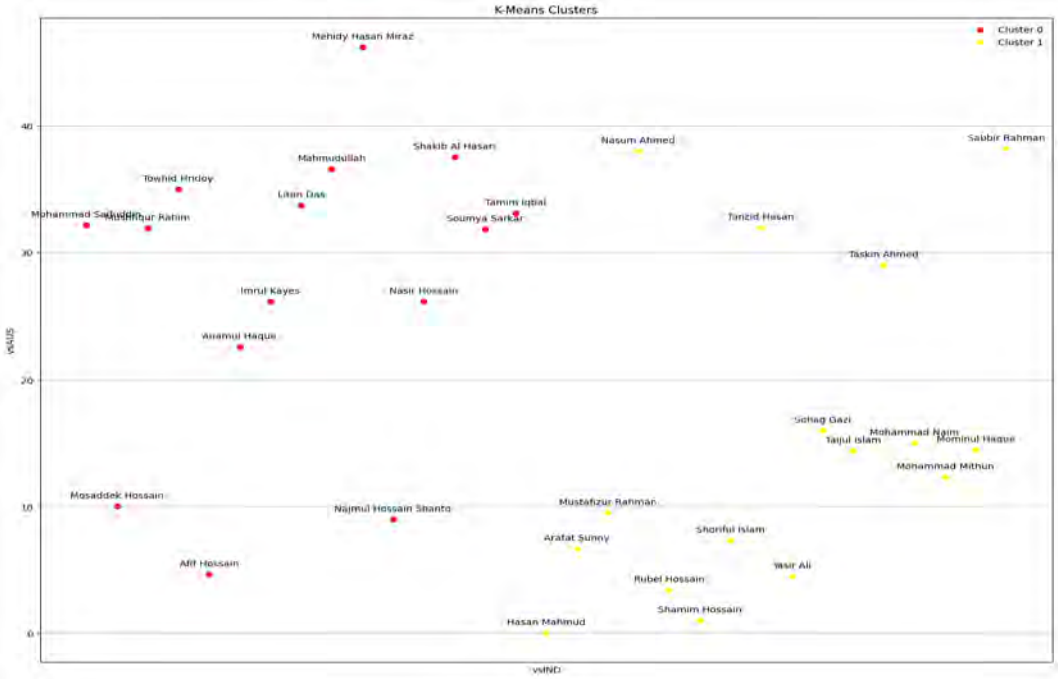

Figure 5.1: K-Means Clustering result

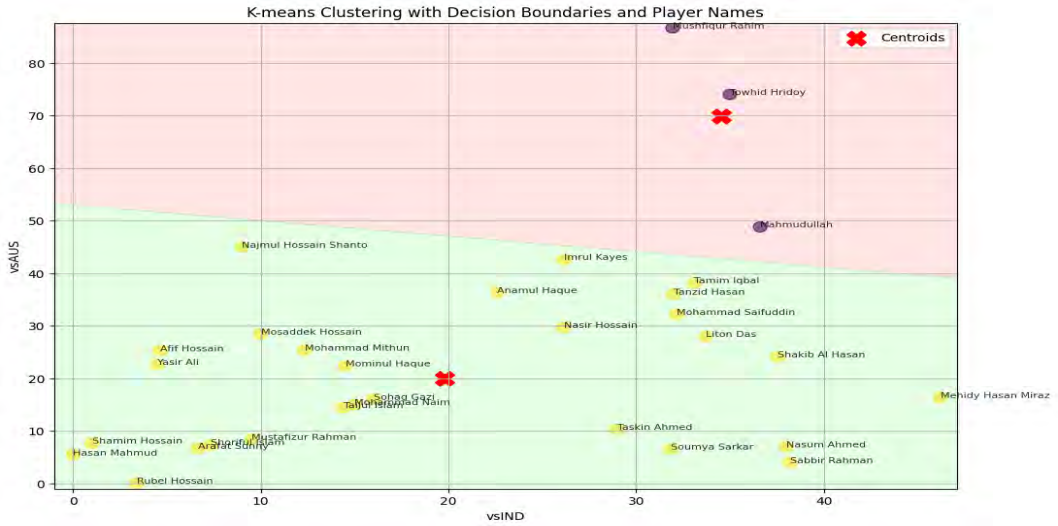

Figure 5.2: K-means Clustering based on two features

These figure shows the centroids of the clusters. The bellow part shows the players who are not in good form in batting. And the above part shows the player who are comparatively in good form in batting. Here the clustering work only for two feature.

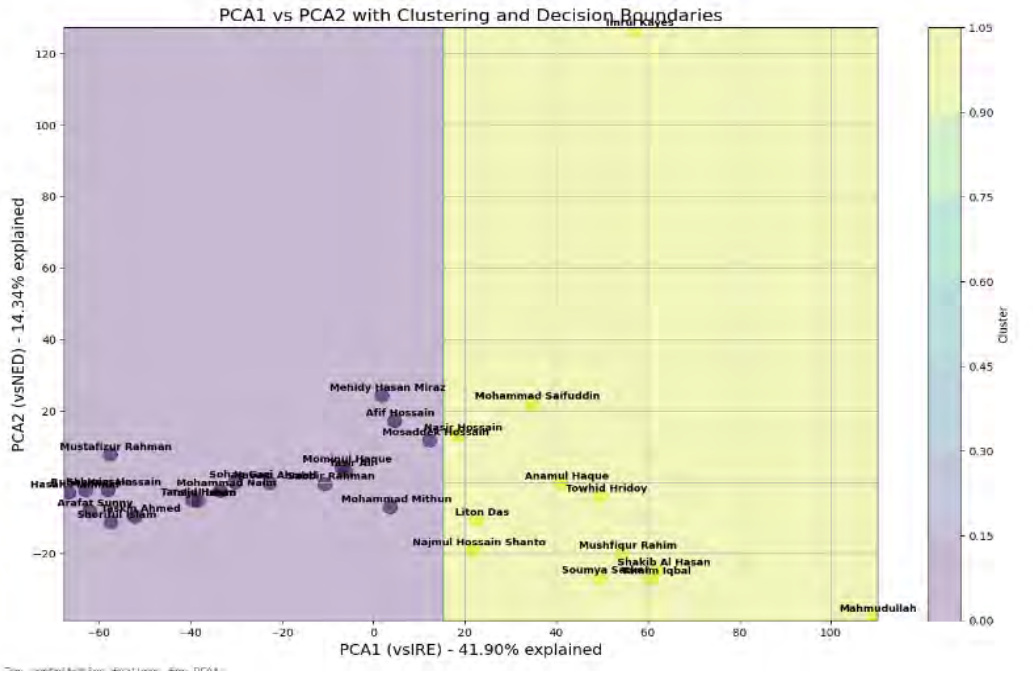

Figure 5.3: Using all features (PCA)

Here using K-means, all the features are merged into two features and the plot shows the clusters among those players based on the Principle Component Analysis that the dimension we reduced.

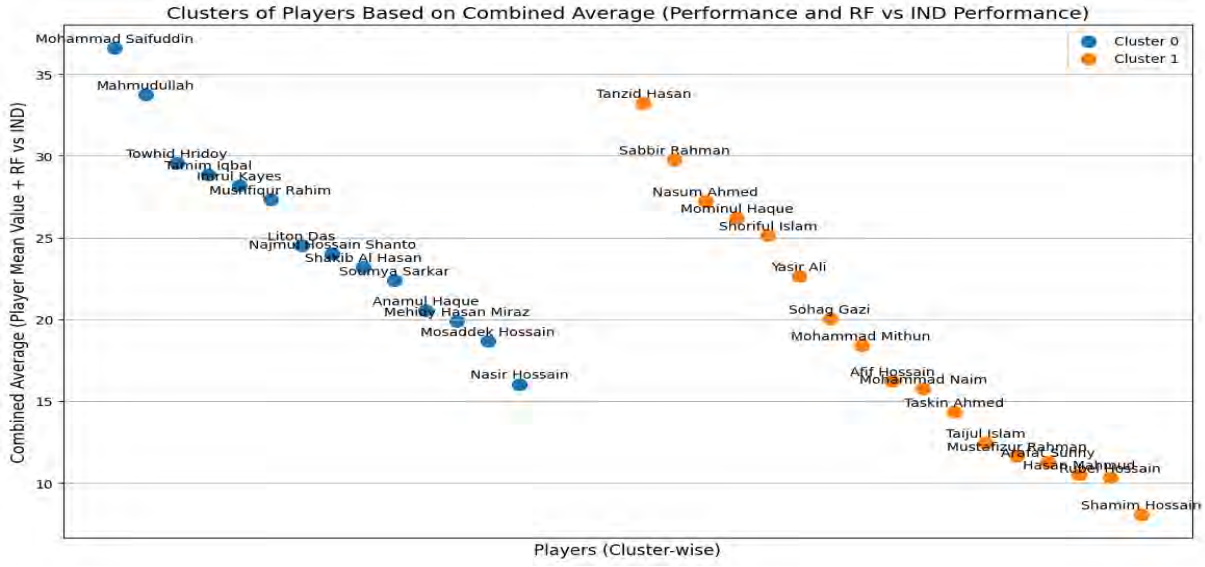

Figure 5.4: Combined average of players based on recent form using K-means.

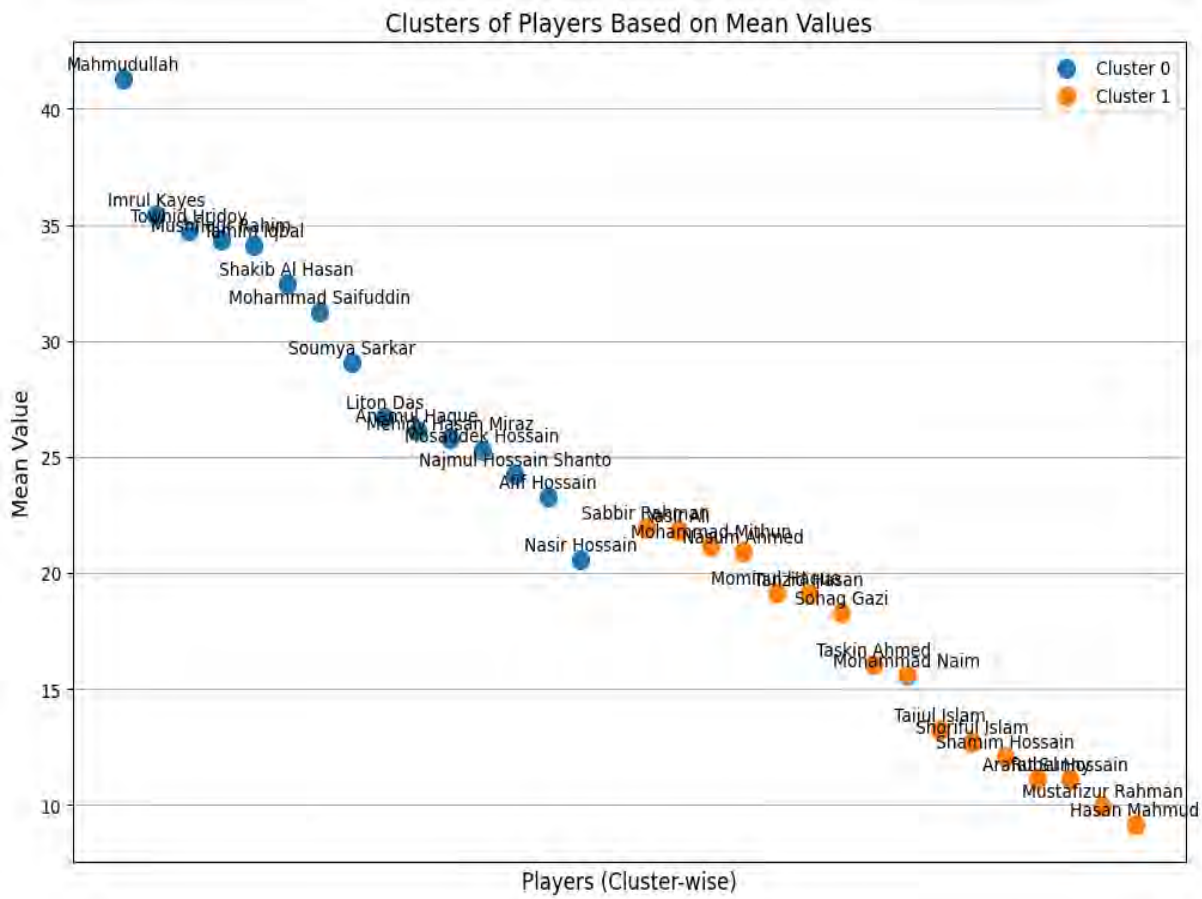

Figure 5.5: Combined average of players form using K-means.

