Predicting an Optimal Sri Lankan Cricket Team for ODI Matches According to the Nature of the Game

A dissertation submitted for the Degree of Master of Science in Computer Science

W.A.S.C. Perera

University of Colombo School of Computing

2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name:

Registration Number:

Index Number:

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

Certified by:

Supervisor Name:

Signature: Date:

Abstract

This paper focuses on predicting an optimal Sri Lankan cricket team for One Day International (ODI) matches according to the nature of the game. In general, the team selection process in One Day International is based on performance measures such as batting and bowling averages. These measures have several numbers of limitations. The number of runs scored by batsmen and wickets taken by bowlers serves as a natural way of quantifying the performance of a cricketer. However, the factors such as scoring runs against a strong bowling line-up or delivering a brilliant performance against a team with a strong batting line-up, etc. deserves more credit. This research presents a method of prediction by scanning the dependencies applied in the game such as the average performances of the players, the ground, the opposition team and the match outcome. Due to the complexity in the data set in size and the dimension, and analysis required, advanced analysis techniques such as Clustering and Association Rule Mining has been used to predict the players. The study concludes by predicting teams (eleven players per each match) for thirty-five matches played in between 2013-2018. The final outcome shows that the Sri Lankan cricket team can win the match with 88% by predicting players using our system.

Keywords: Association Rule Mining, Clustering, Cricket, Game conditions

Acknowledgment

First and foremost, I would like to thank my supervisor, Dr. H.A. Caldera for his assistance and feedback during the past few months. I consider myself enormously lucky that I had the opportunity to work with him.

In addition, I am grateful for the support of the academic and non-academic staff of the University of Colombo School of Computing, Sri Lanka.

I would like to extend my gratitude to all my colleague in Vavuniya Campus, University of Jaffna for giving me huge support during my master's degree.

Especially, I must express my enormous gratitude to my husband and my brother as they initially proposed this research when I had a little idea about the topic. This accomplishment would not have been possible without them. Finally, I would like to thank the rest of my family for their unwavering support and continuous encouragement throughout the research process.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction

Cricket is considered as the most popular game in Sri Lanka. The Sri Lankan national cricket team has gained vital importance and prestigious recognition in the country. The Board for Cricket Control (BCCSL) is the governing body for Sri Lankan cricket. The BCCSL operates the Sri Lankan national cricket team and first-class cricket within Sri Lanka [17].

It is the responsibility of the Sri Lankan Cricket Selection Committee to rank the players and select the national team as well as the required squads. The policy is to have an honest, open, transparent and consistent selection process that selectors, administrators, and players fully understand. Selectors are required to attend the first class, domestic One day and Twenty20 matches in order to determine the next representatives of the national team. The general criteria on which the selectors will be considered are, current form, past performances (batting average, strike rate, bowling average, and economy rate), balance of the team, health/fitness, contribution to the team environment and investing in youth development

This manual process consumes more effort and time. The selectors are not retained on a full-time basis. Though they make every effort, it is impossible to attend at all the matches or at every day of matches [4]. This could be a disadvantage for the players. In the other hand, there is a confusion with the transparency of the selection process in certain situations. More importantly, selectors just consider three or four factors only for the selection process. It is accepted that these measures have severe limitations in assessing the true performances of players. The number of runs scored by batsmen and wickets taken by bowlers serves as a natural way of quantifying the performance of a cricketer. It is accepted that these measures have several numbers of limitations in assessing the true performances of players. When selecting a team, consists of eleven players, plenty of information should be considered: the performances, the ground, the opposition team, the match outcome and etc. This each element can make a big difference to the final outcome.

1.2 Background

Cricket is the second most popular game in the world which has 2.5 billion fans approximately. Cricket is initiated in England in the 16th century and later spread to most of the Commonwealth countries. The governing body of cricket is the International Cricket Council (ICC) which controls the cricketing events around the globe. Although the ICC includes 104 member countries, only 12 countries with test status.

An oval-shaped playing field is used to play the game and does not define an exact size for it. The playing field contains a rectangular 22-yard area called the pitch which is in the middle. The main actions take place on the pitch [1]. A cricket pitch is showing in Figure 1.1.

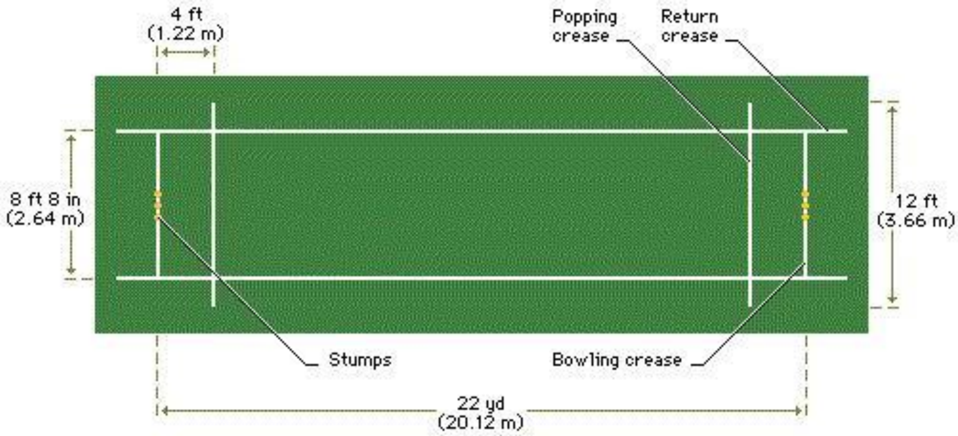

Figure 1.1: The layout of the pitch.

Generally, there are three forms of the game, Test, One Day International (ODI) and Twenty20 (T20). Test is the longest format of the game which lasts for five days involving 30-35 hours. Test is known as the standard format of cricket. But even after playing in five days, the match could end without any winner (draw). The Shorter format is the ODI, lasting almost 8 hours. One Day cricket is a form of limited overs cricket in which each team faces a maximum of 50 overs. The game needs an entirely different set of skills compared to the other formats. In here, players need to be very precise about their starting speed and should maintain it till the very end. During late 2000, the ICC introduced the shortest format called Twenty20 (later abbreviated to T20) cricket which lasts approximately for 3 hours. In Twenty20, each team has a single inning, which is restricted to twenty overs.

Other than the above-mentioned categories, there is another form known as First Class matches. Usually, first-class cricket is played at the national level between inter-zonal teams in order to get selected for the national team. According to the ICC definition, a cricket match is a first-class match if [16],

- the match is scheduled for three or four days
- Each team has eleven players
- each side may have two innings

- the match is played on natural, and not artificial, turf
- the match is played at a standard venue
- the match conforms to the Laws of Cricket
- ICC or the sport's governing body in the country recognizes the match as first-class.

1.3 Statement of the Problem

The research is focused on predicting an optimal Sri Lankan cricket team for One Day Internationals according to the game’s nature. When selecting a team, consists of eleven players, plenty of information should be considered: the performances, the layout of the ground, the opposition team, the match outcome and etc. This each element can make a big difference to the final outcome (win, lose or draw).

The layout of the ground is directly affected by team selection. A pitch consisted of loose clay or sand (dusty pitch) favors the spinners to get a good amount of spin and bounce from the pitches. Green pitch is a challenge for even the best batsmen as they have to judge the movement of the ball after pitching in a short time. Dead pitches favor batsmen a lot [7]. So, the state of the pitch is one of the primary considerations that should be taken into account.

More importantly, the team has an advantage when the pitch is domestically located. Researches have shown that a home-field is affected by the home teams to win 57% of all matches [28].

The team selection should be always depended on the opponent. Different players are familiar with opponent teams in a different manner. They show the performances against different teams in a different manner. The personal experiences of each player are highly affected to the above point.

Choosing a team is more than just picking the best players from a pool. The team should be balanced and the balance should reflect closely the tactics having for winning the match. So, there should be an advanced analysis technique that checks all the dependencies for the selection. The current process [4], which is based on a few factors, is not being able to produce a standard team since the most important factors are hidden. These concealed factors might be able to change the modern game strategies totally.

1.4 Aims and Objectives

The aim of this research is to develop a new model by scanning the dependencies applied in the game and select the best team to represent the country and support the overall development of the game indirectly.

The objectives of this research are,

- collect the dataset regarding the players performances and games' nature.
- Analyze the dataset by using the data mining algorithms.
- predict the optimal team according to the nature of the game.

1.5 Scope of the Research

The research considers only the Sri Lankan cricket players who played more than twenty matches as members of the Sri Lankan national cricket team (ODI matches) or who play for the first class matches domestic or internationally. Further, those who are retired from the game are not considered. Current players, who play for the matches at the end of 2018, are taken into account.

Only One Day International (ODI) cricket matches, played in between the year 2013-2018, are considered for both collecting data and predicting teams.

This research is based on the discipline of Data Mining which extracts or mines knowledge from large amounts of data and data is collected mainly from the espncricinfo (http://www.espncricinfo.com/) website.

1.6 Research Contribution

A certain number of researches have been done to analyze the performances by using data analyzing techniques. New performance measures have identified. Optimal batting orders were recognized. The match outcomes are predicted. But no research found to identify the new attributes and predict a team according to the condition of the match by utilizing the advanced analysis techniques.

Chapter 2: Literature Review

2.1 Introduction

As mentioned in the previous chapter, this research explores the prospective efficiency of advanced analysis techniques for the dimension of team selection. Several Studies that addressed different research issues related to various dimensions of the cricketing sport can be found in the literature. Some of these issues are analyzing about individual players performances, rating players, the ranking of teams, finding the best batsmen and bowlers, developing the strategies for winning games and tournaments, predicting the final outcome of a match, predicting an optimal team and etc.

2.2 Related Works

A comprehensive review of the literature regarding the performance analysis of the players reveals the following findings. Lemmer [9] has shown that, in order to be fair, the calculation formulas using for batting and bowling such as batting averages and bowling averages cannot be used in the case of a small number of matches played. Saikia and Bhattacharjee [18] compared the performance of both Indian and foreign cricketers in the Indian Premier League (IPL). They showed the differences between the player performance when they played the IPL and the national team. They have proposed a model by considering the characteristics such as the number of innings, the strike rate, and the batting average to measure the player performance. Both Van Staden [22] and Bracewell and Ruggiero [23] used graphical measures to illustrate player performances. These researches are based on some mathematical or graphical models. None of these researches have focused on the data mining technique for assessing the players' performances. However, Iyer and Sharda [21] used a neural network approach to predict each cricketer's future performance based upon their past performance.

