**Team Recommendation System**

**A Project Report**

***Submitted by***

**Ayush Chuarasia**

**Anish Srivastava**

**Anurag**

***in fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

**IN**

COMPUTER SCIENCE AND ENGINEERING



**University Institute of Engineering and Technology**

**C. S. J. M. University, Kanpur**

MAY 2020

# **University Institute of Engineering and Technology**

**C. S. J. M. University, Kanpur**

# **BONAFIDE CERTIFICATE**

Certified that this project report **“TEAM RECOMMENDATION SYSTEM”** is the bonafide work of “**AYUSH CHAURASIA (CSJMA1600139016), ANISH SRIVASTAVA (CSJMA1600139008), ANURAG (CSJMA1600139009)”** who carried out the project work under my supervision.

Signature of the Supervisor

**SIGNATURE**

Dr RENU JAIN

CSE DEPARTMENT

UIET CSJMU KANPUR UP

**Abstract**

This report is about our B-tech project titled  “**Team Recommendation System**” on the basis of individual performance of players. Using players strength and weakness against the players of the opponent team by considering the statistics of a set of matches played by players helps captain and coaches to select the team and order the players. In this project, we propose a Unsupervised learning method using Clustering and   others method  to select Players depending upon the historical performance of the player .The comparison among different groups of players at the same level gives the order of groups which contributes to winning probability. We will also propose to develop a system which recommends a player for a specific role in a team by considering the past performances.

Player selection is one the most important tasks for any sport and cricket is no exception. The performance of the players depends on various factors such as the opposition team, the venue, his current form etc. The team management, the coach and the captain select 11 players for each match from a squad of 15 players. They analysed different characteristics and the statistics of the players to select the best playing 11 for each match. Each batsman contributes by scoring maximum runs possible and each bowler contributes by taking maximum wickets and minimum runs.

Cricket has become the most popular game not only in India but also around the world. Day by day it is gaining more people’s attention. The name and glory of team India has trespassed the country’s border and it is creating a huge impact in the world of cricket. To make sure the continuous development of the team, a more professional approach and help from the **IT sector** is needed. Remembering this fact, we used **machine learning** to select the best players from the standby list based on the previous playing statistics and then from that players we have found out the winning team combination. We have collected the data from websites that offer trustable sports statistics. Machine learning methods such as KNN(K-nearest neighbour) , Clustering was used to make cluster from batsman category to bowler category and then we have selected the players for the team according to which cluster players belong. Our goal was to form a well-balanced team through our approach. In this process we have used **Unsupervised machine learning approach**.

**TABLE OF CONTENTS**

ABSTRACT ……….. …………………………………………………………………………..4

TABLE OF CONTENTS…………………………………………………………………………. 5

LIST OF FIGURES ……………………………………………………………………………. 6

ACKNOWLEDGEMENTS ……………………………………………………………………..7

**Chapter 1** Introduction………………………………………………………………………….. 8

1.1 Problem………………………………………………………………………………………. 8

1.2 Background…………………………………………………………………………………… 8

1.3 Scope and Objectives……………………………………………………………………….. 8,9

1.4 Document Overview…………………………………………………………………………. 10

**Chapter 2** Literature Review……………………………………………………………………. 11

2.1 literature Review……………………………………………………………………………11

2.2 Batting and Bowling performances……………………………………………………… 11,12

**Chapter 3** Methodology and Implementation…………………………………………………… 13

3.1 Methodology…………………………………………………………………………………. 14

3.1.1 Batting measures and batting index……………………………………………………14,15

3.1.2 Bowling measures and bowling index…………………………………………………16

3.2 Implementation……………………………………………………………………………. 17-21

**Chapter 4** Results………………………………………………………………………………. 21-23

**Chapter 5** Conclusions and Future Work……………………………………………………….. 24

5.1 Conclusions………………………………………………………………………………….. 24

5.2 Future Work ………………………………………………………………………………… 24

REFERENCES…………………………………………………………………………………….25

**List of figures**

|  |  |
| --- | --- |
| **Fig No.** | **Name of the Figure** |
| 1. | PLAYER RECORD |
| **2**. | SERIES-WISE RANKING INDEX |
| 3. | AVERAGE RANKING INDEX |
| 4. | MAX INDEX VALUES |
| 5. | MAX INDEX VALUES |
| 6. | SERIES-WISE RANKING INDEX (STATE PLAYERS) |
| 7. | PREDICTED TEAM |
| 8. | CLUSTERS OF STATE LEVEL PLAYERS |

# 

# **Acknowledgement**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and supervisors. I would like to extend my sincere thanks to all of them. I am highly indebted to our mentor **Prof. Dr. Renu Jain** for her guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. I would like to express my gratitude towards my member of our team for their kind cooperation and encouragement which help me in completion of this project. My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

**CHAPTER-1**

**INTRODUCTION:**

**1.1 Problem:** To apply a best suitable model for selection of team of players in cricket match

.