The results of target variables K-means showing for INDIA: Sorted by Performance Against vsIND (Descending)

Player Name	**Cluster**	**Target variable**	**Mean Value**
Mehidy Hasan Miraz	0	46.2000	25.0161
Shakib Al Hasan	0	37.5500	32.2832
Mahmudullah	0	36.6100	41.4370
Towhid Hridoy	0	35.0000	34.7633
Liton Das	0	33.7200	26.4622
Tamim Iqbal	0	33.1100	34.1545
Mohammad Saifuddin	0	32.1667	31.2067
Mushfiqur Rahim	0	31.9500	34.4597
Soumya Sarkar	0	31.8500	28.9721
Imrul Kayes	0	26.1600	35.7918
Nasir Hossain	0	26.1400	20.3431
Anamul Haque	0	22.5700	26.2827
Mosaddek Hossain	0	10.0000	25.8482
Najmul Hossain Shanto	0	9.0000	24.7822
Afif Hossain	0	4.6600	23.9431
Sabbir Rahman	1	38.2000	21.3363
Nasum Ahmed	1	38.0000	20.2777
Tanzid Hasan	1	32.0000	18.6058
Taskin Ahmed	1	29.0000	15.5534
Sohag Gazi	1	16.0000	18.3854
Mohammad Naim	1	14.9967	15.5919
Mominul Haque	1	14.5000	19.3077
Taijul Islam	1	14.3980	13.1687
Mohammad Mithun	1	12.3300	21.4959
Mustafizur Rahman	1	9.5000	9.9687
Shoriful Islam	1	7.3325	12.9336
Arafat Sunny	1	6.6667	11.2509
Yasir Ali	1	4.5000	22.4624
Rubel Hossain	1	3.4000	11.3667
Shamim Hossain	1	1.0000	12.4944
Hasan Mahmud	1	0.0000	9.4666

And the result of target variables K-means showing for Recent Form

Player Name	**Cluster**	**Target Variable**	**Mean Value**	**Combined Average**
Mohammad Saifuddin	0	42.3141	30.8308	36.5725
Mahmudullah	0	25.6280	41.8438	33.7359
Towhid Hridoy	0	23.9419	35.1728	29.5574
Tamim Iqbal	0	23.1890	34.5219	28.8555
Imrul Kayes	0	20.4110	36.0047	28.2079
Mushfiqur Rahim	0	19.7113	34.9130	27.3121
Liton Das	0	22.1597	26.8903	24.5250
Najmul Hossain Shanto	0	23.8160	24.2334	24.0247
Shakib Al Hasan	0	13.1780	33.1859	23.1819
Soumya Sarkar	0	15.1720	29.5898	22.3809
Anamul Haque	0	14.5244	26.5807	20.5525
Mehidy Hasan Miraz	0	13.5380	26.2258	19.8819
Mosaddek Hossain	0	11.4966	25.7927	18.6447
Nasir Hossain	0	11.1870	20.8969	16.0420
Tanzid Hasan	1	48.4736	17.9957	33.2347
Sabbir Rahman	1	38.1840	21.3369	29.7605
Nasum Ahmed	1	34.0370	20.4245	27.2307
Mominul Haque	1	33.8470	18.5911	26.2191
Shoriful Islam	1	38.4728	11.7802	25.1265
Yasir Ali	1	23.4944	21.7589	22.6267
Sohag Gazi	1	21.8681	18.1681	20.0181
Mohammad Mithun	1	15.4191	21.3814	18.4003
Afif Hossain	1	8.6197	23.7965	16.2081
Mohammad Naim	1	15.9736	15.5557	15.7647
Taskin Ahmed	1	12.4770	16.1654	14.3212
Taijul Islam	1	11.5925	13.2726	12.4325
Mustafizur Rahman	1	13.4846	9.8211	11.6529
Arafat Sunny	1	11.5084	11.0716	11.2900
Hasan Mahmud	1	11.9946	9.0224	10.5085
Rubel Hossain	1	9.4522	11.1426	10.2974
Shamim Hossain	1	3.7028	12.3943	8.0486

For these two types of result we choose batsman recent form because a batsman recent form is very much important for a team to face their opponent team.

5.1.2 For Bowling

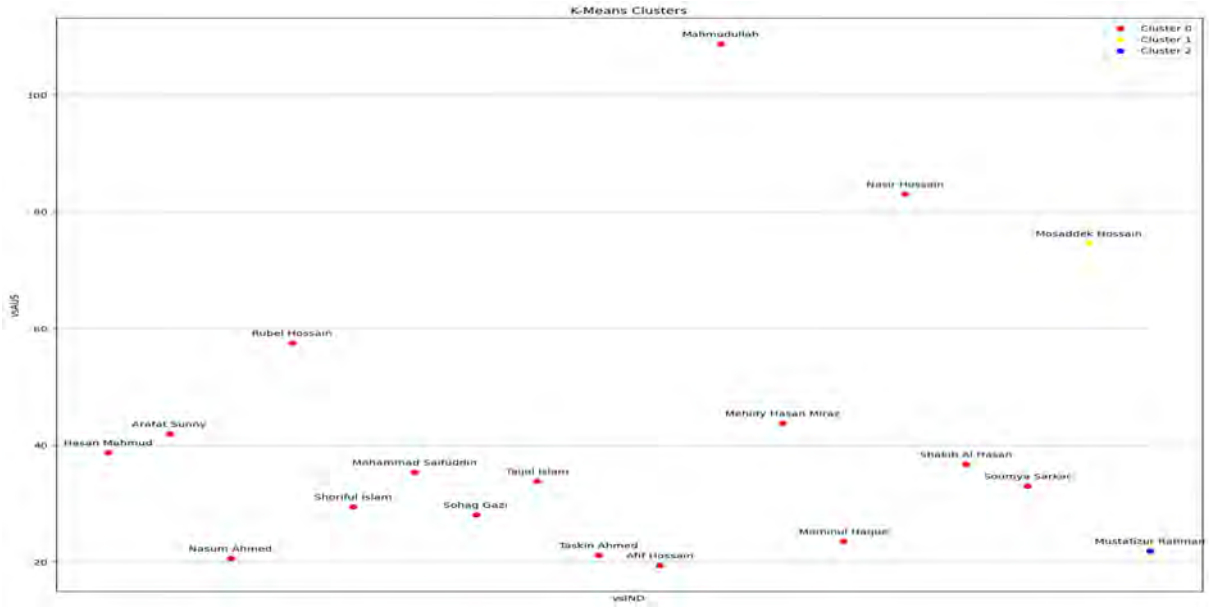

Figure 5.6: K-means Clustering result for bowling

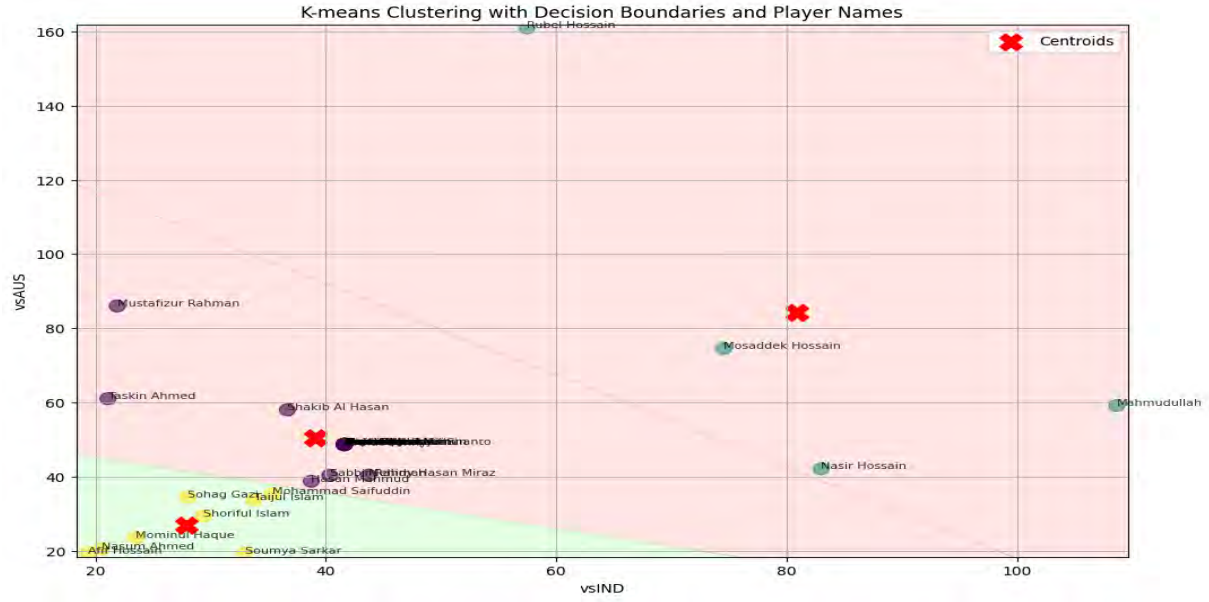

Figure 5.7: K-means clustering based on two features for bowling

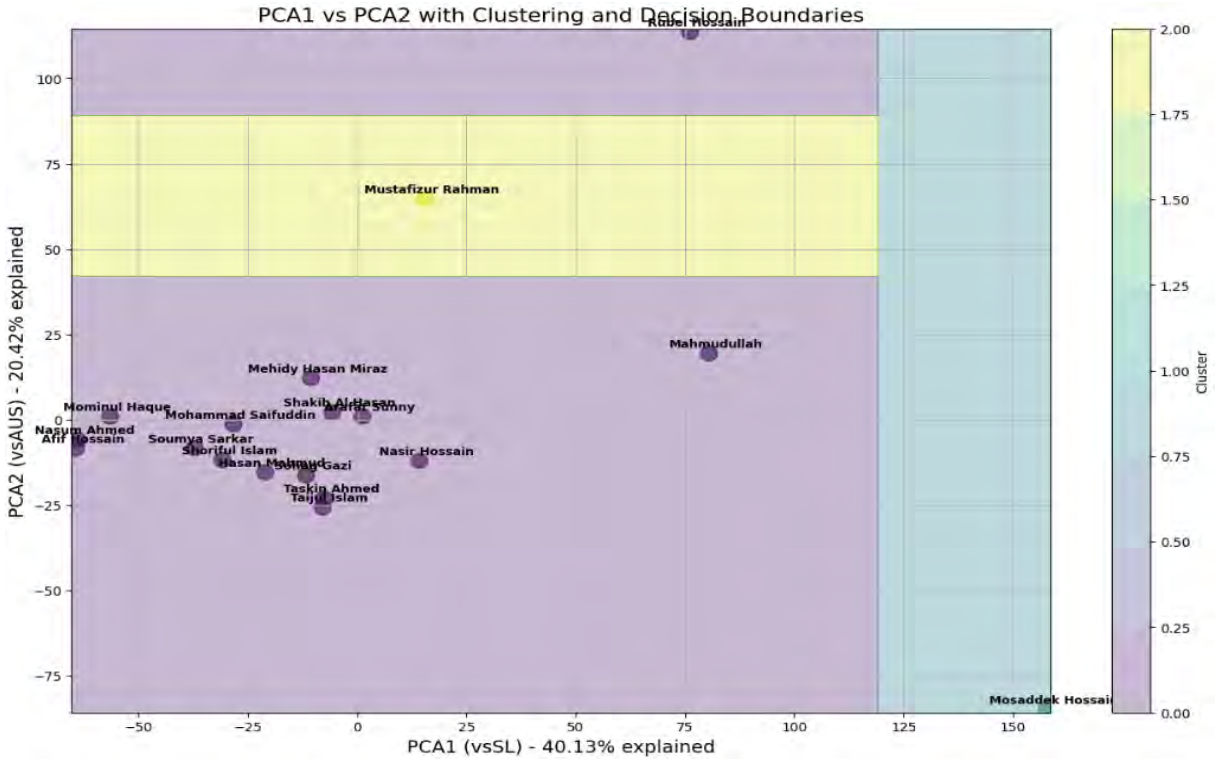

Figure 5.8: K-means clustering based on two features for bowling

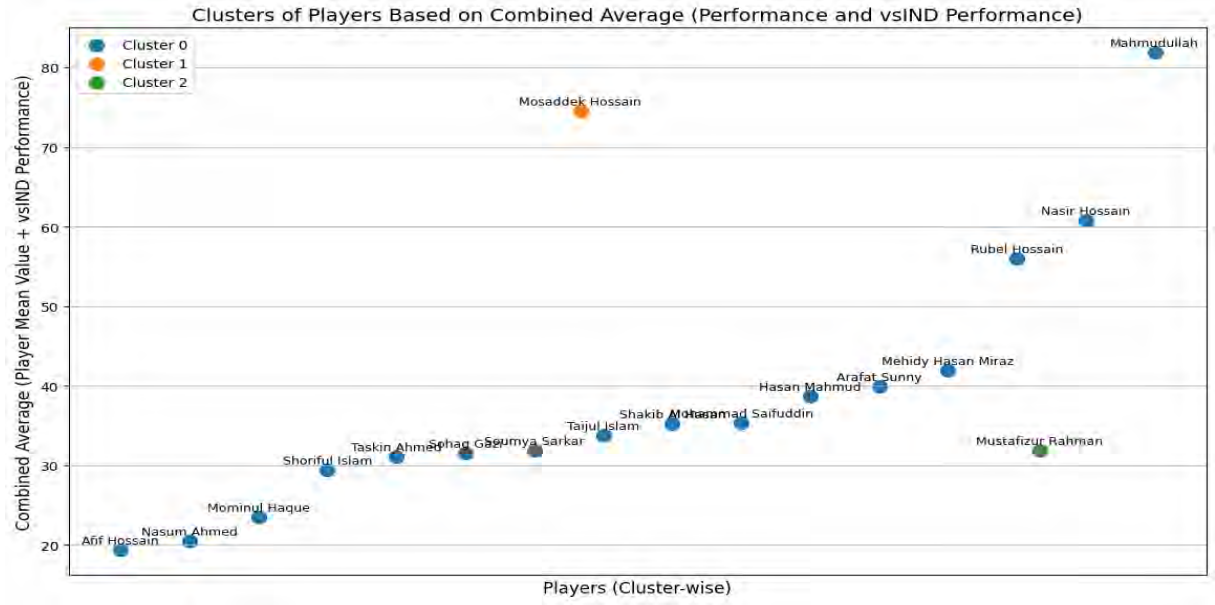

Figure 5.9: K-means clustering based on two features for bowling

Target variable result against country

Player Name	Cluster	Target Variable	Mean Value
Afif Hossain	0	19.3333	19.3308
Nasum Ahmed	0	20.5167	20.5192
Taskin Ahmed	0	21.0700	41.0550
Mominul Haque	0	23.5000	23.5000
Sohag Gazi	0	28.0000	35.0427
Shoriful Islam	0	29.3750	29.3750
Soumya Sarkar	0	33.0000	30.8119
Taijul Islam	0	33.7350	33.7375
Mohammad Saifuddin	0	35.3322	35.3318
Shakib Al Hasan	0	36.6500	33.7375
Hasan Mahmud	0	38.7250	38.7250
Sabbir Rahman	0	40.3333	40.3308
Najmul Hossain Shanto	0	41.5991	38.1267
Mohammad Mithun	0	41.5991	36.7934
Liton Das	0	41.5991	38.1267
Imrul Kayes	0	41.5991	38.1267
Mohammad Naim	0	41.5991	38.1267
Towhid Hridoy	0	41.5991	38.1267
Mushfiqur Rahim	0	41.5991	38.1267
Yasir Ali	0	41.5991	38.1267
Tanzid Hasan	0	41.5991	38.1267
Shamim Hossain	0	41.5991	38.1267
Arafat Sunny	0	41.5991	38.1128
Anamul Haque	0	41.5991	38.1267
Tamim Iqbal	0	41.5991	38.1267
Mehidy Hasan Miraz	0	43.7100	40.0988
Rubel Hossain	0	57.4400	54.6026
Nasir Hossain	0	83.0000	38.6778
Mahmudullah	0	108.6600	55.0398
Mosaddek Hossain	1	74.5229	74.5217
Mustafizur Rahman	2	21.8800	41.8567

