

Selection of Best Players in ODI Cricket using Ranking Based Indexing Method

MSc Research Project
MSc Data Analytics

Bhaskar Reddy Yenuga
Student ID: x21150567

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name:Bhaskar Reddy Yenuga.....

Student ID:x21150567.....

Programme:MSc Data Analytics..... **Year:**2022-23....

Module:MSc Research Project.....

Supervisor:Cristina Hava Muntean.....

Submission

Due Date:21-12-2022.....

Project Title:Selection of Best Players in ODI Cricket using Ranking Based Indexing Method

Word Count:21..... **Page Count:**.....7515.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:Bhaskar Reddy Yenuga.....

Date:21-12-2022.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Selection of Best Players in ODI Cricket using Ranking Based Indexing Method

Bhaskar Reddy Yenuga
x21150567

Abstract

In an effort to gain priority, cricket players are rated to show their superiority over rivals to acquire a position. Numerous scholars have proposed various statistical techniques to evaluate teams and their players in studies on this topic. The ability to compare a player's bowling and batting results have been established. In contrast to previous studies, it is difficult to achieve the goal of player and team selection. In the context of One-Day International (ODI) cricket, the research focuses on the prediction of the best player's selection for team-making using a ranking-based indexing method based on linear programming. Additionally, the dataset contains information from the ESPNCRICINFO website by web scrapping about seven important features, namely batting, bowling, fielding, partnership, all-rounder, batting from first innings, and bowling from first innings records. All the characteristics and several attributes are well analysed through the linear discriminate method to obtain the rank of each player and lastly the final rank for the selection of the 11 best players from a team bench of 15. Each rank of the cricketer's role for the selected players suggests which player plays which team's role along with their final rank estimation. In comparison, it has been found that 9 players from the predicted players from selection were selected for one ODI match played against Australia. It is calculated that the accuracy rate of the findings is close to 82%.

1 Introduction

Without using the best performance indicators available, selecting a cricket squad cannot be done reasonably. The conventional bowling average can be wildly inaccurate, particularly when there are few runs scored but a large percentage of them are not outs. The current study focuses on selecting a team after a limited number of games have been performed using the most appropriate metrics. It is anticipated that game circumstances may affect the rate at which batters score. To choose a squad from the athletes that competed in the ODI Series for the International Cricket Council (ICC) Champions Tournament, these criteria are utilized as an example. The use of a mixed integer linear technique in the selection procedure is demonstrated (Shah 2017). The idea is that a cricket team should consist of a variety of specialties, including wicketkeepers, all-rounders, bowlers, and batters. Ranking batters by batting aptitude and bowlers by bowling capability may be possible for a selection panel, but ranking all-rounder is more difficult. The ranking all-rounder is challenging because they must be both a good batter and a decent bowler by necessity. In addition, how numerous of every category of representative should be chosen? Integer performance tuning, a scientifically valid approach, is utilized in this research to demonstrate how cricket squad

selection might benefit from its application. There are also recommendations provided for choosing a squad if professional measured values must be used (Singla and Shukla, 2020).

International games, Limited - overs games, and ODI games are some of the forms used in these sports. In addition to this, there are several major tournaments performed at the club and county levels around the nation. An appropriate squad of performing 11 and 4 extra individuals must be chosen to create a side for any such competitions, series, or World Cup. One wicketkeeper with the ability to bat makes up a cricket team together with a group of batters and bowlers. The team captain and selectors must choose the bowlers and batsmen for the squad, along with a specialist batsman. The squad has a variety of spinners, including fast, moderate fast, and spin spinners, and each hitter is trained to bat at a certain position within the playing eleven. Based on the opposition team the team is competing against, team members must be chosen from a group of players (Prakash, Patvardhan and Singh, 2016). The chosen team may also be examined for its ability to outperform or set goals higher than the team they are playing. Therefore, choosing such a squad is prejudiced and prone to error; this choice is made based on the individual's expertise and analytical skills when analyzing player key metrics and data.

Managers and captaincy may choose the most effective squad with the best expectation of winning the match by automating the recruitment process by considering quality metrics and predicting the match outcome. The absence of data for new players who haven't participated in many international series might be another obstacle during choosing. In this case, it is crucial to compare the new player's numbers from his club matches and a few international matches to those in the player dataset already in place (Vetukuri, Sethi, and Rajender, 2021). In this, a team selection technique and game consequence prediction tool with player characterization capability to systematize this procedure and provide the new player's set of metrics. Numerous individuals have researched squad selection, team placement, and cricket game victory predictions. The 15-member cricket team selection process is automated using optimization algorithms (11 players and 4 additional players: A total of 15 players for a team bench). This method considers individual players as a bit more information in a sting, representing the squad as a thread. These characteristics are later turned into fitness by applying an optimal solution to assess them. To choose the Indian World Cup team, this variation was created by cross-over, recombination, and substitution processes (Perera 2021).

Utilizing a brief cricket tournament, a team is chosen using linear mathematical programming. The measurement that is produced by contrasting team players with the team's success in the series is suggested after a detailed discussion of the multiple measures utilized in the research. The best batsmen order and match result prediction are carried out either before or after the team is chosen to project the team's success (Jha et al. 2022). By considering, the cricket and bowling traits of the chosen players, an ideal team composition is generated using the stochastic optimization process from the vast combinatorial field. By combining a supervised and unstructured technique, closest neighbor grouping and linear regression are utilized to predict the result of the one-day worldwide game. One of the key elements in forecasting the result of the match is the maximum number of runs that will be scored in it, which may be anticipated using these approaches. To make predictions, historical information gathered from prior matches is merged with current match data, such as the number of catches and batting average. While the game is taking place, the logistic regression algorithm is utilized to extract characteristics from the one-day cricketers, drastically reducing the variables (Jha et al. 2022). The characteristics that must be utilized for the models are chosen using the passing approach. It is possible to use multiple models to forecast the performance of the team batted first along with the win margins in ODI since the

scores achieved by each side and the differential between the points of the two groups are roughly distributed according to the normally distributed. Using the match situation at the beginning of the session, the pre-match team quality, and other aspects of test cricket, multivariate logistic modeling is used to estimate the result of the test series as a multinomial answer like win prediction, loss prediction, and draw prediction (Jayanth et al. 2018).

