K2136854_sports_analytics_dissertation.docx

by Yash Jadwani

Submission date: 19-Sep-2022 03:49PM (UTC+0100)
Submission ID: 186097905
File name: K2136854_sports_analytics_dissertation.docx (2.49M)
Word count: 17805
Character count: 94802

FACULTY OF SCIENCE, ENGINEERING AND COMPUTING

School of Science, Engineering and Computing

MSc DEGREE
IN
DATA SCIENCE

Submitted by: Yash Jadwani

Student Id: K2136854

A new Machine Learning based approach to analyse player performance in T20 Internationals

Date: 19th September 2022

Supervisor: Dr James Denholm-Price

Kingston University London

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

Acknowledgement

First and foremost, I'd like to thank my patient and supportive supervisor, Dr James Denholm-Price, who has helped me throughout this research project. I am grateful for our friendly conversations at the end of our meetings, as well as your support in my academic endeavours.

I thank the Almighty for providing us with the fortitude and tenacity required to finish the challenging task. This project is meant to act as a thank you to everyone who has helped in any manner to make this research a huge success.

I appreciate Dr Gordon Hunter for exposing me to the exciting research on utilising machine learning to analyse player performance in T20 Internationals.

Yash Jadwani

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

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

1. Web scraping website and building a clean dataset.
2. Define new Key Performance Indicators (KPIs) for batters and bowlers.
3. Performing Exploratory Data Analysis and building a new dataset for Analysis.
4. To analyse a player's past performance based on predefined performance indicators as well as newly defined KPIs.
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 bowling performance evaluation discussed in this chapter:

Method	Bowling performance indexes
Lemmer (2002)	CBR = 3* Runs Conceded / (Wickets taken + overs bowled + wickets taken (Runs Conceded / Bowls bowled))
Basevi et al. (2007)	(Average * Economy) / 6
Iyer et al. (2009)	Neural Network-based approach to classifying bowlers and helps predict team selection.
Deep et al. (2016b)	KPIs: Economy, Wicket Taker, Consistent, Big Wicket taker, short performance Index Methodology: • Calculate MVPI index • Apply Random Forest Algorithm • 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 Methodology: • Calculate MVPI index • Categorise players according to the standard role • 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 Bowled Index, Economy, Big Wicket Index, Short Performance Index. Methodology: • Use K-means to classify batsman • Performed one vs rest classification using Random Forest Algorithm. • Calculate Feature importance and multiply it with the normalised KPI value.
Manage et al. (2013)	KPIs: Wickets, Bowling Average, Strike rate, Economy Methodology: • Apply PCA while using the correlation matrix. • Multiply the first component's weights by the normalised KPI value.

Table 1: Summary of major bowling performance evaluation discussed

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 Methodology: • Calculate MVPI index • Apply Random Forest Algorithm. • 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 Methodology: • Calculate MVPI index • Categorise players according to the batting order • 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 Methodology: • Use K-means to classify batsman • Performed one vs rest classification using Random Forest Algorithm • Calculate Feature importance and multiply it with the normalised KPI value.
Manage et al. (2013)	*KPIs: Runs, Average, Strike rate, Fours, Six,* *HF = (2*centuries) + half-century* *Methodology:* • Apply PCA while using the correlation matrix. • Multiply the first component's weights by the normalised KPI value.

Table 2: Summary of major batting performance evaluation discussed

3. Contribution Chapters

3.1 Methodology

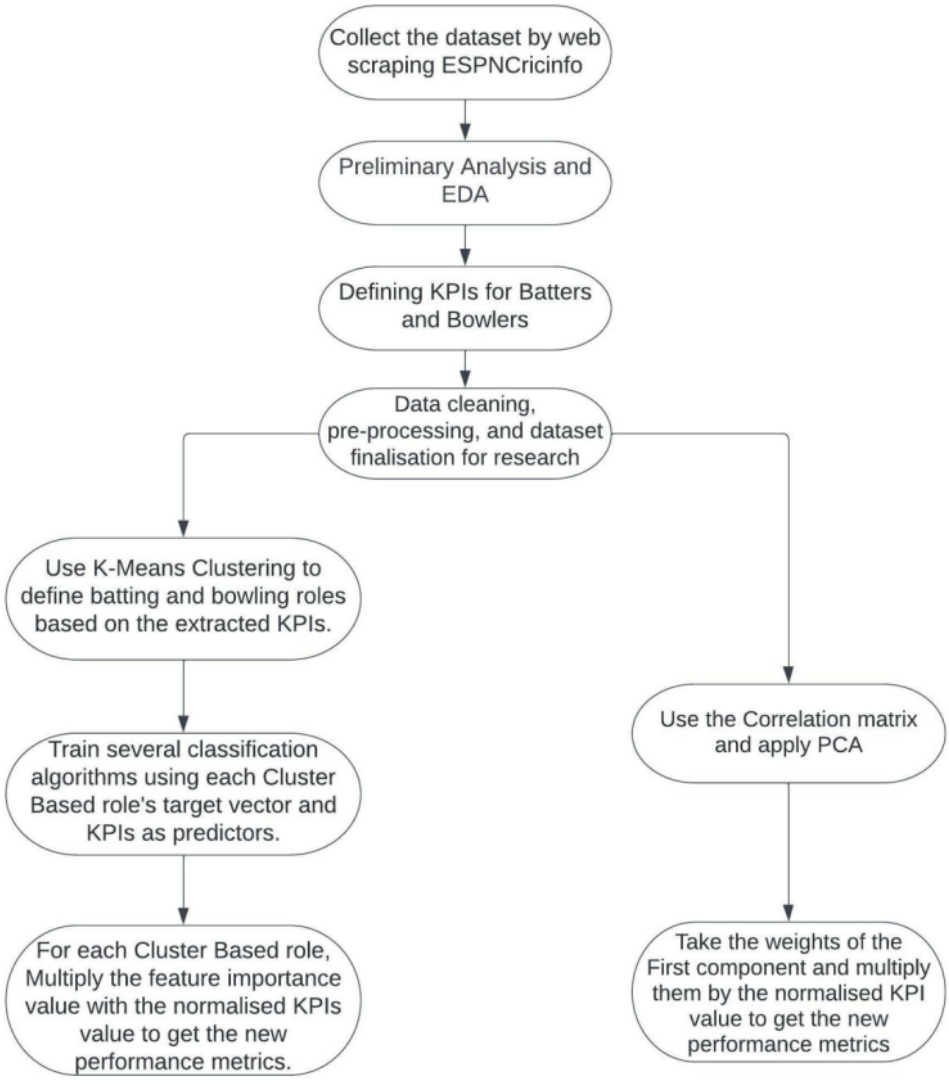

Figure 1: Methodology of this research

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. Apart from Python libraries, this study has used Tableau for creating graphs. More about Tools and technologies can be found in Appendix.

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 April 17th, 2022. 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 and 3, both batter's and bowler's data are 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, as well as any bowlers who have bowled in less than four innings.

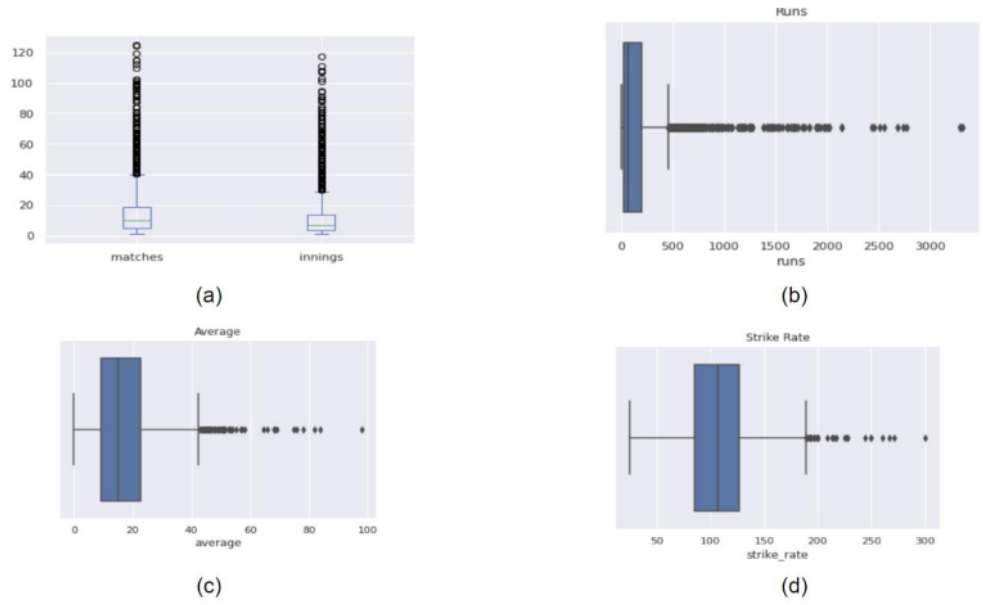

Figure 2: (a) Matches and Innings, (b) Total Runs, (c) Batting Average and (d) Batting Strike rate

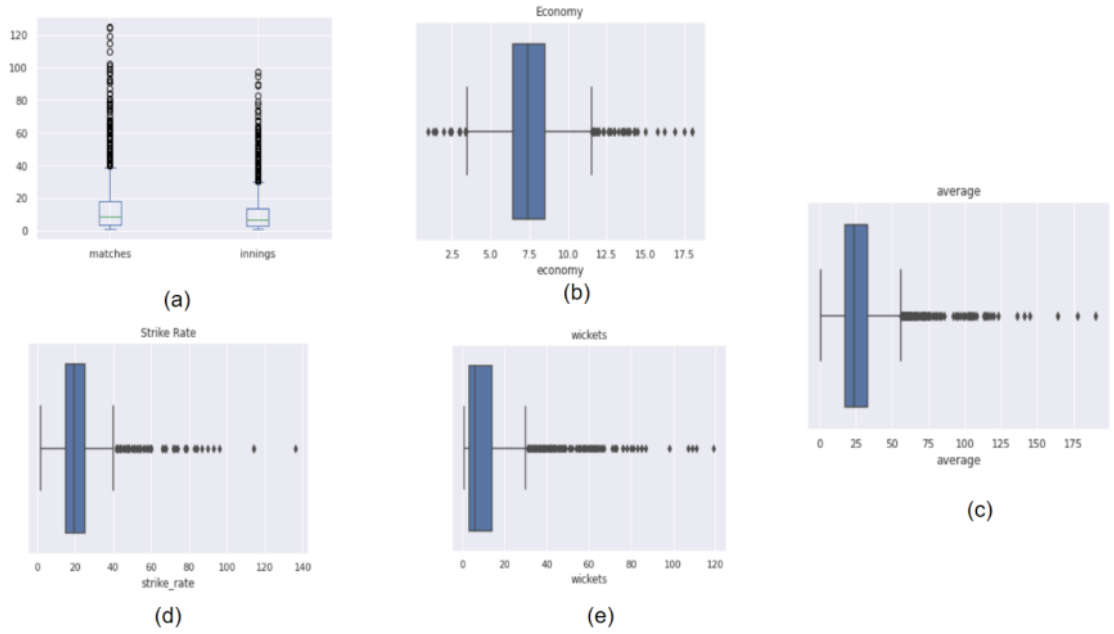

Figure 3: (a) Matches and Innings, (b) Economy, (c) Bowling Average and (d) Bowling Strike rate, (e) Total Wickets

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.

Shakib-al-Hasan (BAN), a highly skilled all-rounder, has 119 wickets in 95 T20 internationals with an average of 19.88, a strike rate of 17.8, and an economy of 6.67. Tim Southee (NZ), one of the most successful T20 international fast bowlers, has 111 wickets with an average of 24.58, a strike rate of 17.9, and an economy of 8.19. There have been nine bowlers who have taken six wickets in a T20 inning. Only BAW Mendis (SL) holds the record for taking six wickets twice in a T20 inning. P Aho, a bowler from Nigeria, has the best bowling figure of taking six wickets while giving only five runs in an innings. According to the ICC Men's T20 international team rankings, R Ravindra (NZ) has the best economy of 4.36, with an average of 13.83 among the players in the Top 10 teams.

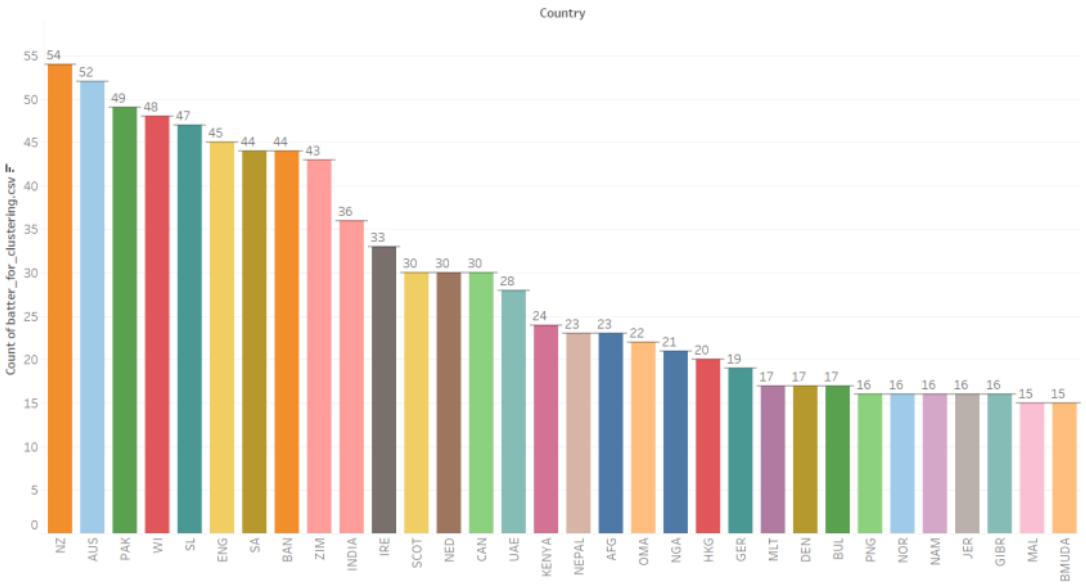

Figure 4: Country-wise batter (min 15 players)

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. The number of bowlers from each nation who participated in a T20 International game with at least 15 bowlers is shown in Figure 4. With 48 bowlers, India has the most, followed by Australia with 47. Afghanistan has the fewest bowler only 20 among the top 10 ranked countries in the ICC Men's T20 international team rankings.