In the literature of team selection, Thakare *et al.,* [8] followed the association rule mining for enhancing the team selection process by considering the attributes such as age, running capacity, experience, and Achievements. But these researchers have done their studies regarding the Handball. Sharp *et al.,* [19] quantifying a cricket player's performance based on his ability to score runs and take wickets. By using these performance measures, they have developed an integer program in order to determine the optimal team. Ahmed, Jindal, and Deb [20] proposed a method to select the team by using a binary integer programming method from the perspective of a multi-objective genetic optimization. Amin and Sharma [24] used data envelopment analysis techniques. The proposed DEA method can be used to select a national cricket team from club players or top-ranked players. None of these researches considered the game's nature. They have found an optimal team which is common to all the matches irrespective of the game's condition.

2.3 Summary

Cricket, as a research area, has continued to evolve as the game itself has evolved. Several kinds of research have been taken place to handle several research issues with regards to the game of cricket. But there are just a few of the publications which analyze cricket data by using the data mining techniques. The researches which predicted the optimal teams have considered the performances of players only. It was not being able to find any research which considers the nature of the game. So, this study focusses on fulfilling this research gap and develop a method by examining the factors which could affect the team selection.

Chapter 3: Methodology

3.1 Introduction

The comprehensive review of the literature showed that a large number of researches have been done to analyze the performances of players by using various techniques like machine learning, data mining, and mathematics. Some of the researches have identified certain factors that affect the players' selection or the match outcome. But no formal study found in predicting optimal cricket teams by considering the games' nature. As a solution, a new method has been developed in this research by using data analysis techniques. This chapter consists of the aspects relating to the proof of concept specification of the solution gained.

3.2 Data Mining Overview

Data Mining (Knowledge discovery) refers to the nontrivial extraction of implicit, previously unknown and potentially useful patterns, associations or interesting knowledge from data in databases. Several industries are maintaining their digital data which have plenty of hidden underlying information. The aim of data mining is to uncover this hidden information and provide the knowledge to the decision makers to make better decisions.

To apply the Data Mining technique, the data set should be complex enough. Complex data refer to the number of measurements, heterogeneity of measurement types, the complexity of descriptors, a variety of descriptor types, inability to formalize the data and etc.

Data Mining is a process that consists of iterative sequence steps as shown in Figure 3.1 [26]. The steps are,

1. Data cleaning (removing noise and inconsistent data)
2. Data integration (combining multiple data sources)
3. Data selection (retrieving relevant data from the database)
4. Data transformation (transforming data into forms appropriate for mining)
5. Data mining (applying intelligent methods to extract data patterns)
6. Pattern evaluation (identifying the interesting patterns)
7. Knowledge presentation (visualizing and representing techniques are used to present the mined knowledge)

Data Mining techniques can be categorized into two approaches as supervised learning and unsupervised learning. The supervised learning approach makes a prediction about data using known results found from different data. Supervised learning includes Classification, Regression, Time Series Analysis, and Prediction. Unsupervised learning identifies patterns and relationships in data by examining the existing properties of data. It includes Clustering, Summarization, Association Rules, and Sequence Discovery.

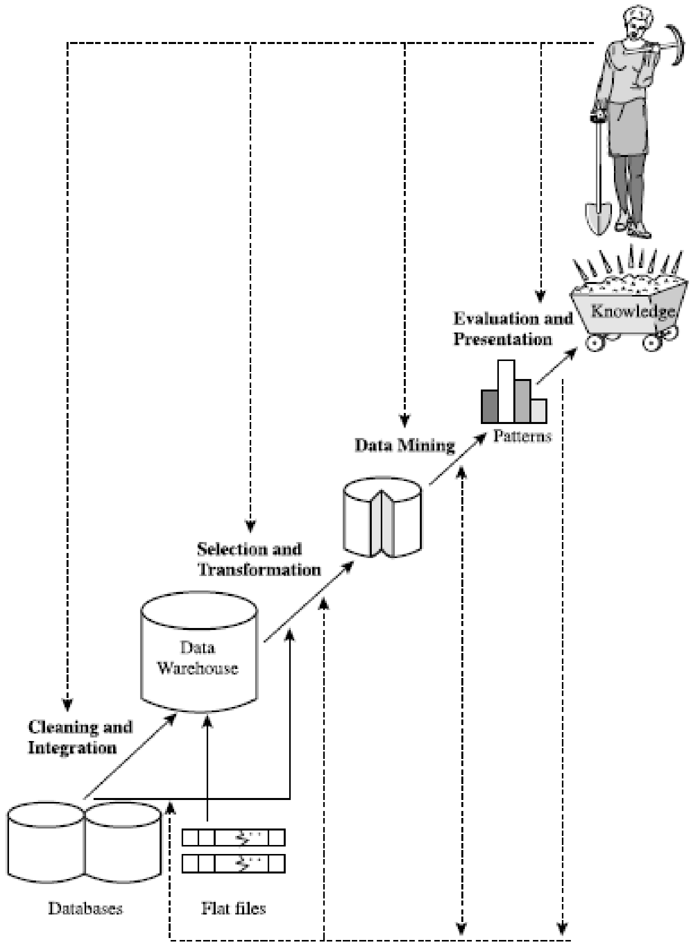

Figure 3.1: Data mining as a step in the process of knowledge discovery [26]

3.3 Complexity of the Problem

The dataset for analysis was collected from espncricinfo (http://www.espncricinfo.com/) website. The ODI matches played in between 2013 January to 2018 December were taken into account. 246 players (46 national players and 200 first class players) and 143 matches' data have been considered. Basically, four data sets have been collected throughout the research as shown in Table 3.1, Table 3.2, Table 3.3 and Table 3.4.

These all four data sets consist of both nominal and numeric data.

Name	Matches	Inns	Runs	Wkts	Ave	Econ	Age	Profile	5w
Akila Dhananjaya	24	22	1027	35	29.34	5.18	24	Allrounder	2
Angelo Mathews	201	154	3901	114	34.21	4.61	31	Allrounder	1
Asela Gunarathne	29	23	678	22	30.81	5.21	32	Batsmen	0
Ashan Priyanjan	23	14	233	5	46.6	5.27	29	Allrounder	0
Dhanushka Gunathilake	33	15	268	1	44.66	5.7	27	Allrounder	0
Dhananjaya De Silva	20	15	291	6	48.5	5.49	27	Allrounder	0
Dushmantha Chameera	20	20	678	15	45.2	5.41	26	Bowler	0
Jeewan Mendis	54	46	1134	28	40.5	5.08	35	Allrounder	0
Lahiru Thirimanne	117	4	94	3	31.33	5.42	29	Batsmen	0
Lasith Malinga	204	198	8705	301	28.92	5.31	35	Bowler	7
Milinda Siriwardhana	26	20	530	9	58.88	5.39	32	Allrounder	0
Nuwan Kulasekera	184	181	6751	199	33.92	4.9	36	Bowler	1
Nuwan Pradeep	28	26	1287	33	39	5.94	31	Bowler	0
Seekkuge Prasanna	38	37	1673	32	52.28	5.41	27	Allrounder	0
Suranga Lakmal	80	78	3275	104	31.49	5.41	31	Bowler	0
Thisara Perera	138	131	4822	156	30.91	5.8	29	Allrounder	3
Chamara Silva	75	2	33	1	33	4.71	39	Batsmen	0

.

Table 3.1: A sample data set which shows the bowling statistics

Name	Matches	Inns	Runs	HighScore	Ave	StrikeRate	Age	Profile	50+100
Akila Dhananjaya	24	19	217	50	13.56	64.01	24	Allrounder	1
Angelo Mathews	201	171	5342	139	42.73	83.75	31	Allrounder	39
Asela Gunarathne	29	23	558	114	29.36	80.17	32	Batsmen	2
Ashan Priyanjan	23	20	420	74	23.33	80.45	29	Allrounder	2
Dhanushka Gunathilake	33	32	957	116	30.87	85.9	27	Allrounder	8
Dhananjaya De Silva	20	19	459	84	27	80.8	27	Allrounder	4
Dinesh Chandimal	139	126	3433	111	32.69	73.82	28	Batsmen	25
Dushmantha Chameera	20	12	90	19	15	65.69	26	Bowler	0
Jeewan Mendis	54	40	604	72	20.13	85.07	35	Allrounder	1
Kusal Mandis	49	47	1325	102	29.44	85.2	23	Batsmen	12
Kusal Perera	78	75	2035	135	29.07	91.66	28	Batsmen	14
Lahiru Thirimanne	117	97	2946	139	34.65	71.33	29	Batsmen	24
Lasith Malinga	204	102	496	56	6.98	75.84	35	Bowler	1
Milinda Siriwardhana	26	23	513	66	23.31	98.46	32	Allrounder	3
Niroshan Dikwella	41	39	1232	116	32.42	90.45	25	Batsmen	8
Nuwan Kulasekera	184	123	1327	73	15.43	81.46	36	Bowler	4
Nuwan Pradeep	28	12	18	7	4.5	34.61	31	Bowler	0

Table 3.2: A sample data set which shows the batting statistics

Table 3.1 consists of the bowling statistics of the players. It includes ten attributes and seven of them are directly related to the bowling performances. Table 3.2 consists of the batting statistics of the players. It also includes ten attributes and seven of them are related to the batting performances directly.

