Replacement for Chris Gayle in IPL based on Machine Learning algorithms (Research Project : 753A01)

Biswas Kumar (bisku859)
Linköping University

Spring 2021

Contents

List of Figures

List of Tables

Abstract

Cricket is a game of uncertainty, and it has the second-highest fan base in the world after Soccer. Many models have been constructed to predict the performance of players and a few about classification techniques as well. This project deals with a specific problem of exploring options for replacing CH Gayle, an explosive batsman cum all-rounder, at the Indian Premier League(IPL). A combination of unsupervised and supervised machine learning algorithms have been implemented to find relevant players who can replace Gayle in the tournament. Moreover, it has been attempted to rank the choice of players in the selection list based on performance and given conditions. Overall, this study tries to build, compare and shortlist a model with higher accuracy, which can group similar players based on past performance and predict their contribution to reason about order of selection.

1 Introduction

1.1 Background

Cricket is a wonderful game played with "bat and ball" on the ground between two teams with 11 players each. The first reference to cricket as a sport could be found in 1611, wherein it was defined as a boy's game by a dictionary [1]. England has a huge influence on the game for its growth and professional recognition, and the first known game in which the teams' used county names goes back to 1709 in England. Cricket has evolved since then and is now played at different regional and international levels. The first-ever international match took place between the United States and Canada in 1844 [2]. Currently, it is one of the most popular sports in the world, which boasts over 2.5 Billion followers worldwide and is only second to Soccer with 3 Billion followers [3].

The aim of cricket is simple: score more runs than the other team within the given limited overs. A team that scores more than the other wins the game. If both teams end up with the same score, the result is a draw. However, there are other special overs (super over or one over eliminator) [4] provided in some games to produce a result. The batsmen of the team batting first try to put up a defendable score on the board, while the bowling team tries to restrict the batsmen while taking wickets or delivering balls difficult to score. Runs could be scored while running between the wickets or by hitting boundaries (sixes or fours), and similarly, wickets can be taken directly, hitting the ball to stumps or through indirect ways such as catch-out.

Similar to cricket as a sport, its game format has also evolved. The Cricket format is primarily categorized as per the duration of the game with adjusted rules and conditions to suit the format. Currently, there are three formats of cricket played at the club, national and International levels – Test, One-Day(ODI), and Twenty20 (T-20) matches.

T-20 cricket is a shortened format wherein each team faces a maximum of twenty overs which was initially introduced to bolster crowds for the domestic game. It was not intended to be played internationally, but the popularity surged, which paved the way for the first T-20 International match on 17 February 2005 between Australia and New Zealand.

The Indian Premier League (IPL) is a professional T-20 cricket league that was founded in 2007 by the Board of Control for Cricket in India(BCCI). IPL conducts a tournament every year, which is contested by eight teams based out of different Indian cities. It is one of the most popular T-20 tournaments, and the brand was valued at a mammoth USD 6.7 billion in 2019 by Duff & Phelps [5]. The popularity could be judged from the recently concluded IPL 2020, which set a massive viewership record with 31.57 million average impressions registering an increase of over 23% from the 2019 season.

1.2 Problem

The teams are owned by different franchisees, and the performance of each player is critical to the performance of the team and the fortune of the franchise. There is so much at stake for a franchise that It becomes critical for them to carefully analyze the replacement of its superstar player who is on the verge of retirement and especially when the player is of the rank of Chris Gayle – the most explosive batsman and a powerhouse of T-20 cricket. Christopher Henry Gayle, who is famously known as Chris Gayle, has set numerous records across all formats of the game and is the first-ever batsman to hit 1000 sixes in T-20 cricket [6]. He was also included in the ICC T-20 Team of the decade in December 2020. There is a famous article published in the leading newspaper of the U.K with the title "The reinvention of Chris Gayle: how West Indies big-hitter became the Don Bradman of T-20" [7], which compared dominance of Gayle in T-20 format to all-time cricket great Don Bradman.

Gayle, who was born on 21 September 1979, is a Jamaican cricketer representing West Indies in international cricket since 1999. Gayle has now turned 41 years of age, and therefore his retirement from cricket is being speculated and talked about in media [8]. In February 2019, Gayle announced that he would retire from ODIs after the 2019 Cricket World Cup but later discarded the news and continued with his international career. In a recent press conference, he made it clear that he has no retirement plans till the 2022 T-20 World cup edition that is to be hosted in Australia [9]. Gayle currently plays for the KingsXI Punjab in the IPL, and amid all the speculations as well as back and forth statements on retirement, the franchisee needs to look out for a strong replacement for Gayle to avoid a decline in team performance.

1.3 Goal

The goal of the study is to explore the players who can replace Gayle in the team and provide reasonable options as a backup plan to avoid any negative impact of the abrupt departure of Gayle from the T20 cricket. The study is based on clustering techniques of unsupervised Machine Learning algorithms to identify similar players based on their performances. Once the clusters are constructed, Machine learning prediction models are built, which are trained and tested on the dataset to pick up the most efficient model based on its accuracy of prediction. The shortlisted model will finally be implemented on the clustered players alongside Gayle to predict scores and wickets as an indicator to reason about the ranking in the selection as a replacement to Gayle in IPL.

1.4 Motivation

Cricket is a game of uncertainty, and people have their own opinions and personal biases for players. There could be various avoidable and unavoidable biases which can be a big issue when it comes to the replacement of players. Machine learning models provide a scientific and reasonable method for player comparison and replacement options. This study aims to avoid human errors and biases in the selection of alternatives to one of the most valuable players in IPL. The selection has to be based on the performance of the players. This is of great importance as a wrong replacement may have a cascading effect on the reputation and commercials of the team and franchise.

2 Literature Review

Cricket is a very popular game that attracts many statisticians and researchers all over the world. The work has been done in different fields of the game. Lakshmi Ajay [10] has performed multiple popular clustering techniques on cricket players using unsupervised learning methodology and also used ipywidgets for UI interactions. Kapadiya and Co.[11] has noted that batting and bowling statistics have the most influence on the outcome of the match. They have highlighted attributes that should be considered for the prediction of the performance of players in ODI games. They have also bundled weather statistics to improve predictions. Passi and Pandey[12] used additional derived attributes such as form and consistency apart from other general attributes for the prediction of a player's performance using ML models. Iyer and Sharda[13] used a neural network model to classify batsmen and bowlers into three categories: performer, failure and moderate. Jhanwar and Paudi [14] predict the outcome of a cricket match by comparing the strengths of the two teams by developing a model where they determine the potential of the player by examining his career and recent performances.

This study is inspired by the real-world problem for the replacement of an IPL player and captures the application insights provided in relevant papers towards features selection, clustering methodology, and modeling techniques. The study uses unsupervised clustering techniques to cluster the IPL players and compare the result. The ideal cluster with the target player is selected for further investigation. The supervised machine learning based model is built to predict the performance of players of the target cluster in a given match or on the relevant dataset for prediction,and to decide ordering of players in the cluster.