Problem Statement - Who will be picked to represent each country's squad as a player? Cricket is a game that needs a lot of human judgments; thus, it is tried to account for as many aspects as it can be used to estimate the best player choice for a game. The main factors affecting a player's performance are their dependability and batting/bowling percentage. To predict players, the researcher used both their overall numbers and current form. The approach has been improved by incorporating numerous different aspects that impact the selection of the players in a team and the formation of a team for a specific match, even though in earlier references metrics were used to forecast the winning potential after carefully examining the results of the variables that influence the Player's selection for a match.

Aims and Objective -

The research's clear objective is to choose the top 15 players for the squad that can prepare for international competition. Finding out all important and irrelevant aspects that contribute to the performance gaps between players is exceedingly difficult. However, it should be overcome through clever reasoning and prediction models based on their success rates. So, to accomplish this goal, one should think about the following goals, which are given below.

1. To gather and construct the database from the reputed website like as ESPN for the Indian country.
2. To visualize the database as a feature importance
3. To show the best possible permutation and combinational approach by linear programming
4. To measure the player's performance by similarity index and rank method.
5. To rank each player before the final selection of a player.
6. To estimate the final rank of the selected 15 players to pick up the best 11 players.

Research questionnaire-

- How to collect and gather the dataset construction from records from websites?
- To rank the players, player performance metrics are quantified using database information. How to give more priority to the player's ranking?
- How to forecast the match before starting from the historical records of both teams' players?
- How to select the final squad from the 15 players for the national team to play in international games.

The document goes as follows, after introducing the research, it follows with some literature review, methodology used to work on the motto of the project, implementation, evaluation of the work, then some cases which show the working of the code, then followed by the results and finally concluding the report with the conclusion and future scope of the project.

2 Related Work

Very few articles about players' prediction problems in the cricket game were found after research. The effectiveness of cricket players has only been the subject of a very modest number of studies. The effectiveness of Indian bowlers versus seven professional players, which even the Indian cricket squad faces the most, was forecasted by the researcher. To forecast how many scores a bowler is likely to give up and the number of wickets a bowler is inclined to take in a specific ODI match, researchers used backpropagation networks and nonlinear activation network functions (Vetukuri, Sethi, and Rajender, 2021). A multilayer linear model was used in some papers to forecast the effectiveness of cricketers in a test series. A standard for comparing and choosing batsmen in restricted over cricket was established by researchers. They employed a two-dimensional visual illustration, with a Scoring Rate on one dimension and chances of getting out on the other, to construct a new metric. The selection process is then established based on the batsmen's striking rate and runs scored. The researcher employed neural network models to forecast player performance, categorizing bowlers and batsmen into three categories: performance, middling, and unsuccessful. They suggest whether a player can be selected in the roster to participate in the World Cup based on how frequently the player has gotten different ratings (Chakraborty, Kumar, and Ramakrishnan, 2018).

The researcher analyzes the characteristics of the two sides to forecast the result of a cricket match. They calculated this by evaluating the individual player abilities of each squad. Researchers created techniques to simulate the abilities of bowlers and batters, and these algorithms evaluate a player's ability by first analyzing his lifetime record and then the most current achievements. The researcher developed a brand-new metric called the Cumulative Bowling Average to evaluate bowlers' effectiveness. The bowling averages, scoring rate, and efficiency are three conventional bowling metrics that make up the total bowling rate. This combined striking rate was to evaluate the effectiveness of the batsmen in the cricket tournament (Bhattacharjee and Saikia, 2016).

2.1 New measurement on Player's evaluation

The new measurement for bowlers also takes into consideration the caliber of each batsman they are bowling too, as well as the caliber of each bowler they are facing. The overall performance index of a batsman is the sum of the individual performances of the batsman against each delivery. The overall quality index of a pitcher is the sum of that bowler's performances against each opponent (Shah and Patel, 2018). For developing algorithms to forecast the results of matches, the investigation established the terms "batting index" and "bowling index" to rank individuals' performances. Using mathematics, the proposed method was evaluated in the meaning of the ideal batted sequences for ODI matches. How statistical simulation might be employed in predictive analysis for various sports was discussed. The related paper reviewed the data mining techniques that had previously been utilized to forecast sports outcomes and outlined the benefits and drawbacks of each approach. The proper investigation was through the multiple records of the cricket players who employed machine intelligence methods to forecast football game results. Neural networks were employed to choose players (Premkumar, Chakraborty, and Chowdhury 2020). The research was likely the first generalized method to forecast the number of runs

batsmen will score and the number of wickets a player will take on a specific match day. It is impossible to generalize for all spinners worldwide because it was restricted to eight Indian players and used neural networks to estimate how so many wickets a spinner will achieve. It has developed a predictive prototype that may be used to forecast every player's success in a particular match using machine intelligence algorithms.

With the emergence of online social networking sites and databases, academics' attention has been drawn to the finding, ratings, and predictions of subject matter experts or contents. The academic world has proposed several efforts to solve these challenges. Among many other complex outcomes, locating Young Stars in scholarly networking, predicting future reference counts, and finding historical experts are some of these proposals. But the sports realm was not considered in these recommendations. Although several concerns with sports prediction are examined while taking the sports domain into account, the researchers do not propose a suitable prediction system. Additionally, none of the plans took cricket into account. The observer determined the top bowlers and batters based on a cricket match. Additional rankings of batsmen based on their productivity metrics and rankings of cricket teams using h-index and Page Rank have been proposed. The following suggestions, meanwhile, did not provide any Young Stars prediction framework for the sports industry (Hussain et al. 2019).