Player	Opposition Team	First batted team	Ground	Runs	Balls	SR	Winner	Performance
Mathewes	Afghanistan	opp	Sheikh Zayed Stadium, Abu Dhabi	22	39	56.4103	opp	low
Mathewes	Australia	opp	Brisbane Cricket Ground, Woolloongabba, Brisbane	0	1	0	SL	low
Mathewes	Australia	opp	Melbourne Cricket Ground	12	14	85.7143	opp	low
Mathewes	Bangladesh	opp	Dubai International Cricket Stadium	16	34	47.0588	opp	low
Mathewes	Bangladesh	opp	Shere Bangla National Stadium, Mirpur, Dhaka	20	26	76.9231	SL	low
Mathewes	England	opp	Kennington Oval, London	18	21	85.7143	opp	low
Mathewes	England	opp	Sophia Gardens, Cardiff	13	15	86.6667	opp	low
Mathewes	India	opp	Barabati Stadium, Cuttack	23	32	71.875	opp	low
Mathewes	India	opp	Brabourne Stadium, Mumbai	3	14	21.4286	opp	low
Mathewes	India	opp	Himachal Pradesh Cricket Association Stadium, Dharamsala	25	42	59.5238	SL	low
Mathewes	India	opp	Khan Shaheb Osman Ali Stadium, Fatullah	6	18	33.3333	SL	low
Mathewes	India	opp	Queen's Park Oval, Port of Spain, Trinidad	10	11	90.9091	opp	low
Mathewes	Pakistan	opp	Pallekele International Cricket Stadium	8	11	72.7273	SL	low
Mathewes	Pakistan	opp	R.Premadasa Stadium, Khettarama, Colombo	4	13	30.7692	opp	low
Mathewes	Pakistan	opp	Rangiri Dambulla International Stadium	0	1	0	SL	low
Mathewes	Pakistan	opp	Sharjah Cricket Stadium	31	32	96.875	opp	low
Mathewes	Pakistan	opp	Sheikh Zayed Stadium, Abu Dhabi	8	31	25.8065	SL	low
Mathewes	Pakistan	opp	Shere Bangla National Stadium, Mirpur, Dhaka	16	13	123.077	SL	low

Table 3.3: A sample data set which shows the match statistics of each batsman

Player	Opposition Team	First Batted team	Ground	Overs	Runs	Wickets	ER	Winner	Performance
Malinga	Afghanistan	opp	Sheikh Zayed Stadium, Abu Dhabi	10	66	1	6.6	opp	low
Malinga	Australia	opp	Melbourne Cricket Ground	10	61	1	6.1	opp	low
Malinga	Australia	opp	Sydney Cricket Ground	10	59	2	5.9	opp	low
Malinga	England	opp	Kennington Oval, London	8	71	0	8.875	opp	low
Malinga	England	opp	Westpac Stadium, Wellington	10	63	1	6.3	SL	low
Malinga	England	SL	Pallekele International Cricket Stadium	4	39	0	9.75	opp	low
Malinga	India	SL	Sophia Gardens, Cardiff	8	54	0	6.75	opp	low
Malinga	India	SL	Sophia Gardens, Cardiff	10	70	2	7	opp	low
Malinga	India	SL	Rangiri Dambulla International Stadium	8	52	0	6.5	opp	low
Malinga	India	SL	Pallekele International Cricket Stadium	8	49	0	6.125	opp	low
Malinga	India	opp	R.Premadasa Stadium, Khettarama, Colombo	10	82	1	8.2	opp	low
Malinga	New Zealand	opp	Hagley Oval, Christchurch	10	84	0	8.4	opp	low
Malinga	New Zealand	SL	Mahinda Rajapaksa International Cricket Stadium, Sooriyawewa, Hambantota	5	42	0	8.4	opp	low
Malinga	New Zealand	SL	Rangiri Dambulla International Stadium	3	31	0	10.3333	SL	low
Malinga	Pakistan	opp	Sharjah Cricket Stadium	10	59	0	5.9	opp	low
Malinga	Pakistan	opp	Dubai International Cricket Stadium	10	78	1	7.8	SL	low
Malinga	Pakistan	SL	Mahinda Rajapaksa International Cricket Stadium, Sooriyawewa, Hambantota	9	60	1	6.66667	opp	low
Malinga	Pakistan	SL	Rangiri Dambulla International Stadium	5	86	0	17.2	opp	low

Table 3.4: A sample data set which shows the match statistics of each bowler

Table 3.3 and Table 3.4 is about the match statistics of the batsmen and bowlers respectively. These tables show the statistics of the nature of each game and the performances of each player in a particular match. Table 3.3 consists of eight attributes and Table 3.4 with nine attributes. Commonly, both tables show the data of the opposition team, first batted team, ground and the winner of each match. Additionally, Table 3.3 consists the data of the runs achieved, balls faced and the strike rate of each batsmen in each match. Table 3.4 consists of the data of number of overs balled, runs conceded, wickets taken and the economy rate of each bowler in each match.

As mentioned earlier, Data Mining is used to find hidden knowledge from complex databases. Therefore, by considering the complexity associated with both the size of the data sets and the content of the data sets, it was decided to apply the Data Mining technique on them.

3.4 Overview of Methodology

As shown in Figure 3.2 and Figure 3.3, the entire research can be mainly divided into two stages. Initially, the data sets contain the (refer Table 3.1 and Table 3.2) basic details of current national players and first-class players. As the first stage, these two data sets are analyzed based on the average performances of each player. As a result, the pool of the most talented batsmen and bowlers can be obtained. This pool again has been analyzed based on the nature of games. For that, another two data sets (refer to Table 3.2 and Table 3.3) have been used. These two data sets contain the selected players' performances in each match. By using the results achieved from the second stage, the optimal teams have been predicted according to the nature of games.

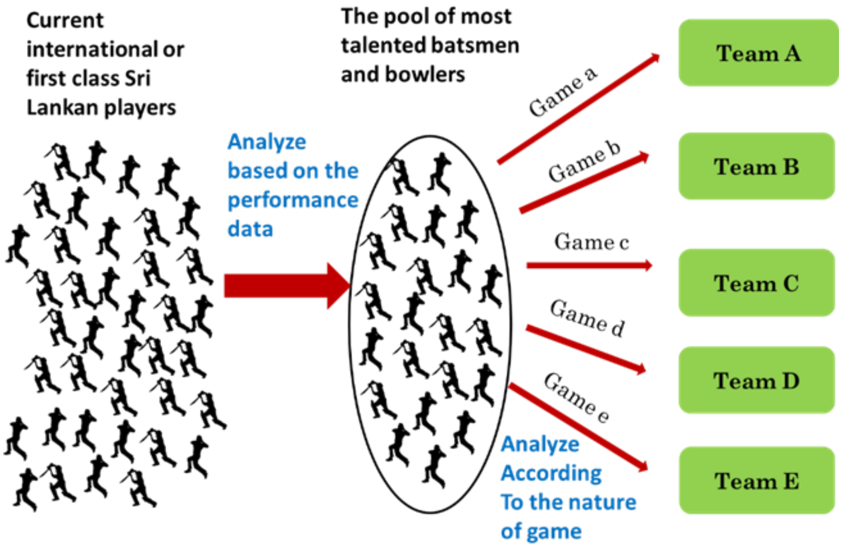

Figure 3.2: An overall image of the research

3.5 Performance Analysis

The initial stage of the study has considered two data sets (for batsmen and bowlers) as shown in Table 3.1 and Table 3.2. Since the clustering algorithm is able to identify the different groups in a dataset, it was decided to conduct the cluster analysis to identify the pool of best batsmen and bowlers. Clustering based on the SimpleKMeans clustering algorithm in WEKA is used as the number of clusters is not clear in this stage.

The biggest challenge with K-Means clustering is to find the most optimal number of clusters. The Elbow finding technique was decided to use for this purpose [27]. In this technique, the sum of squared error is calculated for the different values of k (number of clusters) and select the k value at the elbow as the optimum k.

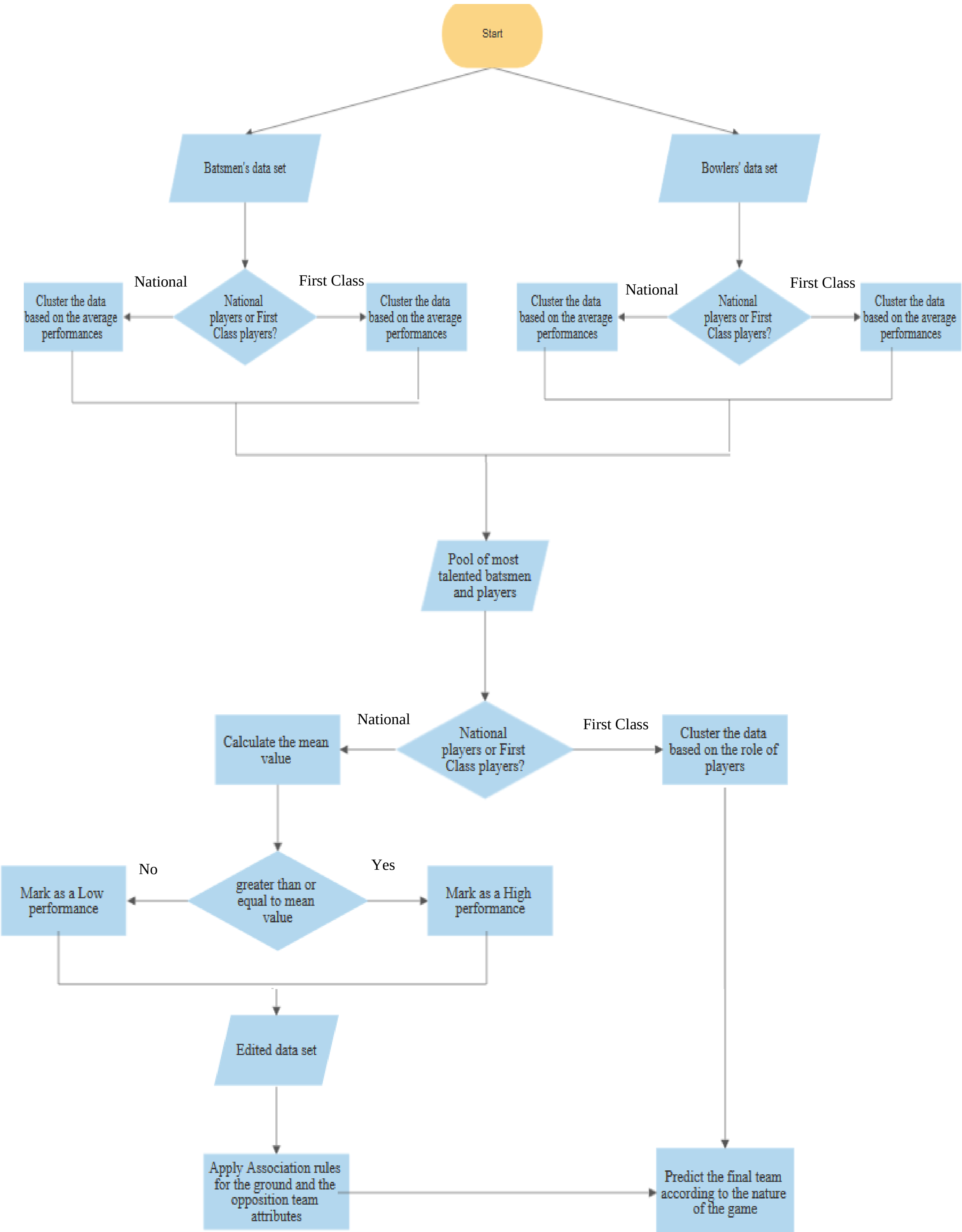

Figure 3.3: The flow chart of the overall research

As illustrated above, this stage is about finding the pool of best batsmen and bowlers. For that, the study decided to consider two batting statistics often used as a measure of a player's performance, the batting average and the batting strike rate. These measures can be defined as follows:

Batting Average=The total number of runs ÷Total number of innings in which the batsman was out

and

Batting Strike Rate=The average number of runs scored per 100 balls faced

According to the elbow finding technique, the abrupt change occurs at four for the batsmen's data set, as shown in Figure 3.4. So, it was decided to use four as the optimized number of clusters that should be used to analyze the performances of batsmen.