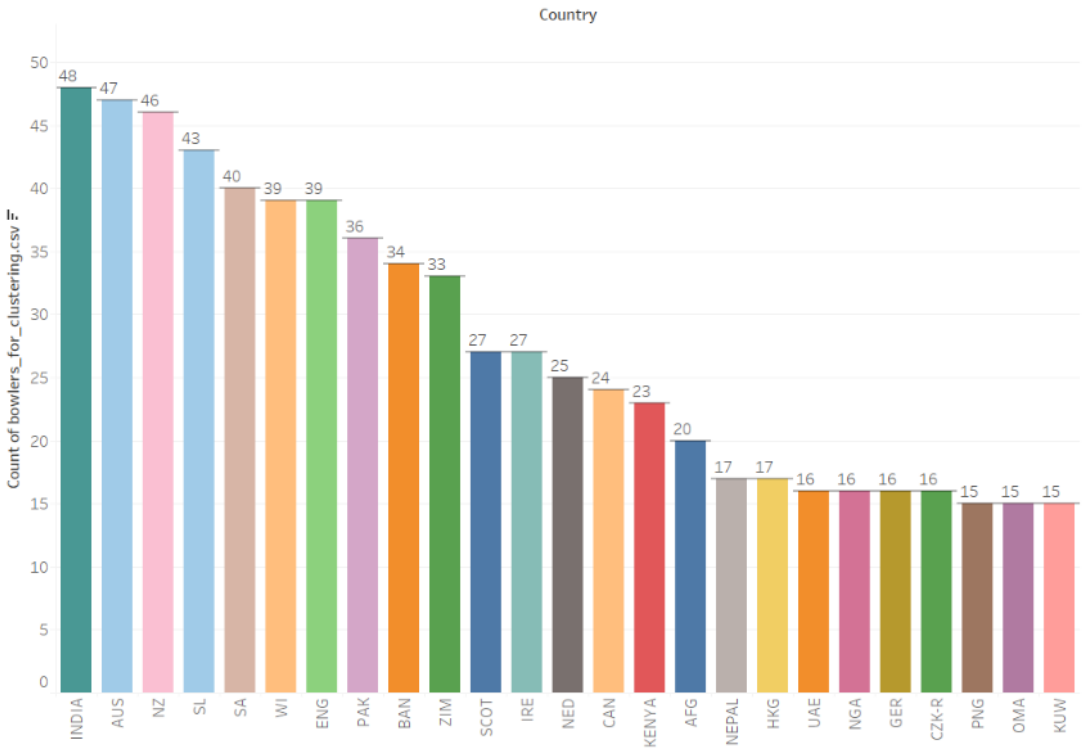

Figure 5: Country-wise bowler (min 15 players)

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, and Prelim Metric (PM) for bowlers as (Average * Economy) / 6. 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 3:PR Criteria for Batters

Criteria	**Ranking**
PM < 30 and Wickets >= 25	Best
PM < 30 and Wickets < 25	Good
PM between 30 and 40	Good
PM between 40 and 65	Average
PM > 65	Poor
Wickets < 5	Poor

Table 4:PR Criteria for Bowlers

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. We took 98 batters' records and 82 bowlers' records. After removing these records, this study had 1302 batter records and 1085 bowler records.

3.3.4 Key Performance index(s) for batters

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.

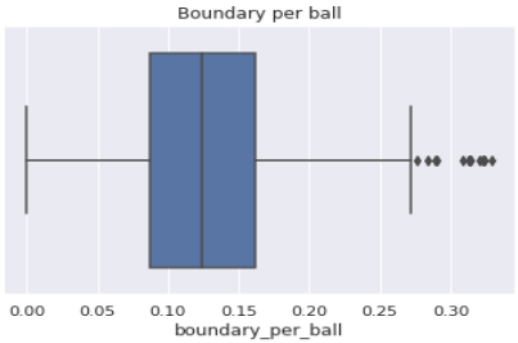

Figure 6: Box plot for Boundary Per Ball KPI

Figure 6 shows that some batsmen have a greater than 30% chance of hitting a boundary on any given ball. In the six innings he has played, F Allen (NZ) has the highest boundary per ball of 0.329 and the highest strike rate of 190.24. He has 19 fours and eight sixes from 82 balls faced.

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

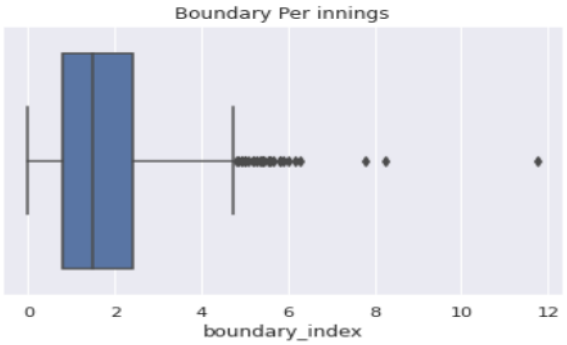

Figure 7: Box plot for Boundary Index KPI

Figure 7 shows that most batsmen can hit more than one boundary in an innings on average. Some exceptions include Zeeshan Kukikhel (Hungary), Azhar Adani (Portugal), and Taranjeet Singh (Romania), who have Boundary Indexes of 11.75, 8.25, and 7.786 respectively.

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.

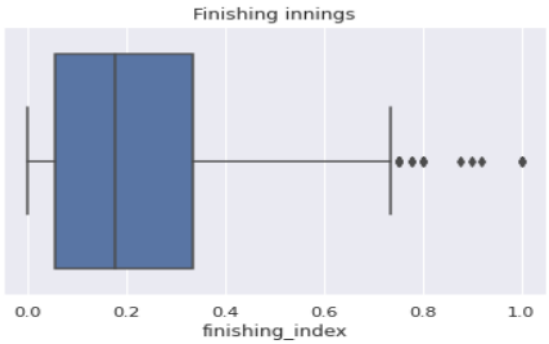

Figure 8: Box plot for Finishing Index KPI

4. **Runs Without Boundary Index:** We attempted to calculate average runs scored by batsmen without fours and sixes in this KPI by subtracting boundaries from total runs and dividing them by total innings.

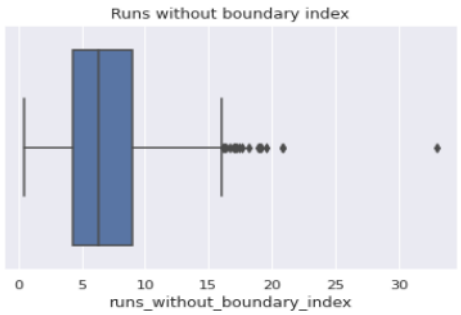

Figure 9: Box plot for Runs Without Boundary Index KPI

5. **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.

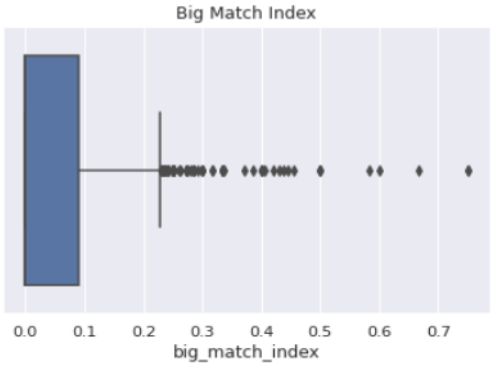

Figure 10: Box plot for Big Match Index KPI

Only 68 players have reached the century mark in T20 internationals while 503 have only reached the half-century mark. Some players have scored a quick 30 but have yet to reach the half-century mark. Babar Azam (PAK) is an active player among the top ten ranked countries in the ICC Men's T20 international team rankings, with a Big Match Index of 0.406 in 69 innings played with 26 fifty and three hundred.

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 5: Summary of Extracted KPIs for Batters

3.3.5 Key Performance index(s) for bowlers

1. **Balls Bowled Per Innings:** In this KPI, we will calculate the average number of balls bowled per inning by each bowler. The higher this KPI value, the more trustworthy the bowler is. The total number of balls bowled divided by the total number of innings will be used to calculate this KPI.

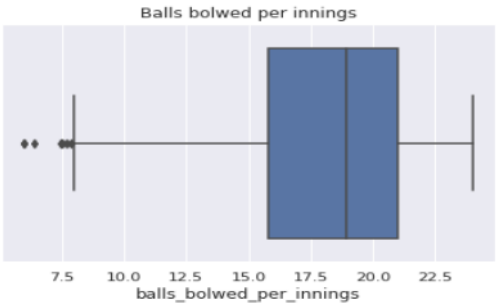

Figure 11: Box plot for Balls Bowled Per Innings KPI

Figure 11 clearly shows that a bowler averages 18 balls per innings (three overs). T Natarajan (IND), M Siraj (IND), and SR Clark (AUS) are among the 25 players who have bowled all their quota overs. Rohit Sharma (IND) is among the bowlers who have bowled less than 7.5 balls per innings and taken a wicket in the nine innings he has bowled.

2. **Wickets Index:** Bowlers' primary goal is to take wickets while remaining economical. We will calculate the average number of wickets taken by a bowler per innings in this KPI. This KPI will be calculated by dividing the total number of wickets by the total number of innings.

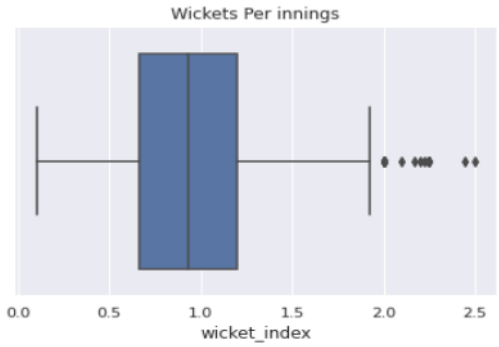

Figure 12: Box plot for Wicket Index KPI

The median line is around 0.92, we can see from the figure 12 boxplot that most bowlers take at least one wicket in an innings on average. Another thing that stands out is the rarity and challenge for bowlers to take at least two wickets per innings. All bowlers with a Wicket Index greater than 2 are bowlers from associate members. Associate Members are countries where cricket is well-established and well-organized but do not meet the requirements for Full Membership (List of International Cricket Council members - Wikipedia, 2022).

3. **Big Impact Index:** In this KPI, we will calculate how many times a player has taken more than three wickets in an innings out of all innings played. By adding the 4-fer and 5-fer wicket innings and dividing it by the total number of innings.

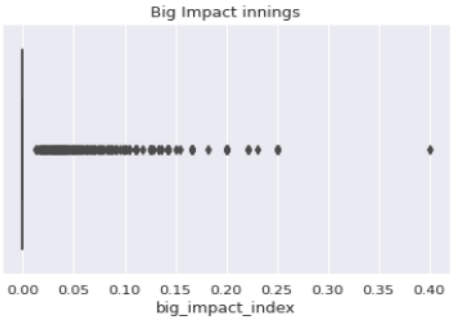

Figure 13: Box plot for Big Impact Innings KPI

According to the box plot above, we can say that there are not many players who have taken a 4-fer or 5-fer in their career. To support the statement, only 277 players have taken at least four wickets in a single inning. One thing is clear from the KPI extraction: as players play more cricket, the KPI value of indexes like the Big Impact Index is likely to decrease. If we look at the Top 10 teams in the ICC Men's T20 international team rankings, MW Parkinson (ENG) has the best Big Impact innings index of 0.25 with six wickets in four innings and his best bowling figures of 4/47.

4. **Short Impact Index:** In this KPI, we will calculate a bowler's total number of wickets without considering 4-fer and 5-fer innings. We will subtract the wickets taken in 4-fer and 5-fer (considering some bowlers have picked 6 wickets in an innings) innings from the total wickets and then divide it by the total number of innings.

Figure 14: Box plot for Short Impact Index KPI

5. **Runs Index:** Bowlers' primary goal is to take wickets while remaining economical. We will calculate the average number of runs conceded by a bowler per innings in this KPI. This KPI will be calculated by dividing the total number of runs by the total number of innings

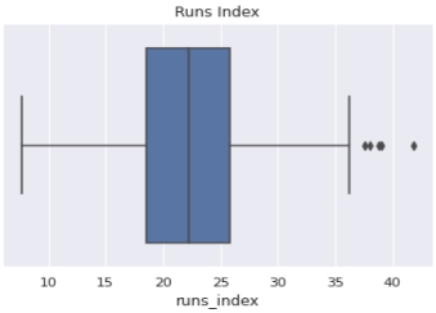

Figure 15: Box plot for Runs Index KPI

The chart above clearly shows that a bowler concedes 22.5 runs per innings on average. Some players give up more than 35 runs per inning. R Shepherd (WI) is one of the bowlers who as bowled 14-inning with a Runs Index of 35.93. M Siraj (IND) has a run index of 41.8 in five innings and an economy of 10.45.