**1.2 Background:**

Cricket is the most popular game in Asian countries, a lot of series are played across the set of different countries during a year and World Cup is played once in four years across all the cricket playing nations. Cricket is played in different formats like one day international (ODI), T20 and Test matches. Apart from this many league matches at club level and national level are played within the country. For any such tournaments, series or World Cup a suitable team of playing 11 and 4 extra players needs to be selected to form a team. Cricket team consists of a set of batsmen and bowlers with one wicket keeper who can also bat or bowl.

The selectors and team captain has to select batsman and bowlers in the team with a wicket keeper. Each batsman in the team will be specialized to bat at a different position in the playing eleven and there are varieties of bowlers like spin, fast and medium fast in the team. Team members need to be selected from a pool of players depending on the opponent team against which the team is playing. The team selected may  also be analyzed on their strength to chase or target setting against the team which they play.

Selecting such a team by coaches and captain is biased and error prone, this decision depends on the experience and analytic skills of the individual by using the performance measures and statistics of the players.

Automating the selection process by considering the performance measures and forecasting the match outcome helps coaches and captains to select the most efficient

team which have higher chances of winning the game. The other challenge which might occur. during selection is the lack of or few number statistics of the new player who have not played many international series. In this scenario, it becomes very important to correlate the new player with the existing player database with the statistics of club level and few international level games he had played. To automate such a process, we propose a team selection and predicting

the outcome of a game with a player profiling facility to assign the performance measure to the new player. Many people have worked on team selection, optimal ordering of the team and winning predictions of the cricket game. A genetic algorithm is used to automate the cricket team selection of 15 members (11 playing and 4 extra players). This algorithm represents the team as a string by considering each player as a bit in a sting. Later these strings are evaluated using objective function and are converted into fitness. This is genetically evolved using the cross over, mutation and replacement techniques to select the Indian squad for the World Cup. To bid players for a league match IPL (Indian Premier League) played in India with a budget constraint, franchises considered both budget constraints and the player performance while selecting the team. With these multiple constraints an optimal team was selected with the best bowling and batsman using a non dominated sorting genetic algorithm (NSGA II) by solving the multi-objective optimization problem Ahmed et al. (2011); Omkar and Verma, (2003). A team selection is done using a short cricket series by using linear integer programming.

 Various measures used in the literature are discussed in detail and the measure is proposed which is derived by comparing the individual performance with the team performance in the series. Lemmer (2013). After or while the team is selected, optimal ordering of the batsmen and predicting the match outcomes are performed to forecast the success of the team selected. An optimal team line up is found from the huge combinatorial space using simulated annealing algorithm by considering the batting and bowling characteristics of the selected players. Parameters like Batting average (BA), Batting strike rate (SR), Bowling average (BA), Bowling economy rate (ER) are used. The combination of a supervised and unsupervised algorithm is used to predict the outcome of the one-day international.

Match by using linear regression and nearest neighbor clustering methods. Using these methods, a total number of runs to be scored in the match is forecasted which is one of the important components in predicting the outcome of the match. Historic features extracted from the previous matches is combined with the ongoing match features like a number of wickets and runs scored are used in prediction. The logistic regression model is used to extract features from the one-day cricket as the match is in progress, which reduces the parameters dramatically. Cross-validation method is used to decide the parameters which need to be used for the model Asif and McHale (2016). A software tool crickAI is developed using the machine learning technique bayesian classifier to predict the outcome of the one-day match.

Factors like scoring, both the team strengths, toss, day-night match effect, home ground advantage are factors used for analysis. The score made by each team and difference between the scores of both the teams are approximated to the normal distribution, which facilitates the use of multiple linear regression to predict a score of the team batting first and victory margining ODI Bailey and Clarke (2006). Multinomial logistic regression is used to predict the outcome

of the test match as a multinomial response (win, draw, loss) using the match position at the session start and pre-match team strength and other features of test cricket Akhtar and Scarf (2012).

**1.3 SCOPE AND OBJECTIVE:**

In this research, we propose a framework which selects a player and performs team analysis and recommends the player role by extracting the statistics about the cricket game and players from various websites.