Since the national players' performances and first-class players performances cannot be comparable, the analysis was carried out independently for each of those two. So, two cluster analysis was performed against the batsmen's data set by using the Batting Average and the Batting Strike Rate attributes as shown in Figure 3.6 and Figure 3.7.

Figure 3.6 shows the clusters of national batsmen's Batting Strike Rate based on Batting Average. A batsman with high values of batting average and batting strike rate is considered to be a good player [19]. Based on that concept, it was decided to select the batsmen who are within cluster 1. Other clusters were ignored. So, nine players have been selected for the best players' pool as batsmen.

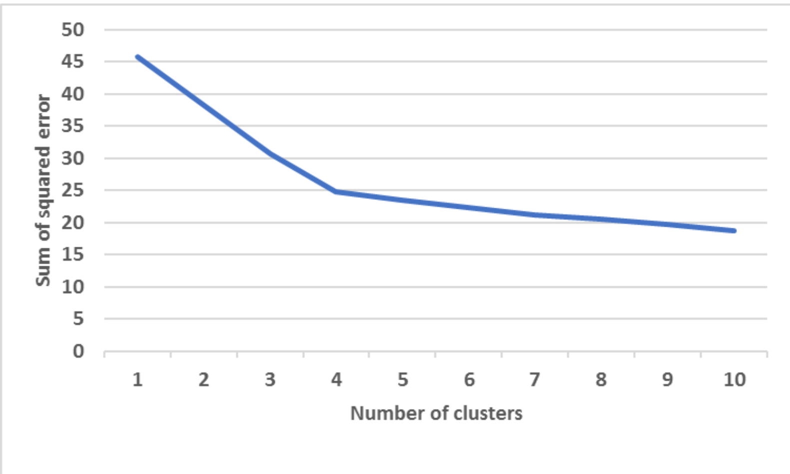

Figure 3.4: Number of clusters vs. Sum of squared error for the Batsmen's data set

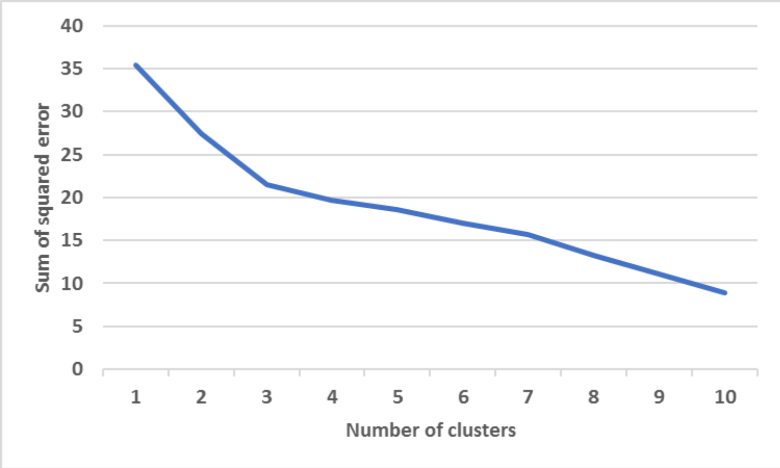

Figure 3.5: Number of clusters vs. Sum of squared error for the Bowler's data set

Figure 3.7 shows the clusters of first-class batsmen's Batting Strike Rate based on Batting Average. As mentioned above a batsman with high values of batting average and batting strike rate is considered to be a good player. Based on that, cluster 0 has been selected. So, six players have been selected for the best players' pool as batsmen. Other players were ignored.

Finally, altogether 15 players have been selected for the best players pool as batsmen as shown in Table 3.5.

To measure the bowling performances, the study decided to consider two bowling statistics often used as a measure of a player's performance, bowling average and the bowling economy rate. These statistics can be defined as follows:

$$Bowling\ Average = The\ average\ number\ of\ runs\ conceded\ per\ wicket$$

and

$$Bowling\ Economy\ Rate = The\ average\ number\ of\ runs\ \ conceded\ per\ over$$

According to the elbow finding technique, the abrupt change occurs at three for the bowlers' data set, as shown in Figure 3.5. So, it was decided to use three as the optimized number of clusters that should be used to analyze the performances of bowlers.

As mentioned earlier, the national players' performances and first-class players performances cannot be comparable. So, the analysis was carried out independently for each of those two. Two cluster analysis was performed against the bowlers' data set by using the Bowling Average and the Bowling Economy Rate attributes as shown in Figure 3.8 and Figure 3.9.

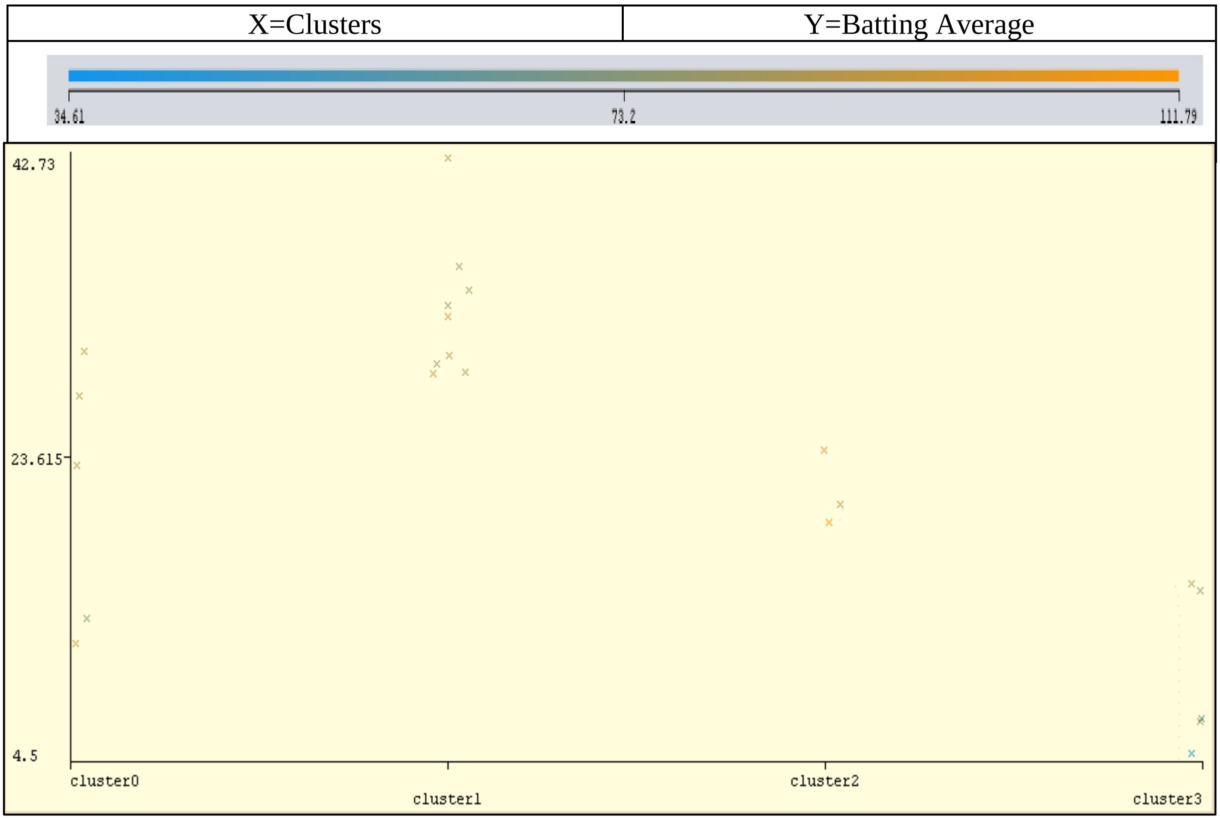

Figure 3.6: Clusters of national players' Batting Strike Rate based on Batting Average.