1	Balls Bowled per Innings	Number of balls bowled / Total innings
2	Wickets Index	Number of wickets taken / Total innings
3	Big Impact Index	Number of times bowler picked more the 3 wickets / Total innings
4	Short Impact Index	Wickets were taken without Big Impact Innings / Total innings
5	Runs Index	Number of Runs Conceded by bowler / Total innings
6	Average	Total Runs / Total Wickets
7	Strike Rate	Balls bowled / Total Wickets
8	Economy	Total Runs / Total Overs bowled

Table 6: Summary of Extracted KPIs for Bowlers

3.3.6 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.

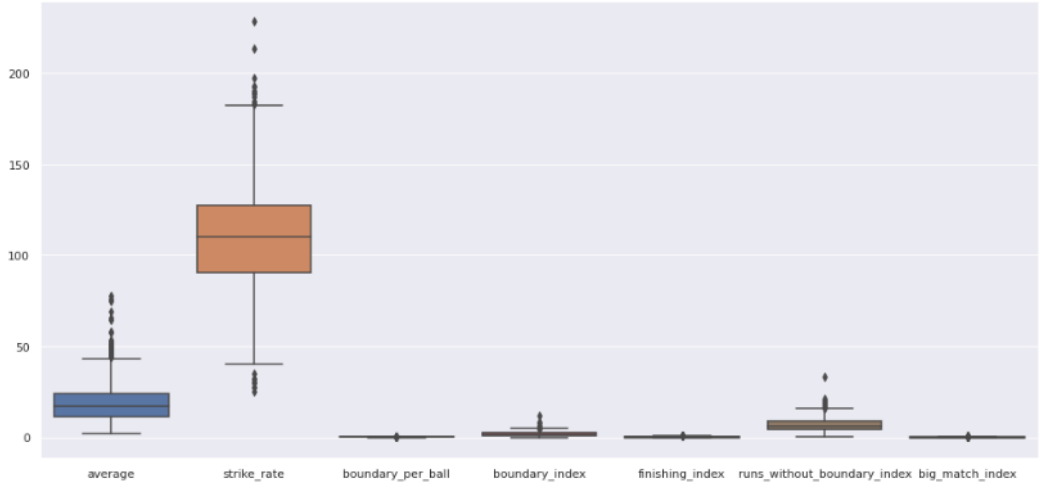

Figure 16: Box Plot for the combined Batters KPIs

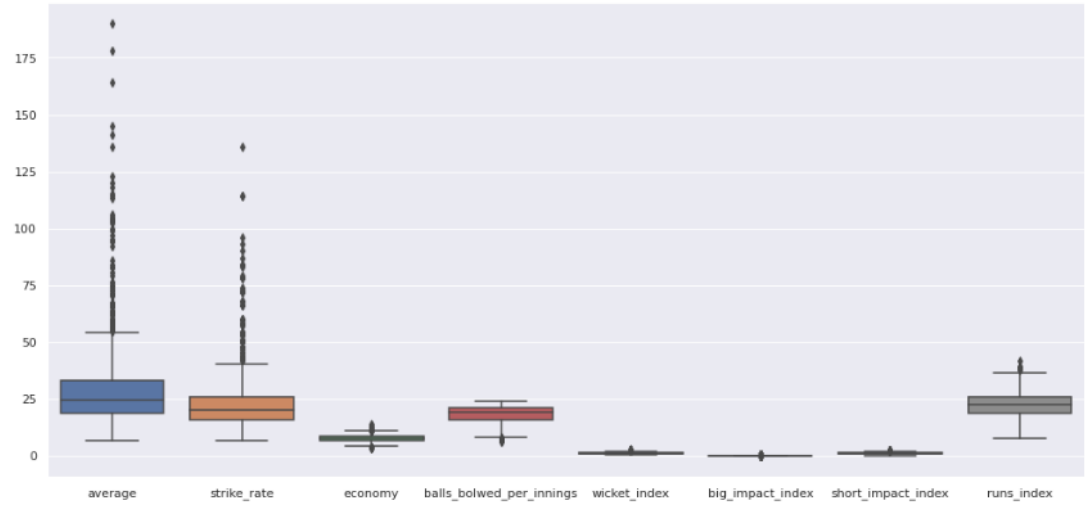

Figure 17: Box Plot for the combined Bowlers KPIs

Figures 16 and 17 show the box plot of the batter's and bowler's KPIs combined respectively. In the case of batter, it clearly shows the gap between the values of strike rate and average compared to other KPIs. In the case of bowlers, the gap between features is also evident. In both the data we can see that the values of our extracted KPI are ranging quite low as compared to traditional KPIs (Average, Strike rate and Economy). That is one of the reasons why Scaling is necessary for our dataset.

- **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 18: Standardisation Formula

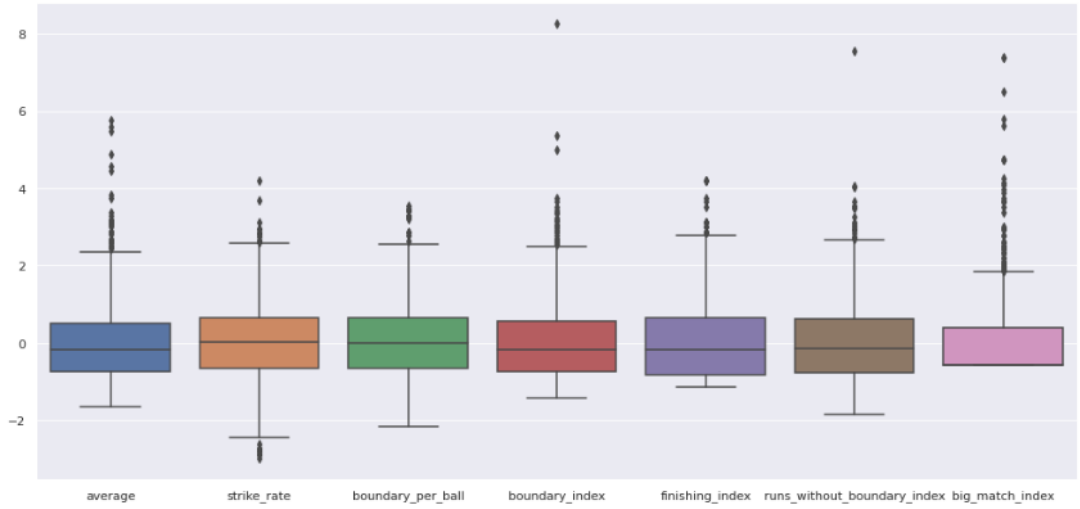

Figure 19: Box Plot for the combined Batters KPIs after Standardisation

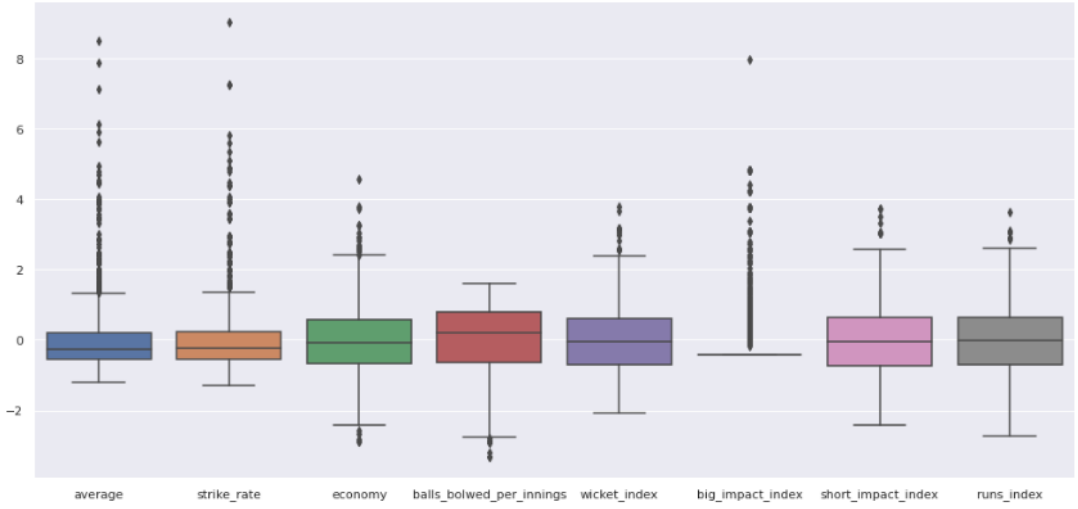

Figure 20: Box Plot for the combined Bowlers KPIs after Standardisation

The box plot of the combined KPIs for batters and bowlers after standardisation is shown in figures 19 and 20, respectively. The contrast between the graphs before and after scaling is obvious. After standardisation, we can see that in both plots, the median line and boxes are aligned on the same line or within the range, suggesting that there is not a significant amount of difference in the values. However, it is working properly for batters because all the values in batters prefer higher numbers

but for bowlers, some KPIs for bowlers, such as average, strike rate, economy, and runs index (runs conceded per innings) expect lower values.

- **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 21: Formula for Min-Max Scaling

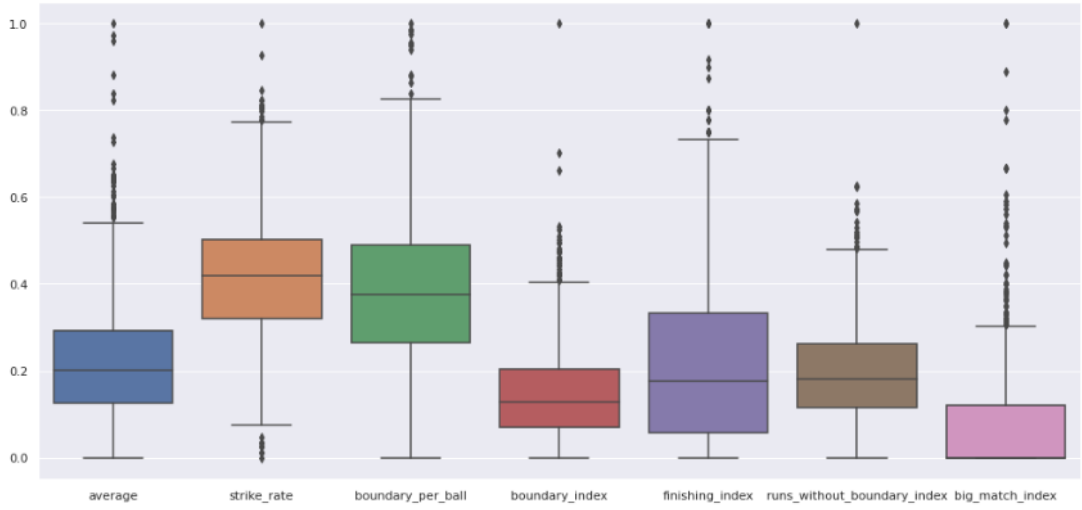

Figure 22: Box Plot for the combined Batters KPIs after Min-Max Scaling

The box plots of the combined KPIs for batters and bowlers after min-max scaling are shown in Figures 22 and 23, respectively. All other boxes in the bowlers start above 0.2, while the average and strike rate boxes are between 0 and 0.2, with many outliers. This study attempted but failed to address the issue of KPIs that favoured lower values using min-max scaling. Two of the scaling

strategies explored were robust scaling and normalisation, but they did not help with our problem of KPIs favouring lower values.

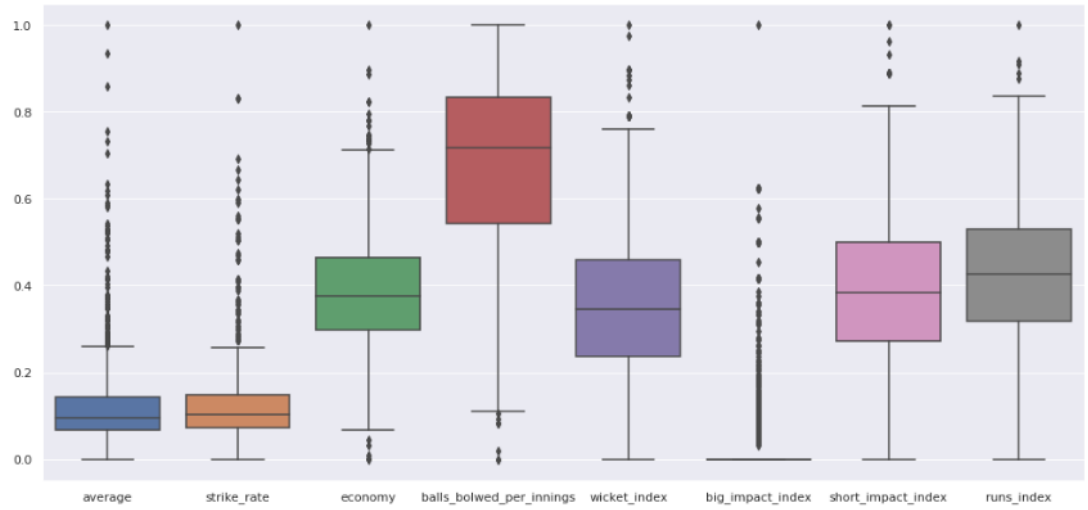

Figure 23: Box Plot for the combined Bowlers KPIs after Min-Max Scaling

- **Formula-Based Scaling for Bowlers**

While researching how to resolve the issues of KPIs that expect a lower value. We came across the formula-based technique in Deep Prakash et al., (2022). Where they have used the modified version of Min-Max scaling by tweaking the formula for KPIs that preferred lower values.