5.1.3 Hierarchical:

For ODI Batting

Here in this models with agglomerative clustering way we used the value of n clusters=2 and got three distinct clusters with the Silhouette Score for 2 clusters: 0.2802

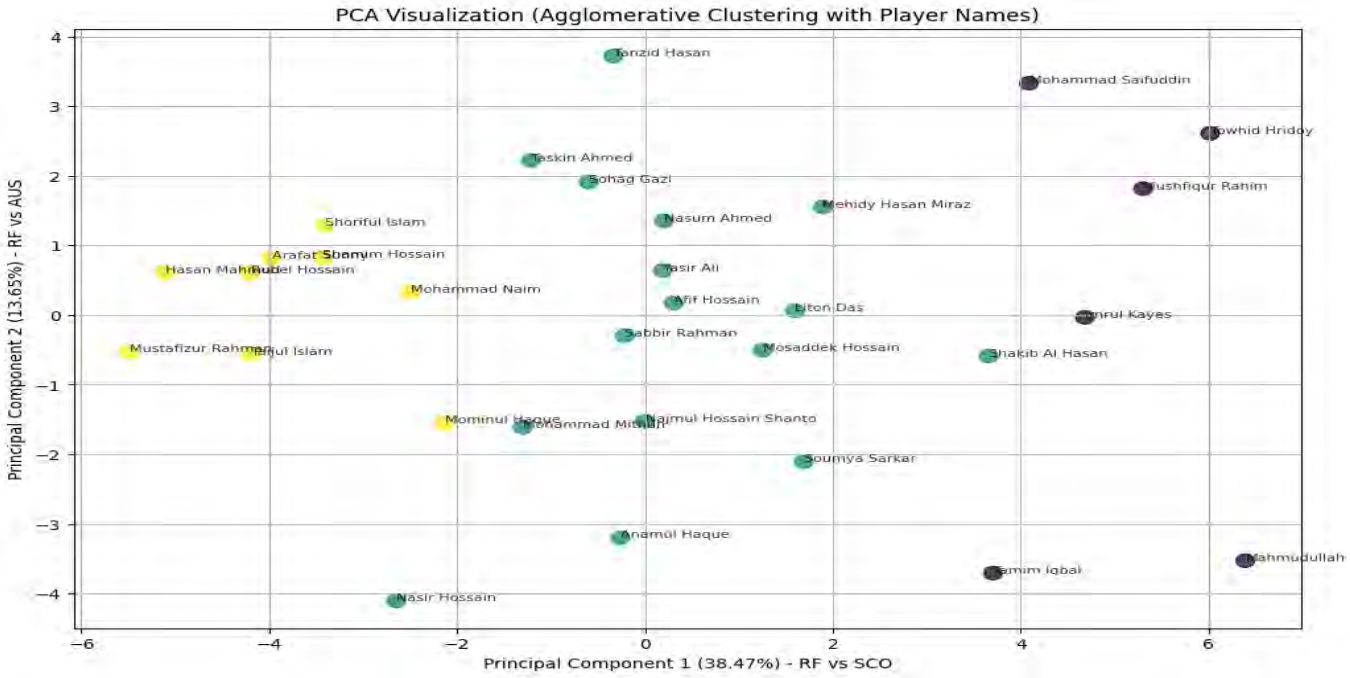

Figure 5.10: Hierarchical Clustering

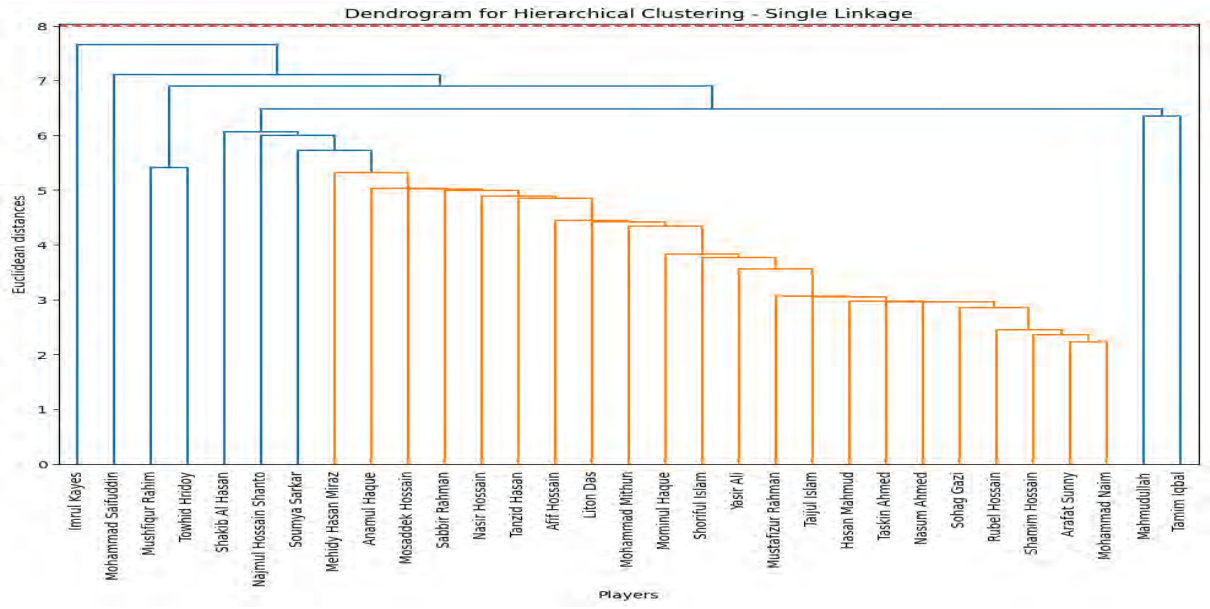

Figure 5.11: Hierarchical Clustering-Single Linkage Dendrogram

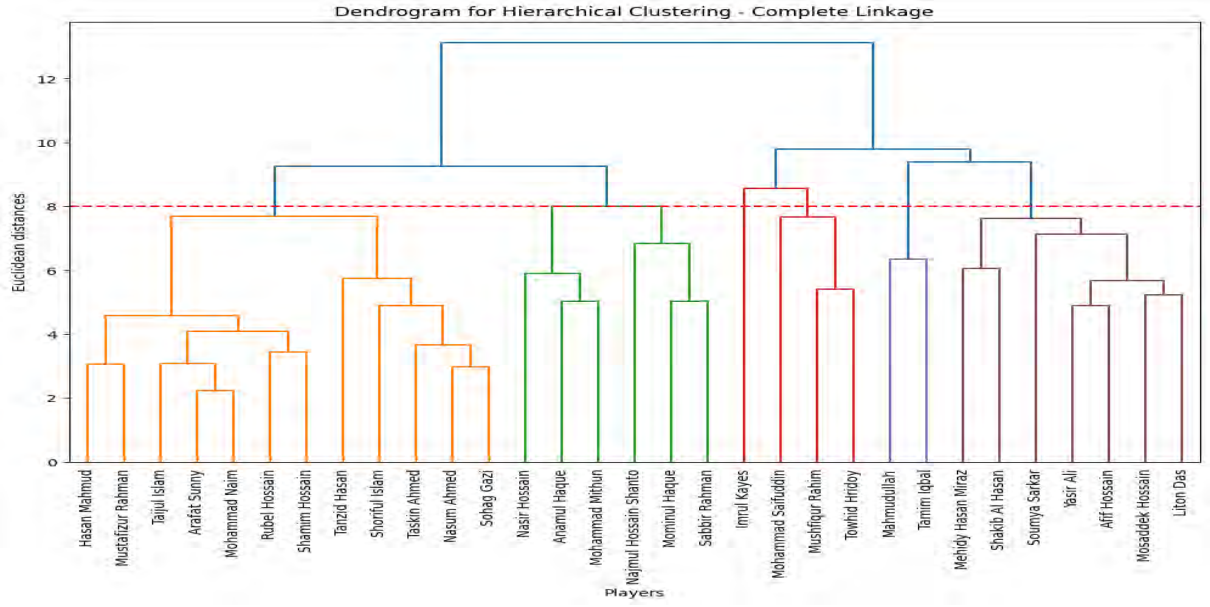

Figure 5.12: Hierarchical Clustering-Complete Linkage Dendrogram

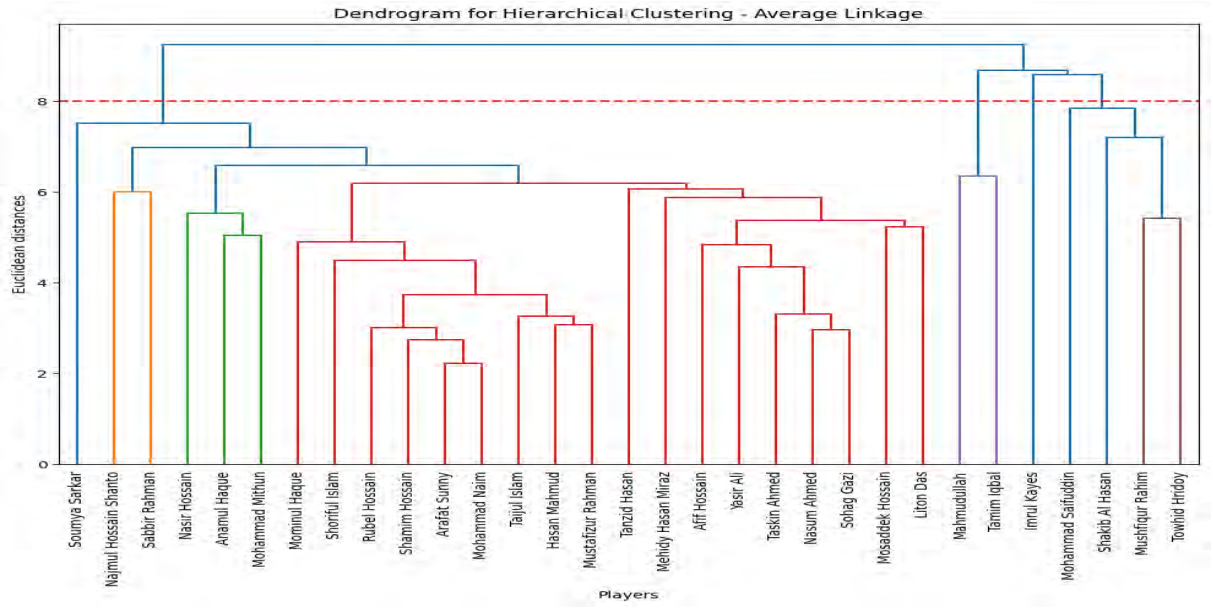

Figure 5.13: Hierarchical Clustering-Average Linkage Dendrogram

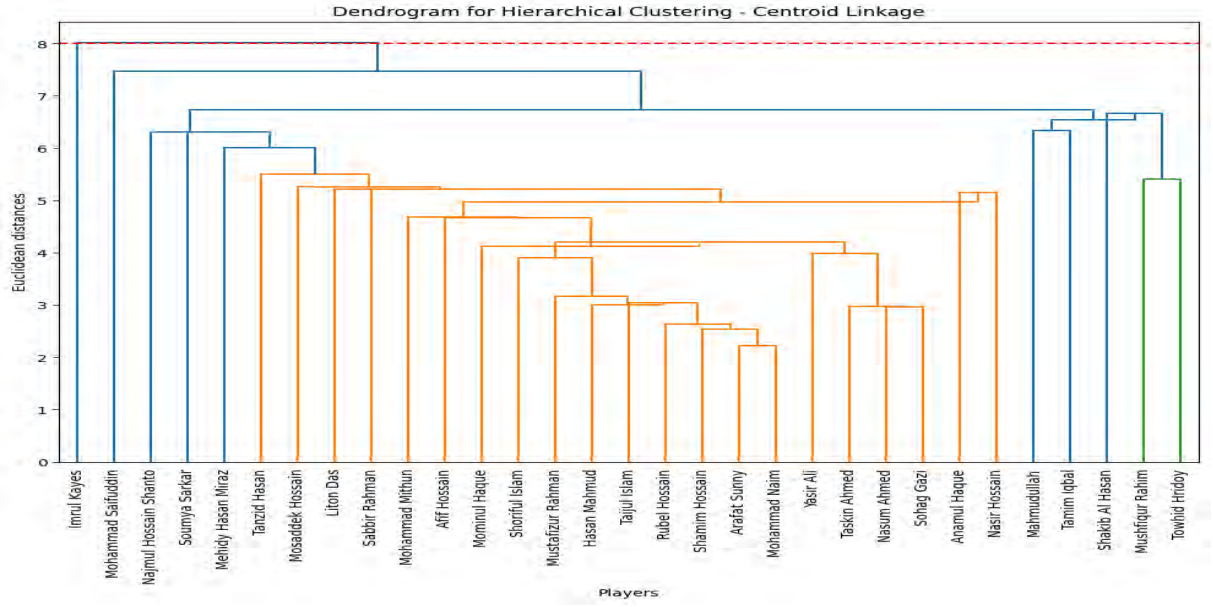

Figure 5.14: Hierarchical Clustering-Centroid Linkage Dendrogram

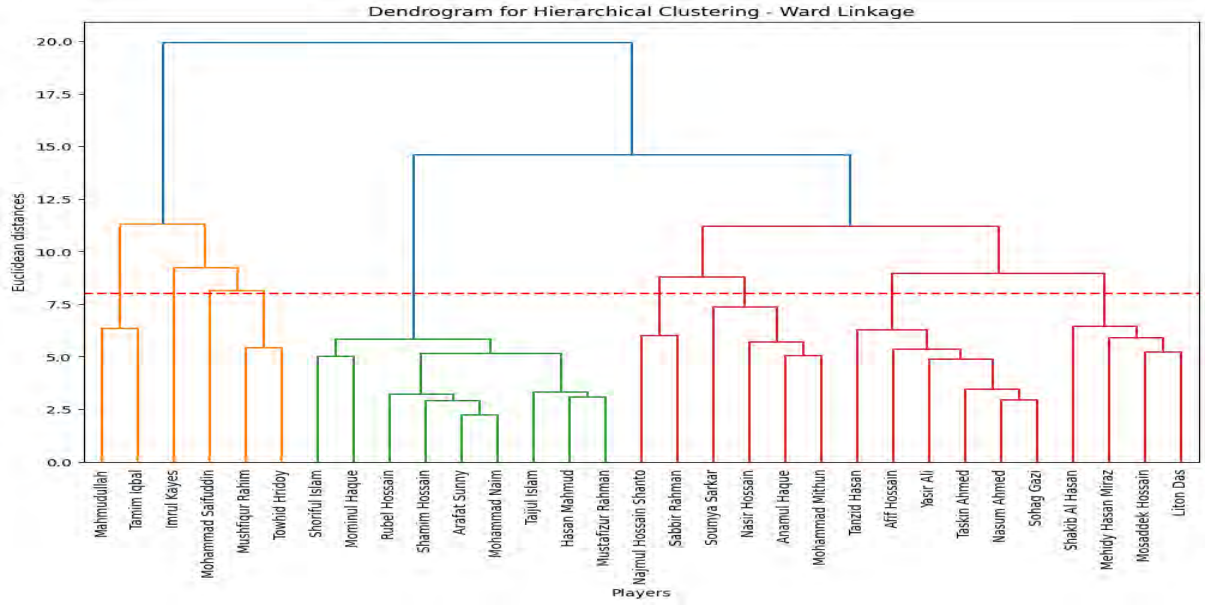

Figure 5.15: Hierarchical Clustering-Ward Linkage Dendrogram

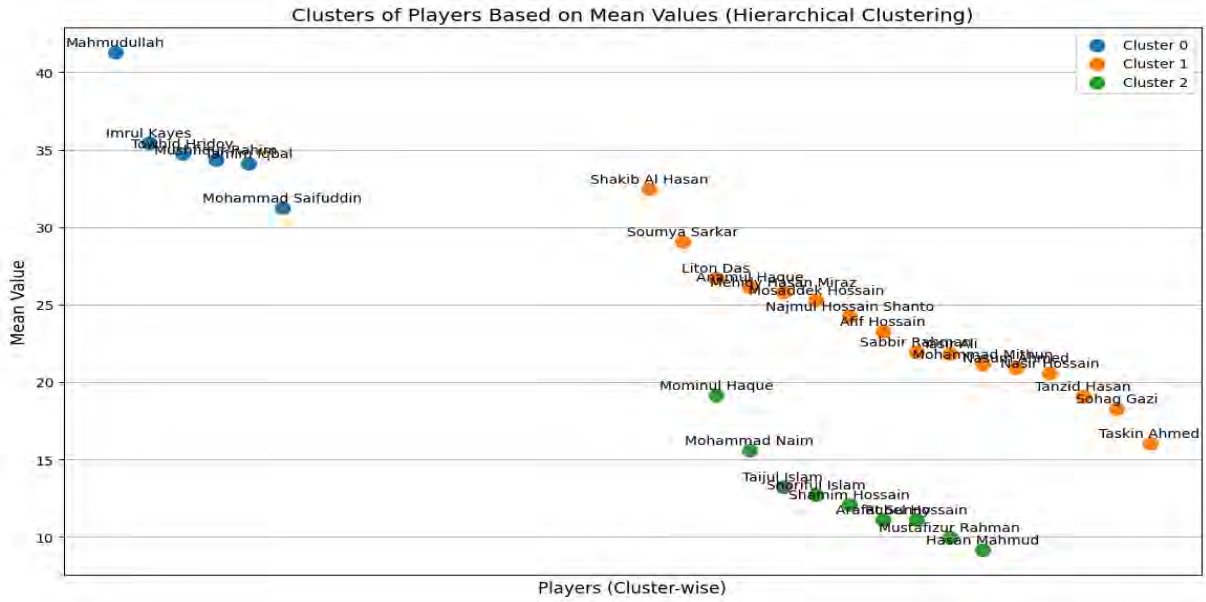

Figure 5.16: Combined Average

Here we choose Ward linkage because as the ward method merges the cluster with the smallest variance which is more statically significant than the complete linkage euclidean distance that is why we chose it. and the result of the ward linkage for recent form We choose the highest cluster score for our selection process.
cluster 0 23.118481
cluster 1: 35.547842
cluster 2: 11.757560

Player Name	**Cluster**	**target Variable**	**Mean Value**	**Combined Average**
Tanzid Hasan	0	48.4736	17.9957	33.2347
Sabbir Rahman	0	38.1840	21.3369	29.7605
Nasum Ahmed	0	34.0370	20.4245	27.2307
Mominul Haque	0	33.8470	18.5911	26.2191
Liton Das	0	22.1597	26.8903	24.5250
Najmul Hossain Shanto	0	23.8160	24.2334	24.0247
Shakib Al Hasan	0	13.1780	33.1859	23.1819
Yasir Ali	0	23.4944	21.7589	22.6267
Soumya Sarkar	0	15.1720	29.5898	22.3809
Anamul Haque	0	14.5244	26.5807	20.5525
Sohag Gazi	0	21.8681	18.1681	20.0181
Mehidy Hasan Miraz	0	13.5380	26.2258	19.8819
Mosaddek Hossain	0	11.4966	25.7927	18.6447
Mohammad Mithun	0	15.4191	21.3814	18.4003
Afif Hossain	0	8.6197	23.7965	16.2081
Nasir Hossain	0	11.1870	20.8969	16.0420
Taskin Ahmed	0	12.4770	16.1654	14.3212
Mohammad Saifuddin	1	42.3141	30.8308	36.5725
Mahmudullah	1	25.6280	41.8438	33.7359
Towhid Hridoy	1	23.9419	35.1728	29.5574
Tamim Iqbal	1	23.1890	34.5219	28.8555
Imrul Kayes	1	20.4110	36.0047	28.2079