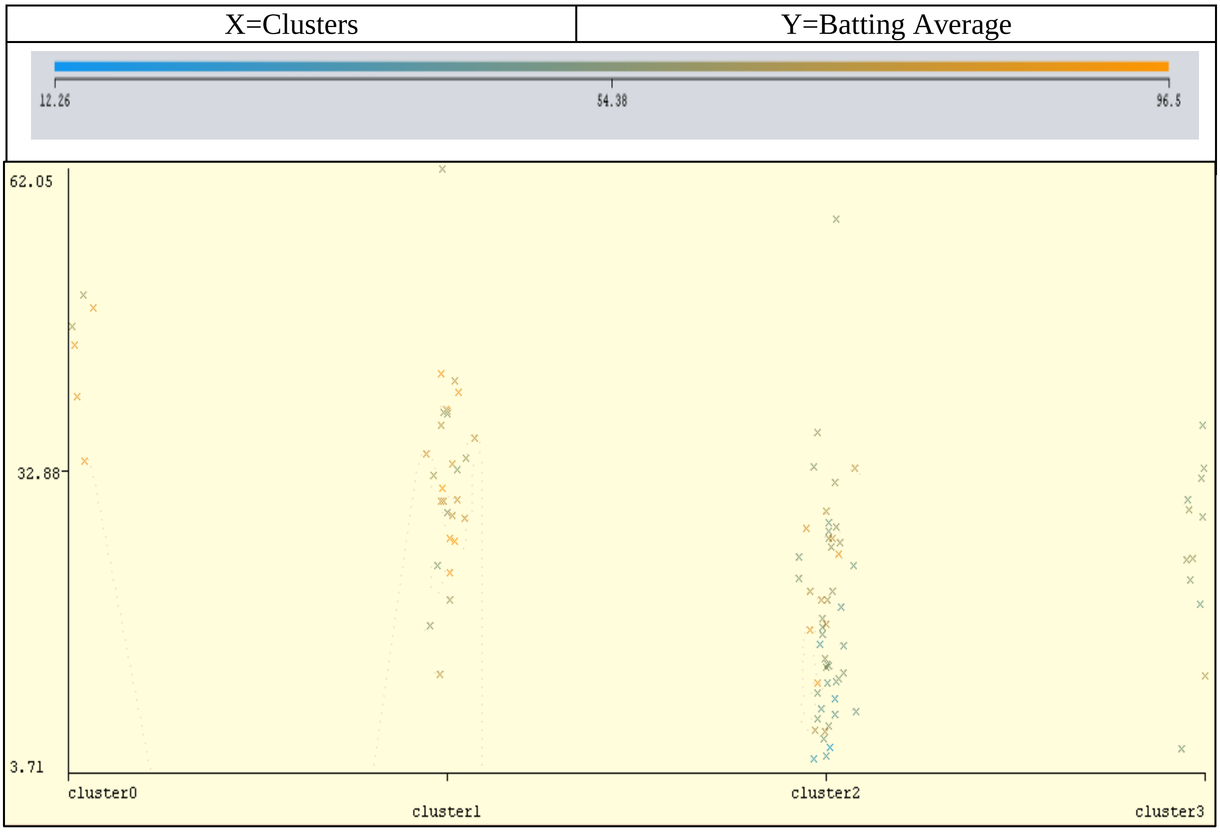

Figure 3.7: Clusters of first-class players' Batting Strike Rate based on Batting Average.

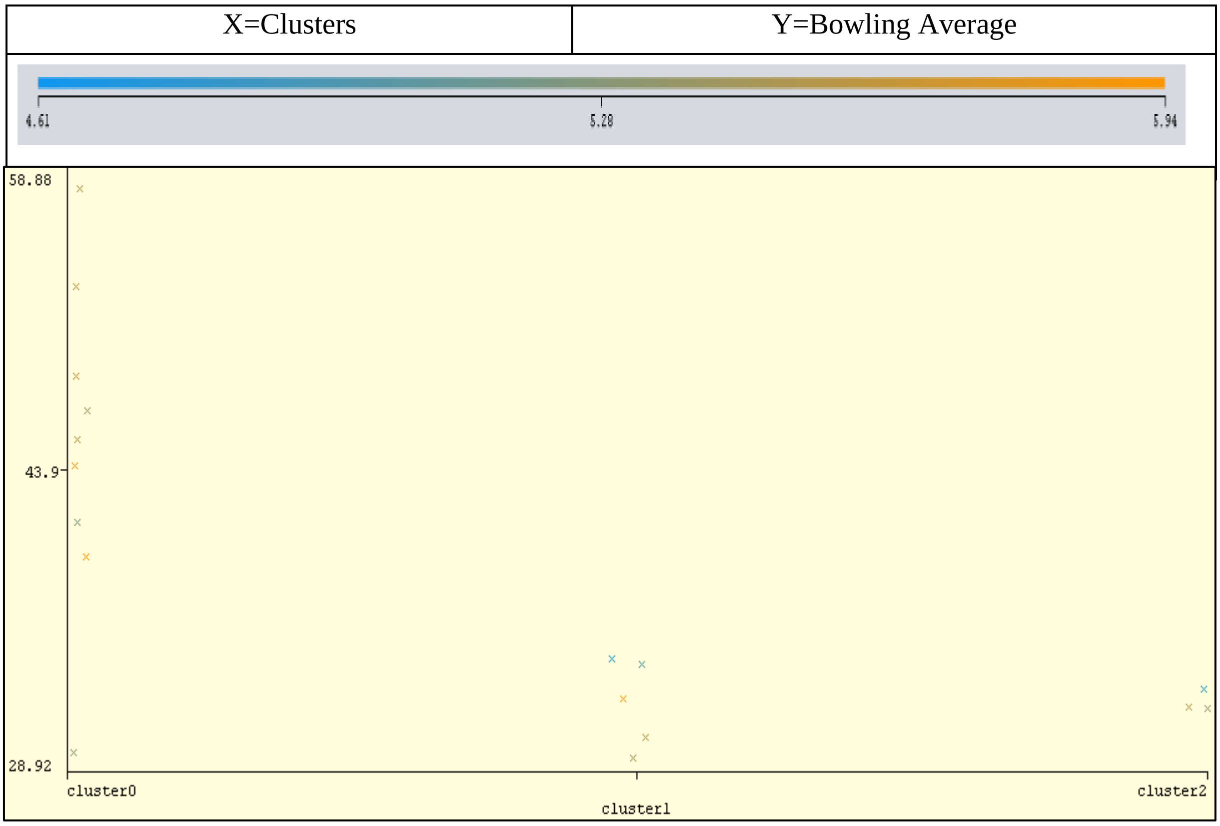

Figure 3.8: Clusters of national players' Bowling Economy Rate based on Bowling Average

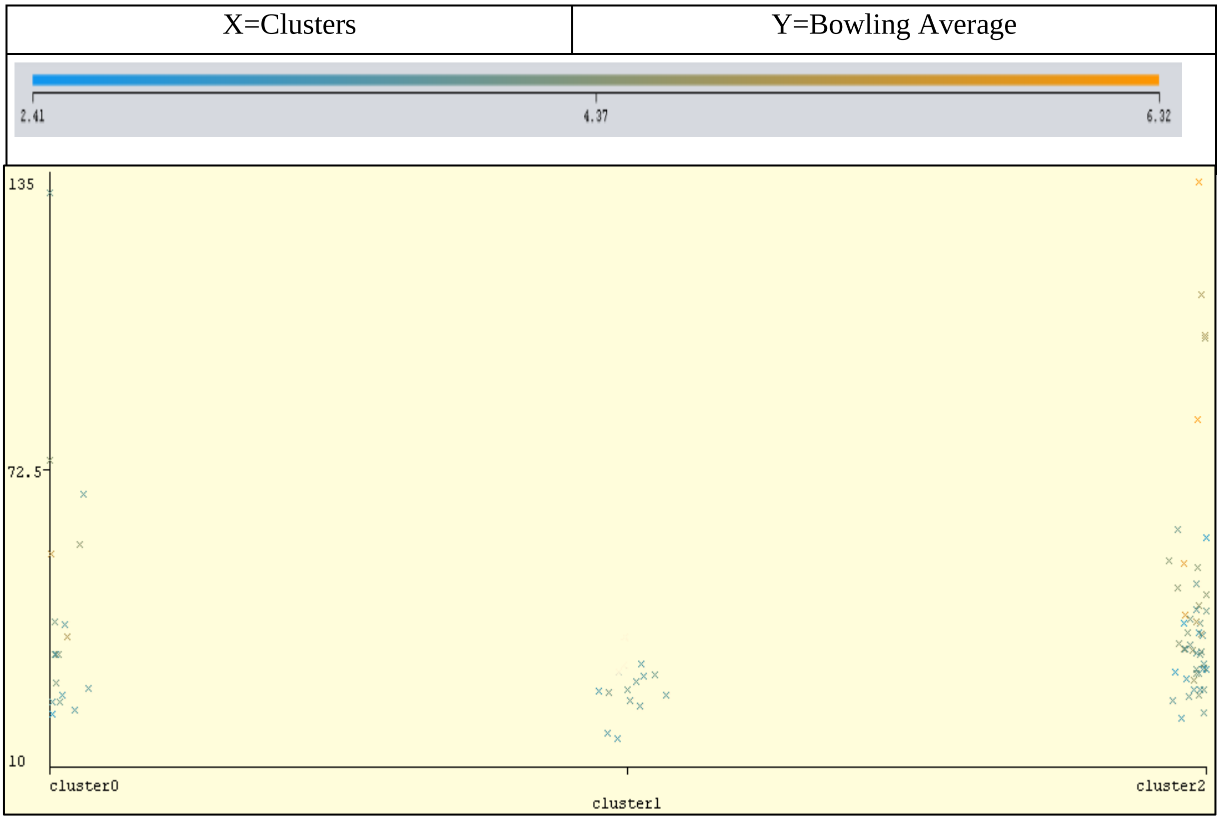

Figure 3.9: Clusters of first-class players' Bowling Economy Rate based on Bowling Average.

Figure 3.8 shows the clusters of national bowlers' Bowling Economy Rate based on Bowling Average. A bowler with low values for bowling average and bowling economy rate is considered to be a good player [19]. Based on that concept, it was decided to select the bowlers who are within cluster 1. Other clusters were ignored. So, five players have been selected for the best players' pool as bowlers.

Figure 3.9 shows the clusters of first-class bowlers' Bowling Economy Rate based on Bowling Average. As mentioned above, a bowler with low values of bowling average and bowling economy rate is considered to be a good player. Based on that, cluster 1 has been selected. So, twelve players have been selected for the best players' pool as bowlers. Other players were ignored.

Finally, all together seventeen players have been selected for the best players' pool as bowlers.

The details of the players who have been selected are shown in Table 3.5 and Table 3.6. Since the Angelo Mathews is selected as both batsman and a bowler, the final team can be concluded with 31 players.

3.6 Team Prediction Overview

The second half of the research considers team prediction. The selected pool of players should be analyzed again based on the nature of the game such as the layout of the ground, the opposition team and the role of the player and predict the most suitable eleven players for the particular game. For this purpose, 143 international ODI matches have been collected, which plays in between the year 2013 January to 2018 December. Each players' individual performances, opposition team, ground, and the match outcome have been collected separately. Table 3.3 and Table 3.4 show a sample of the two data sets (separate data sets for batsmen and bowlers).

For the evaluation purpose, the data set has been divided into four parts and the most recent ¼ of matches data (35 matches) have been separated out and 108 matches data have used for the analysis.

As mentioned early, first-class cricket is played in between inter-zonal teams, basically known as clubs, at the national level. Because of that, there is no way that this study can collect the data regarding the international matches (against international teams in international grounds) played by the first-class players. As a result, again the analysis has to be done separately for the national players and for the first-class players.