The contributions of the proposed framework are as follows:

**1.** Extract the unstructured data about match and players from the sports website or Kaggle

**2.** From the database, statistics are used to quantify the player performance measures to

rank the players.

**3.** Develop a model to predict the outcome of the match, based on the players playing in

both the teams using the historical data of the matches.

**4.** Provides the team structure analysis of the selected team which contributes to win.

**5.** Recommends a preferred role for a given player by finding the set of similar players in

the database.

**6.**This will enable us to select the best suitable team against a particular country.

**1.4 DOCUMENT OVERVIEW:**

In this document there are different sections and the overall document is divided in different chapters chapter 1 give introduction , problem statement and scope of problem.in second chapter literature the review and third chapter contain methodology and implementation .

We have collected the data from websites that offer trustable sports statistics .We have written scrapers to collect the data from websites .After that a lot of data cleaning and other methods are done to scale and normalize the data. Different techniques of statistics are used to calculate performance measures of the players.

 Later on, we have also used a fully connected neural network to find out the performance comparison of different algorithms. We have selected the players for the team according to their performances and experiences . Our goal was to form a well-balanced team through our approach .In this process we have used machine learning methods both supervised and unsupervised learning i.e. k-mean clustering and knn classifier to select appropriate players for the team. Clustering is used to obtain different groups of players according to their past performance. Overall there are two major statistical components for each player first is batting index and second is bowling index .If a player has a higher batting index then he would have been assumed to be a better batsman compared to those who have less batting index. Same holds for bowlers .

After that we have come to the result and conclusion .In this section we also have scope of future work ,correction and advancement.At the end of this document all the references have been mentioned properly.

**CHAPTER 2**

**LITERATURE REVIEW**

An extensive online search produced very few articles related to player’s performance prediction in the game of cricket. A very small number of researchers have studied the performance of cricket players. Muthuswamy and Lam predicted the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. They used back propagation network and radial basis network function to predict how many runs a bowler is likely to concede and how many wickets a bowler is likely to take in a given ODI match [2]

Wickramasinghe predicted the performance of batsmen in a test series using a hierarchical linear model. Using neural networks study of predicting “How many wickets will a bowler take?” can be possible, but their work was limited to only eight Indian bowlers and difficult to generalize for all the bowlers in the world. So some supervised machine learning algorithms to build prediction models that can be used to predict the performance of any player in a given match [3]. Iyer and Sharda[4] used neural networks to predict the performance of players where they classify batsmen and bowlers separately in three categories – performer, moderate and failure. Based on the number of times a player has received different ratings, they recommend if the player should be included in the team to play World Cup 2007. Jhanwar and Paudi[5] predict the outcome of a cricket match by comparing the strengths of the two teams. For this, they measured the performances of individual players of each team. They developed algorithms to model the performances of batsmen and bowlers where they determine the potential of a player by examining his career performance and then his recent performances. Lemmer[6] defined a new measure called Combined Bowling Rate to measure the performance of bowlers.

**2.1 Batting and Bowling Performance**

To select the best combination of players for a team, it is necessary to develop a better selection tool. In the research of Kusiak, Kern, Kernstine and Tseng .They have applied K Nearest Neighbor to find a score which is closer to the standard score or threshold of a player. Mickey Arthur has summed up the power of participation or responsibility in a game victory by addressing some influencing factors of ODI cricket such as physical strength, scoring, home ground advantages, day/night situation, toss, and decision of batting first or second. How these factors will affect the performance of players in a cricket match is determined by understanding and estimation of the values used in Bayesian classifiers .Logistics and normal regression models suggested by Holder and Nevill to find out the ranking and home ground advantages for players according to their research has found that players perform well when they play in their home grounds. Moreover, to determine the weak points of opposite teams, a reverse data mining technique is used, which can plan defensive strategy for the next game the players will play . The form and selection of the cricket players is defined in terms of average score, strike rate and consistent performance. A classification algorithm is used here which assigns classes to the best batsmen accordingly. The combined bowling rate is a combination of three traditional bowling measures: bowling average, strike rate and economy. Bhattacharjee and Pahinkar used this combined bowling rate to analyze the performance of bowlers in Indian Premier League (IPL). They also determined other factors that affect the performance of bowlers and applied multiple regression models to identify the factors that are empirically responsible for the performance of bowlers. Mukharjee. Applied Social Network Analysis to rate batsmen and bowlers in a team performance. He generated a directed and weighted network of batsmen-bowlers using player-vs-player information available for test and ODI cricket. He also generated a network of batsmen and bowlers using the dismissal record of batsmen in the history of cricket. Shah