Mushfiqur Rahim	1	19.7113	34.9130	27.3121
Shoriful Islam	2	38.4728	11.7802	25.1265
Mohammad Naim	2	15.9736	15.5557	15.7647
Taijul Islam	2	11.5925	13.2726	12.4325
Mustafizur Rahman	2	13.4846	9.8211	11.6529
Arafat Sunny	2	11.5084	11.0716	11.2900
Hasan Mahmud	2	11.9946	9.0224	10.5085
Rubel Hossain	2	9.4522	11.1426	10.2974
Shamim Hossain	2	3.7028	12.3943	8.0486

Table 5.1: Players sorted by Combined Average in their respective clusters.

The Recent Form for Complete linkage for each cluster:
cluster 0 : 33.2664
cluster1: 17.5001
cluster2: 24.1012

Player Name	**Cluster**	**target Variable**	**Mean Value**	**Combined Average**
Mahmudullah	0	25.6280	41.8438	33.7359
Tamim Iqbal	0	23.1890	34.5219	28.8555
Imrul Kayes	0	20.4110	36.0047	28.2079
Sabbir Rahman	1	38.1840	21.3369	29.7605
Mominul Haque	1	33.8470	18.5911	26.2191
Shoriful Islam	1	38.4728	11.7802	25.1265
Najmul Hossain Shanto	1	23.8160	24.2334	24.0247
Soumya Sarkar	1	15.1720	29.5898	22.3809
Anamul Haque	1	14.5244	26.5807	20.5525
Mohammad Mithun	1	15.4191	21.3814	18.4003
Nasir Hossain	1	11.1870	20.8969	16.0420
Mohammad Naim	1	15.9736	15.5557	15.7647
Taijul Islam	1	11.5925	13.2726	12.4325
Mustafizur Rahman	1	13.4846	9.8211	11.6529
Arafat Sunny	1	11.5084	11.0716	11.2900
Hasan Mahmud	1	11.9946	9.0224	10.5085
Rubel Hossain	1	9.4522	11.1426	10.2974
Shamim Hossain	1	3.7028	12.3943	8.0486
Mohammad Saifuddin	2	42.3141	30.8308	36.5725
Tanzid Hasan	2	48.4736	17.9957	33.2347
Towhid Hridoy	2	23.9419	35.1728	29.5574
Mushfiqur Rahim	2	19.7113	34.9130	27.3121
Nasum Ahmed	2	34.0370	20.4245	27.2307
Liton Das	2	22.1597	26.8903	24.5250
Shakib Al Hasan	2	13.1780	33.1859	23.1819
Yasir Ali	2	23.4944	21.7589	22.6267
Sohag Gazi	2	21.8681	18.1681	20.0181
Mehidy Hasan Miraz	2	13.5380	26.2258	19.8819
Mosaddek Hossain	2	11.4966	25.7927	18.6447
Afif Hossain	2	8.6197	23.7965	16.2081
Taskin Ahmed	2	12.4770	16.1654	14.3212

Table 5.2: Players sorted in Recent form by Combined Average in their respective clusters.

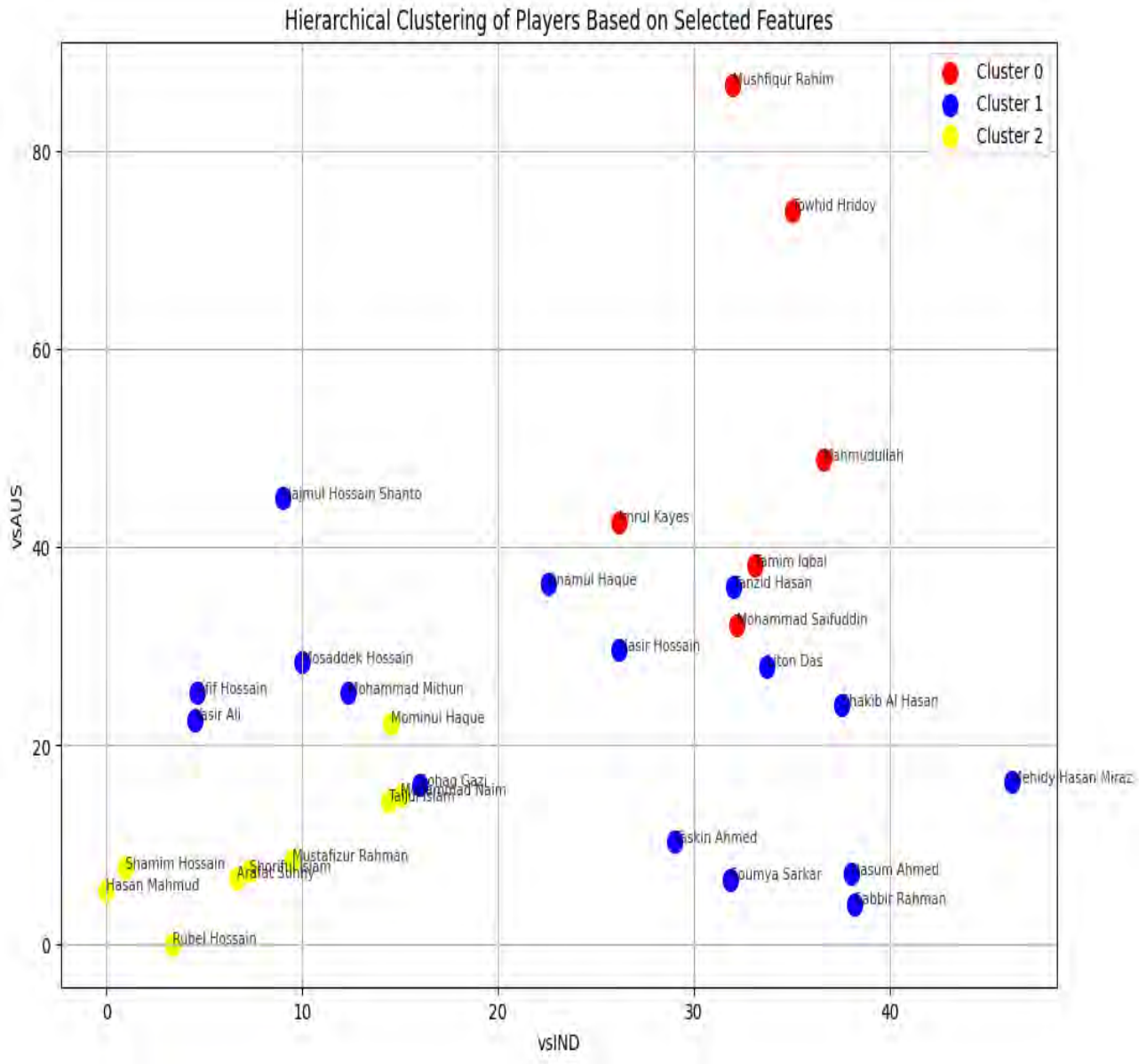

Figure 5.17: Hierarchical Clustering

5.1.4 ListA Batting

Using the Same clustering model the results shows for ListA. We choose the highest cluster score for our selection process.
cluster 0: 921.990542
cluster1: 81.9943
cluster2: 371.253353