3 Data and Tools

The data for the study was obtained from Kaggle [15] as it had the required ball-by-ball and match dataset of IPL since inception. The ball-by-ball dataset enclosed as appendix A, has 193468 balls details and corresponding 18 attributes. Similarly, the match dataset which is also enclosed as appendix A, has details of 816 matches spread over 17 attributes. These datasets were cleaned and grouped into Batsman, Bowler, and Fielders categories with relevant attributes. Finally, all these datasets were merged by the player for input to algorithms and models. The player dataset in appendix A contains details of 170 players propagated over filtered 14 attributes.

This study has used R-studio for data pre-processing, manipulation and implementation of different clustering techniques, Weka [16] for feature analysis, Google Colab and Python for building supervised machine learning models and prediction.

4 Data Pre-processing and Feature description

The attributes that were selected for Players under the different categories are :

4.1 Batsman attributes

The value of runs scored fall under a wide range, and small variations do not discriminate against different batsmen. It is, therefore, the rating under different optimum categories makes sense for quality prediction.

Table 1: Batting attribute table

Attribute	Description
ID	Match number
Runs_ scored	Runs scored per match
Batting_ team	Home team
Bowling_ team	Opponent
No_ 4s	Derived attribute, No. of boundaries (4 runs) scored
No_ 6s	Derived attribute, No. of sixes (6 runs) scored
Balls_ faced	Number of balls played
Strike_ rate	Derived attribute, (No. of runs scored / No. of balls played) x 100
City	Match venue
Man_ of_ Match	Award received for the best performance in the match
No_ 30	Derived milestone, (Scored more than 29 but less than 50 runs)
No_ 50	Derived milestone, (Scored more than 49 but less than 100 runs)
No_ 100	Derived milestone, (Scored 100 or more)
Runs_category	Categories of per runs scored
	1 : 0-9 runs
	2 : 10-29 runs
	3 : 30-49 runs
	4 : 50-79 runs
	5 : 80_plus runs
Date	Year

Weka was used to inspect the batting attributes using the Filter method and Ranker algorithm. The ordered output of Weka is highlighted below :
Ranked attributes: Date, Bowling_team, No_4s, Batting_team, No_100, Runs_scored, No_6s, Balls_faced, Strike_rate, City, No_50, No_30, Man_of_match, ID.

ID was discarded from the final attribute list as it was least important attribute, not making any difference in trial and needed a lot of computational power.

4.2 Bowler attributes

Bowlers have a tough time in T-20 cricket because the batsmen take their chances in the shortened version of the game. Moreover, the maximum number of overs allowed per bowler is 4 overs as against 10 overs in ODIs. Hence, it is necessary to divide the wickets taken into smaller categories to improve prediction results.

Table 2: Bowling attribute table

Attribute	**Description**
ID	Match number
Wickets_taken	Wicket taken
Batting_team	Opponent team
Bowling_team	Home team
Overs_bowled	Number of Overs bowled
Total_runs_conceded	Total number of runs conceded
Bowling_economy	Derived attribute, (No. of runs conceded / No. of overs bowled)
City	Match venue
3_Wickets	Derived milestone, (Wickets >2 but <5)
5_Wickets	Derived milestone, (Wickets $\geq$ 5)
Wickets_category	Category as per wickets taken 1: 0 2: 1-2 3: 3 4: 4 and above
Date	Year

Similarly, for batting, Weka was used to inspect the bowling attributes as well. The ordered output of Weka is as below :

Ranked attributes: Date, Overs_bowled, Wickets_taken, Bowling_team, Batting_team, Total_runs_conceded, 5 Wickets , 3 Wickets, City, Bowling_economy, ID.

ID was discarded from the final attribute list as it was least important attribute, not making any difference in trial and needed a lot of computational power.

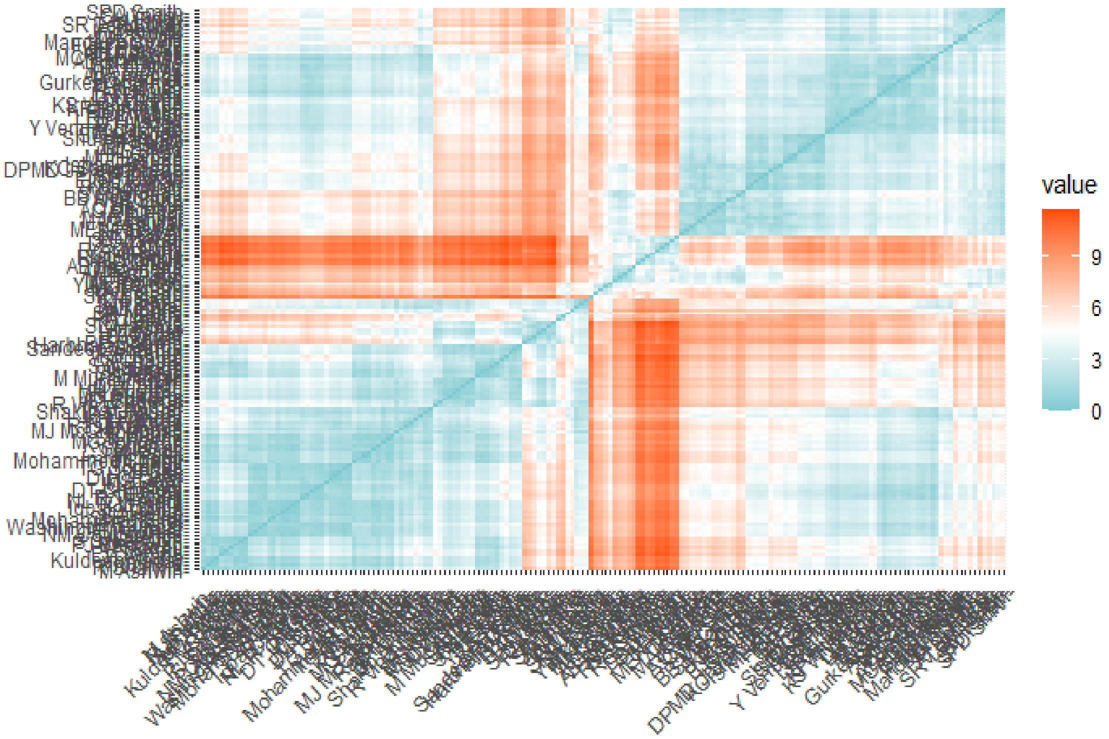
4.3 Player attributes

The player attributes combine both batting and bowling attributes along with general attributes of the fielder as well , such as the number of catches and man of the match. It is essentially combining all attributes at the player level. The ordered output of Weka is as below :

Ranked attributes: No_matches, Batting_avg, No_4s, Runs_scored, Innings_played, No_6s, Strike_rate, Bowling_economy, Balls_bowled, Total_runs_conceded, Overs_bowled, Wickets_taken, Balls_faced.

The dataset was scaled and the distance matrix was visualized ahead of unsupervised clustering implementation. Different clusters can be located based on the color contrast of distance matrix.

