# “CricRank – Revolutionizing T20 Rankings using Machine Learning”

# PROJECT REPORT

Submitted for the course: Decision Support Systems (MAT5024)

By

**20MIY0001 SAMRIDH VIKHAS S**

Slot: G1

**Name of Faculty: NEELABJA CHATTERJEE**

**(SCHOOL OF ADVANCED SCIENCE)**

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ▪ CHENNAI
www.vit.ac.in

24 November, 2023

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| KPI | Key Performance Indicator |
| PCA | Principal Component Analysis |
| EDA | Exploratory Data Analysis |
| ICC | International Cricket Council |
| PM | Prelim Metric |
| PR | Prelim Rank |
| ML | Machine Learning |
| RBML | Role-based Machine Learning |
| PR | Preliminary |

# DECLARATION

I hereby declare that the thesis entitled ***“CricRank – Revolutionizing T20 Rankings using Machine Learning”*** submitted by me, for the award of the degree of Integrated Master of Science in Computational Statistics & Data Analytics to VIT is a record of bonafide work carried out by me under the supervision of Neelabja Chatterjee.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: 24th November, 2023

Samridh

**Signature of Candidate**

# CERTIFICATE

This is to certify that the project work entitled ***"CricRank – Revolutionizing T20 Rankings using Machine Learning"*** that is being submitted by ***"SAMRIDH VIKHAS S(20MIY0001)"*** for Decision Support Systems (MAT5024) is a record of bonafide work done under my supervision. The contents of this Project Work, in full or in parts, have neither be taken from any other source nor have been submitted for any other CAL course.

Place : Vellore

Date : 24th November, 2023

**Signature of Student(s): Samridh Vikhas S (20MIY0001)** - Samridh

**Signature of Faculty: NEELABJA CHATTERJEE (SAS) –**

# ACKNOWLEDGEMENTS

I would like to thank to our Chancellor, Dr.G.Viswanathan, Vice Chancellor, and Registrar for giving me the opportunity to work on this project.

I would like to thank our Dean of SAS, Dr.Arunai Nambiraj N, HOD of Mathematics Department, Dr. Jagadeesh Kumar MS and also the Faculty Members of SAS for their support during the project period.

My profound thanks to my faculty guide and mentor, Dr. Neelabja Chatterjee for his consistent support, guidance and mentoring throughout the study.

My sincere gratitude to my project mentor, Gaurav Sundararaman, Senior Analyst ESPN Cricinfo, my friends and peers for their support throughout the project tenure.

I thank the Almighty and all those who have supported directly or indirectly.

Samridh

**Samridh Vikhas S (20MIY0001)**

# ABSTRACT

The popularity of machine learning algorithms has increased, which is great for sports analytics. The ability to evaluate a player's performance has become simpler for sports analysts thanks to developments in machine learning and data mining. In this research, we have developed a new role-based performance indicator for batters and bowlers using machine learning algorithms. This research allows sports fans and researchers to compare players who play similar roles in different teams.

In this work, we started by collecting data from ESPNCricinfo and extracting meaningful KPIs along with the traditional performance indicators. Data preprocessing, feature selection and feature scaling has been done to apply clustering and classification algorithms. This study used the K-Means algorithm to generate clusters containing players with similar roles and then cluster-based roles were used as target vectors for classification algorithms. To determine the significance of each feature with the roles assigned by the clusters and create a role-based performance indicator, we used the one against-all method with Random Forest classification. To obtain a generalised performance indicator and validate the outcomes of the clustering and classification algorithms, we used PCA. In the end, this study compared scores generated and roles given to a player by this research with the preliminary score generated and categorization of the player using traditional methodology.

# 1. Introduction:

Cricket is a game that is heavily influenced by statistics. Every match is a unique event that generates a lot of data that may be utilised to assess a player’s performance. A global network of coaches, analysts, physiotherapists, dietitians, and trainers supports today's athletes and teams. There are a lot of variables and data to consider. To identify who has given the best performances, scorecards, series averages, and career statistics are examined. As a result, sports analysts are invaluable to players, coaches, the media, and sports fans, who look to them for more relevant and in-depth analysis. Sports analytics has benefited greatly from the development and acceptance of machine learning methods (Deep et al., 2016). Traditional metrics can help with these comparisons, but analytics provides the crucial context.

The game of cricket is a team sport for eleven players in which the team with the most runs scored wins the game. Three main formats of international cricket are played: test matches, one-day internationals, and 20-over matches. Each team bats, bowls, and fields in turn for at least one inning of each game. The batter tries to score as many runs as possible before getting out, while the fielding team utilises a bowler to get the batter out. A cricket inning is further divided into several overs, each of which consists of six valid deliveries delivered to the batter batting at the other end of the cricket pitch from one end. Each team is given a maximum of two innings in a test match, and a test match can last up to five days. 90 overs are bowled each day. In one-day internationals, both teams have a maximum of fifty overs to bat in the ODI format. In contrast, each team has a maximum of twenty overs to bat in T20 cricket. A team's captain may declare an innings over and instruct the opposition to bat, or the batting team may lose ten wickets or play the maximum number of overs. T20 cricket, then, is the quickest and most popular format for cricket fans who like quick action and excitement. Players who can bowl and bat are known as all-rounders, and they contribute by scoring runs and taking wickets. By putting out his best effort in every game, each player helps the team perform (Passi et al., 2018).

Rating individuals in team sports is a more complex task, primarily because of the team structure and some of the rules favour top-order batsmen (batsmen who bat during powerplay/field restrictions). In T20 internationals, the first six overs of an inning will be a mandatory powerplay, with only two fielders allowed outside the 30-yard circle. Beginning with the seventh over, no more than five fielders will be allowed outside the 30-yard circle (Power play - Wikipedia, 2022). The batting order is familiarly divided into three sections: the top order (batters one to three), the middle order (batters four to eight), which is further divided into two sections: the upper middle order (batters four and five), the lower middle order (batters six to eight), and tailenders (batters nine to eleven). The order in which the eleven players bat is usually determined before the start of a cricket match, but it can be changed during play (Batting order (cricket) - Wikipedia, 2022). If a batter plays in the top order, he will have a chance to play more balls with field enforcement so that he can score freely.

In professional team sports, player evaluation is the Holy Grail of analytics. Through player selection, teams are constantly attempting to improve their lineups (Davis et al., 2015). Baseball analytics have been extensively discussed, and Bill James is regarded as a pioneer of sabermetrics. Moneyball (Lewis, 2003), which was later turned into a well-liked Brad Pitt movie, A small-market Major League Baseball team called the Oakland Athletics used advanced analytics to identify and sign undervalued baseball players during the 2002 season. Moneyball may have sparked many of today's advances and interests in sports analytics (Davis et al., 2015). Performance evaluation is an important tool for quantitative analysts and operational researchers. Fund managers, for example, evaluate the performance of traders, and engineers monitor the performance of manufacturing lines. In sports, performance is typically measured using a rating system (in which players receive points for their performances) or a rankings list (in which players' performances are ordered). Rating systems are used by many stakeholders in the sports industry, including teams, fans, and pundits (an expert that comments publicly about a certain topic) (McHale et al., 2012).

In the shortest format of the game, we can use their batting (or bowling) “averages” to compare the performances of different batsmen (and bowlers) across multiple matches. A batter's “average” is calculated by dividing his total runs scored by the number of times he has been out. A bowler's average is calculated similarly by dividing the number of runs allowed by the number of wickets taken. The “strike rate” is another way to compare the performances of different players. A batter's “strike rate” is calculated by dividing the total runs scored by the number of balls faced and multiplying the result by 100. A bowler's strike rate is calculated similarly by dividing the number of balls bowled by the number of wickets taken. The “economy” is another measure available to bowlers for comparing the performances of different bowlers. The number of runs conceded by a bowler in one over is used to calculate his or her "economy".

## 1.1 Aim of the Project:

The purpose of the research is to develop a new role-based performance indicator for a player based on records using a large amount of data available on the internet and machine learning algorithms.

## 1.2 Objectives of the Project:

**1.2.1** Web Scraping Website and building a Clean Dataset

**1.2.2** Define new KPIs for batters

**1.2.3** Performing EDA and building a new dataset for Analysis

**1.2.4** To analyze a player’s past performance based on predefined performance indicators as well as newly defined KPIs

**1.2.5** Developing a new Machine Learning based performance indicator and predicting player performance based on performance indicators and including traditional metrics (strike rate and average).

## 1.3 Ethics:

There are several factors to consider when adhering to ethical principles in a study. First, there is a common desire to study objectives such as information, truth, and error avoidance. For example, prohibitions on fabricating, distorting, or falsifying research data promote reality and prevent errors.

We will only scrape data that is publicly available, and the data will be used for academic research purposes only. We will never claim that the data is ours or that we generated it; it will always be the property of ESPNCricinfo. Whatever information we gather will not be used or published for commercial or monetary gain. During this research, the data will be securely stored. Once the research is over and the data collected, all the models and KPIs will be deleted. The goal of this study's research will always be to discover new performance indicators rather than to mock or replicate any athlete. Any research findings will be presented truthfully and unedited. To reduce the bias produced by ML models, we will use normalisation, use correct learning models, mindful pre-processing, and other methods. We will not alter any datasets to improve a player's performance metrics or to defame any player.

# 2. Literature Review:

This chapter discusses related work on the development of new T20 international metrics to replace traditional metrics like average, strike rate, and economy. Researchers and sports enthusiasts have attempted to invent various methods for developing performance metrics using statistics and numerous machine learning algorithms.

During the past decade or two, many papers have been published on cricket performance measures and prediction methods. Barr et al. (2004) suggested a method based on strike rates and calculating the probability of getting out using the risk-return formula used in financial markets. If a batsman has a relatively high proportion of not-out scores, the batting average can be misleading. According to Lemmer (2008a), Lance Klusener scored 281 runs in eight innings and was only struck out twice during the 1999 World Cup Series. His best score was 52, but he had a 140.5 average! Nobody will believe that this was an accurate prediction of his next score or that his average in the next series will be at the same level.

Lemmer (2002) proposed a combined bowling method that replaces traditional metrics with a mathematical formula known as CBR, which is 3[1/Bowling Average + 1/Economy + 1/strike rate]. Lemmer published a series of papers on bowling and batting performance analysis using averages and strike rates in 2004, 2008b, and 2012. Lemmer (2008b) calculated several performance indicators to evaluate players in T20 internationals because the game's new and shortest format required new metrics, as traditional metrics favoured players who had played more balls. This study calculated the batter’s performance by generating a formula that aggregated the batter’s scores while he was out and while he was not out.

Davis et al. (2015) proposed a measure of expected run differential to evaluate a player in T20 international a measure that will provide how many additional runs a player will contribute to his team depending on the standard role Initially, they calculated the possible outcomes when a batsman faced a bowled ball. There were eight different outcomes ranging from 0 to 6 runs, with the batsman being dismissed on that ball. Following that, this study used mathematical modelling to calculate how many extra runs a player will score if he is replaced in the lineup by a standard batsman.

Eddie Cowan suggested Mike Hussey's number for batting (A Simple Metric to Understand Batting Efficiency in the IPL | The Cricket Couch, 2012). Michael Hussey, also known as Mr Cricket, was a prolific and consistent batsman. That is a player who averages 60 but hits 100 is as valuable to him as one who averages 20 but hits at a higher rate. Teams track statistics but not just traditional metrics like batting and bowling averages, strike rates, economy rates, and so on, but the combination of these to generate “magic numbers” to evaluate player efficiency. After that Eddie Cowan gave the formula of magic numbers by combining average and strike rate. Further research on the same number combination to get the new performance indicator Travis Basevi and George Binoy, Deputy editor, ESPNCricinfo. Basevi et al. (2007) gave the formula for batsmen by multiplying strike rate by average and then dividing the number by 100 and for bowlers by multiplying economy by average and then dividing the number by six.

Damodaran (2006) presented a Bayesian technique for dealing with not-out scores in cricket as an alternative batting average. They calculated the average runs a player has scored when he has previously been out on the occasion by scoring more runs than he has scored while he was not out and used that method to anticipate the number of runs a batter would have scored if he had remained not out during the innings. Machine Learning algorithms are used in all aspects of sports. Researchers, particularly in cricket, have primarily used Machine Learning algorithms to predict match outcomes and evaluate players (Deep et al., 2022).

Supervised Machine learning has been used widely in various sports. Oughali et al. (2019) used Random Forest and XGBoost to analyse player and shot predictions in the regular NBA season. Pugsee et al. (2019) used Random Forest to predict the match outcome in football by collecting the results of the previous three Premier League seasons (Premier League - Wikipedia, 2022). Basit et al. (2020) used Random Forest to predict the winning team of the 2020 T20 World Cup using historical data from ESPNCricinfo.

Based on historical data, Passi et al. (2018) used Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM) to predict how many runs a batsman will score and how many wickets a bowler will take. This study also developed new performance indicators such as consistency, player form, opposition rank, player performance at a specific venue, and more by combining traditional metrics such as average, strike rate, innings, centuries, and ducks of a player in a condition. The weights were calculated using an analytic hierarchy developed by Thomas L. Saaty. Following that, they used that problem to classify the bowler and batsman by assigning a rank from 1 to 5 to each performance indicator.

Iyer et al. (2009) created a neural network-based approach for forecasting athlete performance, which was later applied to predicting team selection. By consulting cricket experts, they provided initial ratings of Performer, Moderate, and Failure. To rank batsmen in one of the categories, they use the number of matches and runs scored, while bowlers use the number of bowls bowled and wickets taken. They employed ranking as a dependent variable after classifying the players. After that, they create a neural network and use a variety of strategies to build numerous models. They then classified their findings into four groups: the player recommended and selected, the player not recommended and selected, neither recommended nor selected, and the player recommended but not selected. The study's drawback was the initial grading system that was based on traditional approaches like runs scored and matches played, which eventually led to the average for batsmen and strike rate for bowlers.

Not only supervised algorithms, but researchers have used unsupervised Machine Learning algorithms also. Jhansi Rani et al. (2020) used a method based on K-means, Hierarchical clustering, and Neural networks to select teams for an Indian premier league match. Spencer et al. (2016) Used k-means clustering and Random Forest to cluster player/team profiles in the Australian Football League. K-Means Clustering is one of the most used clustering methods to classify players in sports due to the collection of data points aggregated together because of certain similarities (Guido and Müller, 2016).

Deep et al. (2016a) created a new Deep Performance Indicator based on machine learning to rank players in the Indian Premier League. They introduced some new key performance indicators (KPIs), such as hard hitter, which is calculated by dividing the number of boundaries a player has hit by the number of balls faced. Finisher, which is calculated by calculating the number of times a player has remained not out, and running between the wickets (RWB), which is the most underrated aspect of this shortest format. They calculated RWB by subtracting boundaries from runs scored and then dividing by the number of balls faced excluding boundaries. Aside from these three new measurements, they included old traditional metrics like average, which is known as consistency, and strike rate, which is known as a fast scorer for batsmen. For bowlers, they used traditional metrics like the economy, strike rate, and average, as well as defined new indicators like the Big Wicket Taker, which is calculated by dividing the number of times a player has taken four or five wickets by the number of balls bowled, and short performance, which is calculated by dividing the number of times a player has taken less than four wickets by the number of innings bowlers have taken at most three wickets. They categorised batters according to batting orders such as openers, middle order, finishers, and inexperience batters. Bowlers categorise them into spinners, pacers, inexperience, spin all-rounders and pacers all-rounders. They used the Caret package from R to get the five important features of each category then they multiplied their KPI value by the feature importance given to get the final rating for each player.