Player Name	Cluster	target variable	Mean Value	Combined Average
Tamim Iqbal	0	114.1578	1769.6672	941.9125
Shakib Al Hasan	0	106.3557	1475.2164	790.7860
Imrul Kayes	0	45.5366	1274.9298	660.2332
Anamul Haque	0	35.9372	1205.9631	620.9502
Nasir Hossain	0	79.8232	996.9602	538.3917
Liton Das	0	50.0969	940.6945	495.3957
Mushfiqur Rahim	0	95.7677	881.5225	488.6451
Mohammad Mithun	0	31.3750	940.7743	486.0747
Soumya Sarkar	0	74.3630	862.7755	468.5693
Najmul Hossain Shanto	0	38.3080	839.9845	439.1463
Sabbir Rahman	0	74.8873	785.4965	430.1919
Fazle Mahmud	0	63.2984	782.1199	422.7091
Mominul Haque	0	27.5297	777.9126	402.7212
Mosaddek Hossain	0	32.4202	763.4440	397.9321
Mahmudullah	0	80.1900	715.0753	397.6327
Saif Hassan	0	58.5078	709.2352	383.8715
Mohammad Naim	0	58.0245	645.4363	351.7304
Nurul Hasan	0	63.4319	622.6948	343.0633
Mashrafe Mortaza	0	83.5835	527.9176	305.7505
Mohammad Saifuddin	1	33.3350	272.2130	152.7740
Muktar Ali	1	44.1053	232.4810	138.2932
Arafat Sunny	1	37.7807	172.3754	105.0780
Sunzamul Islam	1	38.4891	167.8079	103.1485
Abu Hider	1	36.9982	138.0092	87.5037
Taijul Islam	1	16.8531	147.8862	82.3697
Nasum Ahmed	1	25.2653	126.0610	75.6631
Rubel Hossain	1	23.2019	107.3368	65.2694
Nayeem Hasan	1	27.4291	97.7779	62.6035
Taskin Ahmed	1	21.9612	95.1494	58.5553
Saqlain Sajib	1	28.5887	79.0412	53.8150
Kamrul Islam Rabbi	1	25.9741	73.3529	49.6635
Mustafizur Rahman	1	23.8521	55.5808	39.7164
Tanzim Hasan Sakib	1	23.0628	47.6182	35.3405
Rakibul Hasan	1	28.7444	39.5823	34.1633
Sumon Khan	1	21.8701	42.0653	31.9677
Tanvir Islam	1	22.0939	41.5243	31.8091
Shoriful Islam	1	20.1360	42.9357	31.5359
Mrittunjoy Chowdhury	1	22.1801	36.9307	29.5554
Hasan Mahmud (4match)	1	18.7135	36.0009	27.3572
Abu Jayed	1	20.4159	33.5911	27.0035
Hasan Murad	1	14.8181	32.4552	23.6367

Player Name	Cluster	target variable	Mean Value	Combined Average
Subashis Roy	1	15.7778	27.9719	21.8748
Khaled Ahmed	1	16.6114	26.5401	21.5757
Shohidul Islam	1	13.0248	18.9763	16.0006
Ebadot Hossain	1	12.4255	15.3964	13.9109
Zakir Hasan	1	5.9743	7.1870	6.5806
Rony Talukdar	2	57.8664	590.2644	324.0654
Shadman Islam	2	47.8864	505.2411	276.5637
Afif Hossain	2	32.7767	505.5404	269.1586
Towhid Hridoy	2	35.5546	489.3950	262.4748
Tanbir Hayder	2	53.7462	466.6824	260.2143
Ariful Haque	2	55.4550	461.0143	258.2347
Yasir Ali	2	52.9225	445.4975	249.2100
Mehidy Hasan Miraz	2	39.4229	441.9029	240.6629
Shuvagata Hom	2	53.1117	373.5374	213.3246
Sohag Gazi	2	23.6826	366.4468	195.0647
Jaker Ali	2	44.6239	341.8363	193.2301
Tanzid Hasan	2	26.6847	315.4716	171.0781
Mahedi Hasan	2	46.0171	295.0178	170.5175
Parvez Hossain Emon	2	40.9267	255.5678	148.2472
Shahadat Hossain Dipu	2	41.2854	252.3744	146.8299
Mahmudul Hasan Joy	2	37.4679	250.9045	144.1862
Shamim Hossain	2	37.4442	206.4509	121.9476
Rishad Hossain	2	42.7158	29.6537	36.1848

Table 5.3: Players sorted by Combined Average in their respective clusters ListA.

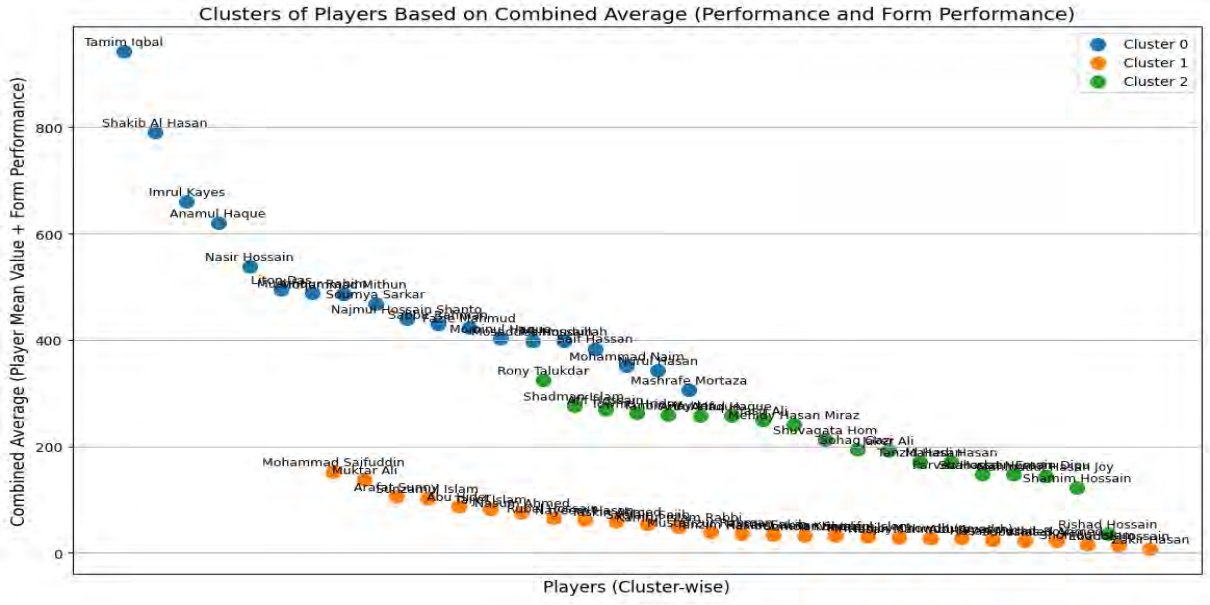

Figure 5.18: List A batting combined average

5.1.5 For Bowling ODI:

Using the Same clustering model the results shows for ODI Bowling. We choose the lowest cluster score for our selection process.