Case 1, if higher values are preferred for the KPI 'i' for a player 'j'

$$KPI_{ij} = (KPI_{ij} - Min(KPI_i))/(Max(KPI_i) - Min(KPI_i)) \quad (1)$$

Case 2, if lower values are preferred for the KPI 'i' for a player 'j'

$$KPI_{ij} = (Max(KPI_i) - KPI_{ij})/(Max(KPI_i) - Min(KPI_i)) \quad (2)$$

Figure 24: Tweaked Formula of Min-Max Scaling for KPIs (Deep Prakash et al., 2022)

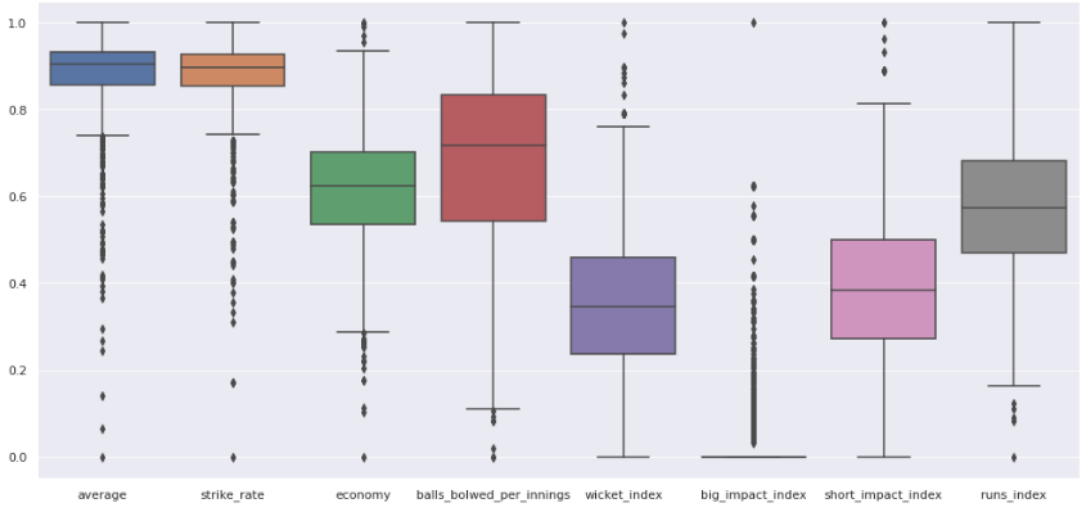

Figure 25: Box Plot for the combined Bowlers KPIs after Formula based normalisation

The box plot of the combined KPIs for bowlers after formula-based normalisation is shown in figure 25. The issue of the KPIs that expects lower value should be scaled appropriately with the KPIs that prefer greater value is resolved in that we can see that the boxes of average, strike rate, and economy flipped and now shifted to the top side of the scale.

In terms of T20 international cricket, Shakib-al-Hasan (BAN) originally had an average of 19.88 and an economy of 6.67, which is considered good. However, following standardisation, his average and economy were adjusted to -0.49 and -0.56, respectively, making him one of the worst bowlers in the dataset. We used formula-based normalisation to address that problem; his average and economy were 0.93 and 0.69, respectively, demonstrating his ability as a wicket-taking bowler. Now, if we use the normalised data for bowlers to train the k-means algorithm, the system will learn correctly, and bias will be significantly decreased. To scale the data for the batter, we use a standardisation approach, but for bowlers, we use formula-based normalisation ahead of the training k-means algorithm.

3.3.7 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).

3.3.7.1 Results of K-Means Clustering for Batters

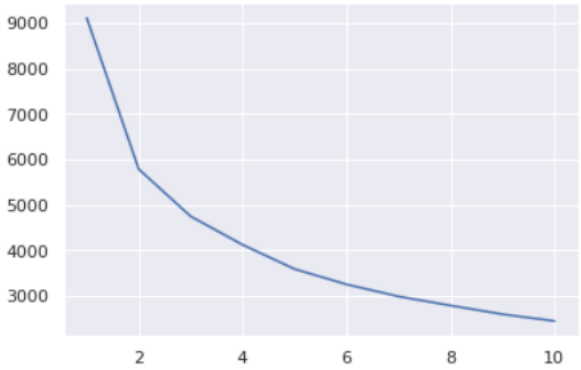

Figure 26: Results of elbow method for batters

```
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 27: 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 26 and 27 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 7: Accuracy score of various Evaluation Index for Batters

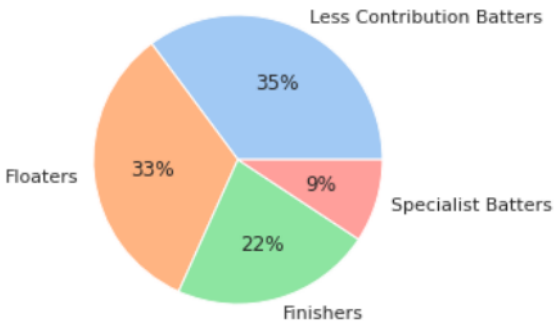

Figure 28: Pie chart of Clustering results for Batters with roles

Figure 28 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 29 displays the value of each centre of the cluster with the number of players in that cluster. Table 8 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 29: 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 8: 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.

3.3.7.2 Results of K-Means Clustering for Bowlers

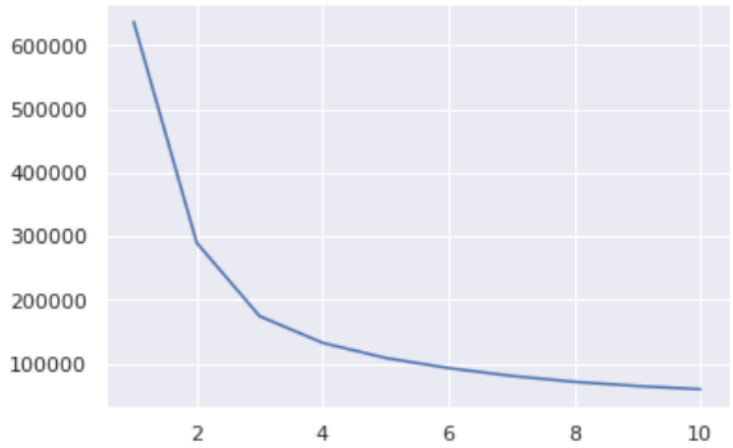

Figure 30: Results of elbow method for bowlers