Deep et al. (2016b) used the same KPIs as Deep et al. (2016a), but instead of categorising players, they used Random Forest to calculate feature importance for each KPI. After multiplying the KPI value with feature importance they get the final Deep Performance Index (DPI) index for each player.

Deep et al. (2022) suggested a new approach based on the Deep Player Performance Index (DPPI), which considers the player's form and role in the team. They used predefined Key Performance Indicators (KPIs) from the previously mentioned paper (Deep et al., 2016a), and added new performance indicators such as the Boundary index number of boundaries players have hit in all the balls the player has faced, and the Big Innings index, which is calculated by dividing the number of big innings played by the number of innings played. They established seven batters KPIs and six bowler KPIs. Then, between 2015 and 2018, they extracted these KPIs for T20 international players, IPL players, and players in Syed Mushtaq

Ali Trophy (SMAT) (Syed Mushtaq Ali Trophy - Wikipedia, 2022), a T20 league like the Vitality T20 Blast (T20 Blast - Wikipedia, 2022) in England. Then, depending on KPIs, they used K-means clustering to determine player roles. Then, using the retrieved KPIs as predictors and cluster-based roles as target vectors, they trained a Random Forest classifier. They then normalised the feature importance value and paired it with the normalised KPI values to produce strengths based on T20 statistics, as well as IPL and SMAT data. To arrive at the DPPI scores, they combined both T20 Internationals and Domestic (IPL/SMAT) information.

The role-based method further divided hitters into roles like floaters (a batter who can bat anywhere depending on the occasion), top order/specialist batsmen, and expert finishers (A batter who comes towards the end of the innings and plays with a good strike rate). Bowlers were further divided into five groups mainly bowlers who can bowl in addition to batting, specialist T20 bowlers, and impact bowlers. Bowling is the most important part of cricket, and it has a much greater impact than batting. Teams often struggle to establish the proper bowling attack based on their roles. For example, in T20, one bowler can be used as a swing bowler (a player who can move the ball in the air after striking the pitch) who will bowl most of his quota in the first seven to eight overs. One bowler may be designated as the death bowler, who will begin bowling after the twelfth over of the game. The same goes for batters: if a top-order batsman gets you off to a good start, you can capitalise on it by sending players who can score with a high boundary index number KPI. One of the study's minor flaws is that they only used half-century and century scores without including strike rate when generating the Big Innings Index. In a T20 contest, if a player scores a half-century in forty balls, his strike rate is 125.0, which is not good, but the context of the innings is crucial in any performance measure. In the preceding situation, where the team only achieved 140 runs for the loss of eight wickets in 120 balls, the player's strike rate of 125 was quite impressive. Obtaining data with that level of granularity, on the other hand, is a difficult task in and of itself. While examining the performance analysis of a player like Axar Patel, who took 27 wickets in six innings against England on the England tour of India 2020-21 with an average of 10.74, a strike rate of 28.3, and an economy of 2.24 (Anthony de Mello Trophy, 2020/21 Cricket Team Records & Stats | ESPNCricinfo.com, 2022), one extra performance indicator based on pitch condition could be added if one can find data with that granularity. Axar player was not included in the reverse fixture of the tour.

Based on the previously mentioned papers, K-means is one of the best techniques for classifying players because it uses a collection of data points combined due to certain similarities. Next, we can combine supervised learning algorithms like Random Forest and XGBoost to categorise data and determine the feature importance of each feature. Below is a summary of one of our strategies. First, we implement the unsupervised learning algorithm K-Means Clustering, which clusters players instead of relying on a predefined set of classes. Then, using a set of labelled players, we implement the random forest supervised classification algorithm to train the model to anticipate unclassified players in role-based clusters. Combining supervised and unsupervised learning allows us to assess the relative value of various traits for each role. It improves upon the player performance evaluation techniques currently used in T20 cricket.

McHale et al. (2012) attempt to assign a single score to every player, regardless of their speciality, based on their contributions to winning performances in the premier league and championship, the top two divisions of English football. Their goal was clear

1). To build a statistical index without any subjective opinions.

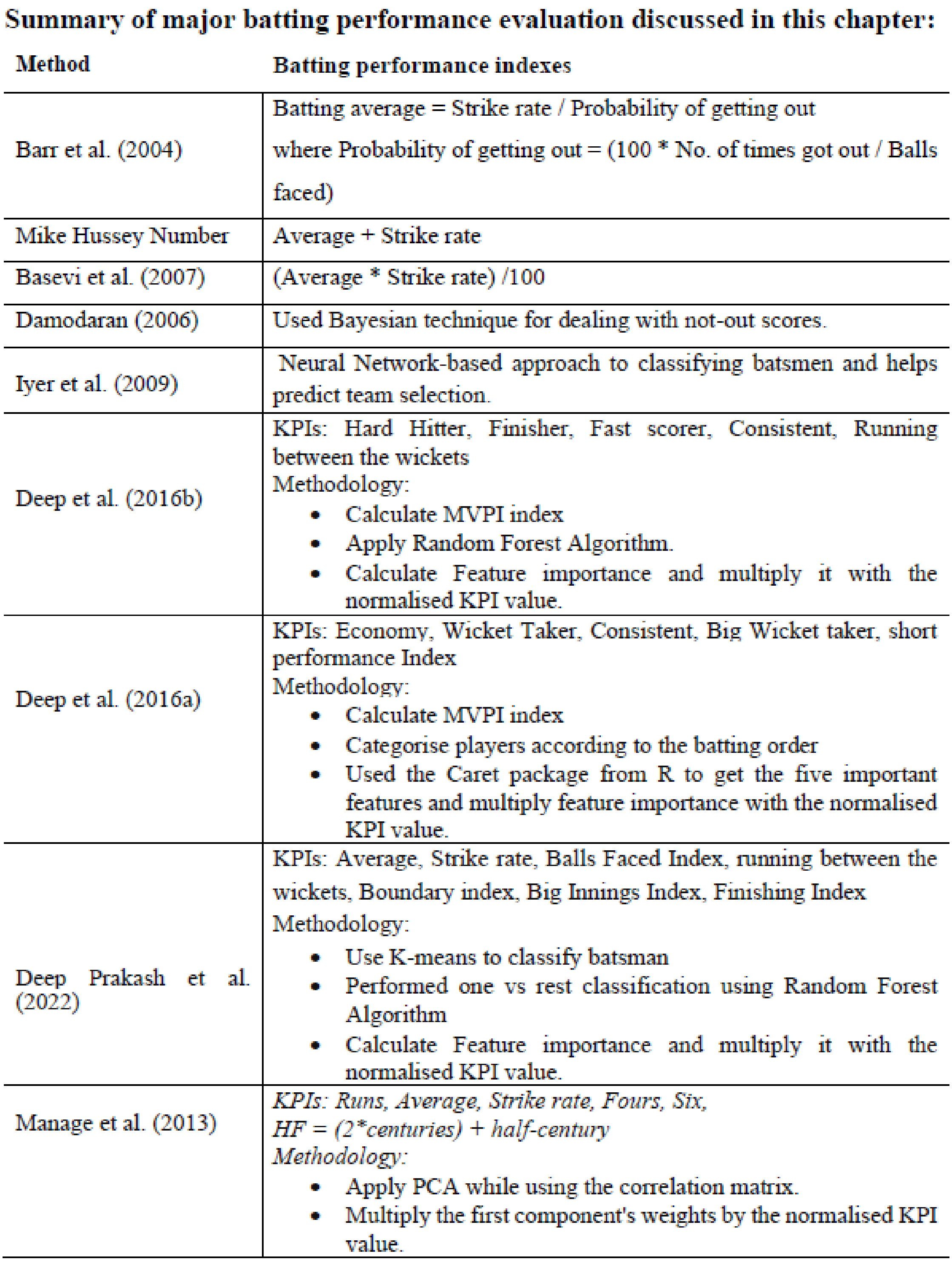
2). Compare players across different positions.

3). A trade-off must be made between the simplicity and complexity of the model.

4). The model should be explainable.

Additional specifications for those metrics included the final essential indicators should include goals scored directly. Players should also receive points if their teams can keep a clean sheet. A player should receive a point if they assist other players to score goals. They divide the entire match into six subindices, such as modelling the match, where they tried to rate the player using a variety of criteria, including crosses, dribbles, passes, interceptions, yellow and red cards, tackle win ratio, and more. Points-sharing index, appearance index, goal scoring index, assist index, and clean sheet index were the other five subindices. The final index is a weighted average of the points earned across all indices. Nevertheless, there are some limitations of this index goal scoring players such as forwards and midfielders will always be rated higher than the defenders and goalkeepers. Because a football team wants to score goals and win the game, the goals scored must be kept in the performance evaluation parameters. Like this, in T20 cricket, important KPIs are runs scored and wickets were taken. Additionally, the flexibility to score runs (how quickly runs are scored, what number of balls the batter faced, and the way many boundaries the batter scored where the ball crossed the playing field's perimeter, which usually yields four or six runs) and therefore the ability to require wickets (how many wickets the bowler took, what percentage runs the bowler conceded, and the way many balls the bowler bowled) also matter looking on the sport situation.

Manage et al. (2013) ranked players who competed in the 2012 Indian Premier League (IPL) were ranked using Principal Component Analysis (PCA). More specifically, the PCA's First Component. This standard used Runs, Batting Average, Strike rate, Fours, Sixes, and HF new indicator based on the number of times the batsmen scored centuries and half centuries. This study used the Bowling average, Strike rate, Economy, and Wickets taken by bowlers. Following PCA analysis, this study determines the calculated eigenvalue and total variability for all PCA components, as well as values multiplied by the coefficient of the first PCA component. One of the study’s drawbacks was they were focused on the traditional metrics such as runs/wickets, strike rate and average rather than finding more KPIs.



**Summary of major batting performance evaluation discussed in this chapter:**

| Method | Batting performance indexes |
|---|---|
| Barr et al. (2004) | Batting average = Strike rate / Probability of getting out where Probability of getting out = (100 * No. of times got out / Balls faced) |
| Mike Hussey Number | Average + Strike rate |
| Basevi et al. (2007) | (Average * Strike rate) /100 |
| Damodaran (2006) | Used Bayesian technique for dealing with not-out scores. |
| Iyer et al. (2009) | Neural Network-based approach to classifying batsmen and helps predict team selection. |
| Deep et al. (2016b) | KPIs: Hard Hitter, Finisher, Fast scorer, Consistent, Running between the wickets<br>Methodology:<br>• Calculate MVPI index<br>• Apply Random Forest Algorithm.<br>• Calculate Feature importance and multiply it with the normalised KPI value. |
| Deep et al. (2016a) | KPIs: Economy, Wicket Taker, Consistent, Big Wicket taker, short performance Index<br>Methodology:<br>• Calculate MVPI index<br>• Categorise players according to the batting order<br>• Used the Caret package from R to get the five important features and multiply feature importance with the normalised KPI value. |
| Deep Prakash et al. (2022) | KPIs: Average, Strike rate, Balls Faced Index, running between the wickets, Boundary index, Big Innings Index, Finishing Index<br>Methodology:<br>• Use K-means to classify batsman<br>• Performed one vs rest classification using Random Forest Algorithm<br>• Calculate Feature importance and multiply it with the normalised KPI value. |
| Manage et al. (2013) | *KPIs: Runs, Average, Strike rate, Fours, Six,*<br>*HF = (2*centuries) + half-century*<br>*Methodology:*<br>• Apply PCA while using the correlation matrix.<br>• Multiply the first component's weights by the normalised KPI value. |

# 3. Contributing Chapters:

## 3.1 Methodology:

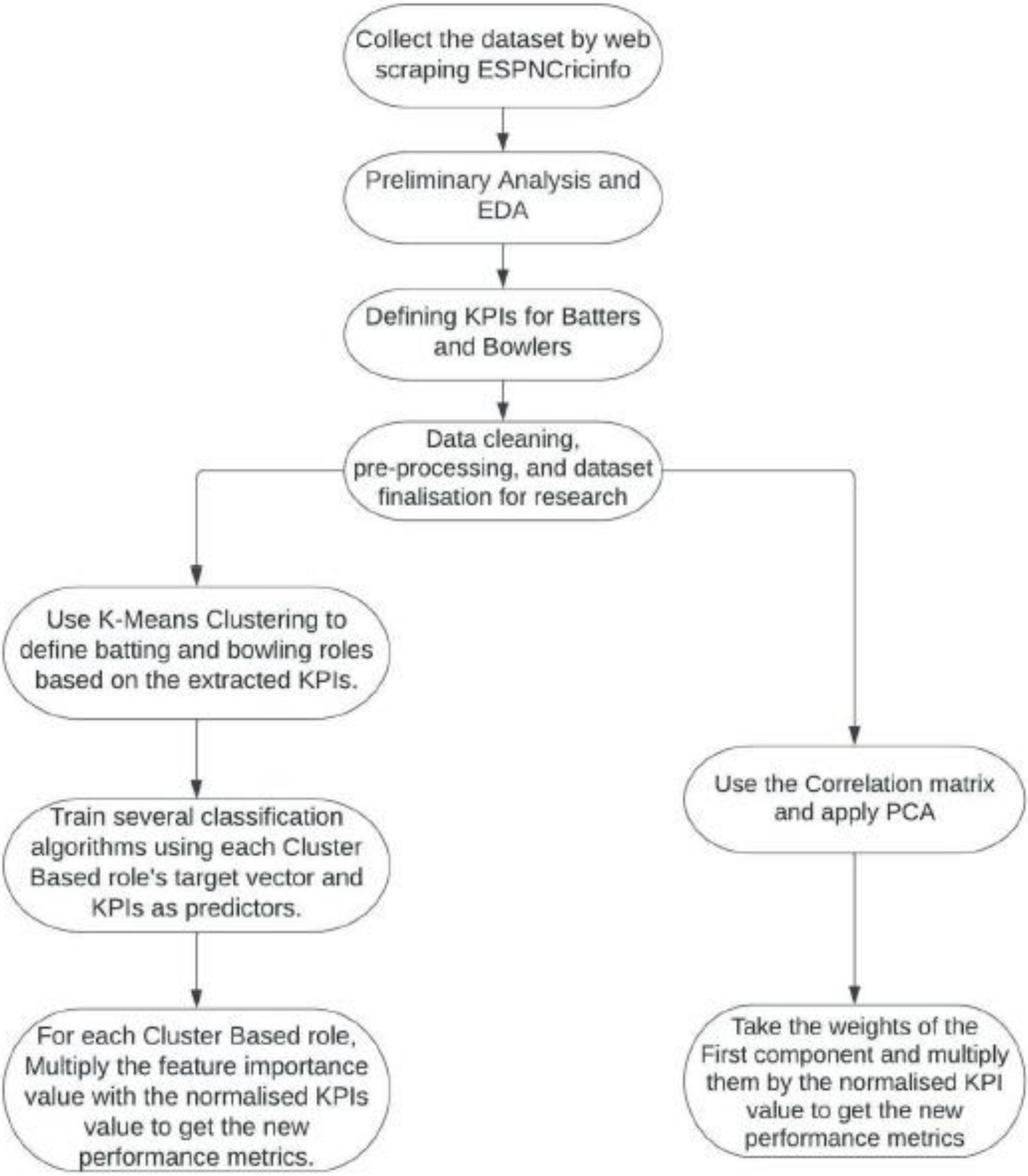

Collect the dataset by web scraping ESPNCricinfo
Preliminary Analysis and EDA
Defining KPIs for Batters and Bowlers
Data cleaning, pre-processing, and dataset finalisation for research
Use K-Means Clustering to define batting and bowling roles based on the extracted KPIs.
Train several classification algorithms using each Cluster Based role's target vector and KPIs as predictors.
For each Cluster Based role, Multiply the feature importance value with the normalised KPIs value to get the new performance metrics.
Use the Correlation matrix and apply PCA
Take the weights of the First component and multiply them by the normalised KPI value to get the new performance metrics

*Figure 1 – Methodology of this research*

Figure 1 depicts the research's methodology. This study considered data from the start of T20 International to April 17th, 2022. Both active and retired players are present in this collection. This study extracted KPIs based on the data available for both batters and bowlers after conducting preliminary analysis, exploratory data analysis, and removing the outliers. Using the KPI values for each data set, we assign each player a particular role. We must strike a balance between the model's intricacy and simplicity. For this reason, rather than setting them specifically for each role, we established these KPIs without taking the roles into account. Before finishing our dataset for further investigation, we will again undertake data preprocessing and cleaning after defining KPIs. Cricket is an eleven-person side game where each player on the team has a distinct responsibility. After that, this study used normalisation techniques to standardise the range of the data so that algorithms can better understand it and cluster it properly without being biased toward larger values like average and strike rate (Why is scaling required in KNN and K-Means? | Medium, 2019).