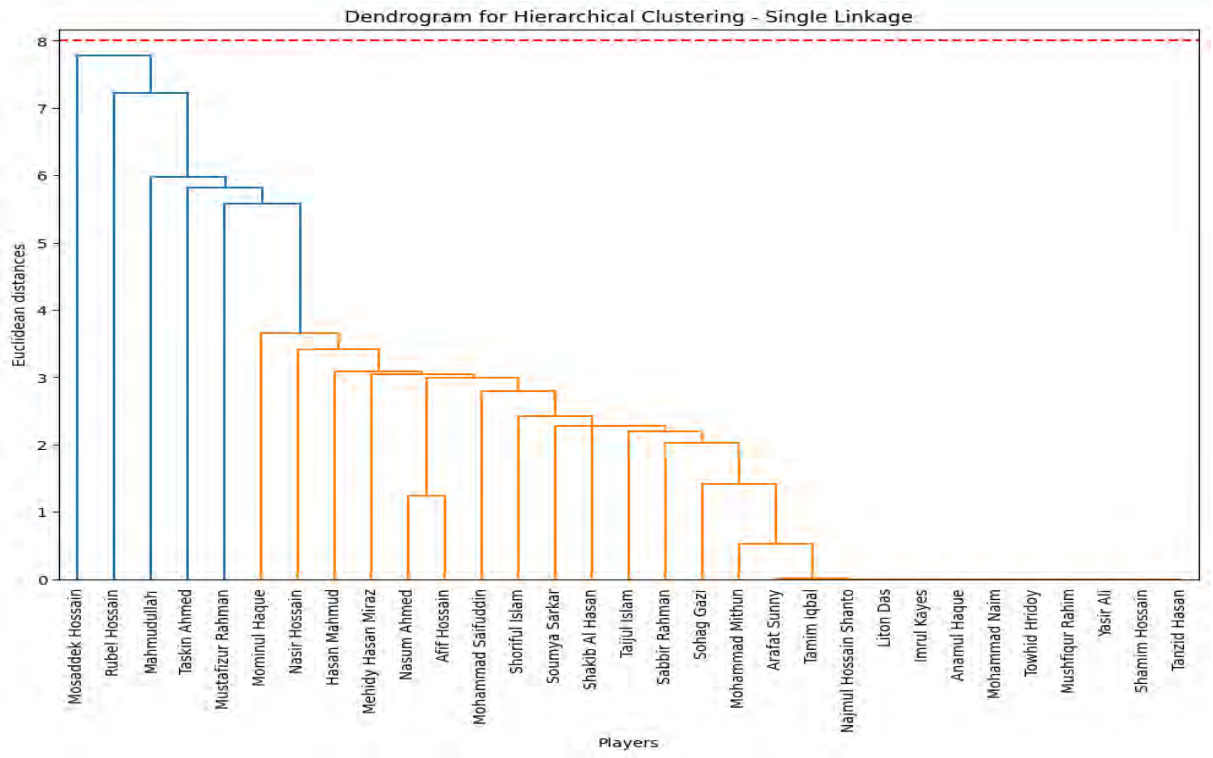

Figure 5.19: Hierarchical Clustering-Single Linkage Dendrogram

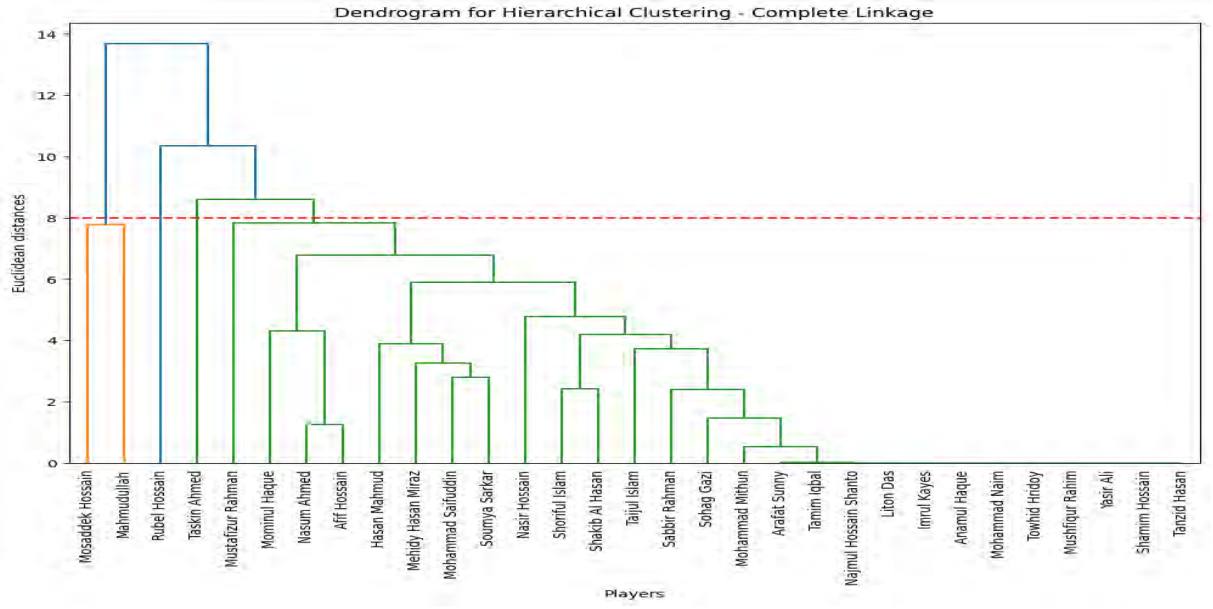

Figure 5.20: Hierarchical Clustering-Complete Linkage Dendrogram

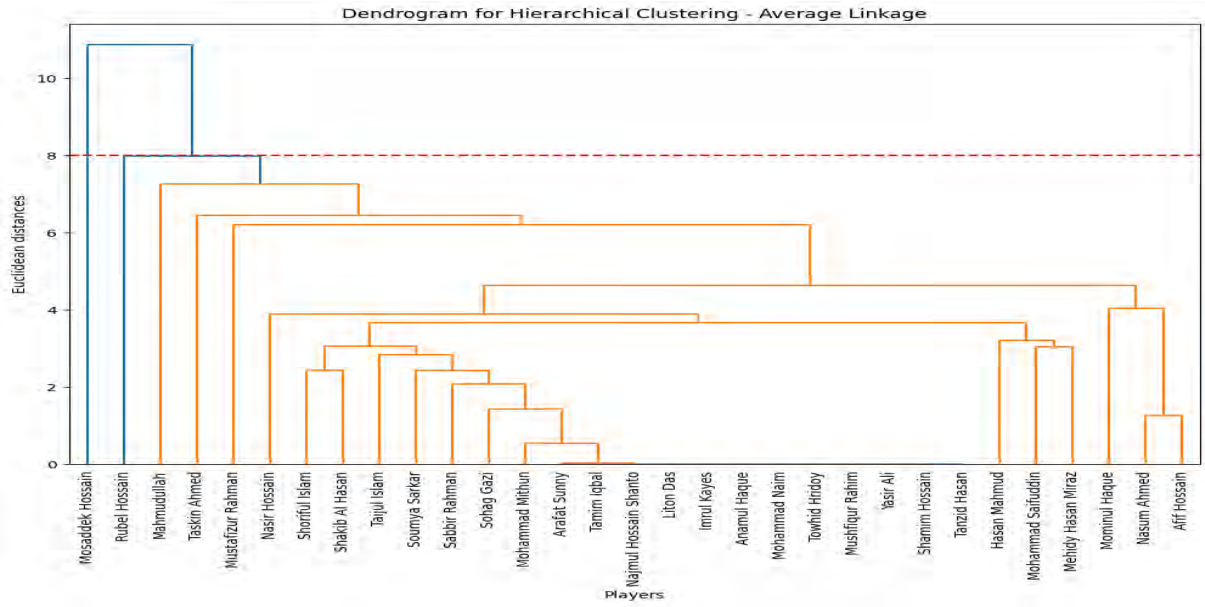

Figure 5.21: Hierarchical Clustering-Average Linkage Dendrogram

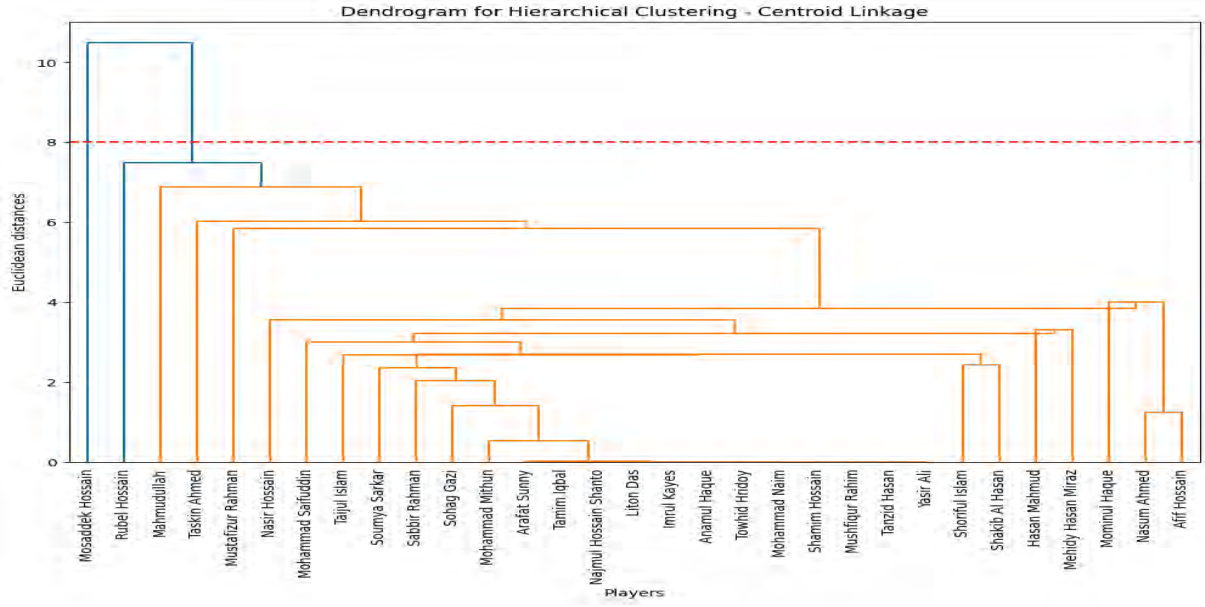

Figure 5.22: Hierarchical Clustering-Centroid Linkage Dendrogram

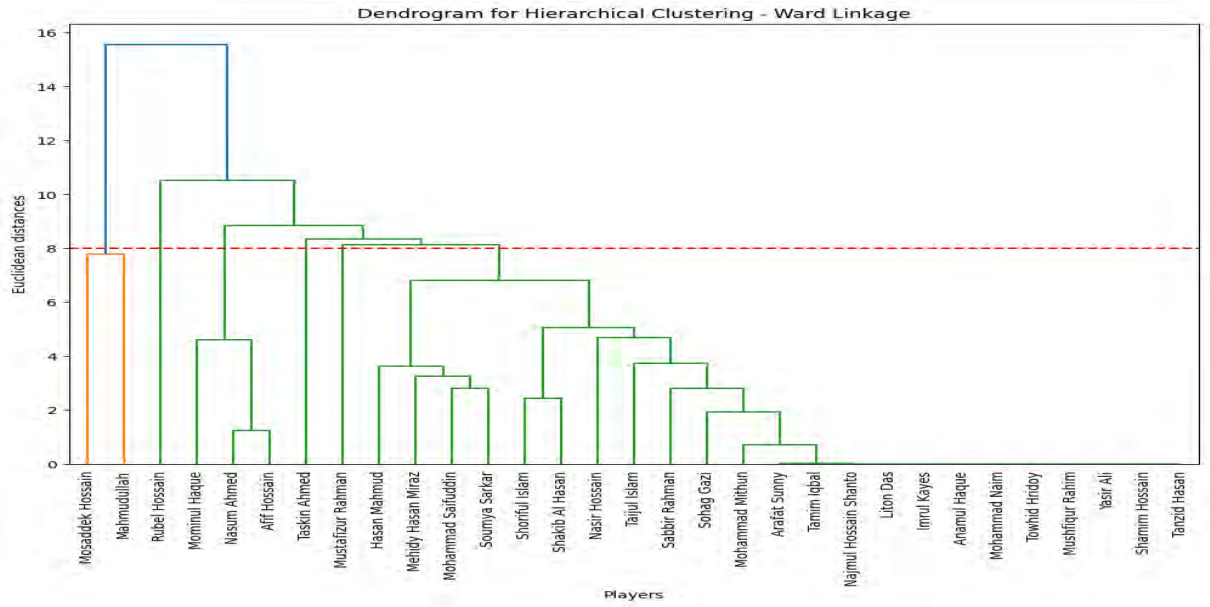

Figure 5.23: Hierarchical Clustering-Ward Linkage Dendrogram

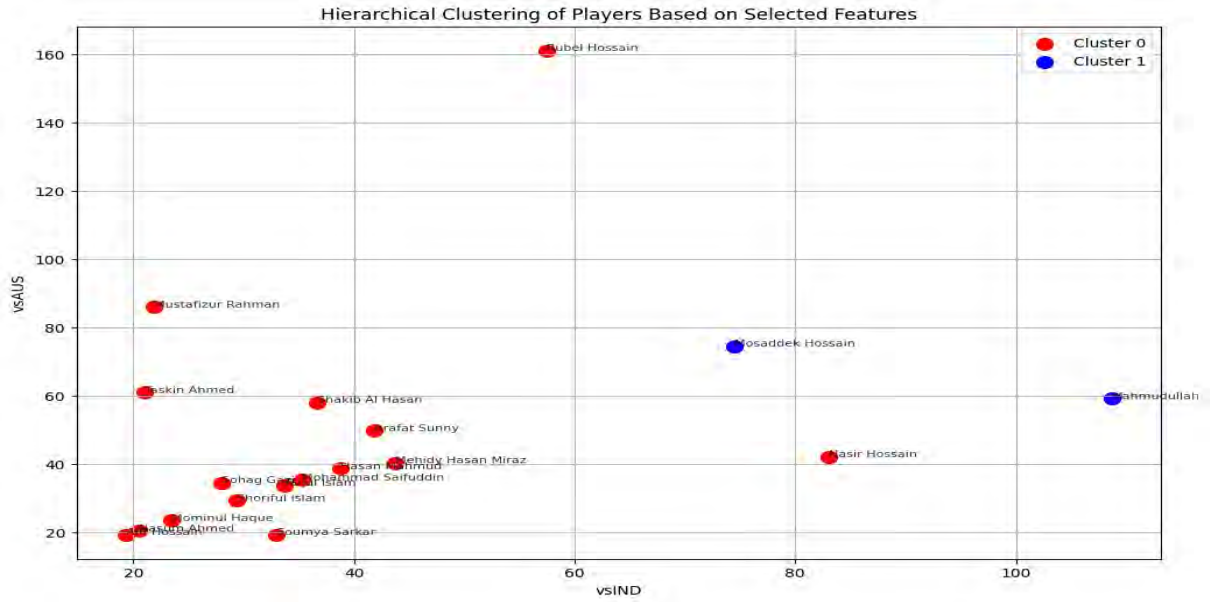

Figure 5.24: Hierarchical Clustering Dendrogram for bowling

Here in the graph Shoriful Islam, Mominul Haque, Nasum Ahmed and Afif Hossain are from cluster 0 and they are all rejected. Because in data preprocessing we see they didn't play any match most of the country. So we calculated all the past matches average set the average to omit the null value.But after the model run there is no changes result for those players. In the table we take lowest cluster result because they are the good bowlers.