```
For n_clusters = 2 The average silhouette_score is : 0.7419608324754727
For n_clusters = 3 The average silhouette_score is : 0.5037514667521731
For n_clusters = 4 The average silhouette_score is : 0.35394619812074773
For n_clusters = 5 The average silhouette_score is : 0.33941729877202964
For n_clusters = 6 The average silhouette_score is : 0.2937039827815529
For n_clusters = 7 The average silhouette_score is : 0.29788831964621837
For n_clusters = 8 The average silhouette_score is : 0.2956955350019551
For n_clusters = 9 The average silhouette_score is : 0.28826508170263787
For n_clusters = 10 The average silhouette_score is : 0.2897275314239923
```

Figure 31: Average silhouette score value based on the number of clusters for bowlers

After looking at the results from the elbow method and average silhouette score for the bowler in figures 30 and 31, We can say that there are three or four clusters in the dataset. This study needed as many roles as possible to do quality research and for that reason went ahead to identify the four clusters in the dataset.

Evaluation Index	**Accuracy Score**
Silhouette Score	0.354
Calinski-Harabasz Index	392.948
Davies-Bouldin Index	1.27

Table 9: Accuracy score of various Evaluation Index for Bowlers

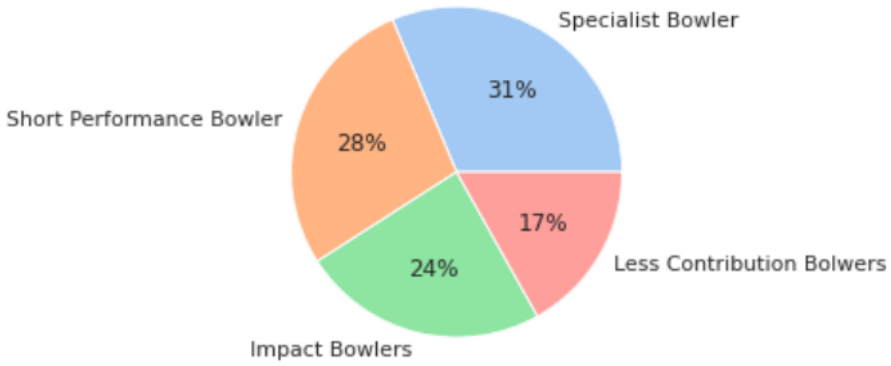

Figure 32: Pie chart of Clustering results for Bowlers with roles

Figure 32 depicts the distribution of participants according to their roles, which were determined by the values of each cluster centre, while Table 9 lists the accuracy scores for several evaluation indices. Figure 33 displays the value of each centre of the cluster with the number of players in that cluster. Table 10 displays some players according to roles assigned by the clusters.

	No. of Bowlers	average	strike_rate	economy	balls_bolwed_per_innings	wicket_index	big_impact_index	short_impact_index	runs_index
Specialist Bowler	340	0.935	0.929	0.658	0.841	0.538	0.107	0.560	0.507
Short Perfromance Bowler	301	0.918	0.899	0.689	0.616	0.342	0.046	0.381	0.674
Less Contribution Bolwers	183	0.780	0.792	0.521	0.350	0.150	0.006	0.200	0.724
Impact Bowlers	261	0.815	0.824	0.527	0.752	0.269	0.016	0.322	0.432

Figure 33: Cluster Centre for Bowlers

Specialist Bowler	Shakib-Al-Hasan (BAN), Tim Southee (NZ), Rashid Khan (AFG), SL Malinga (SL), Adil Rashid (ENG)
Short Performance Bowler	Moen Ali (ENG), Imad Wasim (PAK), D Sammy (WI), A Agar (AUS), Yuvraj Singh (IND)
Impact Bowlers	Chris Jordan (ENG), DJ Bravo (WI), R Jadeja (IND), Sohail Tanvir (PAK), Shane Watson (AUS)
Less Contribution Bowlers	K Pollard (WI), Gulbadin Naib (AFG), Glenn Maxwell (AUS), D Cristian (AUS), J Neesham (NZ)

Table 10: Some Bowler's names based on roles assigned by the K-Means algorithm

The cluster centre chose the names for each cluster, and Specialist Bowlers (SB) refers to bowlers who outperform everyone else in all other KPIs. Short Performance Bowlers (SPB) can bowl with a high average and strike rate, but they concede more runs and have a lower wicket index than Specialist Bowlers (SB). Impact Bowlers (IB) have a better economy than Short Performance Bowlers (SPB), but they bowl fewer overs than Specialist Bowlers (SB) and allow the fewest runs to be scored. Less Contribution Bowlers (LCB) are bowlers who do not bowl as much as expected and do not take as many wickets. One can deduct from table 10 that players are properly categorised by their ability to bowl.

3.3.8 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.

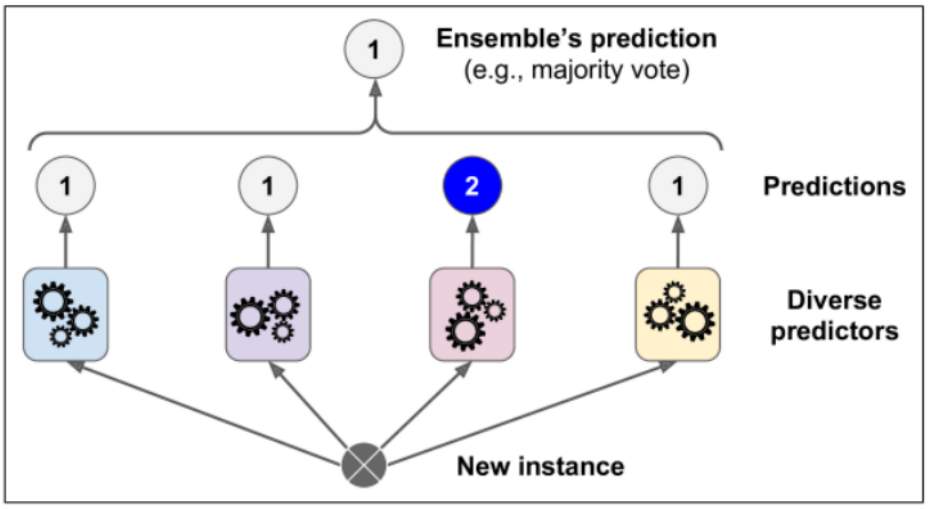

Figure 34: Hard Voting Classifier (Géron, 2020)

3.3.8.1 Results for Batters

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 11: Classification scores from various evaluation metrics 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 11. More about evaluation metrics can be found at scikit-learn Classification Metrics and Evaluation (2022).

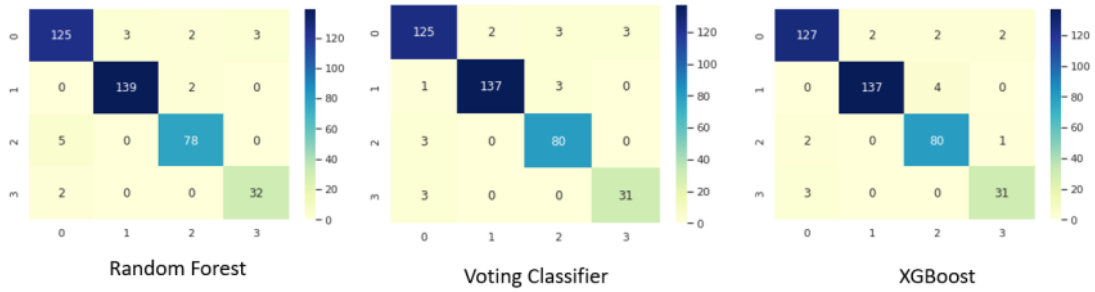

Figure 35: Confusion matrix of each algorithm used (Batters)

Figure 35 and Table 11, show that the clustering by the K-means algorithm did an excellent job of identifying clusters in the data for the batter.

KPIs (Batters)	Feature Importance		
	Random Forest	**XGBoost**	**Voting Classifier**
Average	0.154	0.084	0.11
Strike Rate	0.178	0.243	0.232
Boundary Per Ball	0.153	0.103	0.092
Boundary Index	0.236	0.209	0.247
Finishing Index	0.079	0.063	0.064
Runs Without Boundary Index	0.12	0.124	0.129
Big Match Index	0.08	0.173	0.126

Table 12: Feature Importance by different classification algorithms (Batter)

The weighting factors that various algorithms give to each feature for batters are shown in table 12. Because hitting boundaries at regular intervals is one of the most important tasks for batters in the T20 format, the boundary index was given considerable importance by all algorithms. All algorithms regard strike rate as a desirable thing because it is crucial to score runs quickly in a T20 inning.

3.3.8.2 Results for bowlers

Metrics	Random Forest	XGBoost	Voting Classifier
Balanced Accuracy	0.945	0.925	0.929
F1 Score (Micro)	0.951	0.936	0.939
Recall Score	0.946	0.925	0.929
Precision Score	0.956	0.934	0.941

Table 13: Classification scores from various evaluation metrics for Bowlers

The Balanced Accuracy score, F1 score, recall score, and precision score for each of the classification algorithms used for bowlers are shown in table 13.

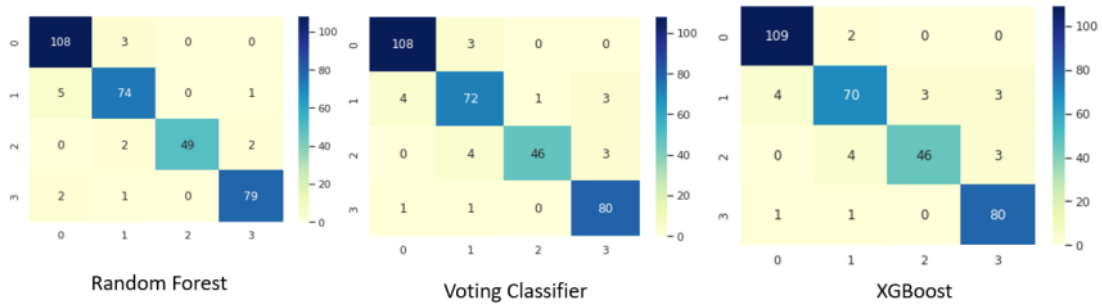

Figure 36: Confusion matrix of each algorithm used (Bowlers)

Figure 36 and table 13 show that the clustering by the K-means algorithm did an excellent job of identifying clusters in the data for bowlers as well. The weighting factors that different algorithms give to each feature for Bowlers are shown in table 14. Taking wickets at regular intervals is one of the most important tasks for bowlers in the T20 format, average and Wicket index has been given significant importance in all algorithms. Additionally, the Big Impact index has been disregarded by all algorithms because few players have ever taken more than three wickets in an innings.

KPIs (Bowlers)	Feature Importance		
	Random Forest	**XGBoost**	**Voting Classifier**
Average	0.163	0.225	0.176
Strike Rate	0.063	0.016	0.033
Economy	0.07	0.022	0.041
Balls Bowled per Innings	0.219	0.178	0.232
Wicket Index	0.171	0.268	0.239
Big Impact Index	0.01	0.003	0.004
Short Impact index	0.131	0.077	0.071
Runs Index	0.173	0.211	0.204

Table 14: Feature Importance by different classification algorithms (Bowler)

Random Forest performs the best on the provided dataset for both batters and bowlers, according to the accuracy and confusion matrix. The feature importance given by each algorithm is generalised not role specific. Rohit Sharma and Hardik Pandya are two different types of batsmen who belong to different roles in a batting lineup. Rohit Sharma opens the batting while Hardik Pandya comes in the middle or lower middle order. So, for Rohit Sharma priority should be to score runs at a steady rate without losing wickets, and for Hardik Pandya, the priority should be to quick runs at the end. The same goes for the bowlers who bowl overs in powerplay and bowlers who bowl towards the end of the innings. They both would have different priorities while bowling. Therefore, we cannot assign the same importance to certain features for all batters and all bowlers.

3.3.9 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.

3.3.9.1 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 15: Role-based feature importance by Random Forest (Batters)

Table 15 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.

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 16: Top 3 Ranks of each role generated by our Role-based ML Score (Batters)

Table 16 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.

3.3.9.2 Results for Bowlers

KPIs (Bowlers)	Feature Importance assigned by Random Forest according to the role			
	Specialist Bowlers (SB)	Short Performance Bowlers (SPB)	Impact Bowlers (IB)	Less Contribution Bowlers (LCB)
Average	0.092	0.128	0.213	0.072
Strike Rate	0.092	0.06	0.104	0.061
Economy	0.038	0.122	0.082	0.052
Balls Bowled per Innings	0.204	0.233	0.141	0.399
Wicket Index	0.279	0.141	0.104	0.156
Big Impact Index	0.027	0.013	0.005	0.01
Short Impact index	0.199	0.085	0.073	0.145
Runs Index	0.068	0.218	0.278	0.105

Table 17: Role-based feature importance by Random Forest (Bowlers)

Each KPI's relative importance is shown in Table 16 based on its function. Big Impact Index is given more weight for Specialist Bowlers (SB), who are the ones who typically take more than four wickets in an innings. For Short performance Bowlers (SPB), the economy and balls bowled per inning are given the most weight. Average, Strike Rate, and Runs Index have all been given more weight for Impact Bowler (IB), while Big Impact Index and Wicket Index have been given less importance than the importance given to the same KPIs in Specialist Bowler (SB) or Short Performance Bowlers (SPB). Less Contribution Bowlers (LCB) have the lowest importance in the Average and Big Impact Index among all the roles because they are expected to take wickets whenever they get the chance to bowl.

Player Name (Country)	**Role given by K-means**	**Preliminary score**	**Preliminary Rank**	**Score generated by a new metric**
L Balaji (IND)	SB	15.246	Good	0.785
T Natarajan (IND)	SB	22.123	Good	0.771
CK Langeveldt (SA)	SB	17.240	Good	0.770
JA Duffy (NZ)	SPB	14.400	Good	0.684
RL Chase (WI)	SPB	13.995	Good	0.683
Fawad Alam (PAK)	SPB	12.523	Good	0.674
M Theekshana (SL)	IB	29.377	Good	0.673
MH Yardy (ENG)	IB	29.455	Good	0.667
Sohail Tanvir (PAK)	IB	32.214	Good	0.653
SR Patel (ENG)	LCB	58.382	Average	0.491
Karim Sadiq (AFG)	LCB	39.468	Good	0.480
JH Kallis (SA)	LCB	33.439	Good	0.479

Table 18: Top 3 Ranks of each role generated by our Role-based ML Metric (Bowlers)

Table 18 provides insights of the score generated by Role-based Machine Learning (RBML) metrics; shows the top 3 bowlers from each role from the Top 10 teams in the ICC Men's T20 international team rankings. If we look at the PR rank generated initially for each bowler in table 18 is 'Good' apart from SR Patel (ENG), but we are rating by their specific roles. PR score and PR rank are based on a calculation using average, economy and wickets taken by the player. If we look at the Less Contribution Bowler (LCB) these are the players who are batsmen or batting-allrounder such as JH Kallis (SA). Specialist Bowlers (SB) are bowler who bowls their full quota of the overs and tries to take as many wickets as possible. L Balaji (IND) with the RBML score of 0.785 highest among the bowlers among all the bowlers from the Top 10 teams in the ICC Men's T20 international team rankings is one of the bowlers who picked two wickets on average. T Natarajan (IND) is one of the bowlers who has bowled his all quota of his overs in a short career he had having second-best RBML score of 0.771. Bowlers in SPB and IB are good bowlers, but all the bowlers have different qualities which this study has tried to explore and categorize through this research.

3.3.10 Principal Component Analysis (PCA)

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.

3.3.10.1 PCA For Batters

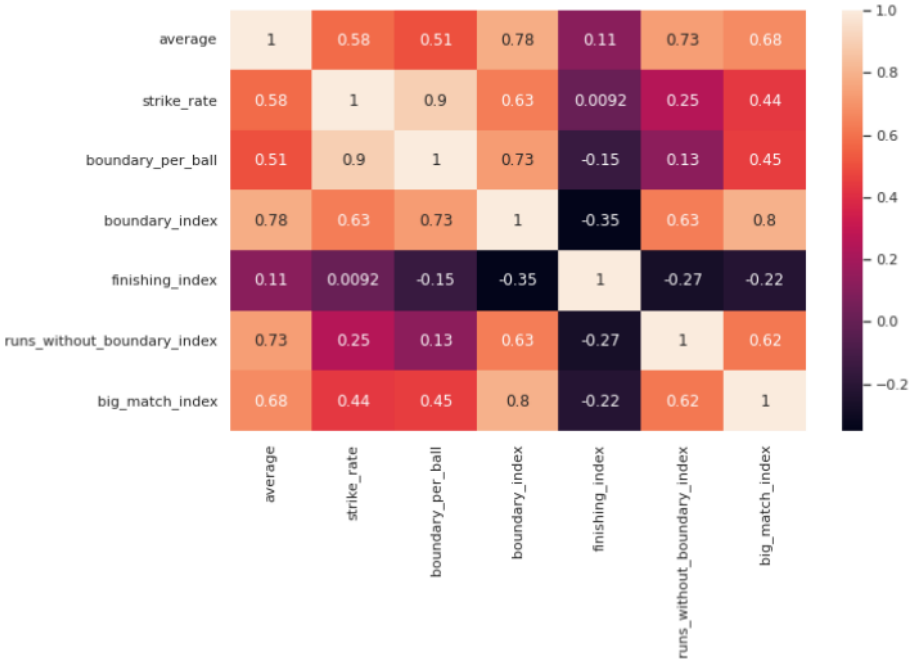