Following normalisation, this study used K-means clustering to obtain the target vector and group players who were playing the same role. This study used the elbow approach and the silhouette score of the algorithm to compute the number of clusters (K) based on different cluster counts. The "Elbow Method," a heuristic used to determine the number of clusters in a data collection, selects the elbow of the curve as the ideal value of n (number of clusters) present in the Dataset (Humaira et al., 2020). The K-Means clustering approach provides the same data division regardless of how the observations are arranged because it is order-independent. After that, to determine the importance of each characteristic for each role, this study utilises classification algorithms like XGBoost, Random Forest, and Voting Classifier. To obtain the new performance measures for players, this study multiplied the feature importance of each cluster-based role by the normalised value of KPIs.

Another strategy will involve performing PCA on the normalised data while using the correlation between the KPIs. In this research, after looking at the eigenvalues and Total Variability values, this study takes the coefficients of the first component as the feature importance value. To obtain the new performance measures for players, this study multiplied the coefficients of the first primary axis with a role by the normalised value of the KPIs.

Following this study, one will have two new performance indicators/formulas to grade players. We have thought about some of the fundamental ideas from (McHale et al., 2012)

When developing this approach,

1. To build a statistical index without any subjective opinions.

2. Compare and rate players based on their roles in the team.

3. A trade-off must be made between the simplicity and complexity of the model.

4. The model should be explainable.

5. When creating KPIs, runs for batters and wickets for bowlers should be prioritised because these are the ultimate goals of each batter and bowler, respectively.

## 3.2 Technologies:

The main programming language we will use to conduct this research is Python. Python is a powerful, adaptable, and simple programming language that is suitable for a wide range of machine learning (ML) and artificial intelligence (AI) projects. There are many machine learning and AI tools and packages to pick from, and Python is one of the most popular programming languages in data science projects. With Python's extensive collection of standard libraries, it is feasible to perform machine learning, image processing, scientific computing, text processing, and other tasks (Paffenroth et al., 2015). The project has used the following Python libraries. Pandas, Numpy, Beautifulsoup, Matplotlib, and Scikit-learn.

## 3.3 Experiments:

### 3.3.1 Web Scraping:

We have done web scraping using BeautifulSoup, one of the widely used libraries for web scraping websites. BeautifulSoup is a library that makes it simple to scrape data from websites. It sits on top of an HTML or XML parser, allowing you to iterate, search, and edit the parse tree using Pythonic idioms. (beautifulsoup4, 2022).

For this study, we have used the dataset of 4000 entries from ESPNCricinfo, 2000 for each batter and bowler. Data was gathered starting with the first T20 International match and continuing through September 25th, 2023. This dataset has included both retired and active players. For batters during their T20 career, Innings, Not Out, Runs, Average, Strike rate, Number of Boundaries, Times a Player Was Out Without Scoring Runs, and Balls Faced are included. The information for bowlers also includes the number of overs bowled, the number of debut overs bowled Best Bowling, statistics, and wickets taken, average, and strike rate of a bowler, as well as how many times they have taken three or more wickets.

### 3.3.2 Preliminary Analysis and Exploratory Data Analysis:

The initial data analysis, pre-processing, and feature engineering will all be covered in this section. In the initial data analysis, we look for null values and missing values. This study made sure that the values are aligned with the feature and that the datatype in Pandas accurately reflects the feature's datatype.

There were 54 batters in the dataset whose average is not defined because they were yet to be dismissed in T20 Internationals. Most of them are lower-order batsmen or tailenders who are bowlers who just come for batting on very few occasions. We replaced their average with the highest score in T20 internationals for all 54 players as all of them have batted for less than 5 innings in T20 internationals. There were a few players whose country was not defined in the dataset and a few players who played for more than one country. We replaced and cleaned the dataset along with the country names.

As you can see in figures 2, the batter's data is skewed towards the left side in all aspects because many players have played few games. To account for this, reduce bias, and ensure that our model functions properly, this study excluded any batters who have played fewer than 25 balls or fewer than four innings of batting.

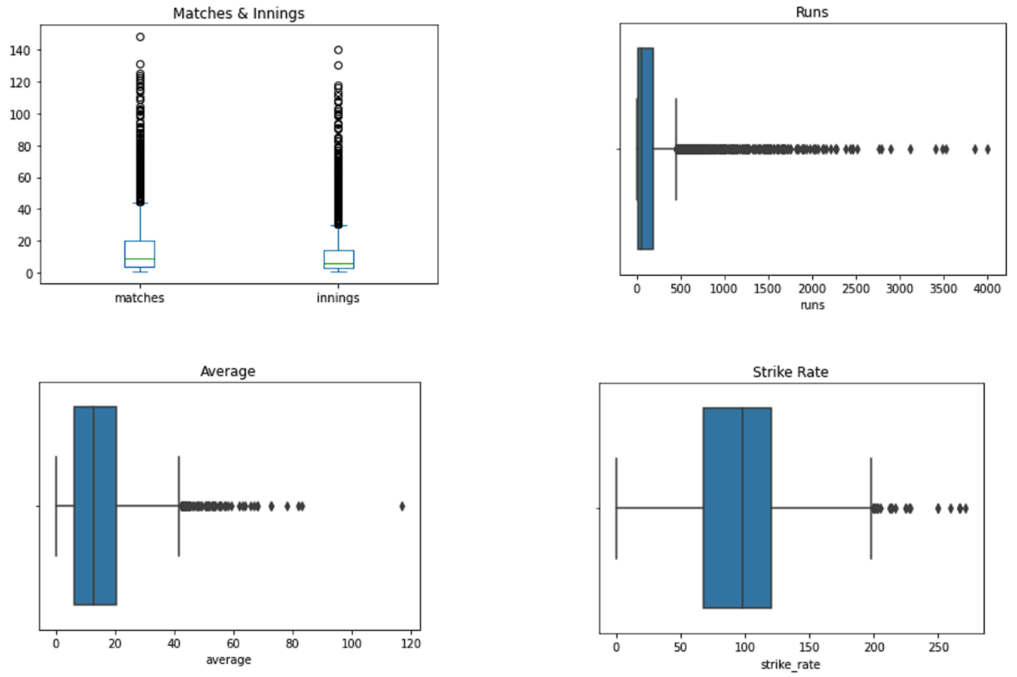


*Figure 2: (a) Matches and Innings, (b) Total Runs, (c) Batting Average and (d) Batting Strike rate*

The International Cricket Council's (ICC) men's T20I team rankings are a global Twenty20 cricket rating system. The two sides involved in each T20I match are awarded points based on a formula following the game developed by David Kendix. All teams are included in a table in order of rating after the sum of each team's points is divided by the total number of matches (ICC Men's T20I Team Rankings - Wikipedia, 2022). As of September 8th, 2022, India is presently leading the ICC Men's T20 international rankings. The top ten countries according to ICC Men's T20 international team rankings are India, Pakistan, England, South Africa, New Zealand, Australia, West Indies, Sri Lanka, Bangladesh, and Afghanistan (ICC Men's T20I Team Rankings | ICC, 2022).

There are only 68 batters who have scored an international century in T20 internationals, and most of the batters who have scored centuries are top-order batsmen. In 117 innings with an average of 32.48 and a strike rate of 139.55, Rohit Sharma (IND) has scored 3,313 runs, including four centuries and 26 half centuries. With 3,299 runs in 108 innings with an average strike rate of 32.66 and 136.71, Martin Guptill (NZ) is not far behind Rohit Sharma in terms of run totals. These two batsmen have one thing in common: they both currently bat as openers, but before 2013 Rohit Sharma began his career as a middle-order batsman. Sami Sohail, a Malawian batsman, has the highest average in our dataset of 78.00 with a strike rate of 114.7, but he has only batted in 11 innings out of 13 matches played so far. Virat Kohli (IND), one of the best T20 international batsmen, has scored 3,296 runs in 89 innings with an average of 51.5 and a strike rate of 137.67. The only batsman to have scored over 3,000 runs with an average above 50.00 and a strike rate above 135.00.

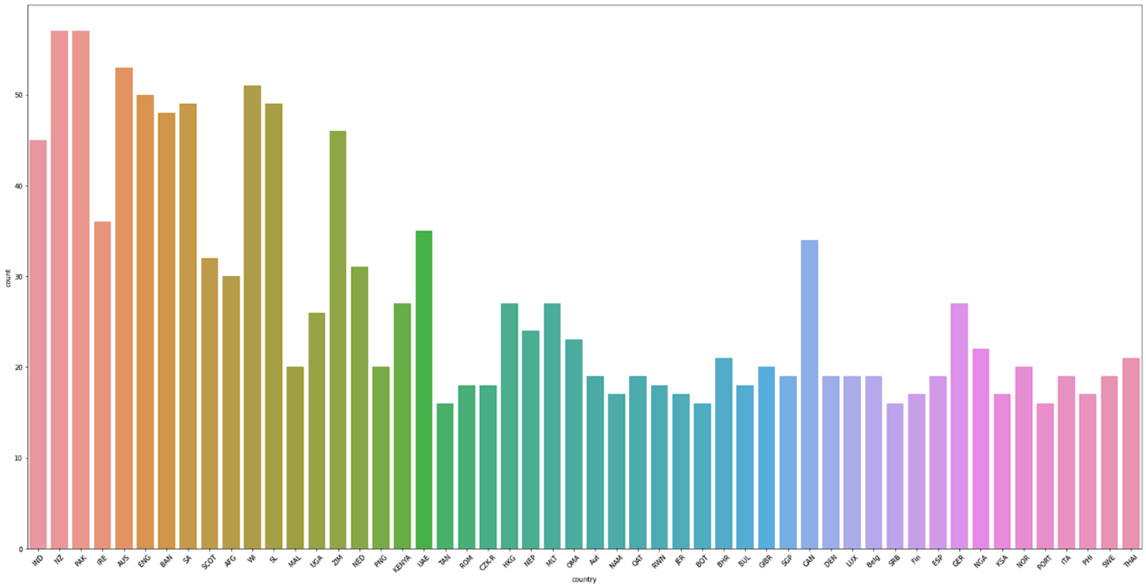


*Figure 3 – Country-wise batter with min 15 players*

Figure 3 displays the number of batters from each nation who had at least 15 batters in a T20 International game. New Zealand has the most batters with 54, followed by Australia with 52 batters. When just the top 10 rated nations in the ICC Men's T20 international team rankings are considered, Afghanistan has the fewest batters, with only 23 batters.

### 3.3.3 Shuffling the Dataset:

This is an important step in this research because it allows you to collect data that can be contrasted with the findings. Dataset had 1400 records for batters and 1167 records for bowlers after preliminary analysis and data cleaning. Before extracting KPIs, we used a formula provided by Basevi et al. (2007) for a new performance indicator to calculate the preliminary metric (PM) for bowlers and batters. Prelim Metric (PM) was calculated for batters as (Average * Strike rate) /100. We used specific criteria to generate Prelim Rank (PR) after calculating the preliminary metrics, such as Best, Good, Average, and Poor.

| Criteria | Ranking |
|---|---|
| PM > 30 and runs >= 500 | Best |
| PM > 30 and runs < 500 | Good |
| PM between 20 and 30 | Good |
| PM between 10 and 20 | Average |
| PM < 10 | Poor |
| Runs < 100 | Poor |

*Table 1: PR Criteria for Batters*

This study shuffled the dataset by sampling and extracting some of the data containing all the Prelim Rank (s). So that, at the end of the process, we can use that data to evaluate our research.

### 3.3.4 Key Performance Indicators for Batter(s):

1. **Boundary per Ball:** We are attempting to calculate the likelihood of a player hitting a boundary on a given ball in this KPI by combining the boundaries and dividing by the total number of balls faced.

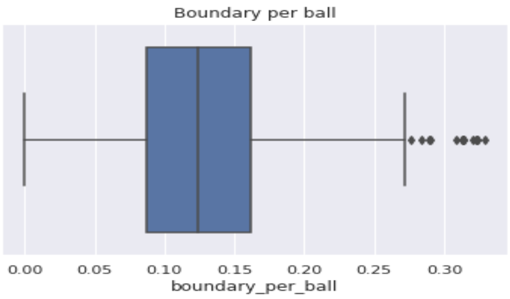


*Figure 4 – Box Plot for Boundary per Ball KPI*

2. **Boundary Index:** In this KPI, we attempted to calculate the average number of boundaries hit by a player during an innings by combining boundaries and then dividing by the total number of innings played.

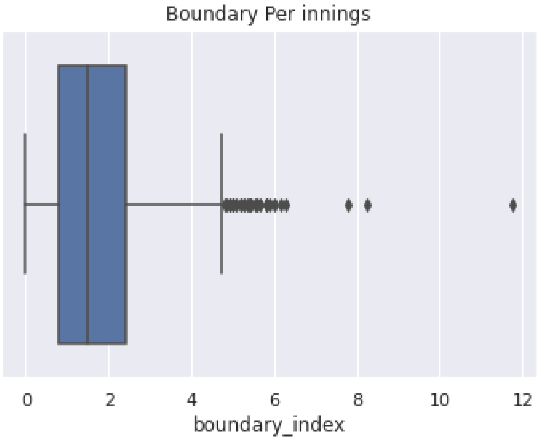


*Figure 5 – Box Plot for Boundary Index KPI*

3. **Finishing Index:** In this KPI, we attempted to calculate how many times a batsman remained not out after coming to bat by dividing the number of times the batsman remained not out by the number of innings.

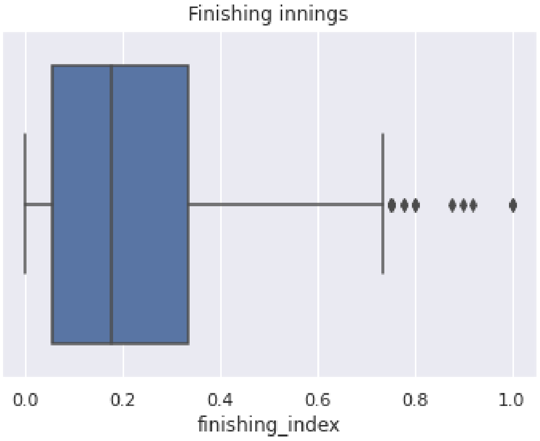


*Figure 6 – Box Plot for Finishing Index KPI*

4. **Big Impact Index:** In this KPI, we attempted to calculate the likelihood of a player scoring a big score in an innings by adding the number of times he has scored century and half-centuries and then dividing by the total number of innings played. In the shortest format of the game, a quick 30 is considered good.