In cricket, evaluating a player's performance is essential for creating a well-balanced squad. Since playing circumstances vary from stadium to stadium, different tours necessitate various player groupings. Thus, in addition to certain other factors like a player's experience, performance in each circumstance, and various other characteristics, selectors must take into account a player's numerous personal traits. One may get this data from a player's overall career (Bhattacharjee and Saikia, 2016). To estimate the value of the qualities of the bowlers and batsmen in the given game and aid in the selection of players for the specified tour, this article discusses the ideas of various unrelated forest regressions. The ODI version of the tournament will employ the concept. The dataset used to create the model is a player's prior performance versus a certain opponent. The model considers the visiting team, the opponent, and the match location. Using the input fields, a rank-wise table of all the batters and bowlers is produced, which the assessors may use to choose the squad according to the preferred mixture.

In the batting and bowling fields, measures are devised specifically to predict coming stars. More specifically, three categories—Co-players, Team, and Opposing Teams—are included, and 9 and 11 attributes, respectively, are established for the forecast of coming stars in batting and bowling. Based on performance evolution measures and weighted averages, two different types of datasets are produced. Utilizing both generating (BN and NB) and discriminatory (SVM and CART) computational methods, the defined features are evaluated. Co-batsmen dominate the other two categories in the batting area, while Team performs best in predicting budding stars in the bowling field. Ultimately, it is shown that NB works better than the other models. Finally, a rankings list of young talents for both domains is offered based on the weighted average, performance progression, and rising star score (Prakash, Patvardhan and Singh, 2016). These rankings are contrasted with the ICC ratings for 2013–16, and it is discovered that the methods, it was provided are useful for predicting rising stars. Therefore, in the ODI and T20 formats, similar traits may also be utilized to forecast rising stars. To get even better outcomes, extra characteristics like opposing team variety, whether home or away, 100s, the 50s for batsmen, and 4, and 5 wickets for bowlers may also be included. Identifying RSs in cricket and other fields is very helpful since it allows the authority like coaches, and managers, to focus their efforts on maximizing the knowledge of such rising stars to achieve the best achievements in the future (Jha et al. 2022).

The prediction of young stars may be implemented using a similar concept in a wide range of sports organizations. The concept for win forecast and team model construction, player assessment, player-specific information gathering, and player performance verification are the four steps of the conceptual methodology for match win predictions, team assessment, and player suggestion. The fragmented matching information is pre-processed and recorded in the data storage during the first stage. The player productivity characterization model utilizes the information of the players recorded in the system to measure and classify the individuals. This information is provided at this stage. The following two steps make use of these player metrics and player characterization information. The player measurement and historical match win or loss data are utilized to develop the SVM for forecasting the victory or defeat percentages in the win forecasting and team structure assessment phases. In the concluding stage, the preferable role for a particular player is suggested using the grouping and k-nearest neighbor approaches (Premkumar, Chakrabarty, and Chowdhury 2020).

2.2 Past related works

S.No.	Reference	Objectives and Outcomes	Critical Analysis
1.	Singla, S. and Shukla, S.S., 2020.	A prospective method of squad selection that suggests a Dream 11 Fantasy squad for the forthcoming match utilizing real-world information gathered from Player achievements in the previous 10 matches. The Python Gurobi package was used to construct the Integer Linear programming method. The Markowitz Optimal solution, which is often used to choose stocks for financial investment, has also been applied to study the squad selection issue.	Integer Linear programming method and Markowitz Optimal solutions are used. As a result of the cricket team selection, there is no consideration for the rising players and new players' chances in the playing 11 from the team squad.
2.	Jayanth, S.B., Anthony, A., Abhilasha, G., Shaik, N. and Srinivasa, G., 2018.	The player rating score is calculated by the investigator utilizing game and player information taken from a certain competition. The n-dimensional information taken into consideration for modelling cannot be sequentially separated, according to test findings. As a result, the asymmetric SVM with RBF kernel performed better than the linear and multimodal kernels. The accuracy, specificity, and recall rate of SVM with RFB kernel are 75, 83.5, and 62.5 respectively. So, for predicting game outcomes, it was advised to use SVM with the Kernel function.	Here predicting game score is specially mentioned but there is no mention of how a new player can show the effect on the highest player's performance.
3.	Hussain, A., Qiang, Y., Bilal, A.Q.M., Ullah, U. and Ullah, N., 2019.	The abilities and flaws of regional teams play a crucial role in choosing the teams. To rank cricket teams in the area, researchers in this work presented the Region-wise Team Rank (RWTR), an adaptation of the PageRank method. It	Here there is no mention of failure rate and combination of old and new players even selection of the best team. Page rank