 also defined new measures to measure players’ performance. The new measure for batsmen takes into account the quality of each bowler he is facing and the new measure for bowlers considers the quality of each batsman he is bowling to.A comprehensive review of the literature regarding the performance of both the player and the game reveals following findings. Stretch (2003) researched about cricketer’s injuries that could affect their performances in the game. In this study, the author applied a hierarchical linear model (HLM) to model the nature of injuries to South African cricketers, including doctors and physiotherapists working with the South African team . Kimber and Hansford (1993) proposed a method, which was based on a nonparametric approach to assess the batting performance of cricket batsmen. Reaction time is regarded as one of the incalculable talents of cricketers in all the departments of the game . Balasaheb, Maman, and Sandhu (2008) attempted to find the impact of visual skills training that could affect the performance of batsmen. In their research, they showed how the visual skills improve the reaction time, depth perception, and eye of the cricketers, which eventually improves the batting performance of the player . While discussing about necessary adjusted measures to analyze the player performance in the game of cricket, Lemmer (2008) investigated the performances of players’ when players participate in a small number of cricket matches.

**Chapter - 3**

**Methodology and implementation**

**3.1 Player performance quantification:**

Performance measures derived for players and tournaments using the players and match statistics helps coaches and captains in team selection, win prediction, team analysis and decide the role for a given player. To create the performance measures and to rank the players, we have used the statistics derived and stored in the database in the previous section. People have worked on different types of game data to derive the performance measures about the players and specific games.

 In cricket, the ranking of the players is done using the batting average, a strike rate of the batsman, average number of wickets taken and the runs conceded by the bowlers. Measuring the

performance of the players with only these measures may not be sufficient. Along with these measures .consistency of the batsman and weights is associated with the strength of opponents to measure the player performance Lemmer (2004, 2007). Similarly, methods are proposed to find one measure for bowling performance and use the type of wickets taken by the bowler(top order batsmen or tail end batsmen) to rank the players Lemmer (2005). To select the player for international cricket matches, performance measures are evaluated for the player with records of the local performances, which could be correlated to international level measures Lemmer (2009). So to find the outcome of the match specific measures of the player needs to be used, which could be compared with the opponent team player measures. so we have used the statistics of batsman and bowlers available on the net to create players ranking for batsmen and bowlers. The details of performance measure evaluations for both batsmen and bowlers are given in the following

**3.1.1 Batting measures and ranking index for batsman:**

To rank the batsman, we have collected the information about particular batsman like total runs scored by the batsman, number of innings played, number of times the batsman is out, number of balls faced , total number of 4’s and 6’s hit by the batsman, number of 100’s and 50’s and highest score of the batsman in the whole tournament. This data is further used to compute the features which aid in quantifying the players .

**Features:**

* BA = Runs scored */* Innings played - NOI……....**(1)**
* BS = Runs scored*/* Balls faced………..................**(2)**
* MRA = 100’s + 50’s  */* Innings played **…………..(3)**
* Outrate = number of times batsman got out*/* number of balls faced by batsman……**.(4)**
* BRPI = 4\*total number of fours+ 6\* total number of six’s */* Innings played………….**(5)**

**The Batting Average (BA)** given in **Equation 1** gives the average runs scored by the batsman in the tournament which considers only the innings played by the batsman and it subtracts it with the number of times batsman was not out during the innings in the tournament (NOI). This is considered because of the assumption that the batsman would have scored more number of runs in case he had a chance of batting. **The Batting strike rate (BS)** given in **Equation 2** provides the information about an average number of runs scored per 100 balls faced by the batsman.

To incorporate the consistency of the batsman, we found that **Milestone reaching ability (MRA)** using the formula given in the **Equation 3**, which uses the total number of 100’s and 50’s made by the batsman, which contributes to the total score of the team. The aggressiveness of the batsman is measured by the capability of the batsman to hit more number of fours and sixes. To calculate the **boundary runs per innings (BRPI)** the **Equation 5** is used which considers the total number of four’s and six’s hit by the batsman in the tournament. As we are quantifying the players and rank them in the order for particular series, We have devised a method in which the overall performance of the tournament is considered. Three measures general **batting average (Bat gen avg)**, general batting **out rate (Bat gen outrate**) and batting **general strike rate (Bat gen sr)** is calculated by using the aggregated measures of the tournament like total number of **runs scored (Tr),** total number of **wickets taken (TW)** and total **number of balls bowled (Tb)**in the entire tournament . These measures are given by the following equations.