	From the National Team	From the First-Class Players
Batsmen	09	06
Bowlers	05	12

Table 3.5: The number of players has been selected for the best players pool from the National Team and the First-Class players

	From the National Team	From the First-Class Players
Batsmen	Angelo Mathews Asela Gunarathne Dinesh Chandimal Kusal Mendis Kusal Perera Lahiru Thirimanne Niroshan Dickwella Upul Tharanga Chamara Silva	Dimuth Karunarathne Roshen Silva Shehan Jayasooriya Angelo Perera Gihan Rupasinghe Kithruwan Withanage
Bowlers	Angelo Mathews Lasith Malinga Nuwan Kulasekera Suranga Lakmal Thisera Perera	Isuru Udana Lahiru Gamage Malinda Pushpakumara Sachith Pathirana Alankara Asanka Silva Chanaka Wijesinghe Charitha Buddika Chathura Randunu Dinusha Fernando Gayan Sirisoma Hasantha Fernando Kosala Kulasekera

Table 3.6: The list of players who have been selected for the best players pool from the National Team and the First-Class players

3.7 National Player Analysis According to the Nature of the Game

When it is the case with national players, the study can use the two data sets (Table 3.3 and Table 3.4) which were collected regarding the international matches.

The runs scored in a particular match is the most crucial factor for a batsman in ODI matches. For bowlers, the most important factor is the economy rate. These measures can be changed for the other types (Test/T20) of matches. By considering the above two facts, it is decided to select the runs and the economy rate as the attributes to measure the performances of batsmen and bowlers in each individual match.

This study has considered the following formula.

$$\overline{x} = \frac{\Sigma x}{N}$$

Where in the case of batsmen x represents runs and in the case of bowlers x represents the economy rate. N represents the number of matches played by each individual player. By applying this formula against each individual player's performance, it is possible to find out the mean value ($\overline{x}$) of performances for each player. This $\overline{x}$ can be used to categorize the players' individual performances as high and low as shown below.

For the batsmen,

If Runs > = $\overline{x}$ *then Performance=high*

Else Performance=low

For the bowlers,

If Economy Rate < = $\overline{x}$ *then Performance=high*

Else Performance=low

This has made another attribute to be added to the two data sets as performance as shown in Table 3.7.

These edited data sets are used to find the association of performance with the opponent team and the ground by using the Association Rule Mining algorithm [15]. Association rule mining algorithms is a data mining technique which used to extract associations among a large set of items. Association Rule Mining based on Apriori algorithm in WEKA is used to discover the associations.

The support for the analysis is varied from 1.0 to 0.01 and the Confidence for the analysis is set at 0.5. The analysis has resulted in some interesting results, contradicting the general notion about the Sri Lankan cricket team. The sample results of the analysis are presented in Table 3.8. Consider that only the interesting rules regarding Angelo Mathews are shown in Table 3.8. These rules reveal that the opposition teams and the ground that he can shows his best performances.

The rules obtained from this stage can be used to predict the national players for a particular game according to the opposition team and the ground of the match.

Player	Opposition	Ist batting	Ground	Runs	Balls	SR	Winner	Performance
Mathewes	Afghanistan	opp	Sheikh Zayed Stadium, Abu Dhabi	22	39	56.41026	opp	low
Mathewes	Australia	opp	Melbourne Cricket Ground	12	14	85.71429	opp	low
Mathewes	Australia	opp	Bellerive Oval, Hobart	67	79	84.81013	opp	High
Mathewes	Bangladesh	opp	Dubai International Cricket Stadium	16	34	47.05882	opp	low
Mathewes	Bangladesh	opp	Shere Bangla National Stadium, Mirpur, Dhaka	20	26	76.92308	SL	low
Mathewes	England	opp	Kennington Oval, London	18	21	85.71429	opp	low
Mathewes	England	opp	Sophia Gardens, Cardiff	13	15	86.66667	opp	low
Mathewes	England	opp	Edgbaston, Birmingham	42	34	123.5294	SL	High
Mathewes	England	opp	R.Premadasa Stadium, Khettarama, Colombo	51	60	85	SL	High
Mathewes	India	opp	Barabati Stadium, Cuttack	23	32	71.875	opp	low
Mathewes	India	opp	Brabourne Stadium, Mumbai	3	14	21.42857	opp	low
Mathewes	India	opp	Khan Shaheb Osman Ali Stadium, Fatullah	6	18	33.33333	SL	low
Mathewes	India	opp	Queen's Park Oval, Port of Spain, Trinidad	10	11	90.90909	opp	low
Mathewes	India	opp	Eden Gardens, Kolkata	75	68	110.2941	opp	High
Mathewes	India	opp	R.Premadasa Stadium, Khettarama, Colombo	70	80	87.5	opp	High
Mathewes	New Zealand	opp	Bay Oval, Mount Maunganui	95	116	81.89655	opp	High
Mathewes	New Zealand	opp	Hagley Oval, Christchurch	46	52	88.46154	opp	High
Mathewes	Pakistan	opp	Dubai International Cricket Stadium	47	44	106.8182	SL	High

Table 3.7: The edited sample data set with the Performance attribute

Angelo Mathews (Bowling Performances)	**Confidence**
Opposition = England ==> Performance = high	81%
Opposition = New Zealand ==> Performance = high	67%
Opposition = Ireland ==> Performance = high	67%
Opposition = South Africa ==> Performance = high	67%
Opposition = India ==> Performance = high	65%
Opposition = Pakistan ==> Performance = high	63%
Opposition = Australia ==> Performance = high	50%
Ground = Rangiri Dambulla International Stadium ==> Performance=high	100%
Ground = Mahinda Rajapaksa International Cricket Stadium, Sooriyawewa, Hambantota ==> Performance = high	100%
Ground = Queen's Park Oval, Port of Spain, Trinidad ==> Performance = high	100%

Ground = Khan Shaheb Osman Ali Stadium, Fatullah ==> Performance = high	100%
Ground = Adelaide Oval ==> Performance = high	100%
Ground = Brisbane Cricket Ground, Woolloongabba, Brisbane ==> Performance = high	100%
Ground = Trent Bridge, Nottingham ==> Performance =high	100%
Ground = Riverside Ground, Chester-le-Street ==> Performance=high	100%
Ground = County Ground, Bristol ==> Performance = high	100%
Ground = Lord's, London ==> Performance = high	100%
Ground = Westpac Stadium, Wellington ==> Performance = high	100%
Ground = Himachal Pradesh Cricket Association Stadium, Dharamsala ==> Performance = high	100%
Ground = Sardar Patel (Gujarat) Stadium, Motera, Ahmedabad ==> Performance = high	100%
. Ground = Punjab Cricket Association IS Bindra Stadium, Mohali, Chandigarh ==> Performance = high	100%
Ground = JSCA International Stadium Complex, Ranchi ==> Performance = high	100%
Ground = Castle Avenue, Dublin ==> Performance = high	100%
Ground = Seddon Park, Hamilton ==> Performance = high	100%
Ground = Dubai International Cricket Stadium ==> Performance = high	100%
Ground = R.Premadasa Stadium, Khettarama, Colombo ==> Performance = high	67%
Ground = Edgbaston, Birmingham ==> Performance = high	67%
Ground = Sheikh Zayed Stadium, Abu Dhabi ==> Performance = high	67%
Ground = Pallekele International Cricket Stadium ==> Performance = high	63%
Ground = Shere Bangla National Stadium, Mirpur, Dhaka ==> Performance = high	50%
Ground = Sophia Gardens, Cardiff ==> Performance = high	50%
Ground = The Village, Malahide, Dublin ==> Performance = high	50%
Ground = Saxton Oval, Nelson ==> Performance = high	50%
Ground = Sabina Park, Kingston, Jamaica ==> Performance = high	50%

Table 3.8: Interesting rules obtained after the analysis by considering the bowling performances of Angelo Mathews

3.8 First-Class Player Analysis

Since the data is not available for the first-class players regarding the international matches, first-class player analysis has to be done separately by using some other attributes. The attribute "Role" has been selected. As shown in Table 3.9, players can be characterized as batsmen, spinner or fast bowler, based on their major talents.

In this stage, another data set has to be used as well, which is shown in Table 3.10. This table consists of different grounds in the world along with the preferences of the role of the player. The data was collected from the espncricinfo website.

It was decided to conduct the cluster analysis to group the players according to the role. Clustering based on the Hierarchical clustering algorithm in WEKA is decided to use as the number of clusters is unambiguous. And also, the hierarchical clustering algorithm is both more flexible and has fewer hidden assumptions.

Figure 3.10 shows the clusters of first-class players' roles. Three clusters were created since three roles as batsmen, spinner and fast bowler, are identified in the data set. These clusters can be used to predict the first-class players to the international matches according to the venue of the match.

Name	Matches	Inns	Runs	Wkts	Ave	Econ	Age	Role	5w
Alankara Asanka silva	108	169	9161	328	27.92	3.5	33	Spinner	19
Angelo Perera	97	54	857	19	45.1	3.67	28	Batsmen	0
Chanaka Wijesighe	134	226	1433	39	36.74	3.72	37	Spinner	0
Charitha Buddika	120	170	6791	256	26.52	3.32	38	Fast Bowler	7
Chathura Randunu	73	124	7580	280	27.07	3.63	34	Spinner	20
Dimuth Karunarathne	151	26	398	3	132.66	3.59	30	Batsmen	0
Dinusha Fernando	153	229	9387	370	25.37	3.44	39	Fast Bowler	19
Gayan Sirisoma	110	159	11401	577	19.75	3.05	37	Spinner	49
Gihan Rupasinghe	97	70	1099	31	35.45	3.3	32	Batsmen	0
Hasantha Fernando	180	290	6737	255	26.41	3.22	39	Fast Bowler	8
Isuru Udana	85	135	5557	190	29.24	3.58	30	Fast Bowler	3
Kithruwan Withanage	84	23	437	4	109.25	4.6	27	Batsmen	0
Kosala Kulasekera	122	188	6179	207	29.85	3.57	33	Fast Bowler	4
Lahiru Gamage	93	153	7667	248	30.91	3.61	30	Fast Bowler	12
Malinda Pushpakumara	117	210	13005	668	19.46	3.17	31	Spinner	54
Roshen Silva	121	18	163	2	81.5	4.07	29	Batsmen	0
Sheahan Jayasooriya	66	110	3733	157	23.77	3.46	26	Batsmen	9
Sachith Pathirana	85	139	8475	296	28.63	3.84	29	Spinner	20

Table 3.9: A sample data set which shows the statistics of the first-class players