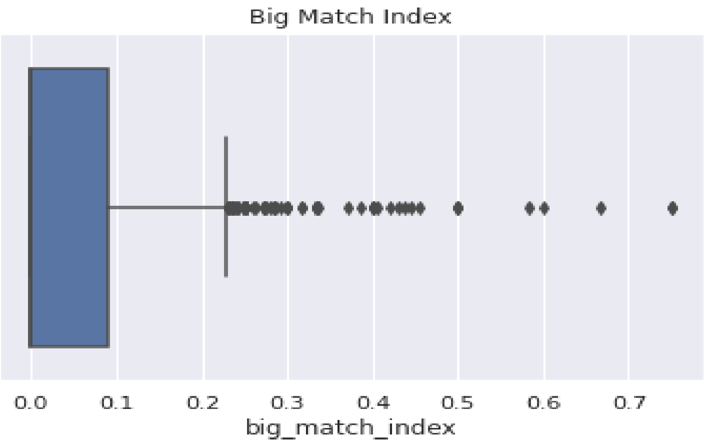
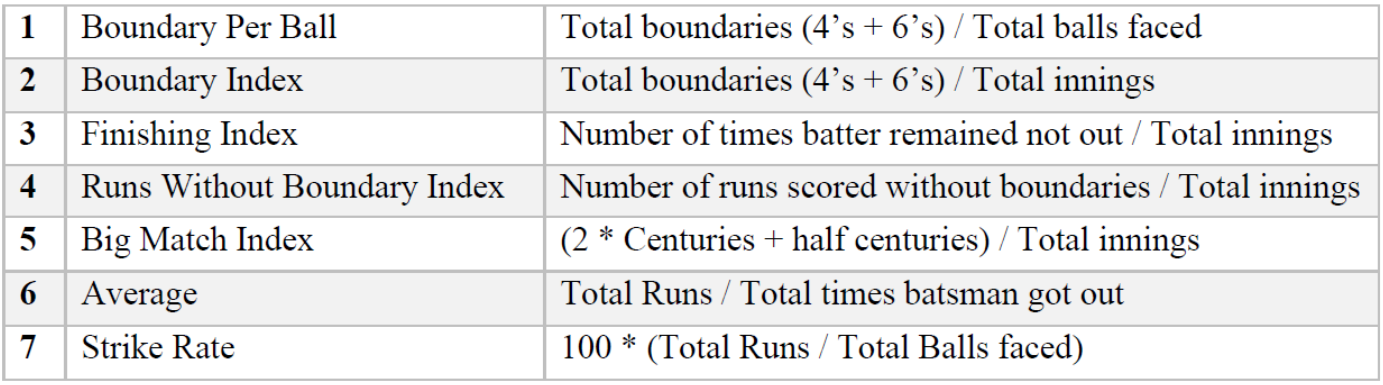


*Figure 7 – Box plot for Big Match Index KPI*



| | | |
|---|---|---|
| **1** | Boundary Per Ball | Total boundaries (4's + 6's) / Total balls faced |
| **2** | Boundary Index | Total boundaries (4's + 6's) / Total innings |
| **3** | Finishing Index | Number of times batter remained not out / Total innings |
| **4** | Runs Without Boundary Index | Number of runs scored without boundaries / Total innings |
| **5** | Big Match Index | (2 * Centuries + half centuries) / Total innings |
| **6** | Average | Total Runs / Total times batsman got out |
| **7** | Strike Rate | 100 * (Total Runs / Total Balls faced) |

*Table 2 – Summary of extracted KPIs for Batters*

### 3.3.5 Feature Scaling:

Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range. It is done as part of the pre-processing of the data to deal with extreme variable magnitudes, values, or units. Machine learning algorithms look at numbers; if there is a significant difference in the range then the algorithm assumes that higher-ranged numbers are somehow superior (All about Feature Scaling | Towards Data Science, 2020). The most common techniques of feature scaling are Normalisation and Standardisation. In this study, you can see that most of the extracted KPIs have smaller values, but the traditional metrics such as strike rate, economy, and average have much higher values as compared to KPIs we have defined. There cannot have a meaningful comparison because the machine learning algorithm computes distance or assumes normality such as K-means clustering, K nearest neighbours, and Principal component analysis.

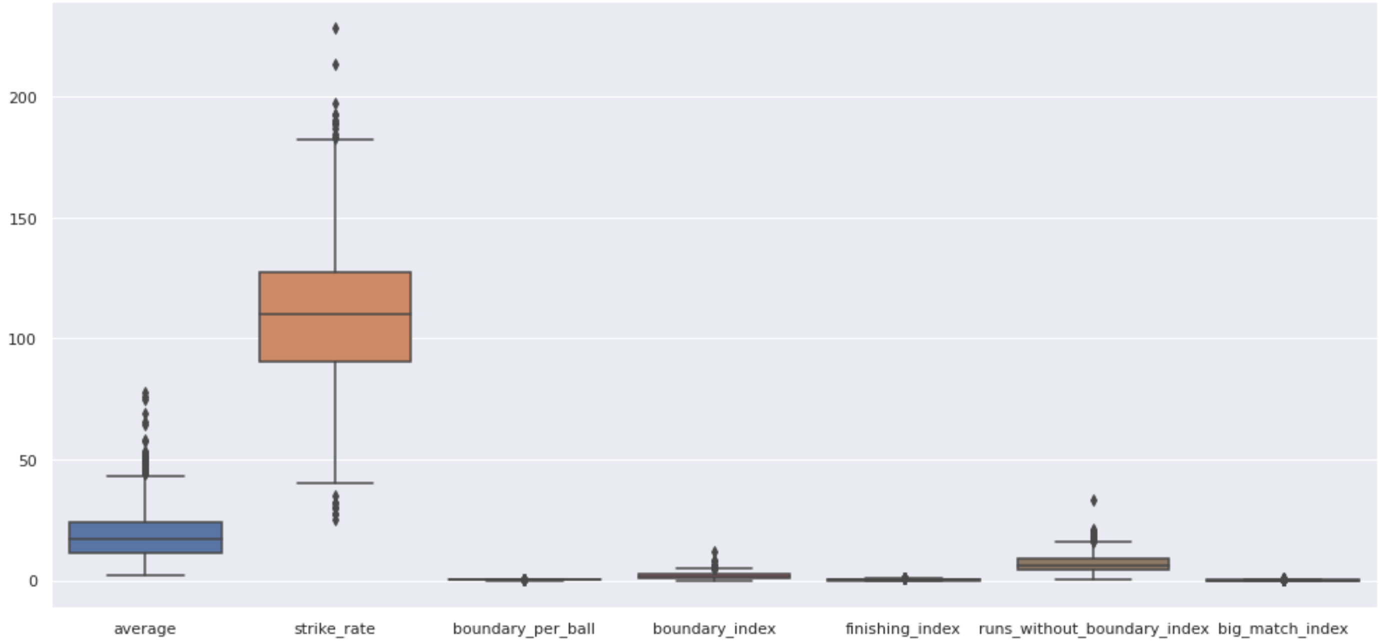


*Figure 8 – Box Plot for combined Batters KPIs*

- **Standardisation:** It is essential to rescale a feature value by standardisation so that its distribution has a mean of 0 and a variance of 1. During standardisation, the data points are rescaled by ensuring that they will adopt a curve-like shape after scaling. We can mathematically represent it as follows,

$$X_{stand} = \frac{X_i - \mu}{\sigma}$$

*Figure 9 – Standardisation Formula*

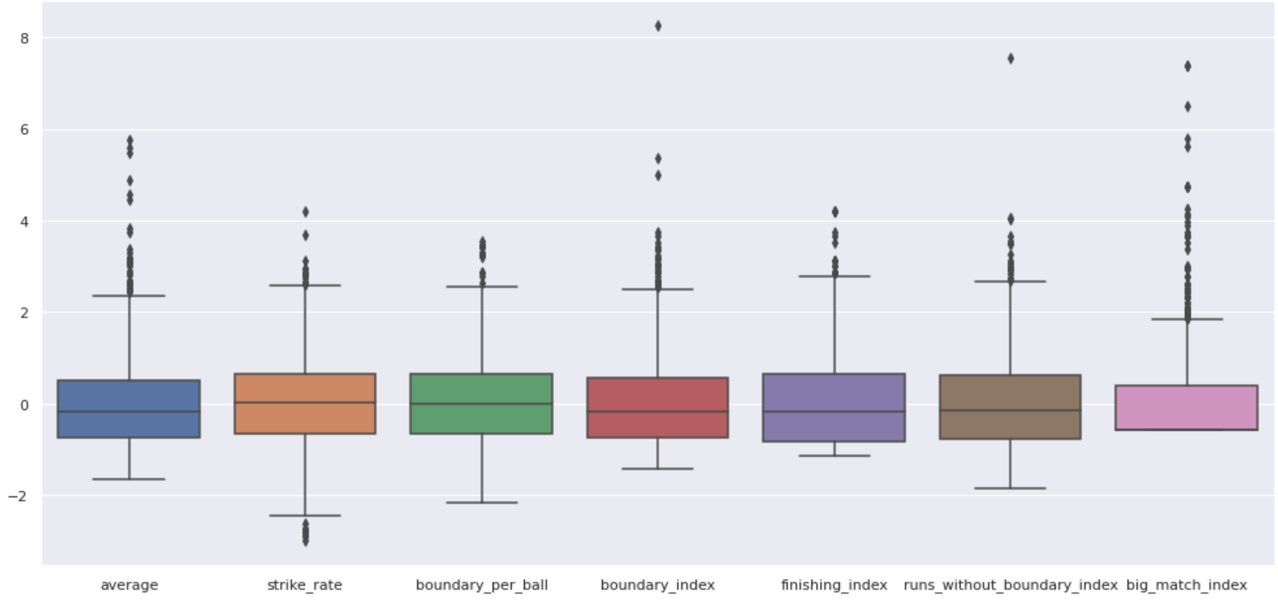


*Figure 10 - Box Plot for the combined Batters KPIs after Standardisation*

- **Min Max Scaling:** Min-Max Scaling simply includes transforming all values measured on many scales into one common scale. However, scaling can signify many other things. Under this process, the difference between any value and the minimum value is divided by the difference between the maximum and minimum values (All about Feature Scaling | Towards Data Science, 2020).

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

*Figure 11 – For Min-Max Scaling*

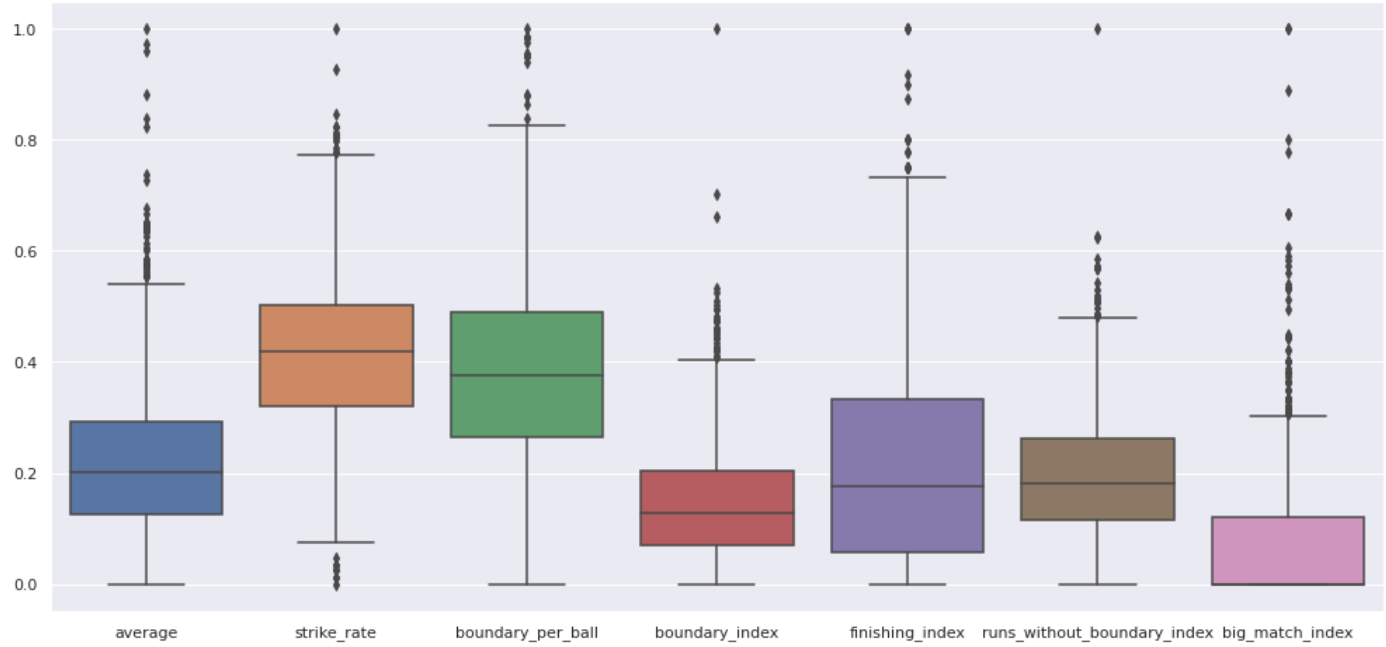


*Figure 12 - Box Plot for the combined Batters KPIs after Min-Max Scaling*

### 3.3.6 K-Means Clustering:

One of the most straightforward and well-liked unsupervised machine learning algorithms is K-means clustering. Unsupervised algorithms typically conclude datasets using only input vectors without considering predetermined or labelled results. A cluster is a group of data points that have been combined due to commonalities (Guido and Müller, 2016). This study used K-means clustering to obtain the target vector and group players who were playing the same role. This study used the elbow approach and the silhouette score of the algorithm to compute the number of clusters (K) based on different cluster counts. The "Elbow Method," a heuristic used to determine the number of clusters in a data collection, selects the elbow of the curve as the ideal value of n (number of clusters) present in the Dataset (Humaira et al., 2020).