**Features:**

* Bat gen avg = Tr */* Tw      ……………**(6)**
* Bat gen outrate =     Tw  */* Tb    ………………**.(7)**
* Bat gen sr =  Tr    */*   Tb………………..**(8)**

Using the generic performance of the overall tournament and the player specific information in the tournament, we find the **AGR** (above generic average runs) for each player and is given by the Equation 9.

* AGR = ((tbatsman - Bat gen sr ∗ nb) + Bat gen avg ∗ nb ∗ (Gen outrate - outrate)),........**.......................(9)**

where **nb** represents the number of balls faced by a batsman and **tbatsman** represents the total runs scored by a batsman in the tournament. Using AGR and team generic batting average the **ranking index (RI)** for the batsmen is calculate using the Equation 10.

* RI =AGR */*  ( 10\* Bat gen avg) …………**.(10)**

**3.1.2 Bowling measures and bowler ranking index system  :**

Similar to the batsman ranking index, the bowler ranking index is calculated. The parameters for the bowlers calculated using the number of runs conceded, a number of wickets taken and the number of overs bowled by the bowler and it is given by the following equations. where **Bowl avg** gives the ratio of runs conceded per wicket taken, **Bowl sr** is the average number of balls bowled per wicket taken and Bowler provides the information about the average number of runs conceded per over for each batsman:

**Features:**

* Bowl avg = Runs conceded */* wickets taken………...**(11)**
* Bowl sr = balls bowled */* wickets taken…………….. **(12)**
* Bowler  = runs conceded  */* overs bowled…………... **(13)**
* Outrate = wickets taken */* balls bowled……………... **(14)**

The generic tournament parameters with respect to bowlers are given by **bowler general average(Bowl gen avg)**, **bowler general outrate (Bowl gen outrate)** and **bowler general strike rate (Bowl gen sr)** by considering the **runs conceded (tc), total wickets (tw)** and **total balls bowled (tb)** in the whole tournament. These measures are given in the following equations.

* Bowl gen avg = tc */*  tw……….., **(15)**
* Bowl gen outrate = tw */* tb………. **(16)**
* Bowl gen sr = tc */* tb …………….. **(17)**

The AGR and Ranking index is given by

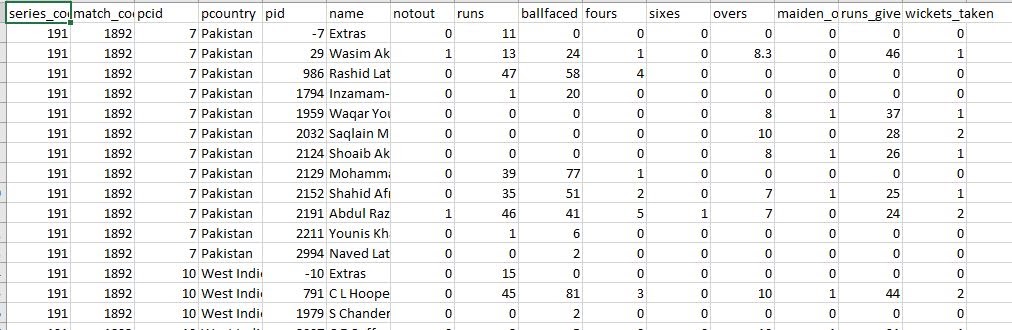
* AGR = (Bowl gen sr ∗ tb – tc) + Bowl gen avg ∗ tb ∗ (Bowl gen outrate − outrate)

where **tc** is total runs given by the bowler, tb total number of balls bowled by the bowler in the tournament, tr total runs given by the bowler.

* RI = AGR  */* (10\* team generic bowling average)………**. (19)**

**3.2 IMPLEMENTATION STEPS**

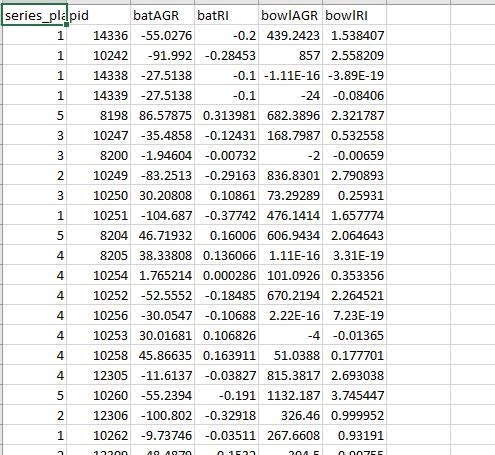
* Collect data of all Players series wise from Different websites like **Crickbuzz,** **howstat.com**, **Espn-Crickinfo** players.