Figure 1: Distance Matrix



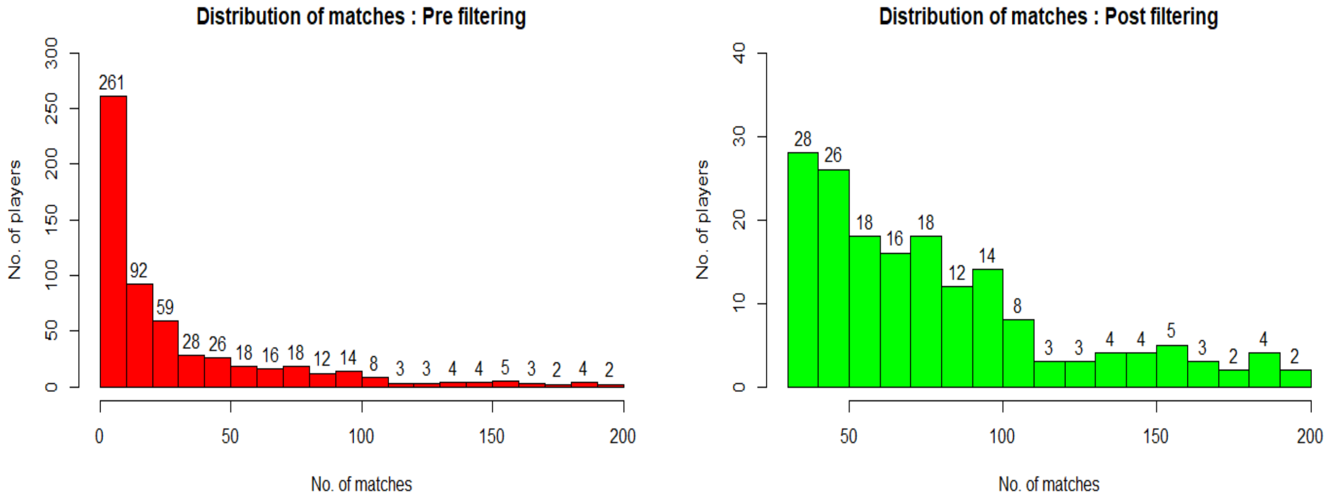
5 Algorithms and its implementation

5.1 Classification (Unsupervised machine learning)

Kamber and Pei [17] has defined cluster analysis or simply clustering as the process of portioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. It is an 'unsupervised machine learning' technique with no defined labels. It essentially looks for patterns in the data and groups them accordingly. There are different algorithms used for cluster analysis. This study has used popular cluster techniques such as K-means clustering and Hierarchical clustering to take an informed decision of the final clusters upon carefully analyzing the respective results.

The Player dataset, which includes batsman, bowler, and fielders attributes, has been used for the clustering purpose to figure the ideal cluster for Gayle. He has played over 140 IPL matches, and therefore the player dataset has been filtered for more than 30 matches for players for reliable comparisons. Thirty matches roughly equate to 2 seasons of IPL, which seems to be a fair cut-off to find players to substitute one of the most experienced IPL batsman.

Figure 2: Pre and Post-filtering of dataset



5.1.1 K-Means Clustering

Most partitioning methods are distance-based. K-means is a distance or centroid-based algorithm wherein a primary partition is created, and then it uses an iterative relocation technique that attempts to improve the partitions by moving objects from one group to another until "K" distinct clusters are obtained.

Three different K-means methods were performed in R using the fviz_nbclut package to find the optimum numbers of clusters.

5.1.1.1 Gap statistic

This is a very powerful method that compares the total within intra-cluster variation for different values of K with their expected values under the null reference distribution of the data. The name comes from the fact that the estimated number of optimal clusters will be a value that maximizes the gap statistic or which yields the largest gap statistic. This means that the clustering structure is far away from the random uniform distribution of points.

5.1.1.2 Silhouette (average silhouette width)

It is a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). The score ranges between -1 to 1, wherein a high value shows that the object is closely matched while low value reflects dissimilarity. The maximum value of the score will give the best value of K.

5.1.1.3 WSS (within-cluster sum of squares)

It reflects the sum of distances between the points and corresponding centroids for each cluster. This is more commonly known as the elbow method as it gives an optimum value of K at the elbow point i.e., the point wherein the distortion score declines the most. It is, therefore, there is a high ambiguity in picking up the value of K.

The optimal number of cluster traced below for all three methods.

The dataset is large, and therefore, the division of 170 players into 9 clusters looks optimum using the Gap statistics method as it does not seem to generalize the players into just a few clusters.

K-9 centres is annexed and the K-9 clusters are represented below :

Figure 3: Optimal number of cluster using Gap_staistics, Silhouttee and WSS methods