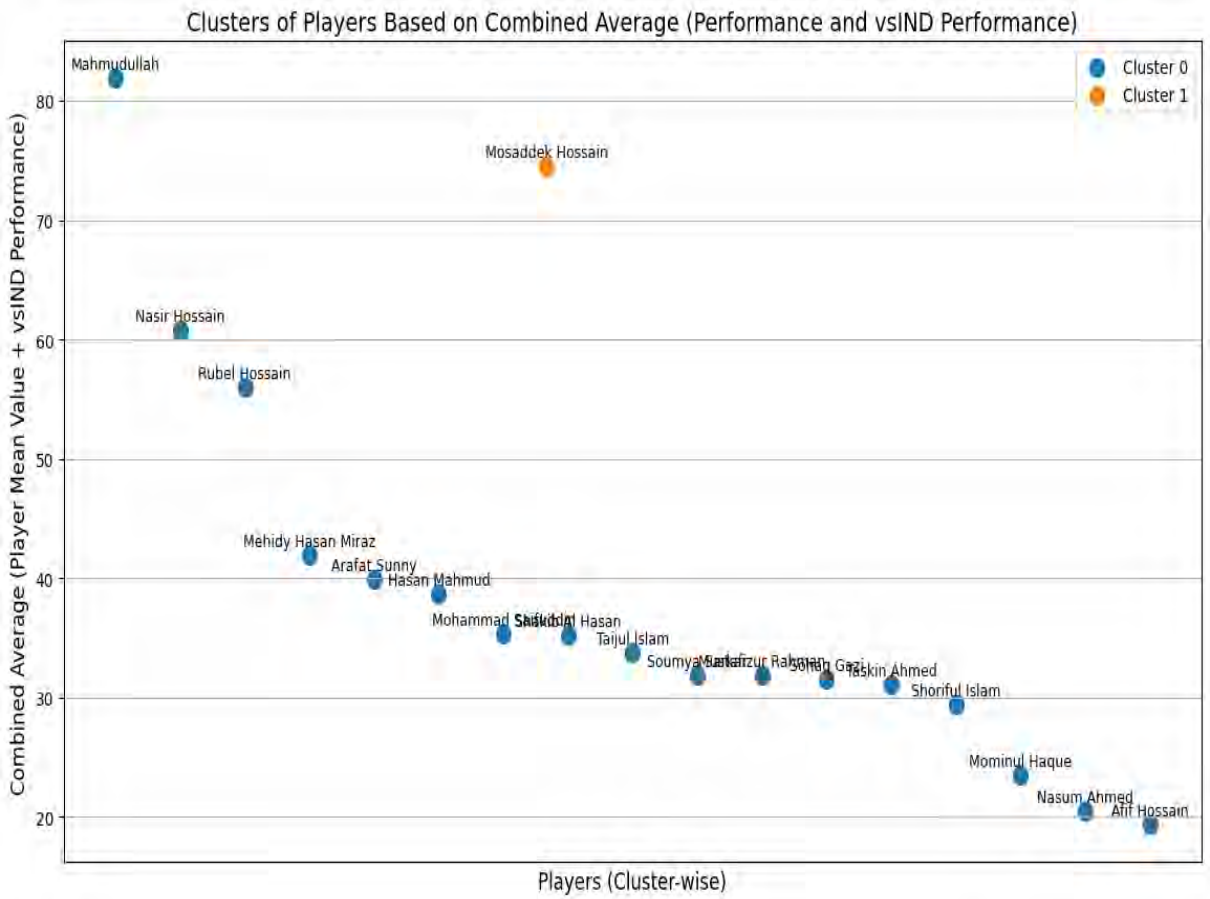

Figure 5.25: Performance based combined average clustering for bowlers

Player Name	Cluster	Target Variable	Mean Value	Combined Average
Mahmudullah	0	108.6600	55.0398	81.8499
Nasir Hossain	0	83.0000	38.6778	60.8389
Rubel Hossain	0	57.4400	54.6026	56.0213
Mehidy Hasan Miraz	0	43.7100	40.0988	41.9044
Arafat Sunny	0	41.8578	37.9524	39.9051
Hasan Mahmud	0	38.7250	38.7250	38.7250
Mohammad Saifuddin	0	35.3322	35.3318	35.3320
Shakib Al Hasan	0	36.6500	33.7375	35.1937
Taijul Islam	0	33.7350	33.7375	33.7362
Soumya Sarkar	0	33.0000	30.8119	31.9060
Mustafizur Rahman	0	21.8800	41.8567	31.8683
Sohag Gazi	0	28.0000	35.0427	31.5214
Taskin Ahmed	0	21.0700	41.0550	31.0625
Mosaddek Hossain	1	74.5229	74.5217	74.5223

Table 5.4: Players sorted by Combined Average in their respective clusters for ODI bowlers.

5.1.6 ListA Bowling

Average means for each cluster:
cluster 0: 289.435371
cluster1: 989.646856
cluster2: 2501.348764
cluster3: 96.955634
For List A bowling we choose the best cluster result .

Player Name	Cluster	Target Variable	Mean Value	Combined Average
Muktar Ali	0	308.4963	658.6074	483.5518
Soumya Sarkar	0	272.7639	576.8447	424.8043
Subashis Roy	0	267.1177	550.9047	409.0112
Nayeem Hasan	0	254.6428	514.2867	384.4647
Abu Jayed	0	234.9823	524.3968	379.6896
Shoriful Islam	0	131.2230	600.5376	365.8803
Sabbir Rahman	0	383.3229	332.4229	357.8729
Ariful Haque	0	208.5938	466.1117	337.3528
Hasan Mahmud	0	211.9922	461.0296	336.5109
Khaled Ahmed	0	196.9669	442.8137	319.8903
Tanbir Hayder	0	192.5966	412.2202	302.4084
Rakibul Hasan	0	198.8161	390.8084	294.8122
R Mondol	0	191.2540	373.1624	282.2082
Tanzim Hasan Sakib	0	168.5819	371.7953	270.1886
Sumon Khan	0	152.0334	333.6137	242.8236
Saif Hassan	0	150.6128	297.2776	223.9452
Ebadot Hossain	0	119.2123	257.4795	188.3459
Fazle Mahmud	0	110.0182	255.1814	182.5998
Mominul Haque	0	27.4362	337.0390	182.2376
Afif Hossain	0	17.0334	306.3346	161.6840
Mrittunjoy Chowdhury	0	96.0229	201.6147	148.8188
Imrul Kayes	0	230.2821	57.1274	143.7047
Mohammad Mithun	0	230.2821	56.1077	143.1949
Shohidul Islam	0	92.7912	193.3557	143.0734
Anamul Haque	0	230.2821	51.4215	140.8518
Tanzid Hasan	0	230.2821	31.5936	130.9378
Najmul Hossain Shanto	0	27.2186	105.1670	66.1928
Rony Talukdar	0	33.0611	57.5339	45.2975
Rishad Hossain	0	28.8907	52.7734	40.8320
Mahmudul Hasan Joy	0	26.1024	40.8302	33.4663
Tamim Iqbal	0	1.7085	42.2071	21.9578
Parvez Hossain Emon	0	6.4580	10.0274	8.2427
Shahadat Hossain	0	1.1529	12.0637	6.6083

Player Name	Cluster	Target variable	Mean Value	Combined Average
Rubel Hossain	1	361.3943	1580.4551	970.9247
Arafat Sunny	1	571.5905	1235.2370	903.4138
Mahmudullah	1	451.2640	1339.5551	895.4096
Sunzamul Islam	1	518.4468	1070.8495	794.6481
Mehidy Hasan Miraz	1	354.5304	1198.2086	776.3695
Nasir Hossain	1	424.2665	1087.3058	755.7862
Saqlain Sajib	1	475.4885	966.0743	720.7814
Mahedi Hasan	1	440.0698	867.2754	653.6726
Mustafizur Rahman	1	382.7871	919.2897	651.0384
Sohag Gazi	1	78.7879	1179.2386	629.0132
Taskin Ahmed	1	265.0676	978.1911	621.6294
Abu Hider	1	381.4359	825.1494	603.2927
Taijul Islam	1	86.3204	1021.6371	553.9787
Kamrul Islam Rabbi	1	327.3571	735.2658	531.3114
Tanvir Islam	1	349.2363	703.7564	526.4964
Mohammad Saifuddin	1	104.5294	908.6635	506.5964
Shuvagata Hom	1	309.2729	634.5665	471.9197
Mosaddek Hossain	1	105.5489	791.6093	448.5791
Nasum Ahmed	1	68.3953	760.9619	414.6786
Mashrafe Mortaza	2	1155.5245	2641.8426	1898.6835
Shakib Al Hasan	2	718.5862	2360.8550	1539.7206
Towhid Hridoy	3	230.2821	47.7232	139.0027
Shamim Hossain	3	94.6350	176.4192	135.5271
Zakir Hasan	3	59.5501	119.3400	89.4451
Yasir Ali	3	31.4512	44.3401	37.8956

Table 5.5: Combined Average of List A bowlers

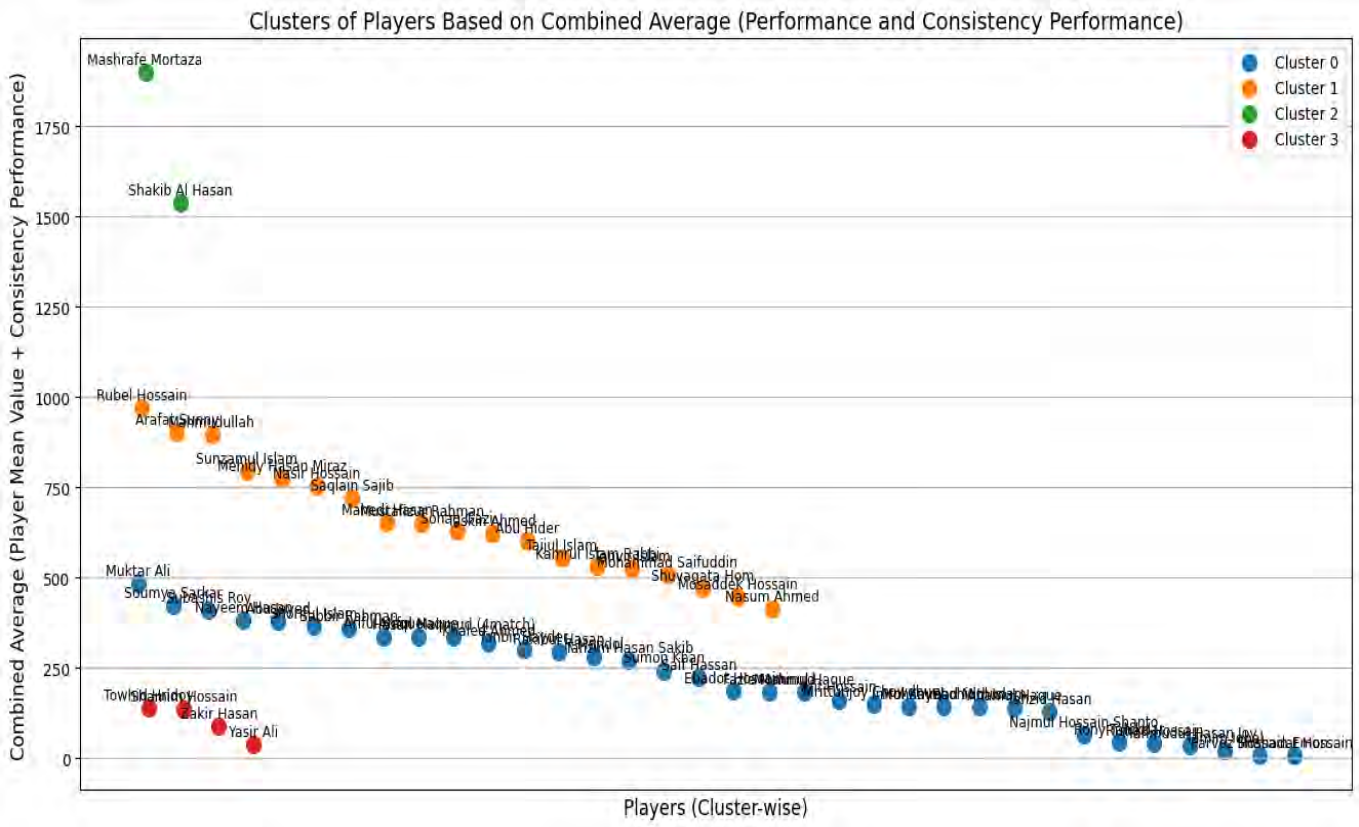

Figure 5.26: Combine Average ListA bowling

5.1.7 DBSCAN:

In DBSCAN at first we did not get any clusters. After tweaking the value of eps=2 and min samples= 5 we got two distinct clusters which are cluster 0 and cluster -1.

Cluster -1: ['Mosaddek Hossain', 'Mushfiqur Rahim', 'Anamul Haque', 'Imrul Kayes', 'Liton Das', 'Mahmudullah', 'Mohammad Mithun', 'Mominul Haque', 'Najmul Hossain Shanto', 'Nasir Hossain', 'Sabbir Rahman', 'Shakib Al Hasan', 'Soumya Sarkar', 'Tamim Iqbal', 'Nurul Hasan', 'Mashrafe Mortaza', 'Rishad Hossain', 'Fazle Mahmud']

Cluster 0: ['Mustafizur Rahman', 'Nasum Ahmed', 'Rubel Hossain', 'Shoriful Islam', 'Tanzid Hasan', 'Mohammad Saifuddin', 'Sohag Gazi', 'Taijul Islam', 'Taskin Ahmed', 'Towhid Hridoy', 'Afif Hossain', 'Mehidy Hasan Miraz', 'Ebadot Hossain', 'Hasan Mahmud', 'Sunzamul Islam', 'Abu Jayed', 'Arafat Sunny', 'Ariful Haque', 'Nayeem Hasan', 'Rony Talukdar', 'Saif Hassan', 'Shamim Hossain', 'Shuvagata Hom', 'Yasir Ali', 'Zakir Hasan', 'Mohammad Naim', 'Khaled Ahmed', 'Mrittunjoy Chowdhury', 'Subashis Roy', 'Tanbir Hayder', 'Shadman Islam', 'Mahmudul Hasan Joy', 'Mahedi Hasan', 'Ariful Haque', 'Abu Hider', 'Kamrul Islam Rabbi', 'Jaker Ali', 'Shahadat Hossain Dipu', 'Parvez Hossain Emon', 'Muktar Ali', 'Tanzim Hasan Sakib', 'Rakibul Hasan', 'Hasan Murad', 'Saqlain Sajib', 'Shohidul Islam', 'Sumon Khan', 'Tanvir Islam']