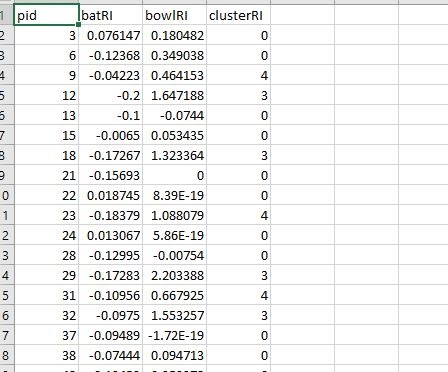
* .Now Calculate batting performance such as **batting average** **, strike rate** , **batting outrate, batting ranking index**.
* Calculate bowling performance such as **bowling outrate** ,**bowling average**

**Bowling index.**

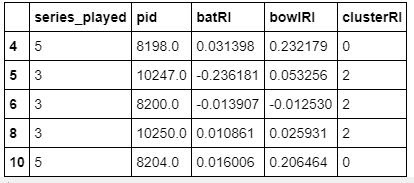
* Calculate generic tournament parameters like **generic bowling outrate**,**average,strike rate.**
* Calculate series wise **batting and bowling index** for all players.



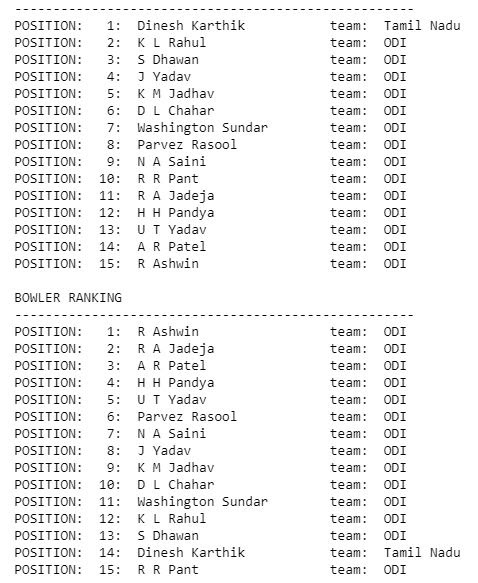
* Now applied **k- mean clustering** to the data after calculating **batting index and bowling index.**
* Train the model with 70% training data and obtain the clusters in which there are 5 clusters



* Cluster (0) has best batsman similarly when we move further from cluster 0 to 4 batting index decreases and bowling index increases this means that middle clustered players are allrounder and cluster(4) players are good bowlers.
* Now we have taken data from Vijai Merchent trophy players dataset and calculate above batting and bowling and batting measures .
* Now we have applied KNN Classifier with weight alpha and 5 neighbours  to predict the clusters for new players .



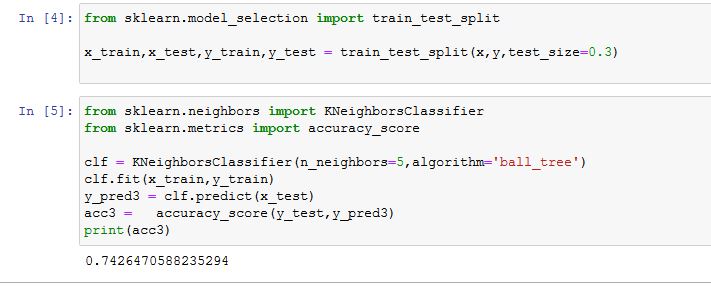
* Taking top 15 players that have higher batting and  bowling index separately we make a cricket team.