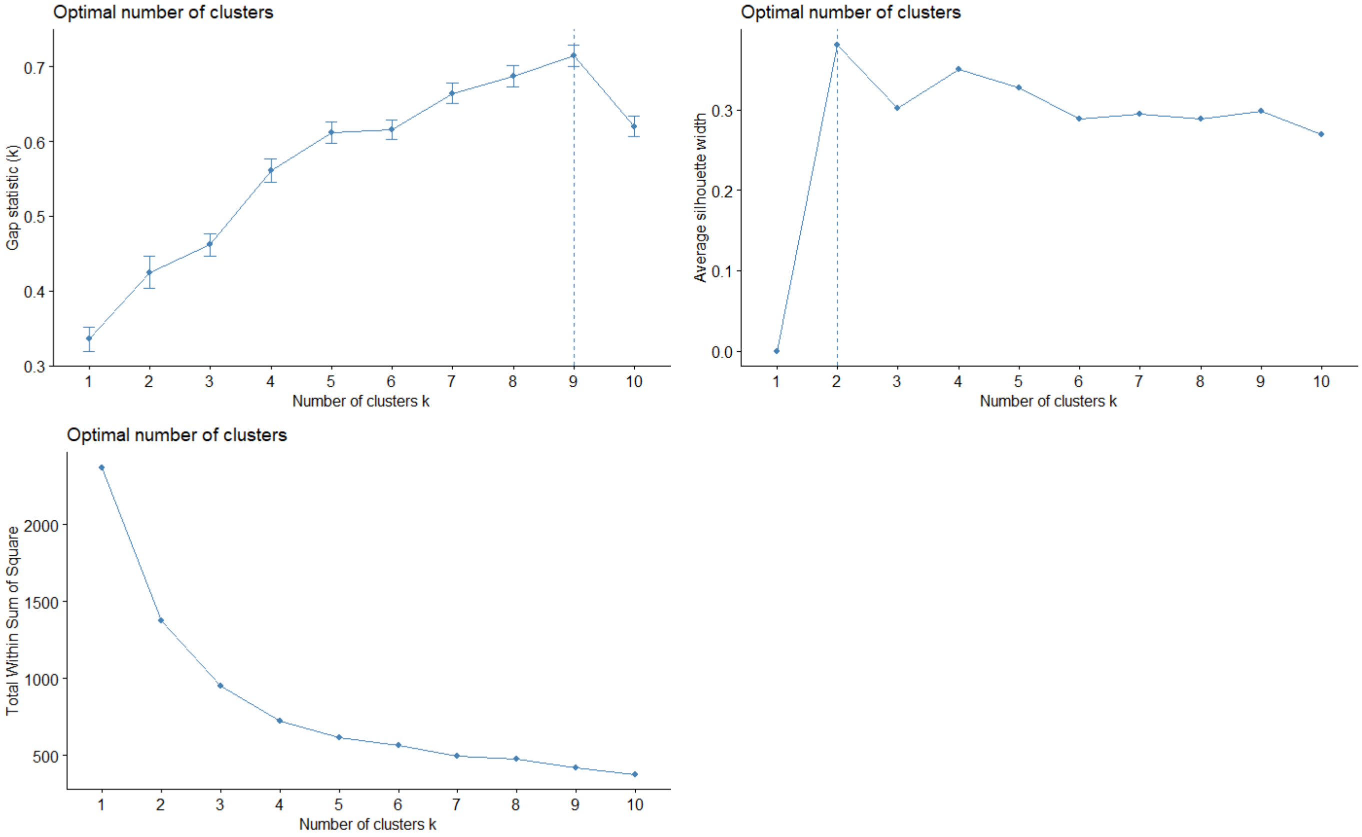


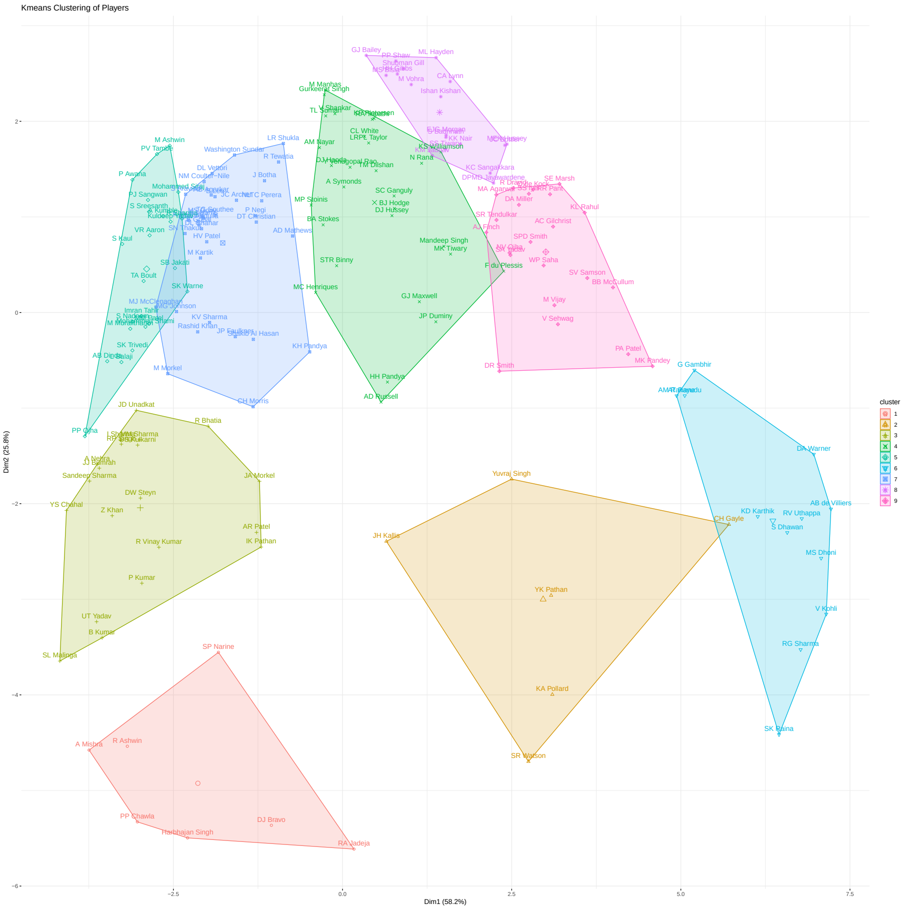
Table 3: Comparison of optimal K-means methods

Method	**No. of Clusters**
Gap_stat	9
Silhouette	2
WSS	4

Table 4: Numbers of Players in each K-9(Kmeans) Cluster

Cluster	**No.of Players**
1	07
2	06
3	20
4	29
5	23
6	12
7	33
8	18
9	22

Figure 4: K-9 Clusters



5.1.2 Hierarchical Clustering

It is a clustering technique that is used to build the cluster tree structure (a dendrogram), a hierarchical series of nested clusters. It starts by treating every data point as a separate cluster and then continue with the merging of the closest cluster to produce a dendrogram.
Agglomerative is a bottom-up hierarchical clustering method that starts with treating every data-points as a separate cluster and merges with closest with every iteration until it reaches the top of the merging point or until some stopping criterion is satisfied [18]

Figure 5: Hierarchical Clustering: Dendrogram

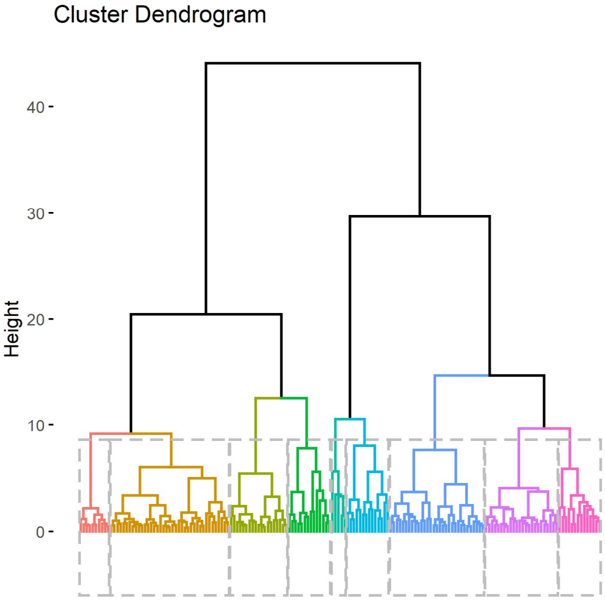


Table 5: Numbers of Players in each Hierarchical Cluster

Cluster	No.of Players
1	10
2	14
3	19
4	24
5	39
6	14
7	31
8	14
9	05

Table 6: Comparison of cluster in which Chris Gayle has been featured

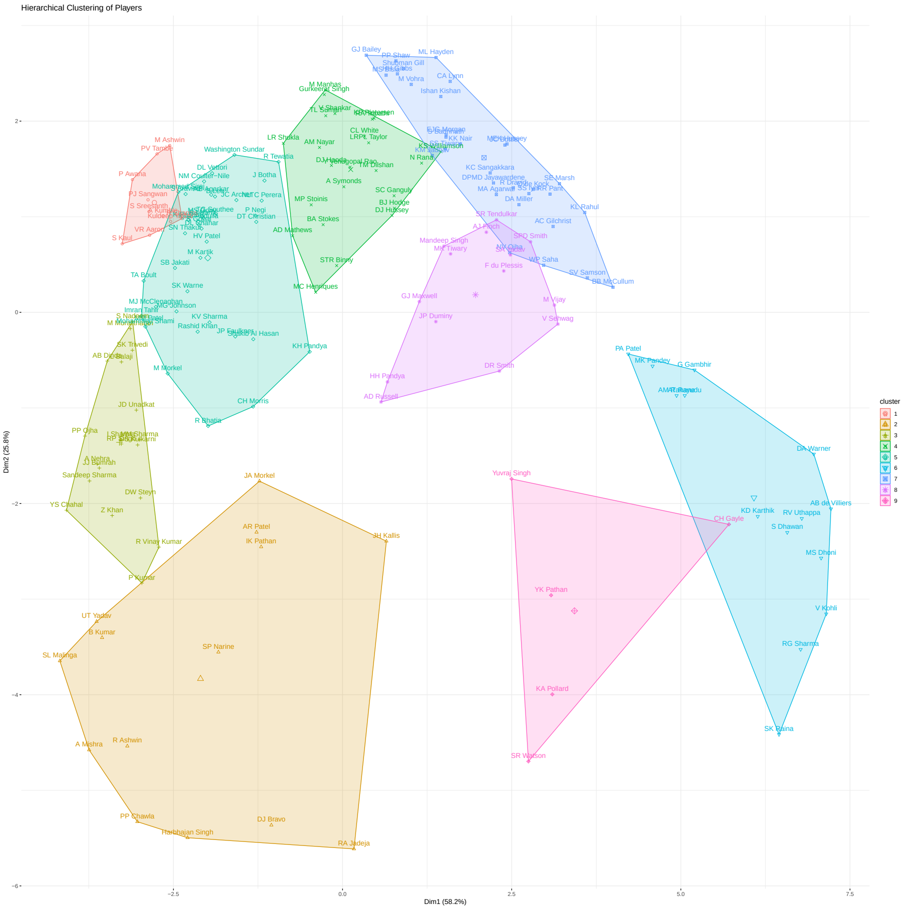
K-Means (Cluster No-2)	**Hierarchical Clustering (Cluster No-9)**
CH Gayle	CH Gayle
JH Kallis	KA Pollard
KA Pollard	SR Watson
SR Watson	YK Pathan
YK Pathan	Yuvraj Singh
Yuvraj Singh	