		makes sense to award more credit to a squad that defeats a tougher opponent. Compared to the standard rating, the suggested ranking more accurately captures the gameplay's flow.	algorithm is used here. This algorithm is used for quick search techniques.
4.	Hussain, A., Yan, Q., Bilal, M.A.Q., Wu, K., Zhao, Z. and Ahmed, B., 2019.	To rate cricket teams, they suggested a Region-wise Team Rank (RTR) and a Region-wise Weighted Team Rank (RWTR). The logic is that a squad that defeats a better team should receive more scores than a team that defeats a weaker team and vice versa. When determining a team's ranking score, the suggested technique considers a team's area-specific strengths and weaknesses in addition to the number of wins and losses in that region. In closing, the teams' rankings are compared to the actual ICC rankings.	This paper is most suitable compared to other papers. But partnership records are not considered here. If considered, then the player's ranking and team selection might be different.
5.	Jha, A., Kar, A.K. and Gupta, A., 2022.	In this essay, a two-step process for selecting teams and evaluating players in fantasy cricket is proposed. Sequential backward reduction in the random forest is used to evaluate individuals while considering situationally player statistics and a revised evolutionary method for player choice. The outcomes demonstrate that the suggested strategy outperforms the conventional hit-and-miss approach used in fantasy athletics to choose teams.	In the paper, the player's ranking was not made by all the necessary features. But the sequential backward reduction in the random forest is used here. It is very useful for each player's comparison in depth.
6.	Premkumar, P., Chakrabarty, J.B. and Chowdhury, S., 2020.	Here, new factors and improvements to current variables have been incorporated based on specific productivity indicators, which will determine a player's ranking in cricket. The former online leaderboards, notably the most popular International Cricket Council scoring system, overlooked many of these factors. This article evaluates batsmen and bowlers who've already played One Day International (ODI) cricket throughout the calendar year 2015 using a dynamic instead of a static technique of calculating factor scores using the factor analytical method, on a match-by-match premise.	Here factor analytical method is used to determine the factor score but there are missing steps for the player's analysis in depth.
7.	Agrawal, P. and Ganesh, T., 2020, April.	The selection procedure for the Indian cricket team's ODI roster is described in this essay. The examination of many criteria that might be used in the selection process is provided. The rankings of the reliable and unreliable players are	Integer optimization is used but no proper attributes like player's partnership and all-rounder fields are missing for

		described in this examination. The participants will be picked in a range of categories, including wicketkeeper, spinner, and batsman. The player's constancy was the major factor considered in the rating. Coding for integer optimization has been used for that measurement.	justification.
8.	Shah, P. and Patel, M.N., 2018.	In this study, they used Principal Component Analysis to rank captains according to several criteria. Additionally, the author has added a weighted average way to evaluate captains depending on the team's rating scale (Z score), as well as the captain's skill as a batsman and bowler.	A comparison between batsman and bowler was made but no other role was mentioned.
9.	Prakash, C.D., Patvardhan, C. and Singh, S., 2016.	Throughout this study, a novel indicator called the Deep Performance Index (DPI), which represents the effectiveness of the batsmen and bowlers on a closer examination of the needs of T20 cricket, is created using a network having to learn methodology. When developing the DPI, significant characteristics are extracted and their respective value is determined using the Recursive Feature Elimination technique, which is based on learning algorithms. It is demonstrated that, when contrasted to other well-known ranking factors for T20 cricket, DPI is better capable of collecting achievement information for both bowlers and batsmen.	Only DPI was based on bowlers and batsmen.
10.	Chakraborty, S., Kumar, V. and Ramakrishnan, K., 2018.	Separate rankings are first created from the extensive dataset usually containing the performances of various Test cricketers, comprised of controllable figures of applicant alternative solutions while trying to impose some restrictions on the least number of overs did play (for batsmen), a limited number of tests did play (for wicketkeepers and pacemen), as well as the minimum figures of points scored, and pitches chosen to take (for all-rounders). Later, the TOPSIS approach is used to rate those bowlers who had made the summary and determine which players will perform well in the projected World XI Test squad.	Isolate ranking was made but no actual evaluation for ODI cricketers.

Table 1: Related work

3 Research Methodology

The dataset was collected from the website ‘ESPN Cricinfo’, so it fully belongs to secondary data analysis. Web scraping coding is one of the best choices for selecting the data directly from the website. There is a reduced percentage chance of an error occurring while fetching data from the live websites. In the project, analysis has been done for the data collected by this method. A total of 7 datasets are collected from the ESPNCRICINFO website. Some of them are in different lengths of dimension. The datasets are selected by querying from the main website and the span of the dataset is up to the last 4 years i.e., Dec 2018 to Dec 2022. Most datasets are in 50 rows with different attributes. Each dataset is described in the below section in table format. In the below section, they have 50 rows each up to 4 sections, player-partnership with 446 rows, and 62 rows for the last two sections.

Table formats of dataset: *Note that: (for setting weight)*

M	H	VH	L
Moderate	High	Very High	Low

i. Batting Section

Player	Span	Mat	Inn	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
			<i>H</i>		<i>H</i>		<i>H</i>	<i>M</i>	<i>VH</i>	<i>M</i>	<i>H</i>	<i>VH</i>	<i>H</i>	<i>H</i>

ii. Bowling Section

Player	Span	Mat	Inn	Overs	Mdns	Runs	Wkts	BBI	Ave	Econ	SR	4	5
			<i>H</i>	<i>H</i>	<i>VH</i>	<i>H</i>	<i>H</i>		<i>H</i>	<i>VH</i>	<i>VH</i>		

iii. All-rounder

Player	Span	Mat	Runs	HS	Bat_Av	100	Wkts	BBI	Bowl_Av	5	Ct	St	Ave_Diff
		<i>H</i>	<i>H</i>		<i>VH</i>	<i>VH</i>	<i>M</i>		<i>H</i>	<i>L</i>	<i>H</i>	<i>H</i>	<i>H</i>

iv. Fielding

Player	Span	Mat	Inns	Dis	Ct	St	Ct_WK	Ct_Fi	MD	D/I
			<i>H</i>	<i>VH</i>	<i>VH</i>	<i>VH</i>	<i>H</i>	<i>H</i>	<i>H</i>	

v. Player-partnership

Player1	Player2	Wkt	Runs	Overs	RR	In	Out	Inns	Opp	Ground	Start Date
			<i>VH</i>	<i>VH</i>	<i>H</i>			<i>H</i>			

vi. Batting in 1st innings

Team	Score	Overs	RPO	Inns	Result	Opposition	Ground	StartDate
------	-------	-------	-----	------	--------	------------	--------	-----------

vii. Bowling in 1st innings

Team	Score	Overs	RPO	Inns	Result	Opposition	Ground	StartDate
------	-------	-------	-----	------	--------	------------	--------	-----------

The methodology followed here to analyse and predict the squad for the country in ODI is **KDD** (Knowledge Discovery Data). At the beginning of the data assessment, there are many more steps required to apply in the data pre-processing steps.