Figure 37: Correlation matrix for Batters data after standardisation

Figure 38: 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 37. 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 38 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 39. 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%.

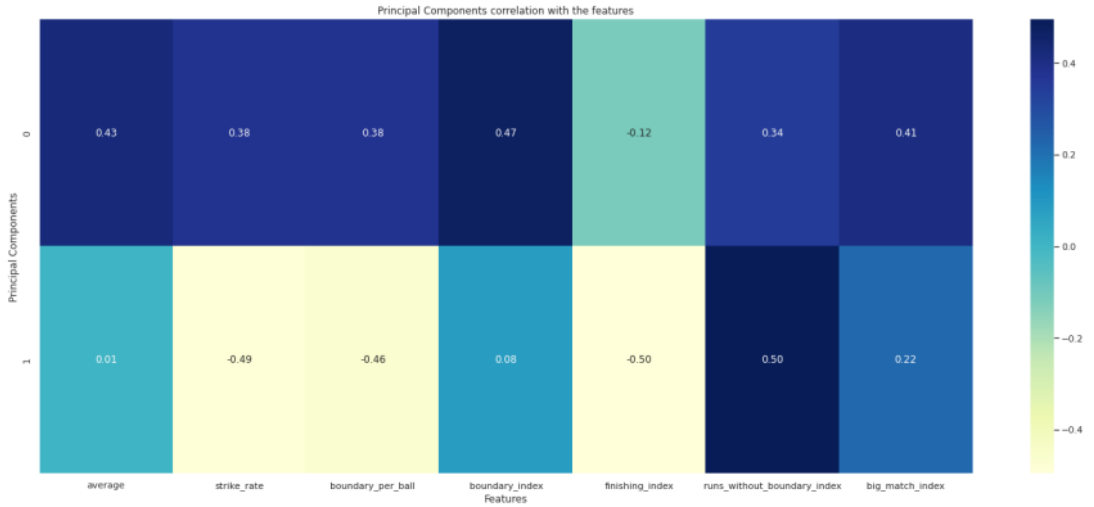

Figure 39: Each feature's coefficients for both PCA components (Batter)

3.3.10.2 PCA For Bowlers

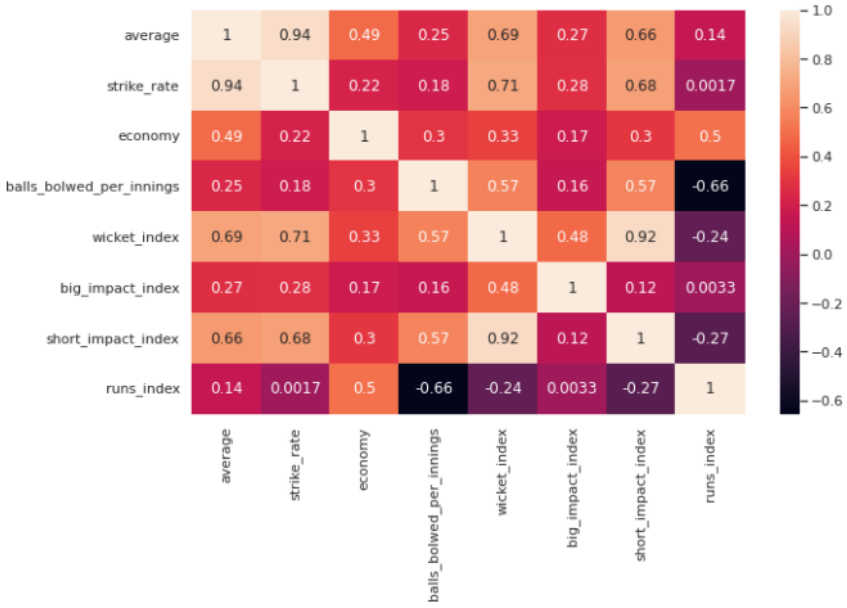

Figure 40: Correlation matrix for Bowler's data after Formula-Based Normalisation

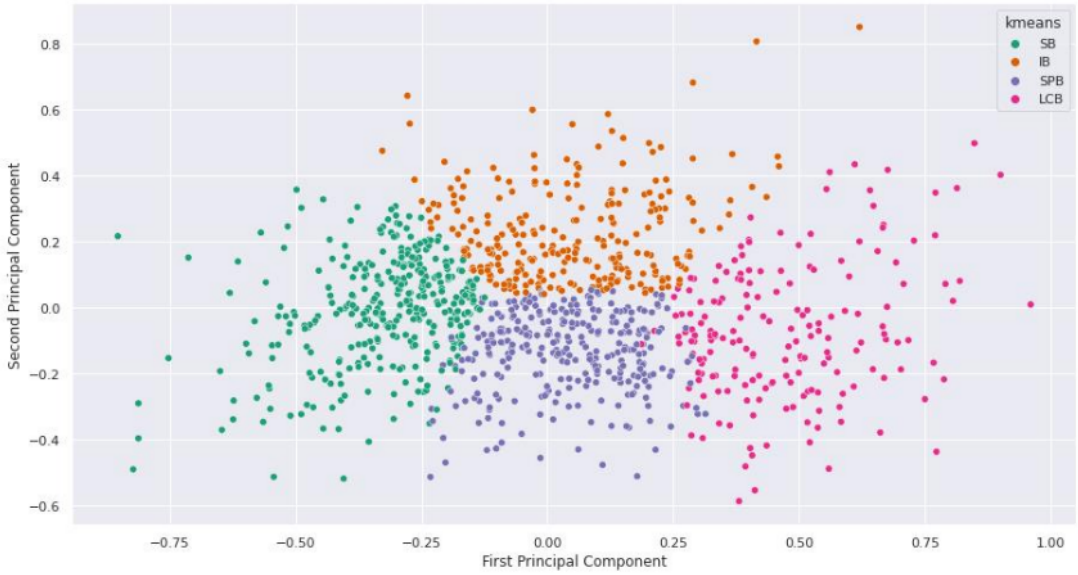

Figure 41: Scatter plot of PCA with points separated by role-based cluster (Bowler)

Figure 40 displays the correlation between the characteristics in the bowler's data. Since taking wickets is crucial in the game's shortest format, average, strike rate, and wicket index are all highly correlated. Balls Bowled per Innings are loosely correlated with the Runs Index as more balls a bowler balls he is likely to concede more runs. Short Performance Index (Wicket less than four in an innings) is strongly correlated with the Wicket index. Big Impact is also correlated with the Wicket index as both KPIs are related to taking wickets in the innings.

In this instance, PCA was used to reduce the number of dimensions in the data to only two. The scatter plot in figure 41 includes both PCA components, and the points are divided into groups according to the roles that the clustering assigned them. The lower the value of the points on the first component, the better the bowler, according to the scatter plot and coefficients of each component in figure 42 and good bowlers are expected to have lower strike rate, average and economy. We can see all the clusters in the plot, which also shows that K-Means did a good job of grouping the points for bowlers as well. The values of Variance Percentages (Total Variability percentage) for both the components are 52.28% and 24.10%. To get the final PCA score for bowlers we multiplied our original PCA score (KPIs and Coefficient of First Component) by -1 to get accurate results.

Figure 42: Each feature's coefficients for both PCA components (Bowler)

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 43: Top 15 Batters according to PCA score

name	country	average	strike_rate	economy	balls_bolwed_per_innings	wicket_index	big_impact_index	short_impact_index	runs_index	prelim_metric	pre-rank	kmeans_role	pca_score
T Natarajan	INDIA	17.42	13.7	7.62	24.000	1.750	0.000	1.750	30.500	22.123	Good	SB	1.712
CK Langeveldt	SA	14.17	11.6	7.30	22.000	1.889	0.111	1.625	26.778	17.240	Good	SB	1.678
Sayed Shirzad	AFG	16.57	12.8	7.73	22.500	1.750	0.000	1.750	29.000	21.348	Good	SB	1.656
Rashid Khan	AFG	13.00	12.7	6.12	22.741	1.787	0.098	1.509	23.246	13.260	Best	SB	1.643
L Balaji	INDIA	12.10	9.6	7.56	19.200	2.000	0.000	2.000	24.200	15.246	Good	SB	1.642
Kuldeep Yadav	INDIA	14.75	12.5	7.07	22.252	1.783	0.087	1.524	26.304	17.380	Best	SB	1.631
Zulfiqar Babar	PAK	15.41	13.0	7.11	22.286	1.714	0.000	1.714	26.429	18.261	Good	SB	1.626
PWH de Silva	SL	14.14	12.8	6.61	22.067	1.722	0.056	1.588	24.361	15.578	Best	SB	1.610
AB Dinda	INDIA	14.41	10.5	8.16	20.000	1.889	0.111	1.625	27.222	19.598	Good	SB	1.607
RP Meredith	AUS	23.50	14.1	9.98	22.200	1.600	0.000	1.600	37.600	39.088	Good	SB	1.593
LB Williams	SA	23.11	15.7	8.78	23.400	1.500	0.000	1.500	34.667	33.818	Good	SB	1.586
Fazalhaq Farooqi	AFG	14.88	16.0	5.58	24.000	1.500	0.000	1.500	22.333	13.838	Good	SB	1.578
RP Singh	INDIA	15.00	13.2	6.81	22.000	1.667	0.111	1.375	25.000	17.025	Good	SB	1.570
BAW Mendis	SL	14.42	13.4	6.45	22.662	1.692	0.128	1.235	24.410	15.502	Best	SB	1.570
Imran Tahir	SA	15.04	13.4	6.73	22.184	1.658	0.105	1.324	24.947	16.870	Best	SB	1.558

Figure 44: Top 15 Bowlers according to PCA score

Figures 43 and 44, show the top 15 batters and bowlers from the Top 10 teams in the ICC Men's T20 international team rankings. We can say that all the Specialist Batter and all the Specialist Bowler have got the higher PCA score. All the players who are shown in the figures above are either ‘Good’ or ‘Best’ in PR rank as well. When comparing players from the Top 10 teams in the ICC Men's T20 international team rankings, ML Hayden (AUS) has the highest RBML and PCA scores among batters, T Natarajan (IND) has the highest PCA and second highest RBML ranking among bowlers, and L Balaji from (IND) has the highest RBML ranking.

3.4 Results and Discussion

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.

3.4.1 For Batters

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 45, 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
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 45: Role given by clustering to Batsmen's Classified as Best by PR rank

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 46: PR rank of Batsmen's Classified as SB by k-means Algorithm

Figure 46 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 For Bowlers

In the beginning, this study kept 82 records for the bowlers separately containing all the PR ranks. In those 82 records, we had 43 bowlers classified as Good, 21 classified as Poor, 11 classified as Best, and seven classified as Average bowlers according to PR rank. After putting those 82 bowlers through our algorithm, we discovered that, following the roles suggested by clustering, 31 were Specialist Bowlers (SB), 20 were Low Contribution Bowlers (LCB), 20 were Impact Bowlers (IB), and 11 were categorised as Short Performance Bowler (SPB). Nine bowlers were categorised as LCB, but they were not categorised as 'poor' by PR rank most of them were categorised as 'Good' by PR rank. Ten bowlers were categorised as 'Poor' by PR rank, but they were not categorised LCB according to the model. The 31 bowlers classified as SB originally have 'Good' (22) and 'Best' (9) PR ranks. In figure 47, you can see the 11 bowlers who were given the PR rank of 'Best', nine of them are labelled as SB, one as SPB, and one as IB by our model. Among all the bowlers in this dataset classified as 'Best', DP Nannes has the best RBML score of 0.797 and PCA score of 1.788. while Shagharai Sefat has the best RBML score of 0.90 and PCA score of 1.853 and was classified 'Good' by PR rank and SB by K-Means cluster.

name	country	average	strike_rate	economy	balls_bolwed_per_innings	wicket_index	big_impact_index	short_impact_index	runs_index	prelim_metric	pre-rank	kmeans_role	rbml_score	pca_score
Mudassar Bukhari	NED	18.13	16.3	6.64	20.091	1.229	0.086	0.969	22.286	20.064	Best	SB	0.619	1.312
DP Nannes	AUS/NED	16.39	13.0	7.52	22.875	1.750	0.062	1.600	28.688	20.542	Best	SB	0.797	1.788
Waseem Abbas	MLT	18.94	15.3	7.39	21.429	1.393	0.071	1.154	26.393	23.328	Best	SB	0.667	1.470
Amir Hamza	AFG	25.06	22.2	6.77	20.182	0.909	0.000	0.909	22.788	28.276	Best	IB	0.590	1.076
Shadab Khan	PAK	21.79	18.3	7.11	22.330	1.217	0.033	1.121	26.517	25.821	Best	SB	0.630	1.393
Shoriful Islam	BAN	18.40	14.8	7.45	19.389	1.316	0.000	1.316	24.211	22.847	Best	SB	0.643	1.350
Umar Gul	PAK	16.97	14.1	7.19	20.030	1.417	0.100	1.093	24.050	20.336	Best	SB	0.663	1.429
J Botha	SA	22.24	20.9	6.37	19.846	0.949	0.000	0.949	21.103	23.611	Best	SPB	0.623	1.103
L Ngidi	SA	19.61	13.1	8.95	20.478	1.565	0.043	1.455	30.696	29.252	Best	SB	0.696	1.555
DT Johnston	IRE	19.87	18.5	6.42	21.214	1.143	0.036	1.037	22.714	21.261	Best	SB	0.612	1.299
AL Phehlukwayo	SA	21.15	15.1	8.40	18.825	1.250	0.031	1.161	26.438	29.610	Best	SB	0.594	1.269

Figure 47: Role given by clustering to Bowler's Classified as Best by PR rank

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 48: 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 48. 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.

ICC_rank	name	country	prelim_metric	prelim_rank	pre-rank	kmeans_role	rbml_score	RBML_rank	pca_score	PCA_rank
1	Umar Gul	PAK	20.336	5	Best	SB	0.663	5	1.429	5
2	S Badree	WI	21.667	7	Best	SB	0.620	9	1.309	9
3	DL Vettori	NZ	18.696	4	Best	SB	0.642	6	1.361	6
4	T Shamsi	SA	25.118	9	Best	SB	0.624	7	1.345	7
5	SP Narine	WI	21.285	6	Best	SB	0.621	8	1.314	8
6	Rashid Khan	AFG	13.260	1	Best	SB	0.769	1	1.643	1
7	Shahid Afridi	PAK	27.006	10	Best	SB	0.598	10	1.285	10
8	JR Hazlewood	AUS	21.684	8	Best	SB	0.693	4	1.530	4
9	PWH de Silva	SL	15.578	2	Best	SB	0.752	2	1.610	2
10	Imran Tahir	SA	16.870	3	Best	SB	0.723	3	1.558	3

Figure 49: Analysis of Top 10 ICC All-Time Bowlers

Figure 49 compares the Top 10 Bowlers from the ICC Men's All-Time T20I Rankings. All the bowlers fall under the category of Specialist Bowlers. Every bowler has received a "Best" rating based on PR rank. According to the various scores/indexes used in this study, Rashid Khan (AFG), who is the fastest player to take 100 T20 International wickets (Fastest to 100 wickets | ESPNCricinfo.com, 2022), has achieved the top ranking followed by PWH de Silva (SL) in second place. One thing to take note of is that eight of the top ten bowlers in the ICC Men's All-Time T20I Rankings are spinners, while the remaining two are pacers, showing that spinners, who typically bowl in the middle over of the game, have a better chance of changing the game on their own.

ICC_rank	name	country	batting_pre_rank	bowling_pre_rank	batting_role	bowling_role	combined_rbml_Score	RBML_rank	combined_pca_score	PCA_rank	hussey_index(bat)	hussey_rank
1	SR Watson	AUS	Best	Good	SB	IB	1.892	2	4.216	2	174.56	2
2	Shahid Afridi	PAK	Good	Best	F	SB	0.992	9	2.208	9	167.92	3
3	Shakib Al Hasan	BAN	Good	Best	FL	SB	1.054	7	2.591	7	142.21	10
4	GJ Maxwell	AUS	Best	Good	SB	LCB	1.895	1	4.268	1	184.96	1
5	Yuvraj Singh	INDIA	Best	Best	FL	SPB	1.354	5	3.097	4	164.40	6
6	ST Jayasuriya	SL	Best	Good	FL	SPB	1.172	6	2.861	6	152.44	7
7	Mohammad Nabi	AFG	Best	Good	FL	IB	1.004	8	2.352	8	165.20	5
8	DJ Hussey	AUS	Good	Good	FL	SPB	0.951	10	1.911	10	144.24	9
9	MN Samuels	WI	Best	Good	FL	SPB	1.382	4	2.977	5	145.52	8
10	CH Gayle	WI	Best	Good	SB	LCB	1.729	3	3.736	3	165.42	4

Figure 50: Analysis of Top 10 ICC All-Time All-Rounders

The Top 10 All-Rounders from the ICC Men's All-Time T20I Rankings are compared in Figure 50. We can see in that figure the batting and bowling roles that our player was assigned based on our research. Numerous batting and bowling combinations can be used to categorise an all-around player. As a Specialist Batter and Impact Bowler, SR Watson (AUS), who has the highest career rating of any All-Rounder in the game, qualifies as a Batting All-Rounder, which he was in fact but by RBML score (1.892), PCA score (4.216), and Hussey rank with a Hussey index of 174.56, he was placed in second place Our study classified GJ Maxwell (AUS) as a Specialist Batter and Low Contribution Bowler, but his combined RBML Score (1.895), PCA Score (4.268), and Hussey Index of 184.96 are the best. CH Gayle (WI), who has an equally impressive RBML and PCA score, is another player who has been assigned the same role as GJ Maxwell (AUS). There are also some bowling all-rounders, such as Shakib-Al-Hasan (BAN) and Mohammad Nabi (AFG) are categorised as Floaters when they bat and Specialist Bowler and Impact Bowler, respectively, when they bowl.