It may be noted that JH Kallis [20], who is featured in Kmeans cluster No-2, has retired from International Cricket. It is therefore, he was not included for further analysis in this study.

These two clusters (Kmeans and Hierarchical) looks identical and list out the same players after exclusion of JH Kallis. It can be concluded that the clustering techniques performed really well and did not provide any conflicting result.

Hierarchical clusters are represented below :

Figure 6: Hierarchical Clusters



These seem to be ideal clusters as all the clubbed players are known for their power-hitting performances. They are equally capable of taking the game away from their opponents at any stage of the game. Moreover, they all are useful bowlers too. This seems to be a great list of batting all-rounders.

Annexure B includes a enlarged complete list of players included in different clusters for K-means and Hierarchical clustering.

5.2 Learning Algorithms (Supervised Machine Learning)

The supervised machine learning models were used for generating the prediction models. The labeled dataset is used to train the model, and then it predicts results for the input features. In this study, the target datasets are Runs_category and Wickets_category used to find two different models for batting and bowling, respectively. The study used popular algorithms such as Naïve Bayes, decision trees and random forest to find the best-suited model for the purpose.

5.2.1 Naive Bayes

Naïve Bayes is essentially a probabilistic classifier based on the application of Bayes' theorem, which assumes strong independence among the features, given the class. If this holds, the Naïve Bayes is the most accurate classifier in comparison. Bayes models often provide generalization performance that is slightly worse than that of linear classifiers [18].
The naive Bayes classifier combines the Bayes probability model with a decision rule and the common rule is to choose hypothesis that is most probable[19].

$$Classify(f_1, \dots f_n) = argmaxp(C = c)) \prod_{p} (F_i = f_i | C = c)$$

Where,

argmax: maximization function,$f_1, ...f_n$:features, $p(F_i \mid C)$: independent probability distributions

P(C): prior probability of class

5.2.2 Decision tree

A decision tree is like a tree structure that reflects a flowchart pattern wherein each node denotes a test on an attribute value; each branch represents an outcome of the test, and leaves represent classes. Essentially, they learn a hierarchy of if/else questions, leading to a decision[18]. The classification is simple yet efficient and works with both numerical as well as categorical features.

5.2.3 Random Forest

Random forest is an ensemble method that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The decision tree tends to over-fit the training data, which is averaged out in the random forest as all trees work well and over-fit in different ways, but Random forest reduces the amount of over-fitting by averaging the different decision tree results[18].

5.2.4 Support Vector Machines (SVM)

SVM is a binary classifier that generates a decision boundary in multi-dimensional space using an appropriate form of a set of vectors(also known as support vectors) that belongs to the training set of vectors. SVM can perform both linear and non-linear classifications. SVM learns the linear discriminants of the form Wt + b that maximizes the margin wherein W represents the weight vector, and b is the threshold value. It maps training examples to points in space to maximize the width of the gap between two categories to have a distinct division of classes [19].

6 Results and Discussion

To find optimum models for batting and bowling predictions, a different set of training and test data was used to check the accuracy and precision of all the models in the experiment. The accuracy results are tabulated below for Naïve Bayes, Decision Tree, Random Forest and SVM classifiers.

The table highlights the fact that the highest accuracy for runs_scored is obtained as 98.21% with two different classifiers i.e., Random Forest and Decision tree, upon using 90% train and 10% test dataset . The precision% helps to understand the quality of classifier. Higher the precision%, better the classifier. In this case, the precision of Random Forest (98.29%) is similar to Decision Tree (98.29 %). In this scenario, it is wise to go ahead with Random Forest because the classifier is much more stable than Decision Tree in general. Overall, the models seem to respond very well to the dataset for prediction. Even with 70% of training and 30% test data sets, none of the models closed below the 91% accuracy mark.

Table 7: Comparison of performance of Batting Model Accuracy, Precision for various test-train combinations of the dataset

	Accuracy%, Precision %		
Classifier	**70% train - 30% test**	**80% train - 20% test**	**90% train - 10% test**
Decision Tree	92.26, 92.07	96.42, 96.66	98.21, 98.29
Naïve Bayes	95.23, 95.46	96.42, 96.63	96.42, 96.80
Random Forest	94.64, 94.74	97.32, 97.54	98.21, 98.29
SVM	91.66, 92.20	95.53, 95.70	94.64, 94.94

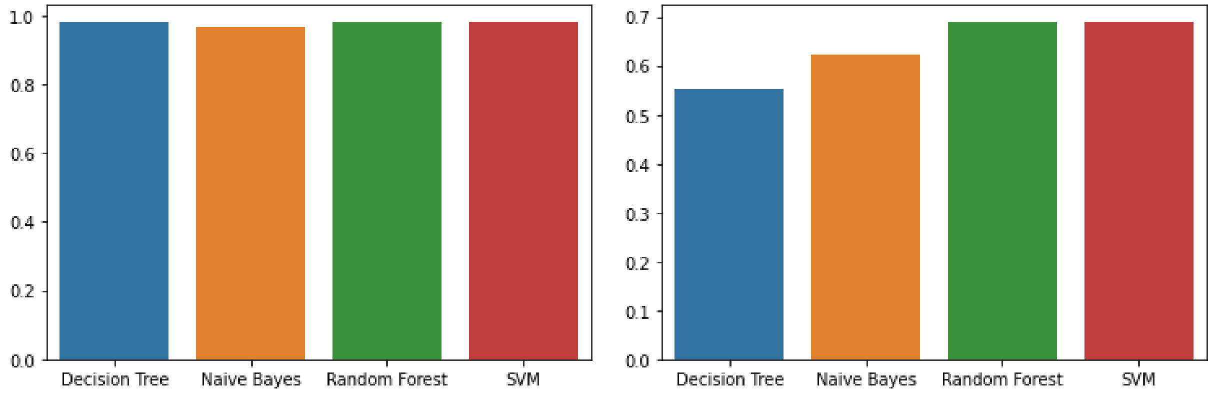
Similarly, the accuracy of bowling prediction models for a combination of the train-test datasets was observed, and the result is tabulated below for easy reference. It was observed that the accuracy level is low in comparison to batting models. The reason could be the lack of huge bowling datasets, and the selected players for prediction are majorly part-time bowlers.