### 3.3.6.1 Results of K-Means Clustering for Batters:

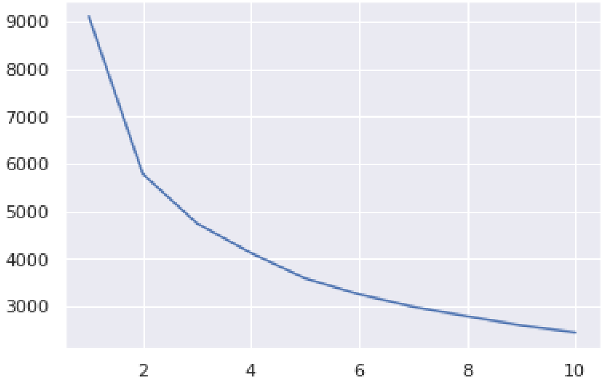
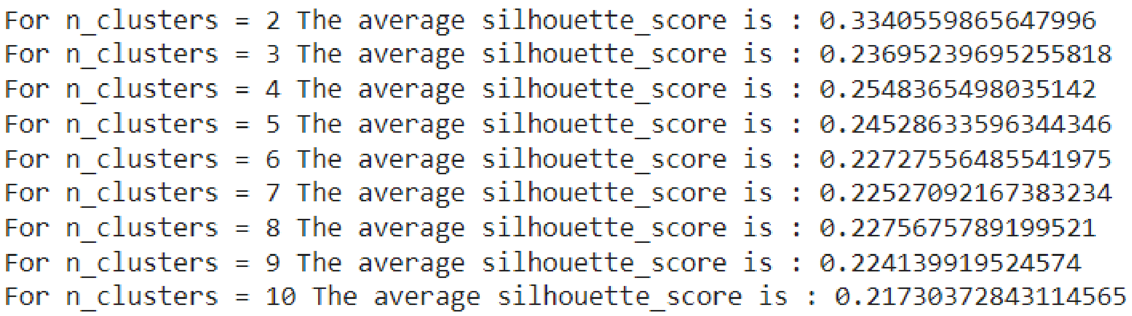


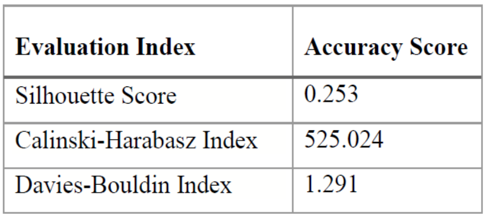
*Figure 13 - Results of elbow method for batters*



```
For n_clusters = 2 The average silhouette_score is : 0.3340559865647996
For n_clusters = 3 The average silhouette_score is : 0.23695239695255818
For n_clusters = 4 The average silhouette_score is : 0.2548365498035142
For n_clusters = 5 The average silhouette_score is : 0.24528633596344346
For n_clusters = 6 The average silhouette_score is : 0.22727556485541975
For n_clusters = 7 The average silhouette_score is : 0.22527092167383234
For n_clusters = 8 The average silhouette_score is : 0.2275675789199521
For n_clusters = 9 The average silhouette_score is : 0.224139919524574
For n_clusters = 10 The average silhouette_score is : 0.21730372843114565
```

*Figure 14 - Average silhouette score value based on the number of clusters for batters*

After looking at the results from the elbow method and average silhouette score values for batters from figures 13 and 14 respectively. We can say that there are four clusters in the dataset, and we went ahead to identify the four clusters in the dataset.



| **Evaluation Index** | **Accuracy Score** |
|---|---|
| Silhouette Score | 0.253 |
| Calinski-Harabasz Index | 525.024 |
| Davies-Bouldin Index | 1.291 |

*Table 3 - Accuracy score of various Evaluation Index for Batters*

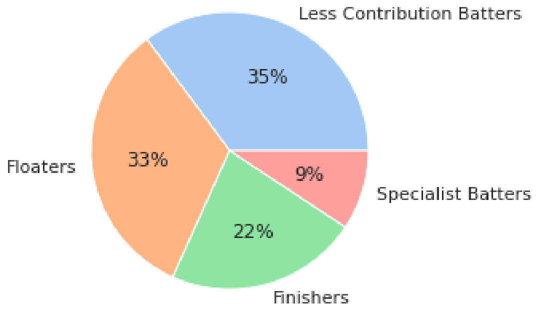
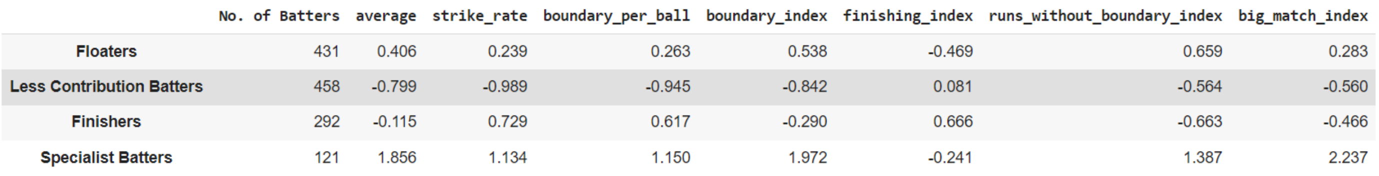


*Figure 15 - Pie chart of Clustering results for Batters with roles*

Figure 15 depicts the distribution of participants according to their roles, which were determined by the values of each cluster centre, while Table 7 lists the accuracy scores for several evaluation indices. Figure 16 displays the value of each centre of the cluster with the number of players in that cluster. Table 4 displays some players according to roles assigned by the clusters.



| | No. of Batters | average | strike_rate | boundary_per_ball | boundary_index | finishing_index | runs_without_boundary_index | big_match_index |
|---|---|---|---|---|---|---|---|---|
| **Floaters** | 431 | 0.406 | 0.239 | 0.263 | 0.538 | -0.469 | 0.659 | 0.283 |
| **Less Contribution Batters** | 458 | -0.799 | -0.989 | -0.945 | -0.842 | 0.081 | -0.564 | -0.560 |
| **Finishers** | 292 | -0.115 | 0.729 | 0.617 | -0.290 | 0.666 | -0.663 | -0.466 |
| **Specialist Batters** | 121 | 1.856 | 1.134 | 1.150 | 1.972 | -0.241 | 1.387 | 2.237 |

*Figure 16 – Cluster Centre for Batters*

| | |
|---|---|
| Specialist Batters | Rohit Sharma (IND), Martin Guptill (NZ), Virat Kohli (IND), Babar Azam (PAK), Jos Buttler (ENG) |
| Finishers | S Afridi (PAK), A Russell (WI), M Ali (ENG), D Sammy (WI), H Pandya (IND) |
| Floaters | E Morgan (ENG), Shoaib Malik (PAK), Ross Taylor (NZ), Shakib-Al-Hasan (BAN), D Miller (SA) |
| Less Contribution Batters | Nathan McCullum (NZ), D Vettori (NZ), Mark Boucher (SA), James Faulkner (AUS), GD Elliot (NZ) |

*Table 4 - Some Batters names based on roles assigned by the K-Means algorithm*

The name of each cluster was chosen by the cluster centre, and Specialist Batters (SB) refers to batters who, although having a poor finishing index, outperform everyone else in all other KPIs. Finishers (F) exhibit the ability to score runs fast and remain not out at the end of an inning by having a high Finishing Index value and a high strike rate. Floaters (FL) are batsmen who are skilled and have a high average, good strike rate, and runs without boundary index, as well as a high big impact innings index. Less Contribution Batters (LCB) are batters who do not make a significant contribution to the bat since they do not always receive the opportunity to bat. One can deduct from table 8 that players are properly categorised by their ability to bat.

### 3.3.7 Classification:

The first technique used in this study was the unsupervised learning method K-Means Clustering, which does not rely on a predefined set of classes and clusters players together. After determining the roles of each batter and bowler, this study used classification algorithms such as XGBoost, Random Forest, and Voting Classifier to determine the importance of each feature for each role. Classification constructs a model from a set of labelled players and predicts unclassified players into several role-based groupings. The combination of supervised and unsupervised learning allows for precise determination of the proportional value of various features for each role. To categorise the data and determine the algorithm's accuracy, we divided data into 70% training data and 30% testing data.

XGBoost is an adaption of the Gradient boosting algorithm (Beginner’s Guide to XGBoost for Classification Problems | Medium, 2021). A voting classifier is a machine learning estimator that trains multiple base models or estimators and predicts by aggregating their results. The aggregating criteria can be a combined voting decision for each estimator output (Géron, 2020). There are two types of voting classifiers: Hard Voting and Soft Voting. In Hard Voting, Voting is calculated on the predicted output class. In Soft Voting, Voting is calculated on the predicted probability of output class. In this study, we have used three diverse predictors in Voting Classifier 1. Decision Tree, 2. Random Forest and 3. XGBoost with equal weightage.

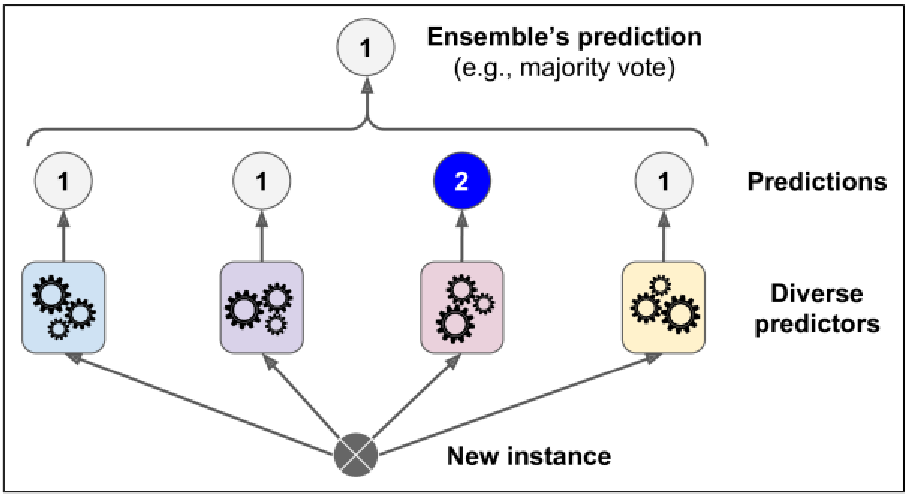


*Figure 17 - Hard Voting Classifier (Géron, 2020)*

#### 3.3.7.1 Results for Batters:

The Balanced Accuracy score, F1 score, recall score, and precision score for each of the classification algorithms used for batters are shown in table 5. More about evaluation metrics can be found at scikit-learn Classification Metrics and Evaluation (2022). Figure 18 and Table 6, show that the clustering by the K-means algorithm did an excellent job of identifying clusters in the data for the batter.

| Metrics | Random Forest | XGBoost | Voting Classifier |
|---|---|---|---|
| Balanced Accuracy | 0.951 | 0.950 | 0.946 |
| F1 Score (Micro) | 0.957 | 0.959 | 0.954 |
| Recall Score | 0.952 | 0.951 | 0.947 |
| Precision Score | 0.948 | 0.947 | 0.944 |

*Table 5 - Classification scores from various evaluation metrics for Batters*

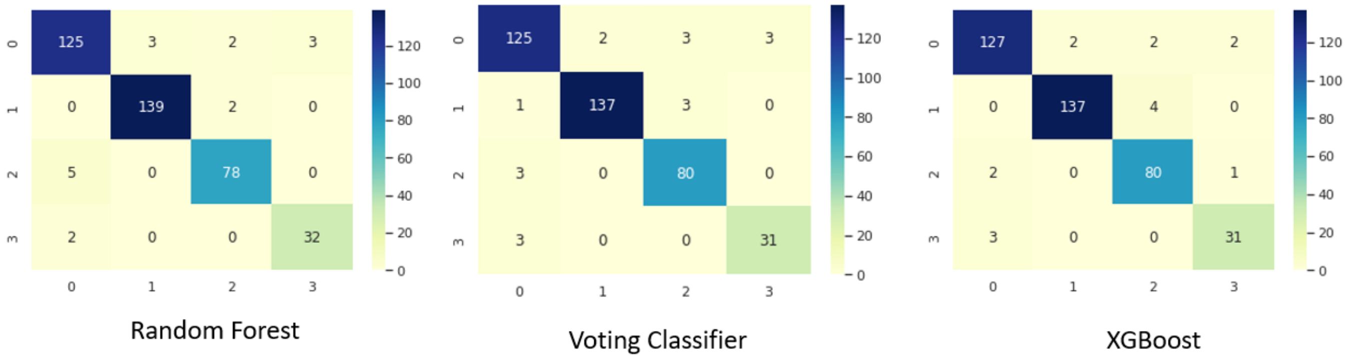
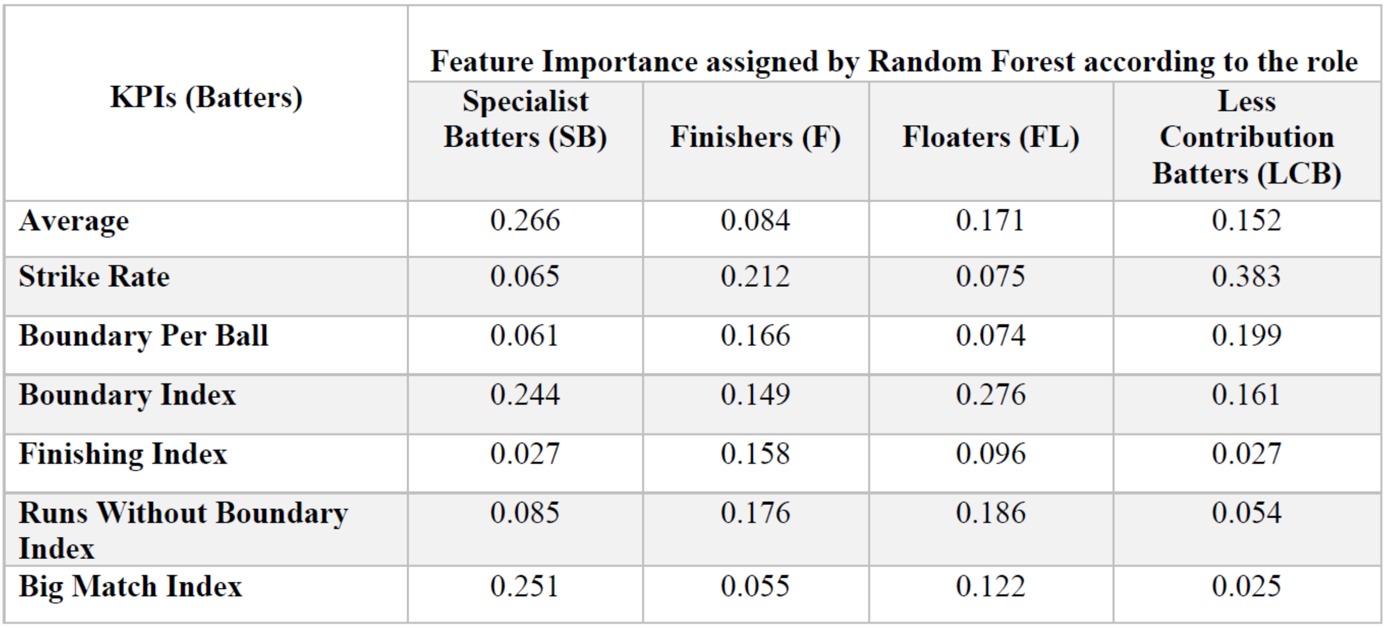


*Figure 18 - Confusion matrix of each algorithm used (Batters)*

### 3.3.8 One vs All Classification:

Instead of using the classifiers and feature significance values as generalisations for the entire dataset in the multiclass label. For each group of players, this study utilised one vs all classification to determine different feature importance for each role or each cluster of the players. The target vector was modified for each group depending on clustering results, and after that entire dataset was combined. This study uses a binary variable (1/0 in the target vector) to indicate whether a player is a part of that cluster-based role for bowling or batting, and learn the feature importance for each of those roles. For instance, players in group 0 will have a target of 1, while players in all other groups will have a target of 0 when we are evaluating the relevance of a feature for them. As we've seen, Random Forest produced the best results for this dataset among the classification algorithms used. As a result, this part of this study uses Random Forest in this classifier because it typically produces better predictive results than other methods. To create a new role-based performance indicator for each batter and bowler, this study multiplied the weighted importance assigned to each feature by the KPIs based on the roles.