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 17th April 2022 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). The same goes for bowlers taking four wickets in four overs is quite tough. 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. When we tried to remove all the players with less than ten innings, we left with 400 data points combining both batters and bowlers.

4.2 Future Work

Player evaluation in Sports analytics is a continuous growing field with new advancements happening as many researchers and fans are working on this topic. This study has only considered role-based clustering and predicted the feature importance of the KPIs, but you cannot pick a player in your team only on past records. One must consider the recent form of the player while doing the player evaluation. Pitch also plays an important role in player performance as the player performs differently in home games and away games depending on pitch and weather conditions. Cricket is one of the most popular sports. In the future, the model developed by this study will remain valid, and one can improve its performance by obtaining more data with easy access to granularity.

References

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,

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.

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 2022].

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),.

Cricinfo. 2022. *Anthony de Mello Trophy, 2020/21 Cricket Team Records & Stats | ESPNCricinfo.com*. [online] Available at: <https://stats.ESPNCricinfo.com/ci/engine/records/bowling/most_wickets_career.html?id=13809;type=series> [Accessed 27 April 2022].

Cricinfo. 2022. Batting Records [online] Available at: <https://stats.ESPNCricinfo.com/ci/engine/stats/index.html?class=3;template=results;type=batting> [Accessed 27 April 2022].

Cricinfo. 2022. Bowling Records [online] Available at: <https://stats.ESPNCricinfo.com/ci/engine/stats/index.html?class=3;template=results;type=bowling> [Accessed 27 April 2022].

Cricinfo. 2022. *Fastest to 100 wickets | ESPNCricinfo.com*. [online] Available at: <https://stats.ESPNCricinfo.com/ci/content/records/1286053.html> [Accessed 15 September 2022].

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.

Davis, J., Perera, H. and Swartz, T., 2015. Player evaluation in Twenty20 cricket. Journal of Sports Analytics, 1(1), pp.19-31.

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

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.

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.

Dutt, A., 2021. A Step By Step Implementation of Principal Component Analysis. [online] Medium. Available at: <https://towardsdatascience.com/a-step-by-step-implementation-of-principal-component-analysis-5520cc6cd598> [Accessed 12 September 2022].

Ecosystem (LEDU), E., 2018. Understanding K-means Clustering in Machine Learning. [online] Medium. Available at: <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1> [Accessed 1 September 2022].

Eddie, 2021. Feature Scaling Techniques | Why Feature Scaling is Important. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/05/feature-scaling-techniques-in-python-a-complete-guide/> [Accessed 10 September 2022].

En.wikipedia.org. 2022. Batting order (cricket) - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Batting_order_(cricket)> [Accessed 2 September 2022].

En.wikipedia.org. 2022. Power play - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Power_play> [Accessed 2 September 2022].

En.wikipedia.org. 2022. Premier League - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Premier_League> [Accessed 2 September 2022].

En.wikipedia.org. 2022. ICC Men's T20I Team Rankings - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/ICC_Men%27s_T20I_Team_Rankings> [Accessed 8 September 2022].

En.wikipedia.org. 2022. List of International Cricket Council members - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/List_of_International_Cricket_Council_members> [Accessed 8 September 2022].

Géron, A., 2020. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. Beijing: O'Reilly, pp.192-200.

Guido, S. and Müller, A., 2016. Introduction to Machine Learning with Python. O'Reilly Media, pp.168-181.

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,.

Icc-cricket.com. 2022. ICC Men's T20I Team Rankings | ICC. [online] Available at: <https://www.icc-cricket.com/rankings/mens/team-rankings/t20i> [Accessed 8 September 2022].

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),

Kumar, S., 2021. Use Voting Classifier to improve the performance of your ML model. [online] Medium. Available at: <https://towardsdatascience.com/use-voting-classifier-to-improve-the-performance-of-your-ml-model-805345f9de0e> [Accessed 11 September 2022].

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).

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.

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).

Lemmer, H., 2008a. Measures of batting performance in a short series of cricket matches. South African Statistical Journal, 42(1): 83-105.

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).

Lewis, M., 2003. Moneyball. New York: W. W. Norton & Co.

Lindgren, I., 2020. Dealing with Highly Dimensional Data using Principal Component Analysis (PCA). [online] Medium. Available at: <https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6> [Accessed 7 September 2022].

Manage, A. and Scariano, S., 2013. An Introductory Application of Principal Components to Cricket Data. Journal of Statistics Education, 21(3).

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.

Medium. 2019. Why is scaling required in KNN and K-Means? | Medium. [online] Available at: <https://medium.com/analytics-vidhya/why-is-scaling-required-in-knn-and-k-means-8129e4d88ed7> [Accessed 11 September 2022].

Medium. 2020. Dealing with Highly Dimensional Data using Principal Component Analysis (PCA) | Towards Data Science. [online] Available at: <https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6> [Accessed 16 September 2022].

Medium. 2020. All about Feature Scaling | Towards Data Science. [online] Available at: <https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35> [Accessed 10 September 2022].

Medium. 2021. Beginner's Guide to XGBoost for Classification Problems | Medium. [online] Medium. Available at: <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390> [Accessed 11 September 2022].

Numpy.org. 2022. NumPy. [online] Available at: <https://numpy.org/> [Accessed 7 September 2022].

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),.

Paffenroth, R. and Kong, X., 2015. Python in Data Science Research and Education. Proceedings of the 14th Python in Science Conference,.

Pandas.pydata.org. 2022. pandas - Python Data Analysis Library. [online] Available at: <https://pandas.pydata.org/> [Accessed 7 September 2022].

Passi, K. and Pandey, N., 2018. Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning. Computer Science & Information Technology.

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,.

PyPI. 2022. beautifulsoup4. [online] Available at: <https://pypi.org/project/beautifulsoup4/> [Accessed 1 September 2022].

scikit-learn. 2022. scikit-learn | Classification Metrics and Evaluation. [online] Available at: <https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics> [Accessed 11 September 2022].

Thecricketcouch.com. 2012. A Simple Metric to Understand Batting Efficiency in the IPL | The Cricket Couch. [online] Available at: <http://thecricketcouch.com/blog/2012/06/15/a-simple-metric-to-understand-batting-efficiency-in-the-ipl/> [Accessed 2 September 2022].

Bibliography

Ahamed, F., 2022. Web scraping Cricinfo data. [online] Medium. Available at: <https://medium.com/swlh/web-scraping-cricinfo-data-c134fce79a33> [Accessed 14 May 2022].

Pandas.pydata.org. 2022. pandas.DataFrame.sample — pandas 1.4.4 documentation. [online] Available at: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html> [Accessed 9 September 2022].

Pandas.pydata.org. 2022. pandas.Series.map — pandas 1.4.4 documentation. [online] Available at: <https://pandas.pydata.org/docs/reference/api/pandas.Series.map.html> [Accessed 12 September 2022].

scikit-learn. 2022. 1.10. Decision Trees. [online] Available at: <https://scikit-learn.org/stable/modules/tree.html> [Accessed 11 September 2022].

scikit-learn. 2022. 1.11. Ensemble methods. [online] Available at: <https://scikit-learn.org/stable/modules/ensemble.html#voting-classifier> [Accessed 11 September 2022].

scikit-learn. 2022. 2.3. Clustering. [online] Available at: <https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation> [Accessed 11 September 2022].

scikit-learn. 2022. Importance of Feature Scaling. [online] Available at: <https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html#sphx-glr-auto-examples-preprocessing-plot-scaling-importance-py> [Accessed 12 September 2022].

scikit-learn. 2022. Selecting the number of clusters with silhouette analysis on KMeans clustering. [online] Available at: <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py> [Accessed 11 September 2022].

scikit-learn. 2022. sklearn.cluster.KMeans. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> [Accessed 11 September 2022].

scikit-learn. 2022. sklearn.ensemble.RandomForestClassifier. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [Accessed 11 September 2022].

Shmueli, B., 2022. Multi-Class Metrics Made Simple, Part II: the F1-score. [online] Medium. Available at: <https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1> [Accessed 12 September 2022].

Xgboost.readthedocs.io. 2022. Introduction to Boosted Trees — xgboost 1.6.2 documentation. [online] Available at: <https://xgboost.readthedocs.io/en/stable/tutorials/model.html> [Accessed 10 September 2022].

Appendices

	Span	Mat	Inns	NO	Runs	HS	Avg	BF	SR	100s	50s	0s	4s	6s
3rd position	2014-2018	2	2	0	**33**	18	16.50	25	132.00	0	0	0	2	2
4th position	2009-2021	57	57	8	**1315**	91	26.83	944	139.30	0	8	3	98	70
5th position	2009-2022	38	38	10	**984**	85*	35.14	739	133.15	0	6	0	75	43
6th position	2017-2022	9	9	3	**109**	26*	18.16	85	128.23	0	0	0	8	5
7th position	2021-2021	1	1	0	**17**	17	17.00	12	141.66	0	0	0	3	0

Fig 1: Position Wise Batting stats for EJ Morgan (ESPNCricinfo)

	Span	Mat	Inns	NO	Runs	HS	Avg	BF	SR	100s	50s	0s	4s	6s
1st position	2016-2016	1	1	1	**145**	145*	-	65	223.07	1	0	0	14	9
2nd position	2016-2016	1	1	0	**66**	66	66.00	29	227.58	0	1	0	7	4
3rd position	2012-2022	14	14	2	**212**	62	17.66	149	142.28	0	1	0	21	9
4th position	2014-2022	41	41	8	**1130**	113*	34.24	732	154.37	2	5	1	90	57
5th position	2013-2022	18	18	2	**403**	75	25.18	294	137.07	0	2	0	29	18
6th position	2012-2020	3	3	0	**57**	27	19.00	39	146.15	0	0	0	3	2
8th position	2012-2012	1	1	0	**4**	4	4.00	7	57.14	0	0	0	0	0

Fig 2: Position Wise Batting Stats for GJ Maxwell (ESPNCricinfo)

- **Technologies**

Pandas - Pandas is the most frequently used open-source Python tool for machine learning and data analytics. It's built on a NumPy module for working with multidimensional arrays (pandas - Python Data Analysis Library, 2022). Your data should be ready to feed into the model directly. To make your dataset ready for analysis a lot of work needs to be done on the pre-processing part. Building your dataset includes most of the work done on Data Wrangling and Data Cleaning.

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)