Table 8: Comparison of performance of Bowling Model Accuracy, Precision for various test-train combinations of the dataset

	Accuracy%, Precision %		
Classifier	**70% train - 30% test**	**80% train - 20% test**	**90% train - 10% test**
Decision Tree	66.27, 66.82	61.40, 61.21	55.17, 55.43
Naïve Bayes	68.60, 68.35	71.92, 70.47	62.06, 70.28
Random Forest	70.93, 69.93	68.42, 68.90	68.96, 70.28
SVM	68.60, 67.05	71.92, 73.99	72.41, 77.87

As can be seen from the table, the SVM classifier has performed the best in accuracy (72.41%) with 90% train-10% test dataset. It is, therefore, SVM is the preferred choice for the bowling model.

Figure 7: Performance of various Batting(L) and Bowling (R)models on 90% train and 10% test dataset



The final summary of the shortlisted model for batting and bowling predictions is tabulated below.

Table 9: Final selected Batting and Bowling model for prediction

Model	**Accuracy%**	**Dataset Ratio**	**Classifier**
Batsman model	98.21	90% train-10% test	Random Forest
Bowler model	72.41	90% train-10% test	SVM

The outcome of the batting model on the selected players on randomly selected 10 matches between 2018-2020 are as highlighted below:

Table 10: Random Forest model: Runs(category) prediction for 10 random matches (cumulative)

Player	**Predicted_runs_category**
Watson	18
Pollard	13
Yuvraj	9
Yusuf	6

The outcome of the bowling model on the selected players on randomly selected 10 matches between 2015-2020 are as highlighted below:

Table 11: SVM model: Wickets(category) prediction for 10 random matches (cumulative)

Player	**Predicted_wickets_category**
Watson	22
Pollard	19
Yuvraj	15
Yusuf	17

Summary of batting and bowling results for the shortlisted players are as below: Final Table (Runs + Wickets Predictions) of randomly selected 10 matches.

Table 12: Random Forest model: Runs and Wickets prediction

Player	Predicted_Runs_category	Predicted_wickets_category	Order
Watson	18	22	1
Pollard	13	19	2
Yuvraj	9	15	3
Yusuf	6	17	4

The above table outline the rank or the preference of players most suited to replace Gayle in IPL. So, In order to respond to the research question, the top three players that can replace Chris Gayle in IPL are Watson, Pollard and Yuvraj in respective rank of choice based on clustering and performance of predictive model.It was a difficult choice for order number 3 but it seems Yuvraj has slight edge over Yusuf.

7 Conclusion and Future Works

The set of players were identified using clustering techniques and the selection seems relevant as all of them were explosive batsmen who can bowl too. The players were also comparative in the sense that they can take away the match single-handedly from the opponent. The order of selection was closely contested among the players. The machine learning algorithm performed extremely well in predicting the batting performances on the random dataset from 2018-2020. The bowling model, however, seems to be less accurate in prediction because of lack of data, and all of them were very unpredictable with their bowling as part-time bowlers. Overall, the final table of combined batting and bowling prediction gives a good insight into the performance ranking.

There could be some improvement in the model if few features could be added as described in some papers [12] [21]. These features could be the current form, weather factor, compatibility with team members and age-based performance. The study could be applied to other players in cricket for whom a franchisee may look out for replacement. This could also extend to other sports as well, such as kabaddi, football, etc.

References

[1] Details on early cricket on ICC official website *"Early Cricket (Pre 1799)"*, https://www.icc-cricket.com/about/cricket/history-of-cricket/early-cricket.

[2] Details on First International on ICC official website *"19th Century Cricket"*, https://www.icc-cricket.com/about/cricket/history-of-cricket/19th-century.

[3] Statistics details on website *"top-10-most-popular-sports-in-the-world"*,https://sportsshow.net/top-10-most-popular-sports-in-the-world/.

[4] Article in leading magazine *"ICC Comes Up With New Regulations For Super Over To Decide Tied T20I Matches - READ With Examples"*, https://www.outlookindia.com/website/story/sports-news-icc-comes-up-with-new-regulations-for-super-over-in-t20i-matches-read-with-examples/347090.

[5] Article in leading newspaper *"IPL brand valuation soars 13.5% to Rs 47,500 crore : Duff & Phelps"*, The Economic Times, September 20, 2019.

[6] Article published in leading newspaper, *"Chris Gayle becomes first to hit 1,000 sixes in T20 cricket"*, The Economic Times, India, January 17, 2021.

[7] Article published in the leading newspaper of the U.K with the title, *"The reinvention of Chris Gayle: how West Indies big-hitter became the Don Bradman of T-20"*, The telegraph, UK, April 14, 2021.

[8] Article published in the leading newschannel,*"Is Chris Gayle hinting at retirement with this cryptic Tweet?"*, Zeenews, November 2, 2020.

[9] Article published in the leading Newspaper *"No retirement plan as of now, two World Cups to go: Chris Gayle"*. The Times of India, January 1, 2021.

[10] Lakshmi Ajay, Article on "Machine Learning to Cluster Cricket Players", https://towardsdatascience.com/, 2021.

[11] Kapadia, Kumash and Abdel-Jaber, Hussein and Thabtah, Fadi and Hadi, Wael,*Sport Analytics for Cricket Game Results using Machine Learning : An Experimental Study*, Applied Computing and Informatics, doi:10.1016/j.aci.2019.11.006, 2019.

[12] K.Passi and N. Pandey, *"Increased Prediction Accuracy in the Game of Cricket Using Machine Learning"*, Int. J. Data Min. Knowl. Manag.Process, vol. 8, no. 2, pp. 19–36, 2018.

[13] S. R. Iyer and R. Sharda, *"Prediction of athletes performance using neural networks: An application in cricket team selection"*, Expert Syst. Appl., vol. 36, no. 3 PART 1, 2009.

[14] M. G. Jhanwar and V. Pudi, *"Predicting the outcome of ODI cricket matches: A team composition based approach"*, CEUR Workshop Proc., vol. 1842, no. September, 2016.

[15] The detailed statistics of all involved players shall be fetched from the websites: https://www.kaggle.com/ and https://www.iplt20.com/.

[16] Weka is open source software issued under the GNU General Public License, *"Machine learning algorithms for data mining tasks"*, www.cs.waikato.ac.nz/ml/weka/index.html.

[17] Han, J., Kamber, M., & Pei, J. *"Data mining: Concepts and techniques, third edition (3rd ed.)"*. Morgan Kaufmann Publishers,(2012).

[18] ANDREAS. MULLER, Andreas C, Sarah Guido, *"Introduction to Machine Learning with Python: A Guide for Data Scientists*, ISBN:9352134575, 9789352134571, O'Reilly Media, Incorporated, 2018.

[19] M. Narasimha Murty, V. Susheela Devi, *"Pattern Recognition: An Introduction"*, ISBN:9788173717253, Universities Press/Orient BlackSwan, 2011.