#### 3.3.8.1 Results for Batters:

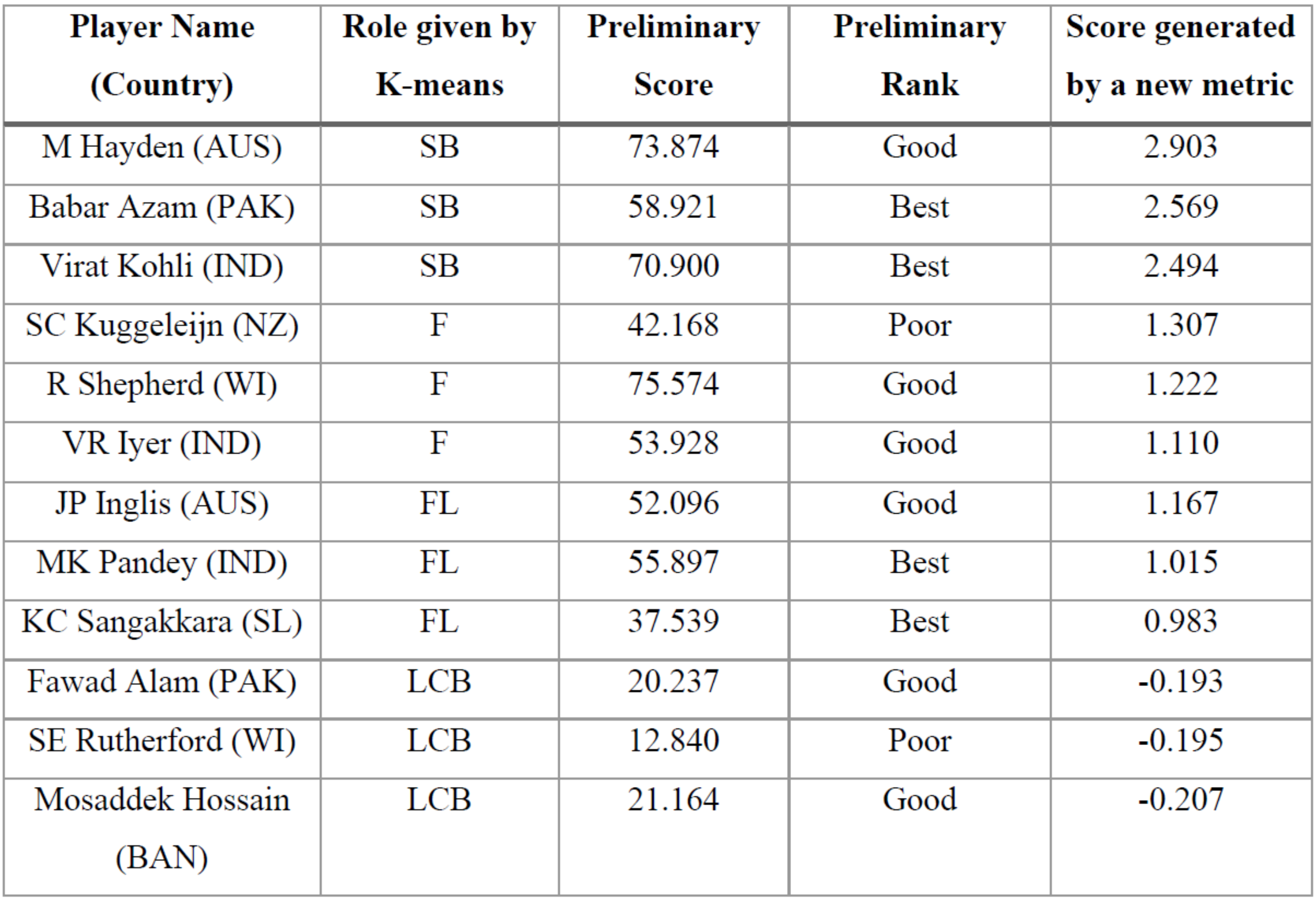


| KPIs (Batters) | Feature Importance assigned by Random Forest according to the role | | | |
|---|---|---|---|---|
| | Specialist Batters (SB) | Finishers (F) | Floaters (FL) | Less Contribution Batters (LCB) |
| Average | 0.266 | 0.084 | 0.171 | 0.152 |
| Strike Rate | 0.065 | 0.212 | 0.075 | 0.383 |
| Boundary Per Ball | 0.061 | 0.166 | 0.074 | 0.199 |
| Boundary Index | 0.244 | 0.149 | 0.276 | 0.161 |
| Finishing Index | 0.027 | 0.158 | 0.096 | 0.027 |
| Runs Without Boundary Index | 0.085 | 0.176 | 0.186 | 0.054 |
| Big Match Index | 0.251 | 0.055 | 0.122 | 0.025 |

*Table 6 - Role-based feature importance by Random Forest (Batters)*

Table 6 displays the importance of each KPI according to its role. For Specialist Batters (SB), Average, Big Match Index, and Boundary Index are given more importance. The Finishing Index, Boundary Per Ball, and Strike Rate are given the most importance for Finishers (F), who must score quickly toward the end of an inning. For Floaters (FL), Average, Big Match Index, Runs without boundary Index, and Big Match Index have all been given higher importance; however, they are all less significant than the significance of the KPIs given to Specialist Batter (SB). Less Contribution Batters (LCB) are expected to end the innings strongly by hitting a few boundaries whenever they get a chance to bat.

Table 7 provides insights of the score generated by Role-based Machine Learning (RBML) metrics; shows the top 3 batters from each role from the Top 10 teams in the ICC Men's T20 international team rankings. If we look at the Specialist Batter most are the batter are batsman who plays in the top order. All the batters were considered to have PM score above 50 but only ML Hayden (AUS) has played a few innings and scored less than 500 runs that is why he has a PR rank ‘Good’ and has the RBML score of 2.903 highest among all the batters from the Top 10 teams in the ICC Men's T20 international team rankings. This also states that RBML does not take how many runs scored by a batsman into account directly. This study has diversified total runs into several KPIs so that our method rewards batter based on skills required in the shortest format of the game. The same goes for SC Kuggeleijn (NZ), He might have not scored many runs but whenever he got the chance to finish, he has done well and can be categorized as a finisher. MK Pandey (IND) and KC Sangakkara (SL) are categorised as “Best” by PR rank and that is correct, but they have not played as top-order batsmen throughout their careers. They have played at many positions in the shortest format of the game. That is one of the reasons why they are categorised as Floaters. Players in LCB are mostly bowlers who did not get much chance to bat but they try to contribute whenever they get a chance.



| Player Name (Country) | Role given by K-means | Preliminary Score | Preliminary Rank | Score generated by a new metric |
|---|---|---|---|---|
| M Hayden (AUS) | SB | 73.874 | Good | 2.903 |
| Babar Azam (PAK) | SB | 58.921 | Best | 2.569 |
| Virat Kohli (IND) | SB | 70.900 | Best | 2.494 |
| SC Kuggeleijn (NZ) | F | 42.168 | Poor | 1.307 |
| R Shepherd (WI) | F | 75.574 | Good | 1.222 |
| VR Iyer (IND) | F | 53.928 | Good | 1.110 |
| JP Inglis (AUS) | FL | 52.096 | Good | 1.167 |
| MK Pandey (IND) | FL | 55.897 | Best | 1.015 |
| KC Sangakkara (SL) | FL | 37.539 | Best | 0.983 |
| Fawad Alam (PAK) | LCB | 20.237 | Good | -0.193 |
| SE Rutherford (WI) | LCB | 12.840 | Poor | -0.195 |
| Mosaddek Hossain (BAN) | LCB | 21.164 | Good | -0.207 |

*Table 7 – Top 3 Ranks of each role generated by our Role-based ML Score*

### 3.3.9 Principal Component Analysis:

Principal Component Analysis (PCA) is a popular technique for dimensionality reduction while keeping most of the crucial data. It operates by calculating the basic factors and changing the basis. The data in the direction of the greatest variance is retained. There is no correlation between the reduced features (Dutt, 2021). PCA's main objective is to condense your model's features into a smaller number of elements to aid in visualising patterns in your data (Lindgren, 2020). This has assisted us in reducing the dimensions of the data, which will ultimately assist us in achieving the goal of this research with another approach. The first component's coefficient values are multiplied by the KPIs to create a new Performance indicator. Despite not being role-based, the performance indicator obtained by PCA allows us to determine whether our experiments of supervised and unsupervised learning were successful.

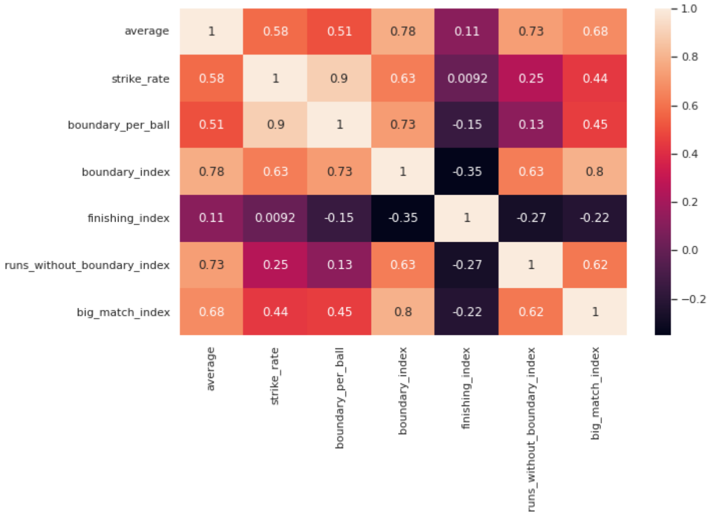


*Figure 19 - Correlation matrix for Batters data after standardization*

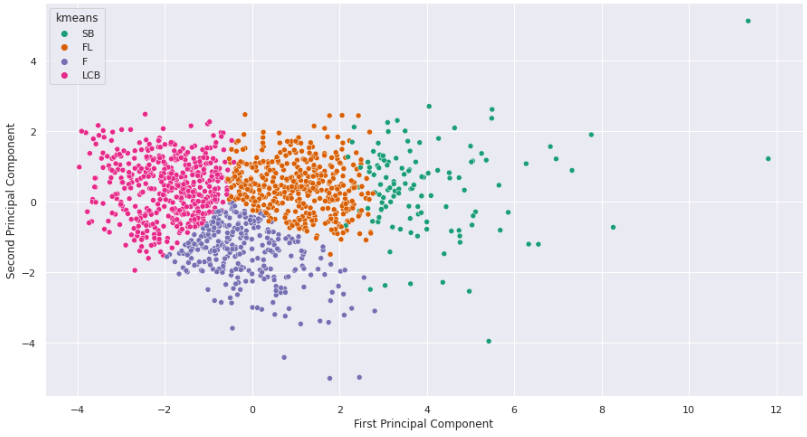


*Figure 20 - Scatter plot of PCA with points separated by role-based cluster (Batter)*

The correlation between the features in the batsman data is shown in figure 19. Except for Finishing Index, every feature correlates with average because scoring runs is one of the key objectives in the game's shortest format. The more boundaries a player hits, the better strike rate he is going to achieve that is why the strike rate is strongly correlated with the number of boundaries per ball. The more boundaries a player hits, the greater their chances are of receiving a larger score, and this is true for both the Boundary Index and the Big Match Index. A batter's average is likely to be high based on the number of times he stays on base. That explains why the finishing index has a weak correlation with each feature, with the average showing the finishing index's strongest correlation.

In this instance, PCA was used to reduce the number of dimensions in the data to only two. The scatter plot in figure 20 includes both PCA components, and the points are divided into groups according to the roles that the clustering assigned them. The higher the value of the point on the First component, the better the batter, according to the scatter plot and coefficients of each component in figure 21. We can see all the clusters in the plot, which also shows that K-Means did a good job of grouping the points. The values of Variance Percentages (Total Variability percentage) for both the components are 57.50% and 18.51%.

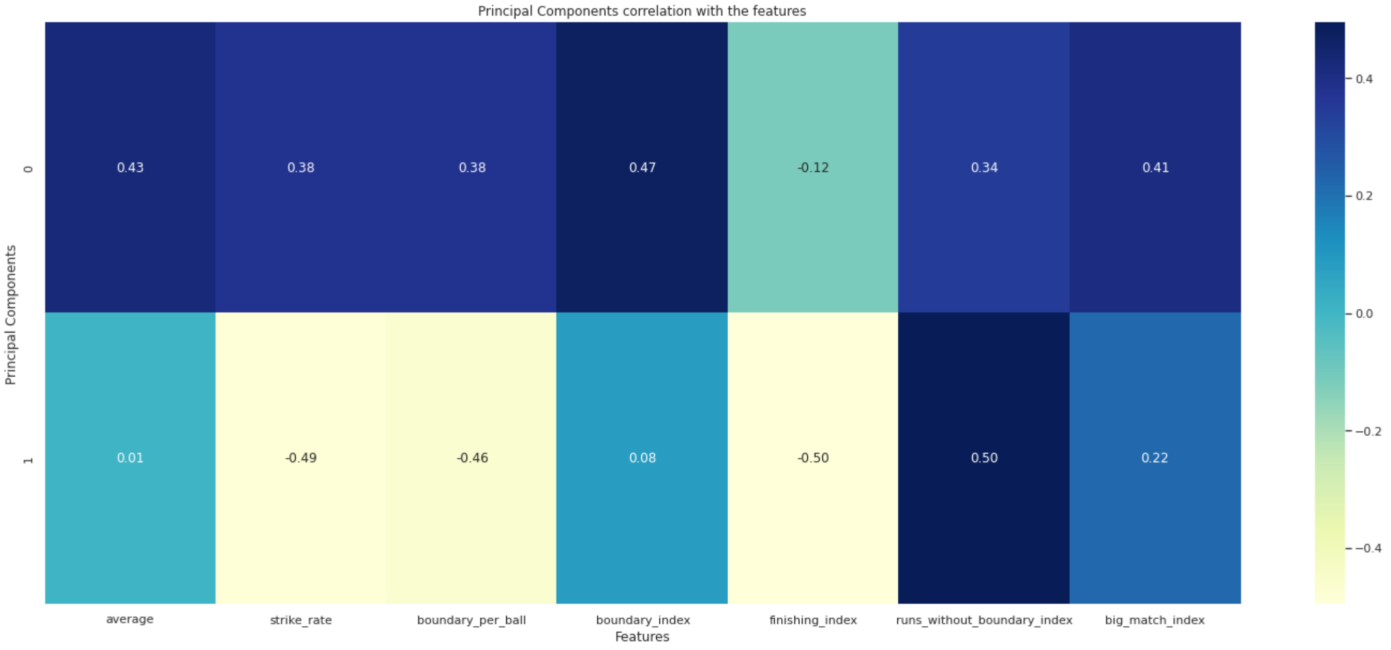
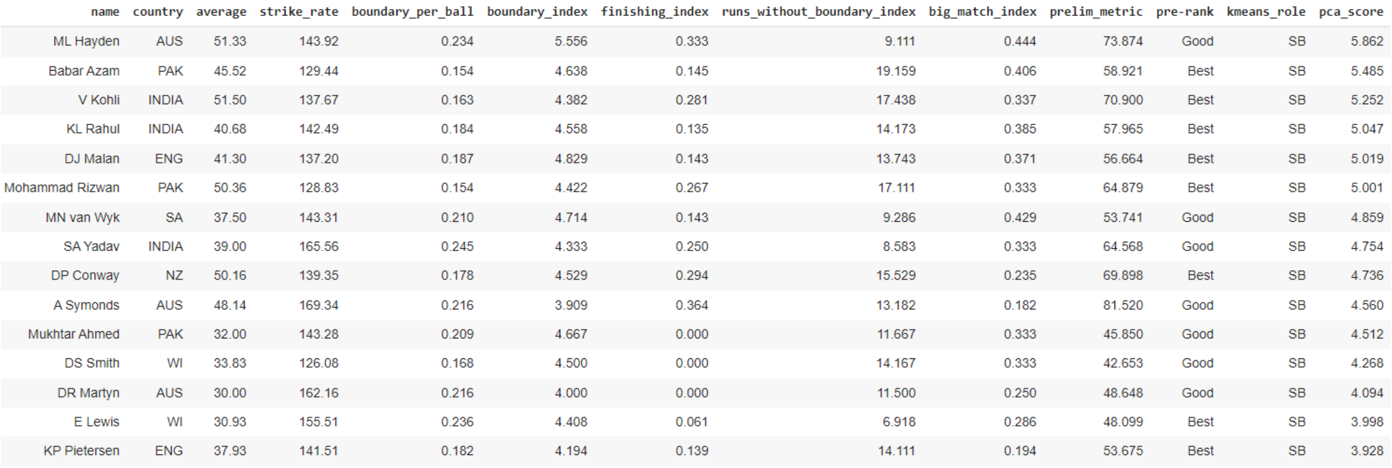