Ground	Best For
Adelaide Oval	Batsmen
Bellerive Oval, Hobart	Batsmen
County Ground, Bristol	Spinners
Dubai International Cricket Stadium	Batsmen
Eden Gardens, Kolkata	Batsmen
Edgbaston, Birmingham	Spinners
Galle International Stadium	Bawlers
Kennington Oval, London	Fast Bawlers
Old Trafford, Manchester	Spinners
Pallekele International Cricket Stadium	Bowlers
R.Premadasa Stadium, Khettarama, Colombo	Fast Bawlers
Rajiv Gandhi International Stadium, Uppal, Hyderabad	Batsmen
Rangiri Dambulla International Stadium	Bowlers
Sabina Park, Kingston, Jamaica	Bawlers
Seddon Park, Hamilton	Bawlers
Sharjah Cricket Stadium	Bawlers
Shere Bangla National Stadium, Mirpur, Dhaka	Batsmen
St George's Park, Port Elizabeth	Batsmen

Table 3.10: A sample data set which shows the ground details

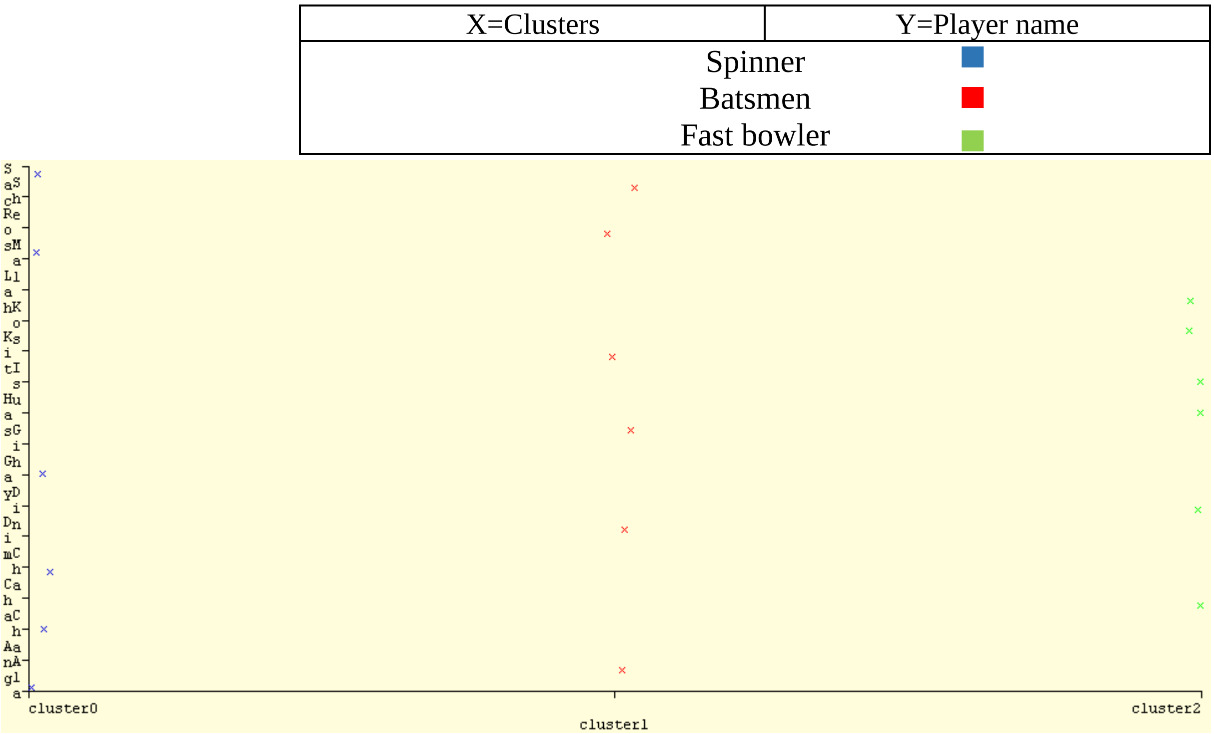

Figure 3.10: Clusters of first-class players' role

3.9 Summary

This chapter has shown the basic steps that have been followed during the development of the model. Basically, the methodology can be divided into two parts as performance analysis and team predictions. Mainly two data analysis algorithms have been used as the clustering algorithm and classification rule mining algorithm. Since the performances of the national players and the first-class players cannot be comparable and some of the data are missing with first-class players, the analysis was carried out separately those two groups.

By using the suggested model, it has predicted teams, each consist of eleven players, for thirty-five matches. The results are quite interesting and different comparing to the conventional team selections.

Chapter 4: Proposed Solution

By using two data mining algorithms, the Clustering algorithm and the Association Rule Mining algorithm the study was able to find a method that can predict optimal teams for matches according to the match condition.

For the evaluation purpose, the data collected regarding the international ODI matches were divided into four parts and the most recent 1/4 of data (thirty-five matches) were kept without used for the analysis. By using the newly developed method, the team members were predicted for these matches and the study was able to find some interesting results. Table 4.1 shows a part of the results that have obtained.

Match No	Date	Opposition Team	Ground	Players
3888	12/6/2017	Pakistan	Sophia Gardens, Cardiff	Anjelow Mathewes
				Asela Gunarathne
				Kusal Mendis
				Niroshan Dickwella
				Lasith Malinga
				Nuwan Kulasekera
				Suranga Lakmal
				Thisera Perera
				Roshen Silva
				Anjelo Perera
				Dimuth karunarathne
3897	30/06/2017	Zimbabwe	Galle International Stadium	Anjelow Mathewes
				Asela Gunarathne
				Dinesh Chandimal
				Kusal Mendis
				Kusal Perera
				Niroshan Dickwella
				Upul Tharanga
				Lasith Malinga
				Nuwan Kulasekera
				Suranga Lakmal
				Thisera Perera
3905	20/08/2017	India	Rangiri Dambulla International Stadium	Anjelow Mathewes
				Asela Gunarathne
				Dinesh Chandimal
				Lahiru Thirmanne
				Niroshan Dickwella
				Lasith Malinga
				Nuwan Kulasekera
				Suranga Lakmal
				Thisera Perera
				Roshen Silva
				Anjelo Perera

3959	19/01/2018	Bangladesh	Shere Bangla National Stadium, Mirpur, Dhaka	Anjelow Mathewes
				Asela Gunarathne
				Dinesh Chandimal
				Kusal Mendis
				Kusal Perera
				Lahiru Thirmanne
				Niroshan Dickwella
				Upul Tharanga
				Lasith Malinga
				Suranga Lakmal
				Thisera Perera
4058	23/10/2018	England	R.Premadasa Stadium, Khettarama, Colombo	Anjelow Mathewes
				Dinesh Chandimal
				Kusal Mendis
				Niroshan Dickwella
				Lasith Malinga
				Nuwan Kulasekera
				Suranga Lakmal
				Thisera Perera
				Roshen Silva
				Anjelo Perera
				Dimuth karunarathne
3909	3/9/2017	India	R.Premadasa Stadium, Khettarama, Colombo	Anjelow Mathewes
				Asela Gunarathne
				Dinesh Chandimal
				Kusal Mendis
				Lahiru Thirmanne
				Niroshan Dickwella
				Lasith Malinga
				Nuwan Kulasekera
				Suranga Lakmal
				Thisera Perera
				Roshen Silva
3955	17/01/2018	Zimbabwe	Shere Bangla National Stadium, Mirpur, Dhaka	Anjelow Mathewes
				Dinesh Chandimal
				Kusal Mendis
				Kusal Perera
				Lahiru Thirmanne
				Niroshan Dickwella
				Upul Tharanga
				Lasith Malinga
				Nuwan Kulasekera
				Suranga Lakmal
				Thisera Perera

Table 4.1: A sample of the results that have obtained as the team predictions

Chapter 5: Evaluation of the Results

The players which have obtained by predicting for the thirty-five international ODI matches were evaluated against the real match outcomes. According to the collected data, it was able to manually calculate the average runs that each batsman can score according to the game's nature. This means, against the particular opposition team in a particular ground, the average score that a particular batsman can score was calculated. By using these average scores, the total marks that every predicted team could score were calculated as well. These total scores were compared with the real scores of the Sri Lankan team in each match. Figure 5.1 shows the comparison and it clearly shows that 88% predicted teams' scores are higher than the real scores.

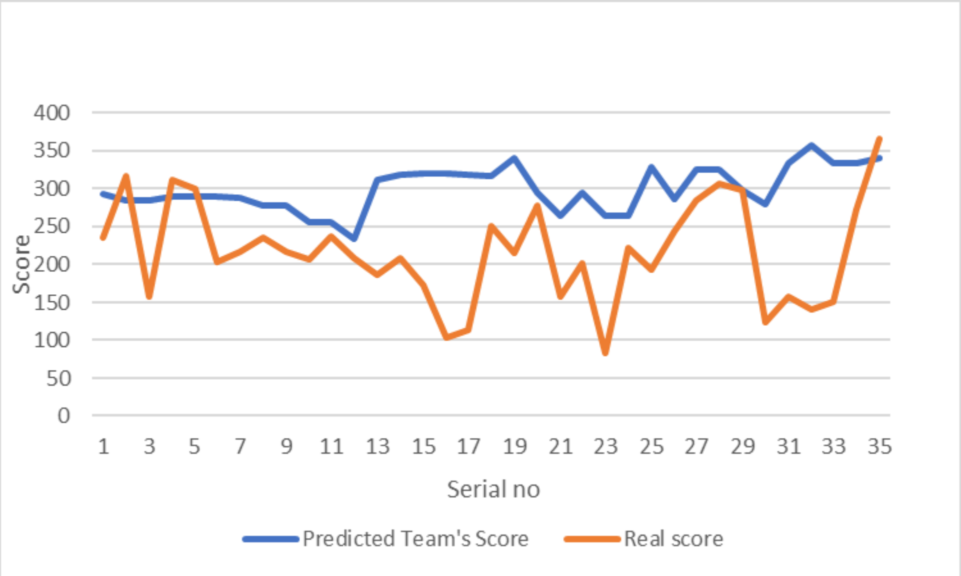

Figure 5.1: The comparison between the predicted Sri Lankan teams' scores and the real scores obtained by the Sri Lankan teams in each match

The same procedure has followed for the bowlers as well. The average economy rate (the average number of runs conceded per over) that each bowler can be obtained according to the game's nature was calculated manually. This means, against the particular opposition team in a particular ground, the average economy rate that a particular bowler can obtain was calculated. The predicted teams consisted of five bowlers and each gets ten overs to bowl. By using these average economy rate, the total marks that every predicted teams' bowlers concede to the particular opposition teams to score were calculated as well. These total scores were compared with the real scores of the opposition teams for each match. Figure 5.2 shows the comparison and it shows that 71% predicted scores are higher than the real scores for the opposition teams.