[20] Article published in leading Cricket News website,*"Kallis retires from international cricket"*, https://www.espncricinfo.com/story/kallis-retires-from-international-cricket-765649, 2014.

[21] Ankit Shah, *"Intelligent Cricket Team Selection by Predicting Individual Players' Performance using Efficient Machine Learning Technique"*, 2020, doi: 10.35940/ijeat.C6339.029320.

Appendices

A The datasets

1. Ball-by-ball dataset:
   The raw dataset downloaded from Kaggle that contains ball-by-ball details of all the matches in IPL(2008-2021) is enclosed as "IPL Ball-by-Ball 2008-2020.csv".

2. Match dataset:
   The raw dataset downloaded from Kaggle that contains details at match level(2008-2021) in IPL (2008-2021) is enclosed as "IPL Matches 2008-2020.csv".

3. Player dataset:
   The processed dataset that contains details of all shortlisted players who had played more than 30 matches in IPL(2008-2021) has been enclosed as "Players_statistics_final.csv " file for references.

4. Batsman train- test dataset:
   The processed train-test dataset (2008-2017) at Batsman level is enclosed as "bat_train_test_dataset.csv". It contains dataset for the clustered players i.e,. Gayle, Pollard,Yuvraj,Yusuf and Watson.

5. Batsman prediction dataset:
   The randomly chosen batting dataset of 10 matches (2018-2020) for each player in shortlisted cluster (i.e,.Pollard, Yuvraj, Yusuf and Watson) is enclosed as dataset for individual performance prediction of batting. This dataset has not been utilized in training or testing the model.

Table 13: Dataset for prediction of Runs

Player	Dataset
Watson	Watson_pred_data.csv
Yuvraj	Yuvraj_pred_data.csv
Pollard	Pollard_pred_data.csv
Yusuf	Yusuf_pred_data.csv

6. Bowler train-test dataset:
   The processed train-test dataset (2008-2014) at Bowler level is enclosed as "train_test_dataset_bowl.csv". It contains dataset for the clustered players i.e,. Gayle, Pollard, Yuvraj, Yusuf and Watson.

7. Bowler prediction dataset:
   The randomly chosen bowling dataset of 10 matches (2015-2020) for each player in shortlisted cluster (i.e,. Pollard, Yuvraj, Yusuf and Watson) is enclosed as dataset for individual performance prediction of bowling. This dataset has not been utilized in training or testing the model.

Table 14: Dataset for prediction of Wickets

Player	**Dataset**
Watson	Watson_pred_data_bowl.csv
Yuvraj	Yuvraj_pred_data_bowl.csv
Pollard	Pollard_pred_data_bowl.csv
Yusuf	Yusuf_pred_data_bowl.csv

B Clustering

1. Kmeans (K-9) centres enclosed as "k_9_centers.csv" which represents scaled cluster centers.

2. The result of Kmeans clustering technique listing all players enclosed as "K_means_cluster.csv".

3. The result of Hierarchical clustering technique listing all players is enclosed as "Hierarchial_Clust.csv".

4. K-9 clusters: Different colors represents different clusters and players are labelled for visual inspection.

5. Hierarchical clusters: Different colors represents different clusters and players are labelled for visual inspection.

C Learning Models

1. Python Raw File:
   The python file with title "Python_File.ipynb" is enclosed. It contains the Machine Learning models, comparisons and final results.

2. Python Pdf File:
   The pdf version with codes and output is also enclosed as "Python_File.pdf" for easy references.

3. Rmd Raw File:
   The R file "R_File.Rmd" is enclosed. It contains codes and results for data pre-processing, filtering and clusters.

4. Rmd Pdf File:
   The pdf version with codes and output is also enclosed as"R_File.pdf"for easy references.

Figure 8: K-9 Clusters (larger image)