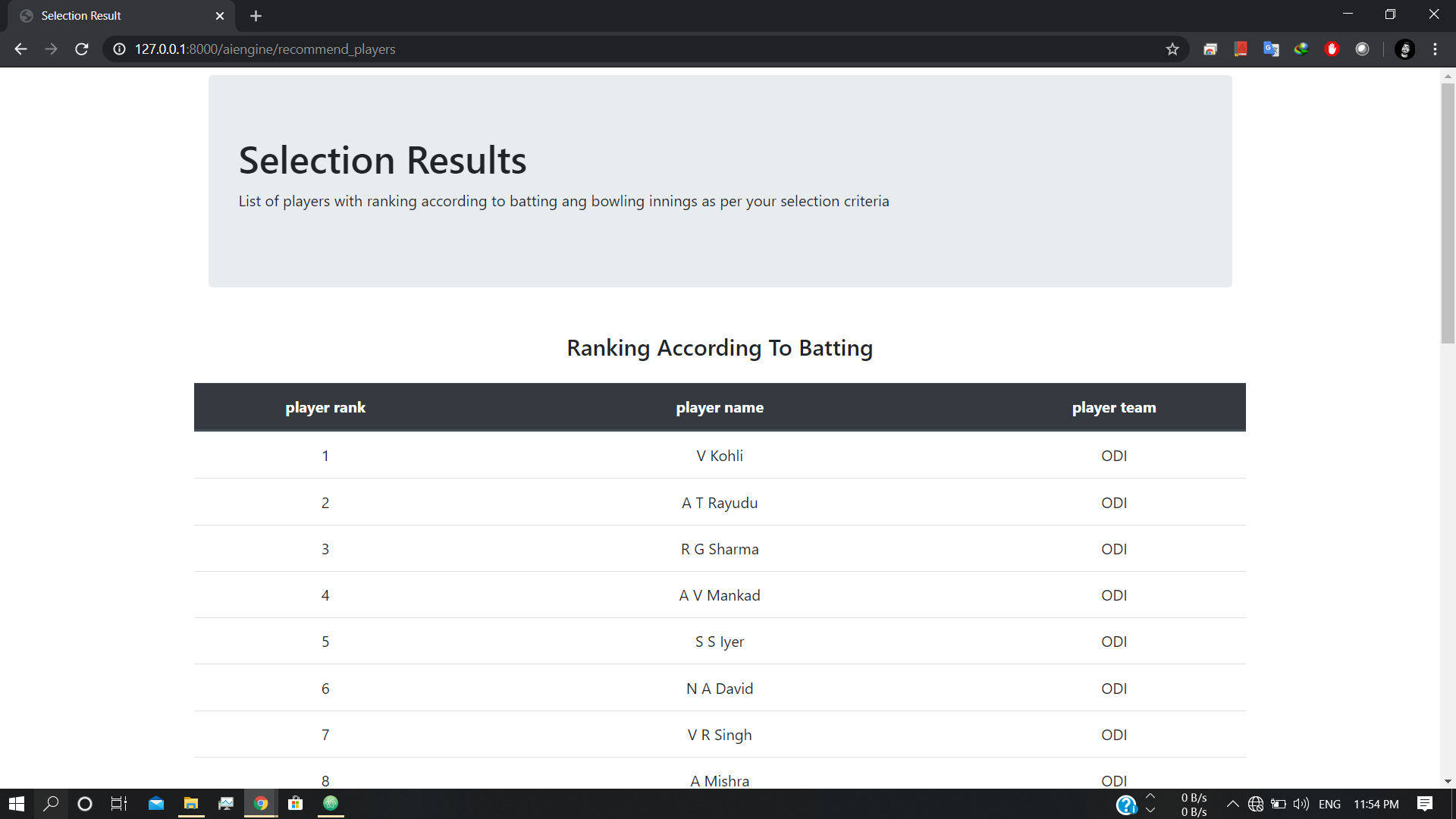


**Chapter-4**

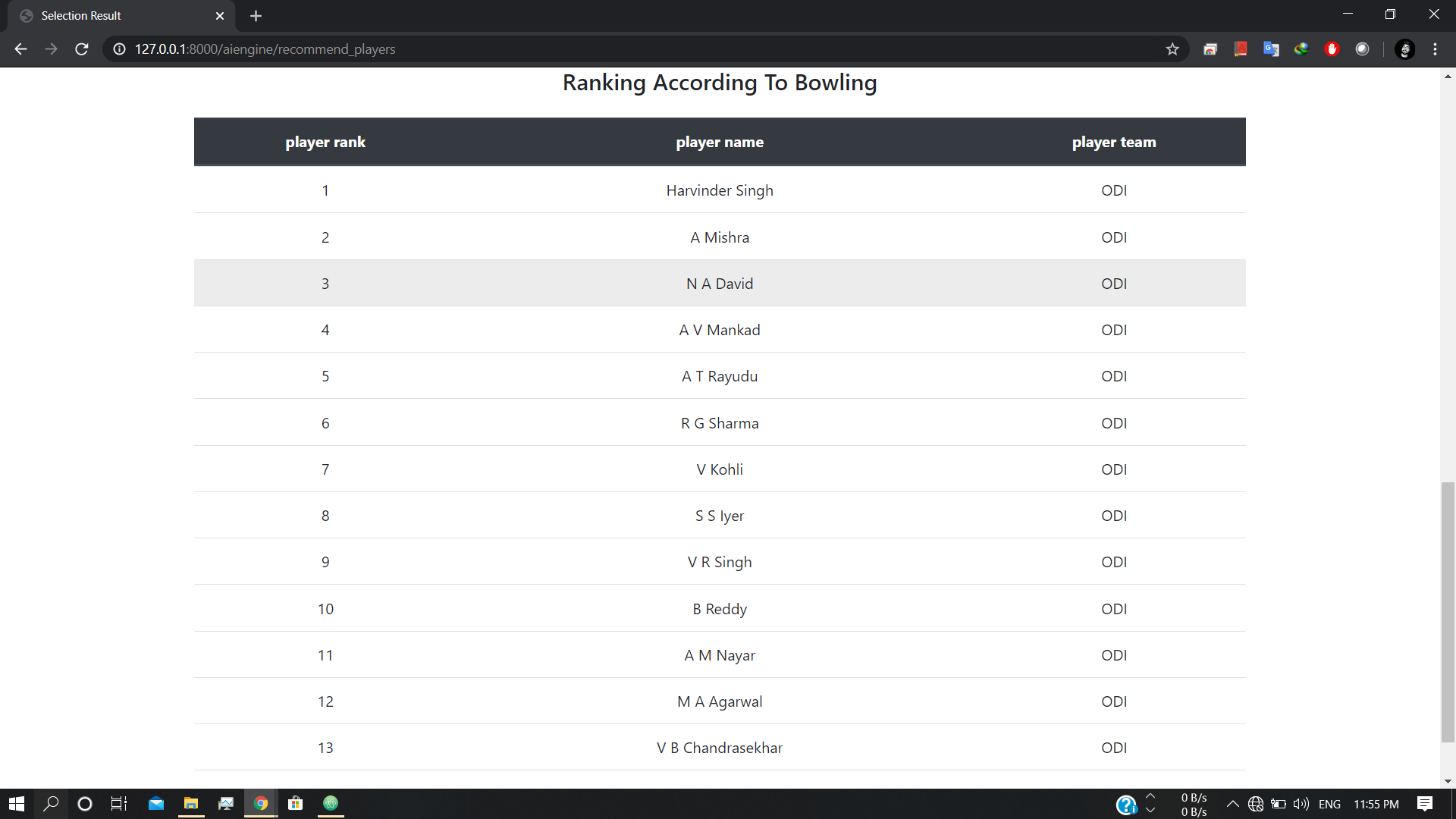
**Results**

Training accuracy of knn classifier is 74%,





.

****

**Chapter-5**

**CONCLUSION AND FUTURE WORK**

“*Our imagination is the only limit to what we can hope to have in the future*.”

- *Charles F. Kettering*

**5.1 CONCLUSION:**

By using Machine Learning Algorithm we converted Manual selection of cricket team into the Automated selection of Cricket team . You need to give players past record and as output you will find that player deserve place in the team or not. The result we got as output was satisfactory and considerable to replace manual system. The team selected through the machine learning algorithm is much closer to BCCIs team. From the result we can also conclude that using ML algorithm it seems that it is impossible to get 100% accuracy .The accuracy of this model is roaming around 70 to 90 percent which is pretty enough.

**5.2 FUTURE WORK:**

Many different areas left for research in this project due to limited time period. Enhancing “Features” for the selection of player is the key concept. For mapping to the real world selection process , more features need to count and research of such model or method which can assign the weightage to the features. Actually a wide area is open for research and modify the current version. More tuples needed which can analyze to get good results and accuracy.

**REFERENCES**

1. Journal of Sports Analytics 4 (2018) 263–273 DOI 10.3233/JSA-170196 IOS Press.
2. www.towardsdatascience.com
3. www.crickbuzz.com
4. www.ESPNcricinfo.com
5. [www.icc-crAhmed, F., Jindal, A. & Deb, K., 2011, Cricket team selection using evolutionary multi-objective optimization, in ‘International Conference on Swarm, Evolutionary, and Memetic](http://www.icc-cricket.com) [Computing’, Springer, pp. 71-78.  icket.com](http://www.icc-cricket.com).
6. www.howstat.com
7. Lemmer, H.H., 2013, Team selection after a short cricket series, European Journal of Sport  Science 13(2), 200-206.
8. Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems Beijing Boston by Aurélien Géron
9. Pattern recognition and machine learning by Christopher M Bishop

10. Machine Learning a probabilistic perspective by Kevin P Murphy