Data Pre-processing Steps:

- **Data quality assessment** - In this step, a check has been done in the dataset if there are any missing values from the dataset. Also, the analysis has checked the data types whether the data is an integer, float, or any string category.
- **Data Cleaning** - In the next phase mainly data errors like non-responsive and observational errors have been checked and repaired or transformed into the data for analysis. Here empty values were changed into zero values.
- **Data Transformation** - In this phase new fields have been estimated by a bunch of numerical techniques. Here feature importance, aggregation, and averaging methods are adopted.
- **Data reduction** - In this phase, the dataset has been evaluated as in short dimension, so it is called "Data reduction". Some attributes are selected to manipulate the player's ranking.
- **Important Feature Selection by weight-based aggregation** – In every dataset of cricket's characteristics there are a lot of attributes for identifying the player's role and performance on the Cricket Ground. But say essentially not all attributes are equally important to use for finding the player's evaluation or selection purpose. And even from all selective attributes it is not required to set equal priority for finding ranking weightage. By based on setting asymmetric weight values to the selective features player's rank was made through this method.

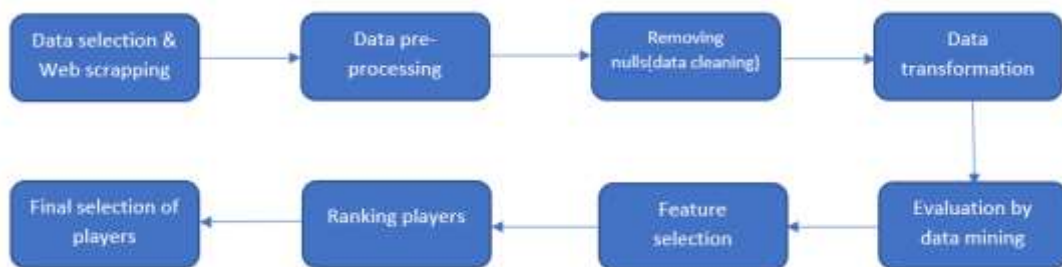


Figure 1: Methodology applied on the data (KDD)

4 Design Specification

After collecting whole datasets to select the final 11 players, it is essential to estimate the ranks for each player from all available records. How rank is estimated in linear programming fashion and what is benefitted against the different classifier's accuracy rate – which is important to find out the novelty of the work with compared to past related works. The advantage of this method over the use of different classifiers is maintaining the static result and not looking for other variation results by the unequal classification rate. Hence, it is a big advantage for the linear discriminate ranking method. In this section, the analysis has been described briefly.

Proposed Diagram:

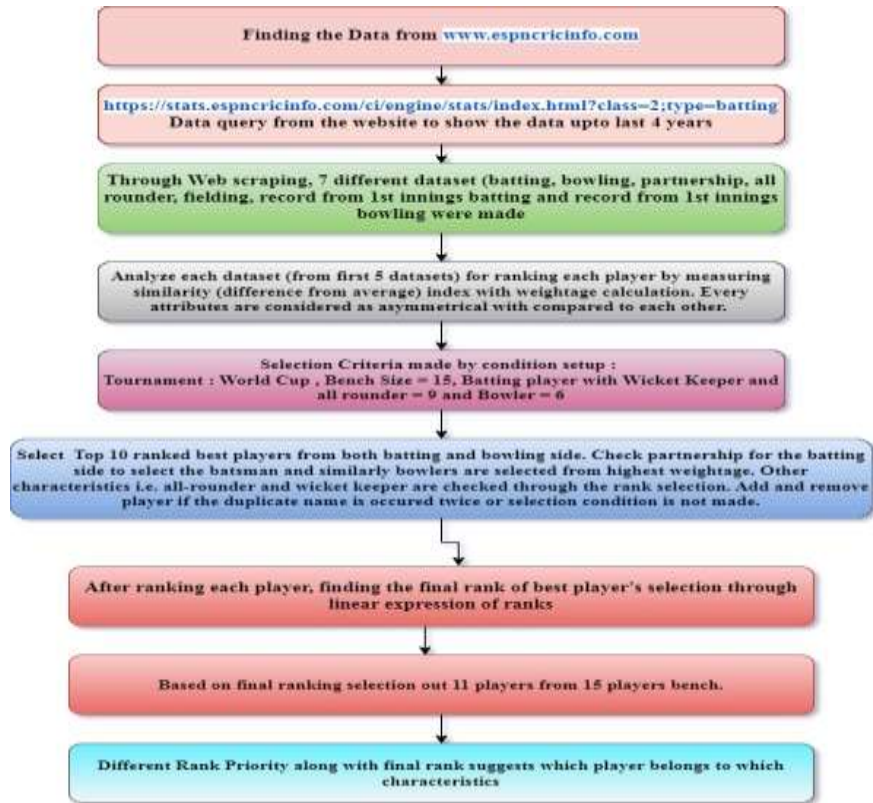


Figure 2: Proposed methodology for selection best players done by linear programming method.