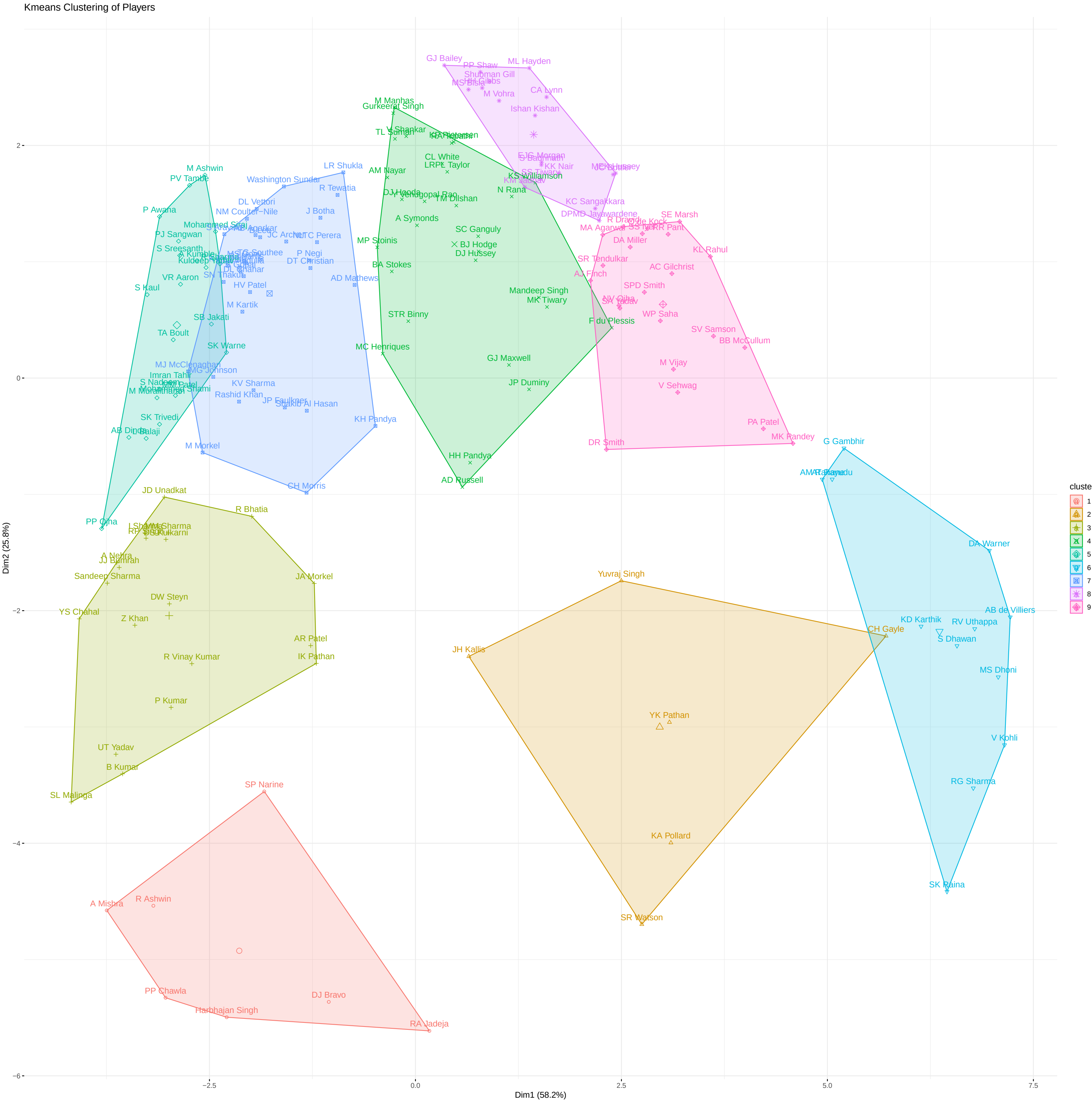
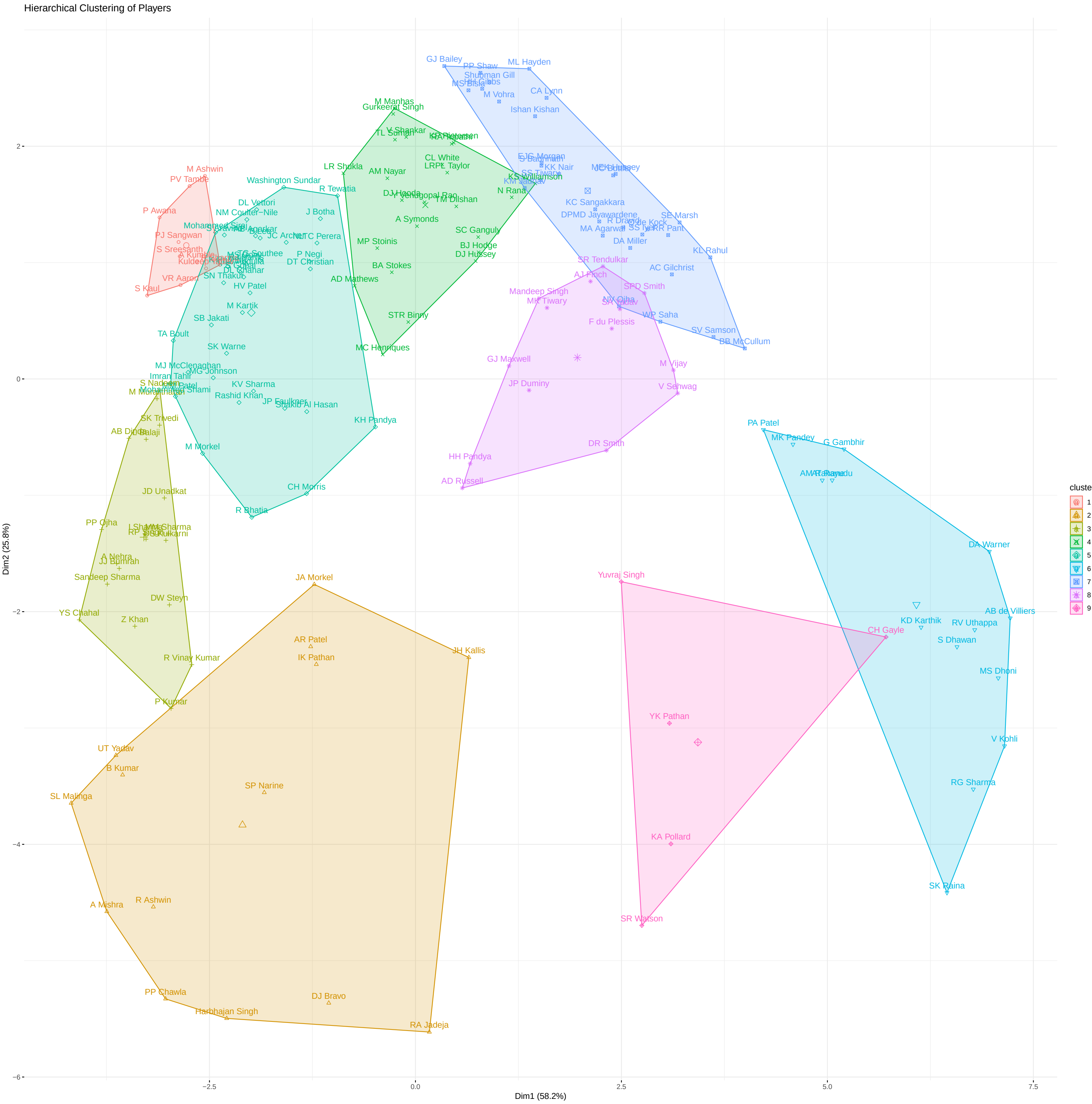


Figure 9: Hierarchical Clusters (larger image)


Hierarchical Clustering of Players
Dim1 (58.2%)
Dim2 (25.8%)
cluste
GJ Bailey
PP Shaw
ML Hayden
Shubman Gill
MS Bisla
M Vohra
CA Lynn
Ishan Kishan
M Manhas
Gurkeerat Singh
V Shankar
TL Suman
KP Pietersen
LR Shukla
AM Nayar
CL White
LRPL Taylor
EJG Morgan
KK Nair
SS Tiwary
KS Williamson
KM Jadhav
MEK Hussey
N Rana
M Ashwin
PV Tambe
Washington Sundar
R Tewatia
DJ Hooda
Y Venugopal Rao
TM Dilshan
P Awana
DL Vettori
NM Coulter-Nile
J Botha
A Symonds
KC Sangakkara
DPMD Jayawardene
SE Marsh
PJ Sangwan
Mohammed Siraj
JC Archer
NLTC Perera
SC Ganguly
MA Agarwal
SS Iyer
RR Pant
DA Miller
S Sreesanth
A Kumble
MP Stoinis
BJ Hodge
DJ Hussey
KL Rahul
Kuldeep Yadav
TG Southee
P Negi
DT Christian
SN Thakur
DL Chahar
BA Stokes
SR Tendulkar
AJ Finch
AC Gilchrist
VR Aaron
AD Mathews
HV Patel
SPD Smith
S Kaul
Mandeep Singh
MK Tiwary
NV Ojha
SA Yadav
M Kartik
STR Binny
WP Saha
SB Jakati
F du Plessis
SV Samson
BB McCullum
TA Boult
MC Henriques
SK Warne
GJ Maxwell
M Vijay
MJ McClenaghan
MG Johnson
Imran Tahir
S Nadeem
KV Sharma
JP Duminy
V Sehwag
M Muralitharan
MM Patel
Mohammed Shami
Rashid Khan
JP Faulkner
Shakib Al Hasan
SK Trivedi
KH Pandya
PA Patel
AB Dinda
L Balaji
MK Pandey
G Gambhir
M Morkel
DR Smith
HH Pandya
AT Rayudu
JD Unadkat
AD Russell
CH Morris
R Bhatia
PP Ojha
I Sharma
MM Sharma
RP Singh
DS Kulkarni
DA Warner
A Nehra
JJ Bumrah
Sandeep Sharma
JA Morkel
Yuvraj Singh
DW Steyn
YS Chahal
Z Khan
AB de Villiers
KD Karthik
RV Uthappa
CH Gayle
S Dhawan
AR Patel
IK Pathan
JH Kallis
R Vinay Kumar
MS Dhoni
P Kumar
YK Pathan
UT Yadav
V Kohli
B Kumar
SP Narine
RG Sharma
SL Malinga
KA Pollard
A Mishra
R Ashwin
SK Raina
SR Watson
PP Chawla
DJ Bravo
Harbhajan Singh
RA Jadeja