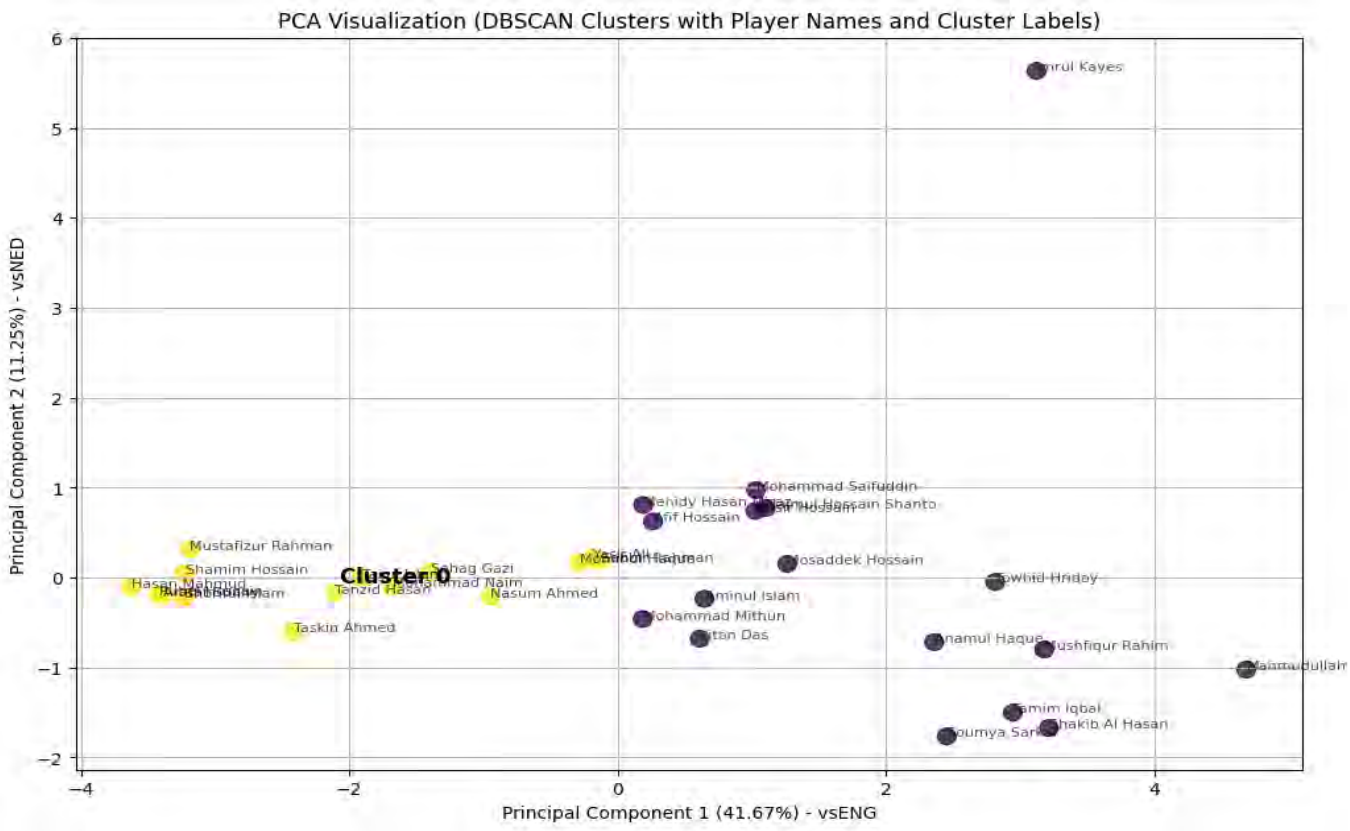

Figure 5.27: DBSCAN Clustering using PCA

5.1.8 ExKMC

Here we just showed how the models are taking decisions based on the k-means models where the model is taking decisions based on the features. Here we have shown two tree where first one has the value of k =2

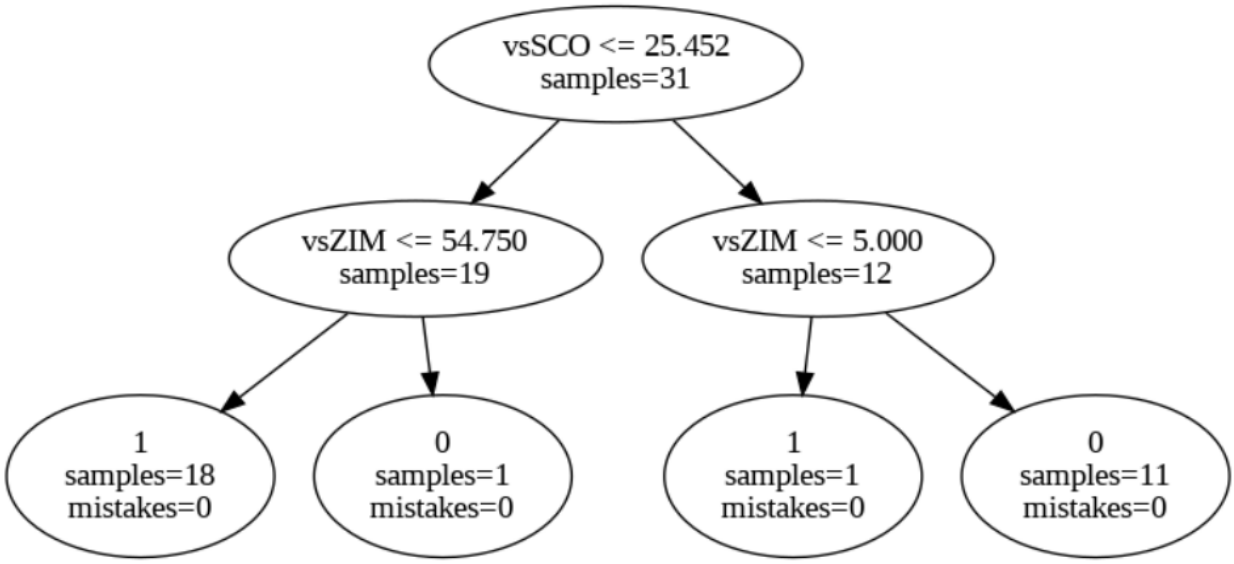

Figure 5.28: ExKMC for ODI Batting

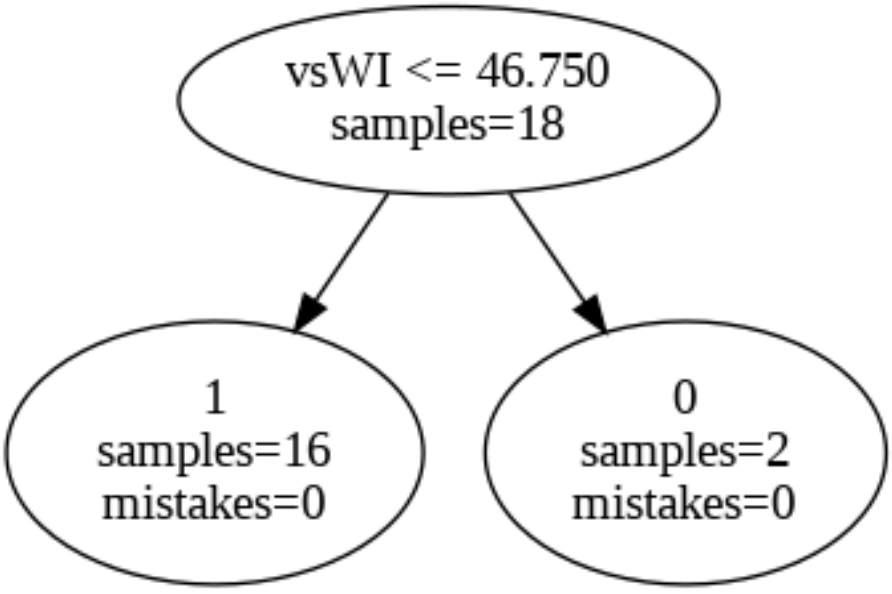

Figure 5.29: ExKMC for ODI Bowling

Chapter 6

Conclusion

6.1 Future Work

This studies focused on selecting optimal cricket team using unsupervised machine learning models. While the current approach provides valuable perception into player selection based on their country-wise total score averages, overall form and their recent form. There are also some several opportunities to make the model more accurate.

6.1.1 Incorporating Pitch and Weather Condition:

Analyse the pitch and weather condition would be one of the most important next step in this model. These factors also creates great impact on the players performance. By including this type of data, the model could show more precious team recommendations based on the expected match condition.

6.1.2 Expanding Different Formats of Tournaments:

Different types of league tournaments like IPL, BPL, Big Bash, which often require varying player skills. Considering the skills of the players in all these league tournament, their current performance and accurate data can be collected. By using this data, the model can give more accurate results.

Including these elements into future versions of the model would enhance more accurate and better-informed decisions in the dynamic world of cricket.

6.2 Conclusion

In conclusion, the use of machine learning in the selection of cricket teams has significant potential and can improve the decision-making process. The application of machine learning to cricket team management is likely to become more common and successful as technology develops and more data becomes available. In terms of predicting player performance in various scenarios and formats, our algorithms have produced encouraging results. This can aid in the selection of individuals who are most likely to perform well in particular circumstances, enhancing overall team performance. To determine who is performing exceptionally well, they use statistics like batting and bowling average,

batting strike rate, and bowling economy rate. Moreover, it will also check player performance against a specific team or opponent and the player's performance and records at that specific venue, Also it will determine these with weather conditions. Moreover, Massive volumes of previous player data can be analyzed by machine learning models, allowing selectors to base decisions on objective performance indicators rather than their subjective opinions. This lessens prejudices and improves the impartiality of team selection. In choosing a cricket squad, human judgment is still essential, thus machine learning should be viewed as a supplement rather than a replacement. Through data-driven decisions, better and more competitive teams may be formed in the future of cricket, which is made possible by this research.

Bibliography

[1] I. B. Mohamad and D. Usman, "Research article standardization and its effects on k-means clustering algorithm," *Res J Appl Sci Eng Technol*, vol. 6, no. 17, pp. 3299–3303, 2013.

[2] G. R. Amin and S. K. Sharma, "Cricket team selection using data envelopment analysis," *European journal of sport science*, vol. 14, no. sup1, S369–S376, 2014.

[3] R. Al-Shboul, T. Syed, J. Memon, and F. Khan, "Automated player selection for a sports team using competitive neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, 2017.

[4] T. Strauss and M. J. Von Maltitz, "Generalising ward's method for use with manhattan distances," *PloS one*, vol. 12, no. 1, e0168288, 2017.

[5] K. Passi and N. Pandey, "Predicting players' performance in one day international cricket matches using machine learning," in *Computer Science & Information Technology*, Feb. 2018, pp. 111–126. DOI: 10.5121/csit.2018.80310.

[6] K. Passi and N. Pandey, "Increased prediction accuracy in the game of cricket using machine learning," *arXiv preprint arXiv:1804.04226*, 2018.

[7] S. Banerjee, A. Mitra, D. Ganguly, R. Majumdar, and K. Chatterjee, "Innovative ranking strategy for ipl team formation," *arXiv preprint arXiv:1908.01725*, 2019.

[8] S. Sharma, N. Batra, *et al.*, "Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering," in *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, IEEE, 2019, pp. 568–573.

[9] P. Agrawal and T. Ganesh, "Selection of indian cricket team in odi using integer optimization," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1478, 2020, p. 012 001.

[10] N. Frost, M. Moshkovitz, and C. Rashtchian, *ExKMC: Expanding Explainable k-Means Clustering*, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.02399.

[11] C. Kapadiya, A. Shah, K. Adhvaryu, and P. Barot, "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, 2020.

[12] N. M. Patil, B. H. Sequeira, N. N. Gonsalves, and A. A. Singh, "Cricket team prediction using machine learning techniques," *Available at SSRN 3572740*, 2020.

[13] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

[14] M. Devanandan, V. Rasaratnam, M. K. Anbalagan, N. Asokan, R. Panchendrarajan, and J. Tharmaseelan, "Cricket shot image classification using random forest," in *2021 3rd International Conference on Advancements in Computing (ICAC)*, IEEE, 2021, pp. 425–430.

[15] M. K. Mahbub, M. A. M. Miah, S. M. S. Islam, S. Sorna, S. Hossain, and M. Biswas, "Best eleven forecast for bangladesh cricket team with machine learning techniques," in *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, IEEE, 2021, pp. 1–6.

[16] 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, 2021.

[17] L. Gunawardhana, "Optimising cricket team selection for one day international series based on match conditions," *Google Scholar*, 2022.

[18] M. Ishi, J. Patil, and V. Patil, "An efficient team prediction for one day international matches using a hybrid approach of cs-pso and machine learning algorithms," *Array*, vol. 14, p. 100 144, 2022.

[19] M. Ramalingam, S. Gokul, L. Mythravarshini, and K. Harine, "Efficient player prediction and suggestion using machine learning for ipl tournament," in *2022 International Mobile and Embedded Technology Conference (MECON)*, IEEE, 2022, pp. 162–167.

[20] N. R. Das, I. Mukherjee, A. D. Patel, and G. Paul, "An intelligent clustering framework for substitute recommendation and player selection," *The Journal of Supercomputing*, pp. 1–33, 2023.

[21] A. Jha, A. K. Kar, and A. Gupta, "Optimization of team selection in fantasy cricket: A hybrid approach using recursive feature elimination and genetic algorithm," *Annals of Operations Research*, vol. 325, no. 1, pp. 289–317, 2023.

[22] G. Saranya, A. Swaminathan, R. Surendran, L. Nelson, *et al.*, "Ipl data analysis and visualization for team selection and profit strategy," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 2023, pp. 592–598.

[23] M. Sumathi, S. Prabu, and M. Rajkamal, "Cricket players performance prediction and evaluation using machine learning algorithms," in *2023 International Conference on Networking and Communications (ICNWC)*, IEEE, 2023, pp. 1–6.

[24] B. Saji, *Elbow method for finding the optimal number of clusters in K-Means*, May 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/.

[25] *Density-based clustering in data minin - Javatpoint*. [Online]. Available: https://www.javatpoint.com/density-based-clustering-in-data-mining.

[26] *Hierarchical clustering in Machine Learning - Javatpoint*. [Online]. Available: https://www.javatpoint.com/hierarchical-clustering-in-machine-learning.

[27] S. KHAN, F. FAISAL, H. SHAH, and K. WAHEED, "Priority-based ranking using optimization model for cricket team performance,"

[28] *Principal Component analysis - JavatPoint*. [Online]. Available: https://www.javatpoint.com/principal-component-analysis.