5 Implementation

Selecting one cricket player from the large set of collections is not made easier because there are a lot of attributes that talk about the cricketer's habit on the ground. Which attribute is importantly observed for the player's selection – it is vital for giving the highest weightage (to the attribute). Henceforth the team selectors never look at all attributes symmetrically. Hence, weight multiplication for finding the total weight is required. The following linear equations are essential to find out the rank of each player.

$$Total_{weight} = \sum_{i=0}^{nos.of\ selected\ features\ (n)} W_i * (X_{Feature_i} - Averagevalue_{Feature_i})$$

$Total_{weight}$ represents the comparative weights' calculation based on selected featured importance. W_i represents the respective weights of the selected attribute 'i'. $X_{Feature_i}$ represents the value of $Feature_i$. $Averagevalue_{Feature_i}$ represents the similarity index (estimated by mean calculation of attribute i).

$$Rank_{Total_{weight}} = Sort\ of\ Total_{weight}\ by\ nos.\ of\ rows$$

$Rank_{Totalweight}$ represents the rank of the player. This rank estimation is the subset of final ranking set.

$$Totalweight_Final_{rank} = \sum_{j=0}^{nos.of\ selected\ characterisitcs} Totalweight$$

$Totalweight_Final_{rank}$ represents the total weights of selecting final rank.

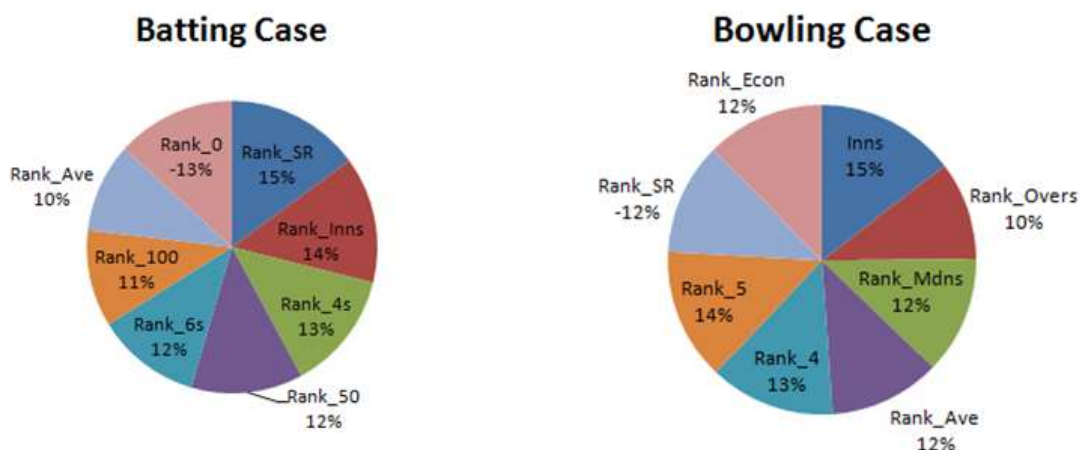
$Final_Rank_{Totalweight_Final_{rank}} = Sort\ of\ Totalweight_Final_{rank}\ by\ nos.\ of\ rows$

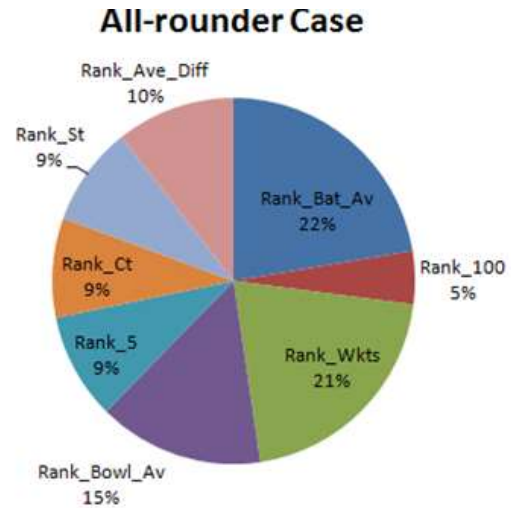
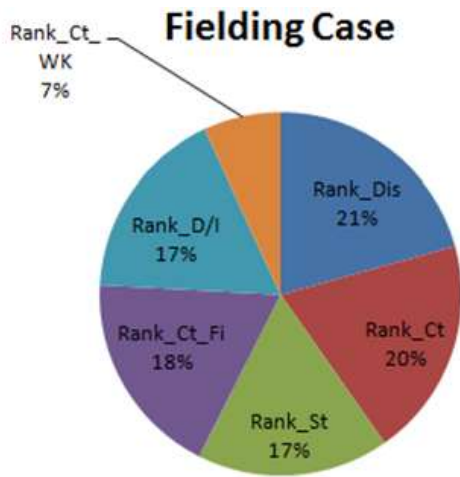
$Final_Rank_{Totalweight_Final_{rank}}$ represents the final rank of selected players.

The following steps are described for the rank analysis as well as team selection.

- **Step 1:** Internet surfing to find the data about the Indian cricketers.
- **Step 2:** Data query from the website for finding the records of batting, bowling, partnership, all-rounder, and fielding, etc.
- **Step 3:** Data collection about the 7 type datasets.
- **Step 4:** from every characteristic find the similarity index by calculating the average value of the whole index and subtracting between each player's record and average value. It provides positive as well as a negative values. the positive value means an upwards difference from the average and the negative value means vice versa. After this finding, each player's rank is estimated.
- **Step 5:** Some selection criteria are first set up before the final team selection.
- **Step 6:** Pick up the top 10 players from the batting and bowling records. And compare each player's statistics with other characteristics like partnership records, all-rounder records, and wicketkeeper records.
- **Step 7:** After ranking each player, find the final rank of the best player's selection through the linear expression of ranks.
- **Step 8:** Based on the final ranking selection out 11 players from 15 players' bench.
- **Step 9:** Finally different rank priority along with final rank suggests which player belongs to which characteristics.
- **Step 10:** According to the tournament style opponent's strength has been evaluated as per high-rank team selection in a round-robin fashion.