*Figure 21 - Each feature's coefficients for both PCA components (Batter)*



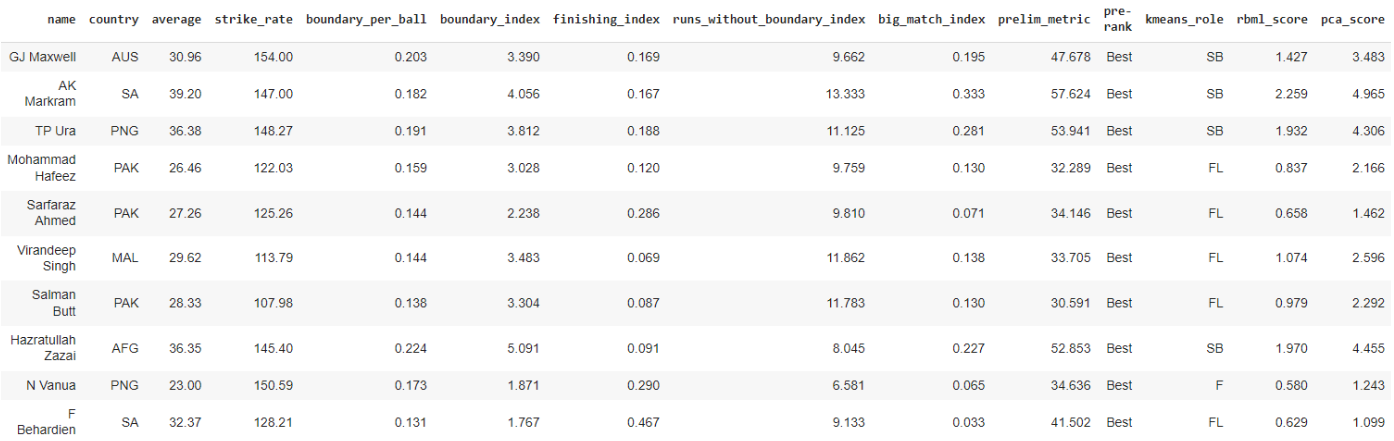
| name | country | average | strike_rate | boundary_per_ball | boundary_index | finishing_index | runs_without_boundary_index | big_match_index | prelim_metric | pre-rank | kmeans_role | pca_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ML Hayden | AUS | 51.33 | 143.92 | 0.234 | 5.556 | 0.333 | 9.111 | 0.444 | 73.874 | Good | SB | 5.862 |
| Babar Azam | PAK | 45.52 | 129.44 | 0.154 | 4.638 | 0.145 | 19.159 | 0.406 | 58.921 | Best | SB | 5.485 |
| V Kohli | INDIA | 51.50 | 137.67 | 0.163 | 4.382 | 0.281 | 17.438 | 0.337 | 70.900 | Best | SB | 5.252 |
| KL Rahul | INDIA | 40.68 | 142.49 | 0.184 | 4.558 | 0.135 | 14.173 | 0.385 | 57.965 | Best | SB | 5.047 |
| DJ Malan | ENG | 41.30 | 137.20 | 0.187 | 4.829 | 0.143 | 13.743 | 0.371 | 56.664 | Best | SB | 5.019 |
| Mohammad Rizwan | PAK | 50.36 | 128.83 | 0.154 | 4.422 | 0.267 | 17.111 | 0.333 | 64.879 | Best | SB | 5.001 |
| MN van Wyk | SA | 37.50 | 143.31 | 0.210 | 4.714 | 0.143 | 9.286 | 0.429 | 53.741 | Good | SB | 4.859 |
| SA Yadav | INDIA | 39.00 | 165.56 | 0.245 | 4.333 | 0.250 | 8.583 | 0.333 | 64.568 | Good | SB | 4.754 |
| DP Conway | NZ | 50.16 | 139.35 | 0.178 | 4.529 | 0.294 | 15.529 | 0.235 | 69.898 | Best | SB | 4.736 |
| A Symonds | AUS | 48.14 | 169.34 | 0.216 | 3.909 | 0.364 | 13.182 | 0.182 | 81.520 | Good | SB | 4.560 |
| Mukhtar Ahmed | PAK | 32.00 | 143.28 | 0.209 | 4.667 | 0.000 | 11.667 | 0.333 | 45.850 | Good | SB | 4.512 |
| DS Smith | WI | 33.83 | 126.08 | 0.168 | 4.500 | 0.000 | 14.167 | 0.333 | 42.653 | Good | SB | 4.268 |
| DR Martyn | AUS | 30.00 | 162.16 | 0.216 | 4.000 | 0.000 | 11.500 | 0.250 | 48.648 | Good | SB | 4.094 |
| E Lewis | WI | 30.93 | 155.51 | 0.236 | 4.408 | 0.061 | 6.918 | 0.286 | 48.099 | Best | SB | 3.998 |
| KP Pietersen | ENG | 37.93 | 141.51 | 0.182 | 4.194 | 0.139 | 14.111 | 0.194 | 53.675 | Best | SB | 3.928 |

*Figure 22 – Top 15 Batters according to PCA Score*

## 3.4 Results & Discussion:

### 3.4.1 Results of Batters:

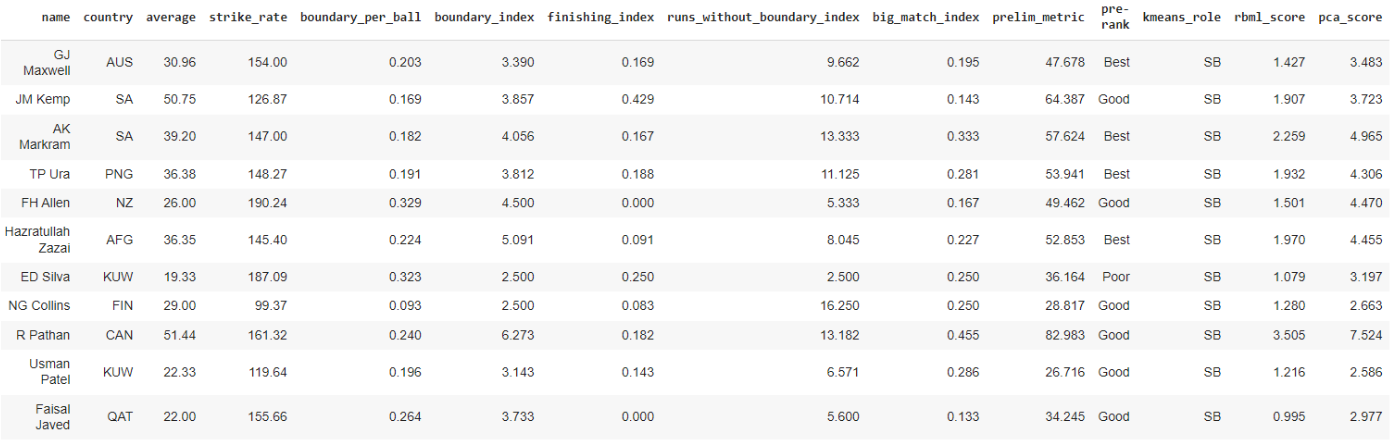
In this section, the whole approach and the results achieved by the experiments performed above are discussed. One of this study's methodologies can be summarized as the unsupervised learning algorithm K-Means Clustering, which does not need a labelled set of classes to group players together, which is the first technique we used. After that, this study used the supervised classification algorithm XGBoost, Random Forest, and Voting Classifier which builds a model from a set of labelled players and predicts the unclassified players into various role-based clusters. We can precisely determine the relative importance of various features for each role thanks to the combination of supervised and unsupervised learning. Our second methodology involved performing PCA on the KPIs we had collected, obtaining generalised metrics for batters and bowlers, and determining whether the findings of K-Means clustering could be understood by PCA. In this section, we will analyse the data we have separated in shuffling the dataset section in chapter 3 and keeping it aside for testing purposes only.



| name | country | average | strike_rate | boundary_per_ball | boundary_index | finishing_index | runs_without_boundary_index | big_match_index | prelim_metric | pre-rank | kmeans_role | rbml_score | pca_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GJ Maxwell | AUS | 30.96 | 154.00 | 0.203 | 3.390 | 0.169 | 9.662 | 0.195 | 47.678 | Best | SB | 1.427 | 3.483 |
| AK Markram | SA | 39.20 | 147.00 | 0.182 | 4.056 | 0.167 | 13.333 | 0.333 | 57.624 | Best | SB | 2.259 | 4.965 |
| TP Ura | PNG | 36.38 | 148.27 | 0.191 | 3.812 | 0.188 | 11.125 | 0.281 | 53.941 | Best | SB | 1.932 | 4.306 |
| Mohammad Hafeez | PAK | 26.46 | 122.03 | 0.159 | 3.028 | 0.120 | 9.759 | 0.130 | 32.289 | Best | FL | 0.837 | 2.166 |
| Sarfaraz Ahmed | PAK | 27.26 | 125.26 | 0.144 | 2.238 | 0.286 | 9.810 | 0.071 | 34.146 | Best | FL | 0.658 | 1.462 |
| Virandeep Singh | MAL | 29.62 | 113.79 | 0.144 | 3.483 | 0.069 | 11.862 | 0.138 | 33.705 | Best | FL | 1.074 | 2.596 |
| Salman Butt | PAK | 28.33 | 107.98 | 0.138 | 3.304 | 0.087 | 11.783 | 0.130 | 30.591 | Best | FL | 0.979 | 2.292 |
| Hazratullah Zazai | AFG | 36.35 | 145.40 | 0.224 | 5.091 | 0.091 | 8.045 | 0.227 | 52.853 | Best | SB | 1.970 | 4.455 |
| N Vanua | PNG | 23.00 | 150.59 | 0.173 | 1.871 | 0.290 | 6.581 | 0.065 | 34.636 | Best | F | 0.580 | 1.243 |
| F Behardien | SA | 32.37 | 128.21 | 0.131 | 1.767 | 0.467 | 9.133 | 0.033 | 41.502 | Best | FL | 0.629 | 1.099 |

*Figure 23 - Role given by clustering to Batsmen's Classified as Best by PR rank*

In the beginning, this study kept 98 records for the batters separately containing all the PR ranks. In those 98 records, we had 50 batters were classified as Poor, 29 classified as Good, ten classified as Best, and nine classified as Average batters according to PR rank. After putting those 98 batters through our algorithm, we discovered that, by the roles suggested by clustering, 39 batters were Low Contribution Batters (LCB), 29 were Floaters (FL), 19 were Finishers (F), and 11 were Specialist Batters (SB). Seven batsmen were categorised as LCB by clustering but were categorised as Average by PR rank. In figure 23, you can see the ten batsmen given the PR rank of 'Best', but only four of them are labelled as SB, five as FL, and one as Finisher by our model. Among the batters in that dataset classified as ‘Best’, AK Markram (SA) received the highest RBML score of 2.259 and the highest PCA score of 4.965. While R Pathan (CAN) has the highest RBML score of 3.505 and PCA score of 7.524 is categorised as Specialist Batter (SB) by model but he has been categorised as ‘Good’ batsmen by PR rank due to less than 500 runs in T20 internationals.



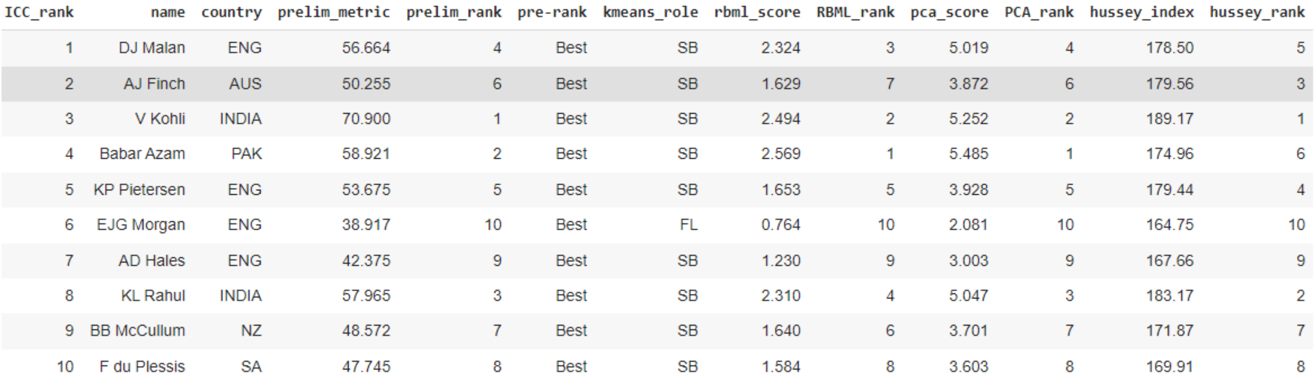
| name | country | average | strike_rate | boundary_per_ball | boundary_index | finishing_index | runs_without_boundary_index | big_match_index | prelim_metric | pre-rank | kmeans_role | rbml_score | pca_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GJ Maxwell | AUS | 30.96 | 154.00 | 0.203 | 3.390 | 0.169 | 9.662 | 0.195 | 47.678 | Best | SB | 1.427 | 3.483 |
| JM Kemp | SA | 50.75 | 126.87 | 0.169 | 3.857 | 0.429 | 10.714 | 0.143 | 64.387 | Good | SB | 1.907 | 3.723 |
| AK Markram | SA | 39.20 | 147.00 | 0.182 | 4.056 | 0.167 | 13.333 | 0.333 | 57.624 | Best | SB | 2.259 | 4.965 |
| TP Ura | PNG | 36.38 | 148.27 | 0.191 | 3.812 | 0.188 | 11.125 | 0.281 | 53.941 | Best | SB | 1.932 | 4.306 |
| FH Allen | NZ | 26.00 | 190.24 | 0.329 | 4.500 | 0.000 | 5.333 | 0.167 | 49.462 | Good | SB | 1.501 | 4.470 |
| Hazratullah Zazai | AFG | 36.35 | 145.40 | 0.224 | 5.091 | 0.091 | 8.045 | 0.227 | 52.853 | Best | SB | 1.970 | 4.455 |
| ED Silva | KUW | 19.33 | 187.09 | 0.323 | 2.500 | 0.250 | 2.500 | 0.250 | 36.164 | Poor | SB | 1.079 | 3.197 |
| NG Collins | FIN | 29.00 | 99.37 | 0.093 | 2.500 | 0.083 | 16.250 | 0.250 | 28.817 | Good | SB | 1.280 | 2.663 |
| R Pathan | CAN | 51.44 | 161.32 | 0.240 | 6.273 | 0.182 | 13.182 | 0.455 | 82.983 | Good | SB | 3.505 | 7.524 |
| Usman Patel | KUW | 22.33 | 119.64 | 0.196 | 3.143 | 0.143 | 6.571 | 0.286 | 26.716 | Good | SB | 1.216 | 2.586 |
| Faisal Javed | QAT | 22.00 | 155.66 | 0.264 | 3.733 | 0.000 | 5.600 | 0.133 | 34.245 | Good | SB | 0.995 | 2.977 |

*Figure 24 - PR rank of Batsmen’s Classified as SB by k-means Algorithm*

Figure 24 displays the PR rank of each batter identified by the model as a Specialist Batter (SB). Strangely, more batters who are rated as ‘Good’ by PR rank are categorised as ‘SB’ rather than ‘Best’. One batter, ED Silva, was given the PR rank of ‘Poor’ because he did not score enough runs, but our algorithm classified him as SB. This shows that a batter should not be judged solely on his average, strike rate, and runs scored.