NumPy - NumPy is an open-source Python module for numerical activities that both simplify and complicate them. Complex arithmetic and vast datasets are required when working with machine learning and deep learning software. NumPy makes these operations easier and more effective to implement than its pure Python counterpart (NumPy, 2022).

Scikit-learn - Scikit-learn(sklearn) is an open-source and powerful library which is built on NumPy and is an efficient tool for predictive data analysis. In the study, we will use the Sklean library for K-Means clustering, Random Forest, XGBoost, Voting classification, feature extraction, and feature selection.

Matplotlib - Matplotlib is a Python library that lets you make static, animated, and interactive graphs. In this study, we will use Matplotlib to generate visualisations that will benefit us in extracting the data set's basic features and outliers.

Apart from Python libraries, we will use the following languages or Technologies / Software:

Tableau - Tableau is a well-known tool for business intelligence and data visualisation. Presenting your analyses' findings is one of any research's most crucial components. For that purpose, it is possible to produce lucrative dashboards that display the research's findings using tableau.

Same as calculating RBML and PCA scores, we calculated scores given by random forest, XGBoost and voting classifier using their generalised feature importance given in the classification section.

	name	country	rfc_score	xgb_score	voting_score	rbml_score	pca_score	kmeans_role
318	ML Hayden	AUS	2.193	2.248	2.224	2.903	5.862	SB
5	Babar Azam	PAK	1.859	1.999	1.987	2.569	5.485	SB
2	V Kohli	INDIA	1.914	1.950	1.971	2.494	5.252	SB
30	Mohammad Rizwan	PAK	1.805	1.833	1.858	2.421	5.001	SB
58	DJ Malan	ENG	1.752	1.867	1.849	2.324	5.019	SB
20	KL Rahul	INDIA	1.738	1.894	1.857	2.310	5.047	SB
420	MN van Wyk	SA	1.667	1.856	1.782	2.285	4.859	SB
148	DP Conway	NZ	1.825	1.755	1.823	2.193	4.736	SB
271	SA Yadav	INDIA	1.791	1.894	1.844	2.080	4.754	SB
284	A Symonds	AUS	1.869	1.803	1.849	1.939	4.560	SB

Fig 3: Top 10 Batters with all the scores

	name	country	rfc_score	xgb_score	voting_score	rbml_score	pca_score	kmeans_role
1293	VD Philander	SA	-1.523	-1.447	-1.472	-1.786	-3.388	LCB
1229	Shapoor Zadran	AFG	-1.308	-1.339	-1.365	-1.684	-3.655	LCB
1289	R Rampaul	WI	-1.199	-1.209	-1.223	-1.622	-3.771	LCB
1117	Abdur Razzak	BAN	-1.280	-1.253	-1.280	-1.603	-3.648	LCB
1259	Rubel Hossain	BAN	-1.196	-1.196	-1.226	-1.541	-3.731	LCB
1098	S Badree	WI	-1.047	-1.044	-1.066	-1.323	-3.292	LCB
1265	SS Cottrell	WI	-1.044	-1.066	-1.100	-1.296	-3.265	LCB
1253	MR Gillespie	NZ	-0.995	-0.973	-0.986	-1.283	-3.203	LCB
1247	Shafiul Islam	BAN	-1.132	-1.095	-1.129	-1.245	-2.777	LCB
1269	D Bishoo	WI	-0.947	-1.000	-1.014	-1.239	-2.844	LCB

Fig 4: Bottom 10 Batters with all the scores

	name	country	rfc_score	xgb_score	voting_score	rbml_score	pca_score	kmeans_role
467	L Balaji	INDIA	0.763	0.766	0.755	0.785	1.642	SB
653	T Natarajan	INDIA	0.751	0.732	0.740	0.771	1.712	SB
289	CK Langeveldt	SA	0.757	0.753	0.753	0.770	1.678	SB
2	Rashid Khan	AFG	0.777	0.770	0.775	0.769	1.643	SB
661	Sayed Shirzad	AFG	0.741	0.727	0.730	0.757	1.656	SB
431	Zulfiqar Babar	PAK	0.751	0.738	0.742	0.754	1.626	SB
80	Kuldeep Yadav	INDIA	0.749	0.742	0.745	0.753	1.631	SB
26	PWH de Silva	SL	0.758	0.748	0.752	0.752	1.610	SB
285	AB Dinda	INDIA	0.724	0.728	0.719	0.745	1.607	SB
510	Fazalhaq Farooqi	AFG	0.774	0.753	0.766	0.742	1.578	SB

Fig 5: Top 10 Bowlers with all the scores

	name	country	rfc_score	xgb_score	voting_score	rbml_score	pca_score	kmeans_role
1077	RG Sharma	INDIA	0.301	0.310	0.304	0.211	0.135	LCB
1070	JL Ontong	SA	0.314	0.329	0.317	0.247	0.324	LCB
1078	JT Smuts	SA	0.269	0.254	0.277	0.280	0.193	LCB
1055	Iftikhar Ahmed	PAK	0.355	0.357	0.359	0.290	0.290	LCB
1082	CL White	AUS	0.423	0.435	0.419	0.302	0.345	LCB
1045	CD Barnwell	WI	0.420	0.430	0.416	0.312	0.392	LCB
1068	Nazmul Hossain	BAN	0.392	0.398	0.390	0.316	0.420	LCB
1066	EMDY Munaweera	SL	0.366	0.361	0.364	0.322	0.402	LCB
861	Hussain Talat	PAK	0.497	0.531	0.493	0.334	0.535	LCB
940	GSNFG Jayasuriya	SL	0.398	0.396	0.396	0.349	0.493	LCB

Fig 6: Bottom 10 Bowlers with all the scores

We can see the main purpose of the role-based score by comparing the generalised score (rfc_score, xgb_score, voting_score) with the role-based (rbml) score. RBML score rewards good players and punishes less contribution players rather than giving generalised feature importance to each player.

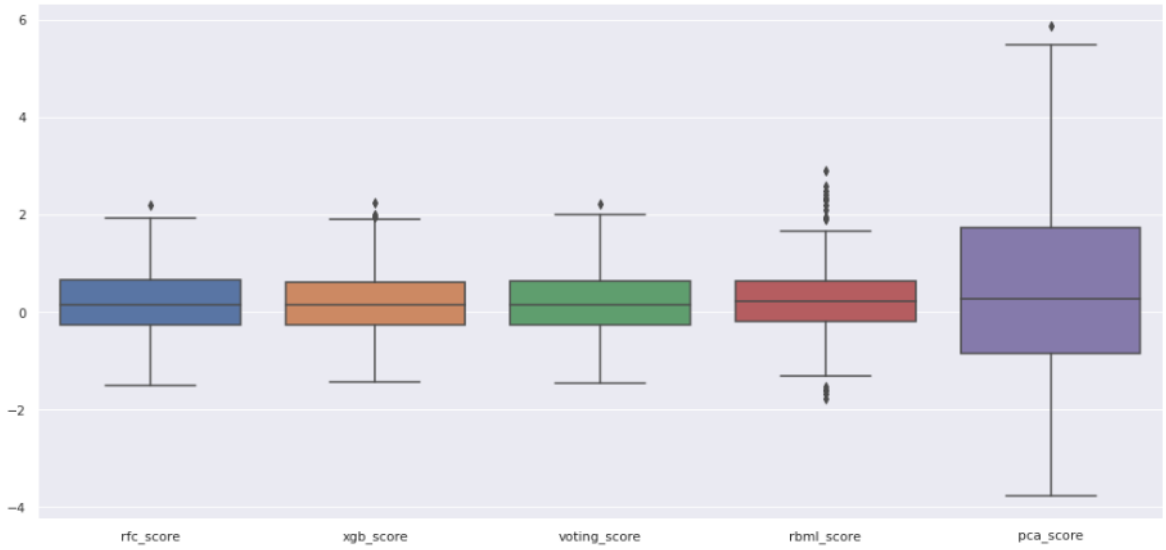

Fig 7: Box plots of all the scores (Batters)

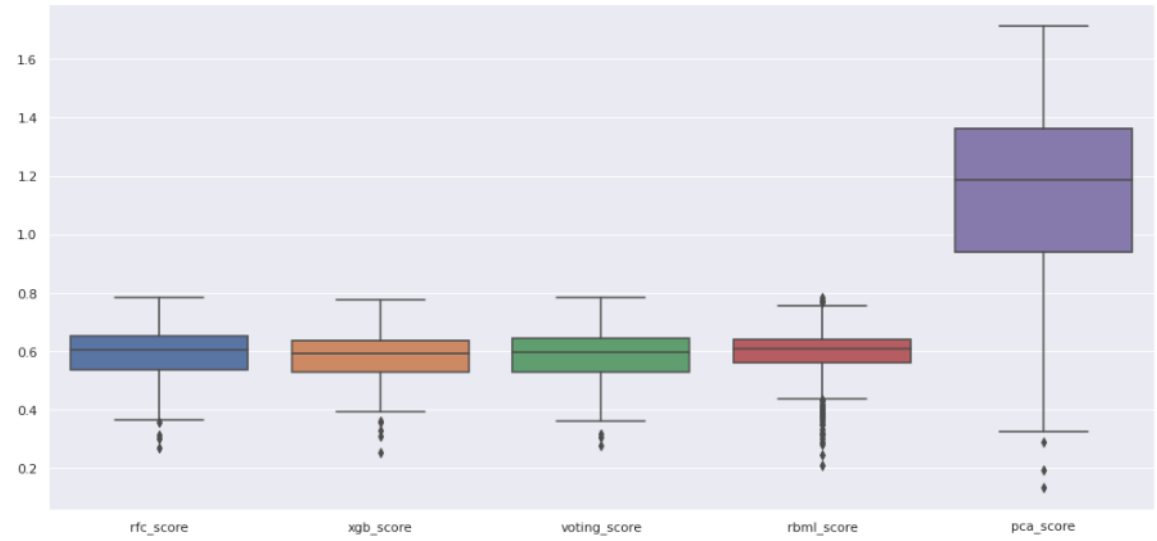

Fig 8: Box plots of all the scores (Bowlers)

K2136854_sports_analytics_dissertation.docx

ORIGINALITY REPORT

19%	13%	11%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	C. Deep Prakash, Sanjay Verma. "A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket", Decision Analytics Journal, 2022 Publication	2%
2	Submitted to Kingston University Student Paper	1%
3	en.wikipedia.org Internet Source	1%
4	www.ncbi.nlm.nih.gov Internet Source	1%
5	Submitted to Liverpool John Moores University Student Paper	1%
6	Submitted to Nottingham Trent University Student Paper	1%
7	content.iospress.com Internet Source	1%
8	pubsonline.informs.org Internet Source	<1%

9	www.researchgate.net Internet Source	<1%
10	Submitted to National College of Ireland Student Paper	<1%
11	Submitted to University of Lancaster Student Paper	<1%
12	Submitted to Loughborough University Student Paper	<1%
13	www.mathsportinternational.com Internet Source	<1%
14	Submitted to City University Student Paper	<1%
15	thecricketcouch.com Internet Source	<1%
16	Submitted to University of Sydney Student Paper	<1%
17	Submitted to Kaplan College Student Paper	<1%
18	Submitted to University of Northumbria at Newcastle Student Paper	<1%
19	dokumen.pub Internet Source	<1%
20	Submitted to Universidade Nova De Lisboa	

	Student Paper	<1%
21	Submitted to Coventry University Student Paper	<1%
22	towardsdatascience.com Internet Source	<1%
23	Submitted to Queensland University of Technology Student Paper	<1%
24	Submitted to University of Ulster Student Paper	<1%
25	mahamediaonline.com Internet Source	<1%
26	www.jaqm.ro Internet Source	<1%
27	Submitted to Asian Institute of Technology Student Paper	<1%
28	researchtrend.net Internet Source	<1%
29	trap.ncirl.ie Internet Source	<1%
30	Submitted to International School of Brno Student Paper	<1%

31	Jack Davis, Harsha Perera, Tim B. Swartz. "Player evaluation in Twenty20 cricket", Journal of Sports Analytics, 2015 Publication	<1 %
32	github.com Internet Source	<1 %
33	Submitted to RMIT University Student Paper	<1 %
34	Submitted to University of Hertfordshire Student Paper	<1 %
35	Submitted to University of Southampton Student Paper	<1 %
36	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
37	Submitted to University of Hull Student Paper	<1 %
38	Submitted to University of Pretoria Student Paper	<1 %
39	Submitted to University of Birmingham Student Paper	<1 %
40	www.mdpi.com Internet Source	<1 %
41	Submitted to 2642 Student Paper	<1 %

42	Submitted to De Montfort University Student Paper	<1 %
43	Submitted to University of Queensland Student Paper	<1 %
44	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
45	www.wikinfo.org Internet Source	<1 %
46	Submitted to Cardiff University Student Paper	<1 %
47	Submitted to University of Liverpool Student Paper	<1 %
48	tesi.luiss.it Internet Source	<1 %
49	core.ac.uk Internet Source	<1 %
50	Vladimir S. Lazarev. "Discipline Impact Factor: Some of Its History, Some of the Author's Experience of Its Application, the Continuing Reasons for Its Use and... Next Beyond", Journal of Data and Information Science, 2020 Publication	<1 %
51	Submitted to Middlesex University Student Paper	<1 %

52	Submitted to Bournemouth University Student Paper	<1%
53	Submitted to University of Bedfordshire Student Paper	<1%
54	repository.up.ac.za Internet Source	<1%
55	Submitted to University of Leeds Student Paper	<1%
56	cashassignment.com Internet Source	<1%
57	www.amstat.org Internet Source	<1%
58	Submitted to Stranmillis University College Student Paper	<1%
59	www.tandfonline.com Internet Source	<1%
60	Submitted to University of Technology, Sydney Student Paper	<1%
61	Chetan Kapadiya, Ankit Shah, Kinjal Adhvaryu, Pratik Barot. "Intelligent Cricket Team Selection by Predicting Individual Players' Performance using Efficient Machine Learning Technique", International Journal of Engineering and Advanced Technology, 2020	<1%

	Publication	
62	Punyaban Patel, Borra Sivaiah, Riyam Patel. "Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques", 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), 2022 Publication	<1 %
63	datagy.io Internet Source	<1 %
64	Ian G. McHale, Philip A. Scarf, David E. Folker. "On the Development of a Soccer Player Performance Rating System for the English Premier League", Interfaces, 2012 Publication	<1 %
65	Submitted to University of Portsmouth Student Paper	<1 %
66	machinelearningmastery.com Internet Source	<1 %
67	Submitted to University College London Student Paper	<1 %
68	artemis.cslab.ece.ntua.gr:8080 Internet Source	<1 %
69	Submitted to Lovely Professional University Student Paper	<1 %

70	Submitted to University of Greenwich Student Paper	<1 %
71	"Social Networking and Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1 %
72	Submitted to Chattahoochee High School Student Paper	<1 %
73	Submitted to University of Westminster Student Paper	<1 %
74	Submitted to Bond University Student Paper	<1 %
75	Submitted to Indian Institute of Technology, Kanpur Student Paper	<1 %
76	pakobserver.net Internet Source	<1 %
77	www.vyaparkendra.com Internet Source	<1 %
78	academic.sun.ac.za Internet Source	<1 %
79	ebin.pub Internet Source	<1 %
80	hdl.handle.net Internet Source	<1 %

81	mdpi-res.com Internet Source	<1 %
82	www.slideshare.net Internet Source	<1 %
83	jcreview.com Internet Source	<1 %
84	jssm.org Internet Source	<1 %
85	pdfs.semanticscholar.org Internet Source	<1 %
86	scholar.ufs.ac.za Internet Source	<1 %
87	silo.tips Internet Source	<1 %
88	velog.io Internet Source	<1 %
89	Dibyojyoti Bhattacharjee, Hemanta Saikia. "On Performance Measurement of Cricketers and Selecting an Optimum Balanced Team", International Journal of Performance Analysis in Sport, 2017 Publication	<1 %
90	Shanu Verma, Vivekanand Pandey, Millie Pant, Vaclav Snasel. "A Balanced Squad for Indian	<1 %

Premier League using Modified NSGA-II", IEEE Access, 2022
Publication

91	Tuck, I.D.. "The impact of water jet dredging for razor clams, Ensis spp., in a shallow sandy subtidal environment", Journal of Sea Research, 200002 Publication	<1 %
92	ficci.in Internet Source	<1 %
93	link.springer.com Internet Source	<1 %
94	researchspace.auckland.ac.nz Internet Source	<1 %
95	sportdocbox.com Internet Source	<1 %
96	timesofindia.indiatimes.com Internet Source	<1 %
97	www.european-language-grid.eu Internet Source	<1 %
98	"ICICCT 2019 – System Reliability, Quality Control, Safety, Maintenance and Management", Springer Science and Business Media LLC, 2020 Publication	<1 %

99 Indika Wickramasinghe. "Bowlers' Performances in 2013 Champions Trophy", Annals of Applied Sport Science, 2014 <1%
Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		