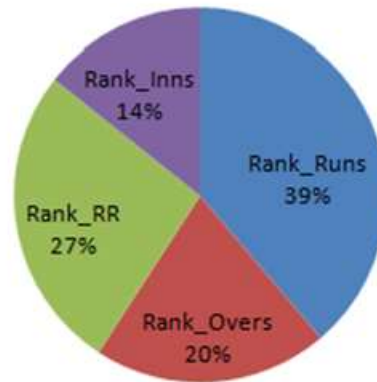
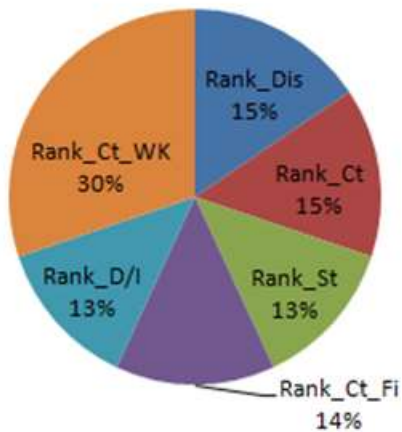
Weight Chart:





Wicket Keeper Case

Player Partnership Case



6 Evaluation

The data analysis and findings determine the actual objectives of the research title. After completion of data collection through web scraping and data pre-processing, one-by-one datasets have been analysed for player evaluation through rank estimation.

6.1 Experiment- Important queries after rank estimation from the datasets

(i) Batting Section:

a. Who made most 6s in the ODI?

```
In [184]: Playerquery_bat_copy.loc[Playerquery_bat_copy['6s'] == Playerquery_bat_copy['6s'].max()]
Out[184]:
```

Player	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	Rank_SR	Rank_Inns	Rank_4s	Rank_50	Rank_6s	Rank_100	Rank_Ave	Rank_0	Total
RG Sharma	2019-2022	40	40	2	1949	159	51.28	2138	91.15	11.762	29.84	175.14	6.08	42.86	7.54	25.6156	1.44	229

4s = 25 columns

In the figure above, it has been observed that after the player's ranking in the batting section a query about who has made the maximum of 6s compared to others was asked for verification of the ranking dataset, and it showed the accurate result.

b. Who made most Runs in the ODI?

```
In [106]: Playerquery_bat_copy.loc[Playerquery_bat_copy['Runs'] == Playerquery_bat_copy['Runs'].max()]
```

```
Out[106]:
```

Player	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	...	Rank_SR	Rank_Inns	Rank_4s	Rank_50	Rank_6s	Rank_100	Rank_Ave	Rank_0	Total_Wei
V Kohli	2019-2022	46	46	2	2121	123	48.2	2295	92.41	...	13.022	35.84	168.14	14.00	7.66	4.54	22.5356	2.44	207.271

rs = 25 columns

In the above screenshot, it has been observed that after the player's ranking in the batting section, a query about who has made a maximum of runs compared to others was asked for verification of the ranking dataset.

(ii) Bowling Section:

a. Who have SR less than 30 but one of first 10 persons wicket takers?

```
In [126]: data_select = Playerquery_bowl_copy.nlargest(10,['Wkts'])
data_select.loc[(data_select['SR'] < 30 )]
```

```
Out[126]:
```

Player	Span	Mat	Inns	Overs	Mdns	Runs	Wkts	BBI	Ave	...	SR	Rank_Overs	Rank_Mdns	Rank_Ave	Rank_Econ	Rank_SR	Rank_4	Rank_5
YS Chahal	2019-2022	34	34	304.4	2.0	1751	62	5/42	28.24	...	1	248.68	0.18	2.4816	1.2844	2.782	3.65	0.94
Mohammed Shami	2019-2022	29	29	254.1	10.0	1456	58	5/69	25.10	...	1	198.38	8.18	-0.6584	1.2644	-0.418	2.66	0.94
M Prasidh Krishna	2021-2022	14	14	112.2	7.0	596	25	4/12	23.92	...	0	56.48	5.18	-1.8384	0.8644	0.282	1.66	-0.06

ows = 23 columns

In the above screenshot, it has been observed that after the player's ranking in the bowling section, a query about who has a strike rate of less than 30 but one of the first 10 persons wicket takers with compared to others was asked for verification of the ranking dataset.

b. Who have maidens greater than 2 but one of first 10 persons wicket takers?

```
In [127]: data_select = Playerquery_bowl_copy.nlargest(10,['Wkts'])
data_select.loc[(data_select['Mdns'] > 2 )]
```

Player	Span	Mat	Inns	Overs	Mdns	Runs	Wkts	BBI	Ave	...	SR	Rank_Overs	Rank_Mdns	Rank_Ave	Rank_Econ	Rank_SR	Rank_4	Rank_5	Total_Weight	Rank
Mohammed Shami	2019-2022	29	29	254.1	10.0	1456	58	5/69	25.10	...	1	198.38	8.18	-0.6584	1.2644	-0.418	2.66	0.94	176.83894	4
Kuldeep Yadav	2019-2022	38	38	351.1	6.0	1965	50	4/18	39.30	...	0	295.38	4.18	13.5416	1.1344	15.482	2.66	-0.06	248.13844	1
JJ Bumrah	2019-2022	28	28	265.5	17.0	1302	43	6/19	30.27	...	1	209.78	15.16	4.5116	0.4344	10.382	0.66	0.94	164.17044	3
B Kumar	2019-2022	25	25	211.5	6.0	1196	42	4/31	27.52	...	0	155.78	4.18	1.7616	0.9944	3.582	1.66	-0.06	136.73444	5
SN Thakur	2019-2022	24	24	184.4	7.0	1167	35	4/7	33.34	...	0	128.68	5.18	7.5816	1.8544	4.982	0.66	-0.06	119.51944	7
M Prasad Krishna	2021-2022	14	14	112.2	7.0	596	25	4/12	23.92	...	0	56.48	5.18	-1.8384	0.8644	0.282	1.66	-0.06	56.94494	10
Hr Pandya	2019-2022	24	20	144.0	5.0	800	23	4/24	34.78	...	0	88.28	3.18	9.0216	1.0944	10.882	0.66	-0.06	82.61744	8

In the above screenshot, it has been observed that after the player's ranking in the bowling section, a query about who has maidens greater than 2 but one of the first 10 persons wicket takers with compared to others was asked for verification of the ranking dataset.