### 3.4.2 Comparing Top 10 Players from ICC Men’s All-Time T20I Rankings:

The Top 10 Batters, Bowlers, and All-Rounders from the ICC Men's All-Time T20I Rankings have been compared in this section along with their RBML, PCA, and Hussey Index (only for batters). We combined the RBML and PCA scores from the batters and bowlers to determine the scores for All-Rounders.



| ICC_rank | name | country | prelim_metric | prelim_rank | pre-rank | kmeans_role | rbml_score | RBML_rank | pca_score | PCA_rank | hussey_index | hussey_rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | DJ Malan | ENG | 56.664 | 4 | Best | SB | 2.324 | 3 | 5.019 | 4 | 178.50 | 5 |
| 2 | AJ Finch | AUS | 50.255 | 6 | Best | SB | 1.629 | 7 | 3.872 | 6 | 179.56 | 3 |
| 3 | V Kohli | INDIA | 70.900 | 1 | Best | SB | 2.494 | 2 | 5.252 | 2 | 189.17 | 1 |
| 4 | Babar Azam | PAK | 58.921 | 2 | Best | SB | 2.569 | 1 | 5.485 | 1 | 174.96 | 6 |
| 5 | KP Pietersen | ENG | 53.675 | 5 | Best | SB | 1.653 | 5 | 3.928 | 5 | 179.44 | 4 |
| 6 | EJG Morgan | ENG | 38.917 | 10 | Best | FL | 0.764 | 10 | 2.081 | 10 | 164.75 | 10 |
| 7 | AD Hales | ENG | 42.375 | 9 | Best | SB | 1.230 | 9 | 3.003 | 9 | 167.66 | 9 |
| 8 | KL Rahul | INDIA | 57.965 | 3 | Best | SB | 2.310 | 4 | 5.047 | 3 | 183.17 | 2 |
| 9 | BB McCullum | NZ | 48.572 | 7 | Best | SB | 1.640 | 6 | 3.701 | 7 | 171.87 | 7 |
| 10 | F du Plessis | SA | 47.745 | 8 | Best | SB | 1.584 | 8 | 3.603 | 8 | 169.91 | 8 |

*Figure 25 - Analysis of Top 10 ICC All-Time Batters*

The Top 10 Batters from the ICC Men's All-Time T20I Rankings are compared in Figure 25. Except for EJ Morgan (ENG), labelled as a Floater by our model, all the batters are classified as Specialist Batters. Looking at EJ Morgan's career batting positions (Appendix), you can see that he has typically batted in the middle of the order (number 4 or number 5) with a good average and strike rate. By PR rank, every batsman has been rated as "Best." According to various scores/indexes used in this study, V Kohli (IND) and Babar Azam (PAK) have shared the top two spots. If we look at traditional rating criteria, then V Kohli (IND) will have an edge over Babar Azam (PAK) among the batters in the figure but according to RBML and PCA score Babar Azam (PAK) has a slightly better score over V Kohli (IND). DJ Malan (ENG), who has the highest career rating of any T20 International batsman, has the third highest RBML score of 2.324 and fourth highest PCA score of 5.019 among the batters in Top 10 Batters from the ICC Men's All-Time T20I Rankings.

## 4. Conclusion:

Rating individuals in team sports is a more complex task and challenging task, primarily because of the team structure. The primary purpose of this research was to develop a new performance indicator for a player based on records using a large amount of data available on the internet and machine learning algorithms. A batting/bowling average, batting/bowling strike rate and bowling economy reveal a great deal about the performance of the player. However, in the shortest format of the game with a less number of balls, one cannot rely only on the traditional performance indicator. In a T20 inning, it is not good enough for a player to have a high batting average and low batting strike rate. The runs scored slowly and will end up in defeat. Machine learning is gaining popularity, and many researchers are working on developing new performance indicators using mathematical modelling or Machine learning.

Firstly, we started collecting the data from ESPNCricinfo from the inception of T20 cricket to 25th September 2023 using BeautifulSoup and Python. After gathering the data, this study needed to remove the outliers from both the datasets (Batter's and Bowler's) due to there were not many players who played more than 15 innings in the shortest format of the game. This study removed the outliers so that model will train correctly and will be able to minimize bias and noise while training the model. This study extracted five KPIs for each bowler and batter based on traditional performance indicators. Feature Scaling was one of the important parts of this research especially when some KPIs preferred lower values this study used formula-based normalisation rather than a standard scaler. To group players into various clusters and identify their roles based on the KPI values, a K-Means Clustering algorithm approach is used. To determine feature importance for various KPIs, a supervised classifier built on the Random Forest Algorithm is used. We can precisely determine the relative importance of various features for each role thanks to the combination of supervised and unsupervised learning. Another method used in this study is a simple and direct technique called principal component analysis, which can be directly applied to correlated, multivariate data, rather than trying to find the role-based cluster.

Looking at the results of k-Means clustering, we can say that clusters were distinguishable in the scatter plot(s) for both batters and bowlers. We can say that the clustering algorithm did well while grouping the data points. Each classification algorithm as Random Forest, XGBoost and Voting classifier supported the role-based clusters by yielding good accuracy scores and predicting almost 95% of data points correctly in the confusion matrix. Random Forest was among the algorithms that had the best accuracy scores. We used one vs all classification methods to predict the importance of each feature for each role-based cluster, and from this methodology, we got RBML scores. While combining the results of K-Means with the PCA results we can say that PCA also did well to reduce the dimensionality of the data preserving the underlying logic of the data which is visible in PCA scatter plots.

Comparing RBML/PCA ranks with roles assigned to players through clustering with PR rank and scores. These demonstrate that a player's average, strike rate, runs scored, and wickets taken should not be considered solely. Many other factors influence player performance, including batting position, player form, the opposition team, and pitch conditions. This study ranks layers based on their roles rather than batting positions because, in the shortest format of the game, each player in the team has a defined strategic role.

This research also tried a couple of new methods as Self Organizing Maps (SOM) to cluster data points, but it created target imbalance by predicting more than 70% of batters in the same group. Another method which we tried was having a Neural Network (NN) to get weights given to each feature after clustering the data points, but the accuracy of the Neural Network (NN) was very low compared to classification algorithms.

## 4.1 Limitations:

Although this study has achieved good results in this research, still there are certain limitations while we think about KPIs and data available. While extracting Big Impact Index for batters this study has considered half centuries and centuries only, but ideally scoring 30 runs is considered a good score for T20 innings. E.g., When comparing EJ Morgan (ENG) and GJ Maxwell (AUS), they both batted at various positions (Appendix), but GJ Maxwell has three centuries while EJ Morgan has none, and Maxwell outperforms Morgan in terms of strike rate. That is one of the reasons Maxwell is classified as a Specialist Batter (SB), while Morgan is classified as a Floater (F). So, ideally having three wickets or more in an innings is considered impactful bowling. We could not find any dataset with these granularities of the requirements, and most of the players have played fewer innings, so the dataset was skewed and created bias while training the model.

## 4.2 Future Work:

Many scholars and enthusiasts are working on player evaluation in sports analytics, which is always improving. This study only analysed role-based grouping and forecasted KPI feature relevance, however prior records cannot be used to identify a teammate. When evaluating a player, recent form must be considered. Players perform differently in home and away games depending on pitch and weather. Popular sports include cricket. This study's model will remain valid, and more data with easy granularity can improve its performance.

## 5. References:

1. B. Spencer, S. Morgan, J. Zeleznikow, S. Robertson, W. Bulldogs, F. Club, 2016. Clustering team profiles in the australian football league using performance indicators, in: The 13th Australasian Conference on Mathematics and Computers in Sport,
2. Barr, G. and Kantor, B., 2004. A criterion for comparing and selecting batsmen in limited overs cricket. Journal of the Operational Research Society, 55(12), pp.1266-1274.
3. Basevi, T. and Binoy, G., 2007. The world's best Twenty20 players. [online] ESPNCricinfo. Available at: <https://www.ESPNCricinfo.com/story/the-world-s-best-twenty20-players-311962> [Accessed 2 September 2023].
4. Basit, A., Alvi, M., Jaskani, F., Alvi, M., Memon, K. and Shah, R., 2020. ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques. 2020 IEEE 23rd International Multitopic Conference (INMIC),.
5. Cricinfo. 2022. Anthony de Mello Trophy, 2020/21 Cricket Team Records & Stats | ESPNCricinfo.com. [online] [Accessed 27 September 2023].
6. Cricinfo. 2022. Batting Records [online] [Accessed 27 September 2023].
7. Damodaran, U., 2006. Stochastic dominance and analysis of ODI batting performance : The Indian cricket team 1989-2005. Journal of Sports Science & Medicine, Vol. 5, pp.503-508.
8. Davis, J., Perera, H. and Swartz, T., 2015. Player evaluation in Twenty20 cricket. Journal of Sports Analytics, 1(1), pp.19-31.
9. Deep Prakash, C. and Verma, S., 2022. A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket. Decision Analytics Journal, 2, p.10002
10. Deep, C., Patvardhan, C. and Singh, S., 2016a. A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers. International Journal of Computer Applications, 5(2), pp.37-47.
11. Deep, C., Patvardhan, C. and Singh, S., 2016b. A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers. International Journal of Computer Applications, 137(10), pp.42-49.
12. Dutt, A., 2021. A Step By Step Implementation of Principal Component Analysis. [online] Medium.
13. Ecosystem (LEDU), E., 2018. Understanding K-means Clustering in Machine Learning. [online] Medium. [Accessed 1 September 2022].
14. Eddie, 2021. Feature Scaling Techniques | Why Feature Scaling is Important. [online] Analytics Vidhya.

15. En.wikipedia.org. 2022. Batting order (cricket) - Wikipedia. [online]
16. En.wikipedia.org. 2022. Power play - Wikipedia. [online]
17. En.wikipedia.org. 2022. Premier League - Wikipedia. [online]
18. En.wikipedia.org. 2022. ICC Men's T20I Team Rankings - Wikipedia. [online]
19. En.wikipedia.org. 2022. List of International Cricket Council members - Wikipedia. [online]
20. Géron, A., 2020. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. Beijing: O'Reilly, pp.192-200.
21. Guido, S. and Müller, A., 2016. Introduction to Machine Learning with Python. O'Reilly Media, pp.168-181.
22. Humaira, H. and Rasyidah, R., 2020. Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia,.
23. Icc-cricket.com. 2022. ICC Men's T20I Team Rankings | ICC. [online]
24. Iyer, S. and Sharda, R., 2009. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), pp.5510-5522. Jhansi Rani, P., Vidyadhar Kamath, A., Menon, A., Dhatwalia, P., Rishabh, D. and Kulkarni, A., 2020. Selection of Players and Team for an Indian Premier League Cricket Match Using Ensembles of Classifiers. 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT),
25. Kumar, S., 2021. Use Voting Classifier to improve the performance of your ML model. [online] Medium
26. Lemmer, H., 2002. The combined bowling rate as a measure of bowling performance in cricket. South African Journal for Research in Sport, Physical Education and Recreation, 24(2).
27. Lemmer, H. (2012), "The Single Match Approach to Strike Rate Adjustments in Batting Performance Measures in Cricket," Journal of Sports Science and Medicine, 10, 630-634.
28. Lemmer, H., 2004. A measure for the batting performance of cricket players : research article. South African Journal for Research in Sport, Physical Education and Recreation, 26(1).
29. Lemmer, H., 2008a. Measures of batting performance in a short series of cricket matches. South African Statistical Journal, 42(1): 83-105.
30. Lemmer, H., 2008b. An analysis of players\' performances in the first cricket Twenty20 World Cup series. South African Journal for Research in Sport, Physical Education and Recreation, 30(2).
31. Lewis, M., 2003. Moneyball. New York: W. W. Norton & Co.

32. Lindgren, I., 2020. Dealing with Highly Dimensional Data using Principal Component Analysis (PCA). [online] Medium.
33. Manage, A. and Scariano, S., 2013. An Introductory Application of Principal Components to Cricket Data. Journal of Statistics Education, 21(3).
34. McHale, I., Scarf, P. and Folker, D., 2012. On the Development of a Soccer Player Performance Rating System for the English Premier League. Interfaces, 42(4), pp.339-351.
35. Medium. 2019. Why is scaling required in KNN and K-Means? | Medium. [online]
36. Medium. 2020. Dealing with Highly Dimensional Data using Principal Component Analysis (PCA) | Towards Data Science. [online]
37. Medium. 2020. All about Feature Scaling | Towards Data Science. [online]
38. Medium. 2021. Beginner’s Guide to XGBoost for Classification Problems | Medium. [online] Medium.
39. Numpy.org. 2022. NumPy. [online]
40. Oughali, M., Bahloul, M. and El Rahman, S., 2019. Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models. 2019 International Conference on Computer and Information Sciences (ICCIS),.
41. Paffenroth, R. and Kong, X., 2015. Python in Data Science Research and Education. Proceedings of the 14th Python in Science Conference,.
42. Pandas.pydata.org. 2022. pandas - Python Data Analysis Library. [online]
43. Passi, K. and Pandey, N., 2018. Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning. Computer Science &amp; Information Technology.
44. Pugsee, P. and Pattawong, P., 2019. Football Match Result Prediction Using the Random Forest Classifier. Proceedings of the 2nd International Conference on Big Data Technologies - ICBDT2019,.
45. PyPI. 2022. beautifulsoup4. [online]
46. scikit-learn. 2022. scikit-learn | Classification Metrics and Evaluation. [online]
47. Thecricketcouch.com. 2012. A Simple Metric to Understand Batting Efficiency in the IPL | The Cricket Couch. [online]

## 6. Bibliography:

1. Ahamed, F., 2022. Web scraping Cricinfo data. [online] Medium.
2. Pandas.pydata.org. 2022. pandas.DataFrame.sample — pandas 1.4.4 documentation. [online]
3. Pandas.pydata.org. 2022. pandas.Series.map — pandas 1.4.4 documentation. [online]
4. scikit-learn. 2022. 1.10. Decision Trees. [online]
5. scikit-learn. 2022. 1.11. Ensemble methods. [online]
6. scikit-learn. 2022. 2.3. Clustering. [online]
7. scikit-learn. 2022. Importance of Feature Scaling. [online]
8. scikit-learn. 2022. Selecting the number of clusters with silhouette analysis on KMeans clustering. [online]
9. scikit-learn. 2022. sklearn.cluster.KMeans. [online]
10. scikit-learn. 2022. sklearn.ensemble.RandomForestClassifier. [online]
11. Shmueli, B., 2022. Multi-Class Metrics Made Simple, Part II: the F1-score. [online] Medium.
12. Xgboost.readthedocs.io. 2022. Introduction to Boosted Trees — xgboost 1.6.2 documentation. [online]