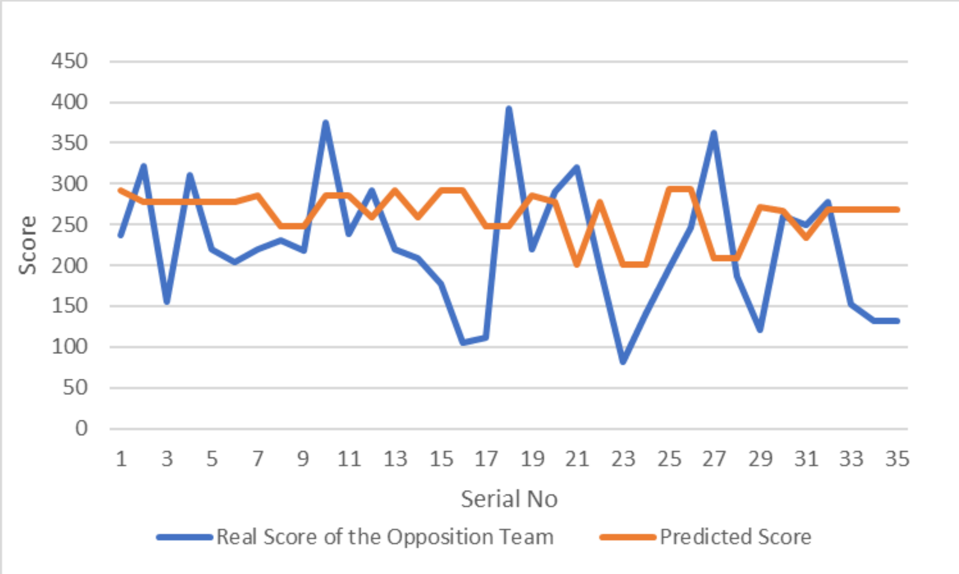

Figure 5.2: The comparison between the opposition teams' scores that could be conceded by the bowlers of the predicted Sri Lankan teams and the real scores obtained by the opposition teams in each match

Finally, both predicted scores (predicted score of the Sri Lankan team that can be obtained and the predicted score of the opposition team which conceded by the Sri Lankan bowlers) for each match were compared with each other. Figure 5.3 shows the comparison and the results are quite interesting. It reveals that 88% of the matches can be won by the Sri Lankan team with the predicted players.

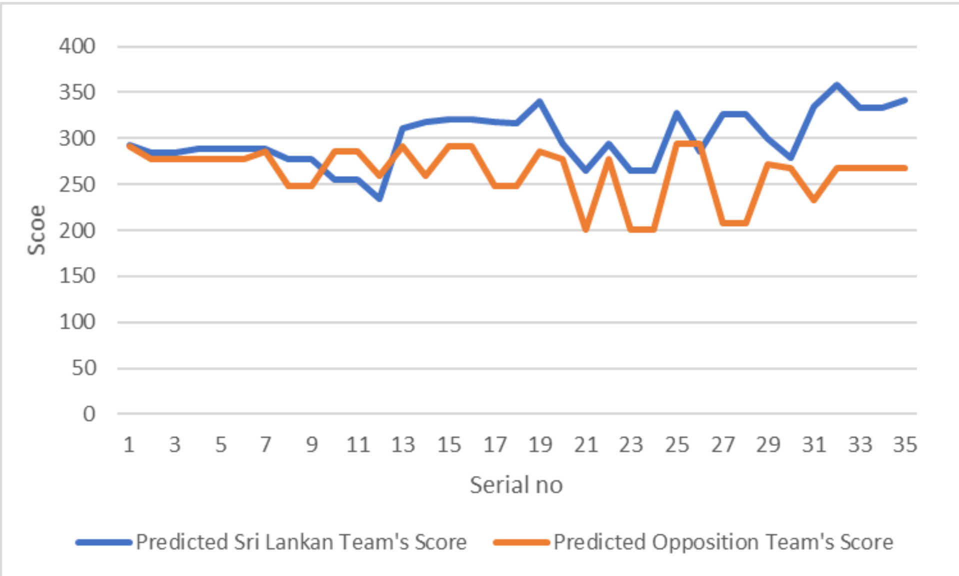

Figure 5.3: The comparison between the predicted Sri Lankan teams’ scores and opposition teams' scores that could be conceded by the bowlers in the predicted Sri Lankan teams in each match

Chapter 6: Conclusion and Future Works

The purpose of this thesis is to predict cricket teams for ODI matches according to the game's nature by using data mining techniques. It has proposed a new model to analyze the performances of players and the nature of games and finally, thirty-five teams have been predicted for thirty-five different games. The final evaluation shows that the Sri Lankan team could be able to win 88% of the matches with the predicted players in teams.

The final teams can be improved by considering the batting line up of the predicted players. It is an important attribute which could be able to change the final outcome of a match.

The predictions are made by examining grounds, opposition teams and the role of the players only. The weather conditions also can be used as an important attribute.

This research did not consider about the opposition teams' players or the batting line up which can be indicated as another important consideration.

References

[1]. H. Perera, “Cricket Analytics,” Ph.D. thesis, Simon Fraser University, 2015.

[2]. “11 Ways of Getting Out in Cricket,” Sportzwiki.com, 2019. [online]. Available: https://sportzwiki.com/cricket/the-eleven-ways-of-getting-out-in-cricket. [Accessed 5 May 2019].

[3]. “Sri Lanka national cricket team,” En.wikipedia.org, 2019. [online]. Available: https://en.wikipedia.org/wiki/Sri_Lanka_national_cricket_team. [Accessed 5 May 2019].

[4]. Sri Lanka Cricket, *Draft Selection Policy Document by Sri Lanka Cricket*, Heads of Cricket Operations and Head of Coaching.

[5]. “Cricket statistics,” En.wikipedia.org, 2019. [online]. Available: https://en.m.wikipedia.org/wiki/Cricket_statistics.4. [Accessed 6 May 2019].

[6]. S. ukherjee, “Quantifying individual performance in Cricket $-$ A network analysis of Batsmen and Bowlers,” *Physica A Statistical Mechanics and its Applications*, 2012.

[7]. “Different types of cricket pitches and their behavior,” Crickcafe.com. [online]. Available: https://crickcafe.com/cricket-pitches/. [Accessed 6 May 2019].

[8]. Thakare, I. Sachin, S.R. Suyal and K.Y. Pandav, “Performance Evaluation for Sports Team Selection Using Data Mining Techniques,” *AADYA-Journal of Management and Technology (JMT) 5*,102-108,2015.

[9]. H. H. Lemmer, “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):71-77, 2008.

[10]. D. Beaudoin and T. Swartz, “The best batsmen and bowlers in one-day cricket,” *South African Statist*, J,37, 203-222, 2003.

[11]. H. Ahmad, A. Daud, L. Wang, H. Hong, H. Dawood, and Y. Yang, “Prediction of Rising Stars in the Game of Cricket,” *IEEE Access (Volume 5)*,2017.

[12]. I.P. Wickramasinghe, “Predicting the performance of batsmen in Test cricket,” *Sport Exercise*, 9(4), pp.744-751, 2014.

[13]. P. Shah and M. Shah, “Pressure Index in Cricket,” *IOSR Journal of Sports and Physical Education (IOSR-JSPE)*,2014.

[14]. S. Akhtar, P. Scarf and Z. Rasool, “Rating players in test match cricket,” *Journal of the Operational Research Society*, 66(4), 2015.

[15]. K. Raj and P. Padma, “Application of association rule mining: A case study on team India,” *International Conference on Computer Communication and Informatics (ICCCI)*, pages 1-6, 2013.

[16]. En.wikipedia.org. (2019). First-class cricket. [online] Available at: https://en.wikipedia.org/wiki/First-class_cricket [Accessed 1 Jun. 2019].

[17]. “Sri Lanka national cricket team", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Sri_Lanka_national_cricket_team. [Accessed: 01- Jun- 2019].

[18]. D. Bhattacharjee and H. Saikia, “On Performance Measurement of Cricketers and Selecting an Optimum Balanced Team,” *International Journal of Performance Analysis in Sport*, 14. 10.1080/24748668.2014.11868720, 2014.

[19]. G. Sharp, W. Brettenny, J. Gonsalves, M. Lourens and R.A. Stretch, “Integer optimisation for the selection of a Twenty20 cricket team,” *Journal of the Operational Research Society*, 62. 10.1057/jors.2010.122, 2011.

[20]. D. Bhattacharjee and H. Saikia, “On Performance Measurement of Cricketers and Selecting an Optimum Balanced Team,” *International Journal of Performance Analysis in Sport*, 14. 10.1080/24748668.2014.11868720, 2014.

[21]. S. Iyer and R. Sharda, “Prediction of athletes performance using neural networks: An application in cricket team selection,” *Expert Syst*, Appl.. 36, 5510-5522. 10.1016/j.eswa.2008.06.088,2009.

[22]. P.J. Van Staden, “Comparison of cricketers’ bowling and batting performances using graphical displays,” *Current Science*, 96:764–766,2009.

[23]. P.J. Bracewell and K. Ruggiero, “A parametric control chart for monitoring individual batting performances in cricket,” *J Quant Anal Sports*, 5(3): 1–19, 2009.

[24]. G.R. Amin & S.K. Sharma, “Cricket team selection using data envelopment analysis,” *European Journal of Sport Science*, 14:sup1, S369-S376, DOI:10.1080/17461391.2012.705333, 2014.

[25]. M.H. Dunham and S. Seshadri, “*Data Mining- Introductory and Advanced Topics*,” 2006.

[26]. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed, Diane Cerra, 2006, p. 6-7.

[27]. R. Tibshirani, G. Walther and T. Hastie, "Estimating the Number of Clusters in a Data Set via the Gap Statistic", *Royal Statistical Society*, 2001, pp. 411-423.

[28]. B.Morley and D. Thomas, "An investigation of home advantage and other factors affecting outcomes in English one-day cricket matches", *Journal of Sports Sciences*, 2005, 23:3, 261-268.