(iii) Fielding Section:

a. Who have most stumping among the wicket keepers?

```
In [150]: M data_select = data_select_extraction.nlargest(10,['St'])
data_select.loc[(data_select['St']==data_select['St'].max())]

Out[150]:
```

Player	Span	Mat	Inns	Dis	Ct	St	Ct_WK	Ct_Fi	MD	Dil	Rank_Dis	Rank_Ct	Rank_St	Rank_Ct_WK	Rank_Ct_Fi	Rank_Dil	Total_Weight	Rank
MS Chori	2019-2019	18	18	19	11	8	11	0	4 (3ct 1st)	1.055	4.566667	-1.0	5.666667	0.833333	-1.833333	0.206667	7.745167	3.0

In the above screenshot, it has been observed that after the player's ranking in the fielding section, a query about who has the most stumping among wicket keepers with compared to others was asked for verification of the ranking dataset.

b. Who have caught most catches?

```
In [151]: M data_select = Playerquery_fielding_copy.nlargest(10,['Ct'])
data_select.loc[(Playerquery_fielding_copy['Ct']==Playerquery_fielding_copy['Ct'].max())]

Out[151]:
```

Player	Span	Mat	Inns	Dis	Ct	St	Ct_WK	Ct_Fi	MD	Dil	Rank_Dis	Rank_Ct	Rank_St	Rank_Ct_WK	Rank_Ct_Fi	Rank_Dil	Total_Weight	Rank
V Kohli	2019-2022	46	46	36	36	0	0	36 (3ct 0st)	3	0.782	30.04	30.32	-0.28	-1.22	31.54	0.40468	77.76751	1.0

In the above screenshot, it has been observed that after the player's ranking in the fielding section, a query about who has caught the most catches with compared to others was asked for verification of the ranking dataset.

(iv) All-rounder Section:

a. List the first 10 ranked Players who's bowling average greater than 30

```
In [166]: M data_select = Playerquery_allround_copy.nlargest(10,['Wkts'])
data_select.loc[(data_select['Bowl_Av'] > 30)]

Out[166]:
```

Player	Span	Mat	Runs	HS	Bat_Av	100	Wkts	BBi	Bowl_Av	...	Rank_Bat_Av	Rank_100	Rank_Wkts	Rank_Bowl_Av	Rank_5	Rank_Ct	Rank_St	F
Kuldeep Yadav	2019-2022	38	65	17	8.12	0	50	4/18	39.30	...	-17.3844	-0.42	41.16	13.9616	0.22	0.32	-0.28	
JJ Bumrah	2019-2022	28	96	14*	12.00	0	43	5/19	30.27	...	-13.5044	-0.42	34.16	4.9316	1.22	-3.68	-0.28	
SN Thakur	2019-2022	24	238	50*	23.80	0	35	4/7	33.34	...	-1.7044	-0.42	26.16	8.0016	0.22	-2.68	-0.28	
HH Pandya	2019-2022	24	716	92*	39.77	0	23	4/24	34.78	...	14.2556	-0.42	14.16	9.4416	0.22	2.32	-0.28	
RA Jadeja	2019-2022	26	465	77	42.27	0	20	2/44	57.65	...	16.7656	-0.42	11.16	32.2116	0.22	6.32	-0.28	

ws x 25 columns

In the above screenshot, it has been observed that after the player's ranking in the all-rounder section, a query about listing the first 10 ranked players who are bowling average greater than 30 compared to others was asked for verification of the ranking dataset.

b. List the first 10 ranked Players who's bowling average as well as batting average greater than 60

```
In [167]: M data_select = Playerquery_allround_copy.nlargest(10,['Total_Weight'])
data_select.loc[(data_select['Bowl_Av'] > 60) & (data_select['Bat_Av'] > 60)]

Out[167]:
```

Player	Span	Mat	Runs	HS	Bat_Av	100	Wkts	BBi	Bowl_Av	...	Rank_Bat_Av	Rank_100	Rank_Wkts	Rank_Bowl_Av	Rank_5	Rank_Ct	Rank_St	Ra
KH Pandya	2021-2021	5	130	58*	65.0	0	2	1/25	111.5	...	39.4956	-0.42	-6.84	85.1516	0.22	-4.68	-0.25	

s x 25 columns

In the above screenshot, it has been observed that after the player's ranking in the all-rounder section, a query about listing the first 10 ranked players whose bowling average as well as batting average greater than 60 compared to others was asked for verification of the ranking dataset.

(v) Player-partnership Section:

a. Who has played 30 overs partnership among most top 10 partnership players

```
In [186]: data_select = Playerquery_partnership_copy.nlargest(10,['Runs'])
data_select.loc[(data_select['Overs'] > 30 )]
```

```
Out[186]:
```

Player1	Player2	Wkt	Runs	Overs	RR	In	Out	Inns	Opposition	Ground	Start Date	Rank_Runs	Rank_Overs	Rank_RR	Rank_Inns	Total_Weight
KL Rahul	RG Sharma	1	227	37.0	6.13	-	1/227	1	v West Indies	Visakhapatnam	18 Dec 2019	190.264574	30.962108	0.397758	-0.461883	392.558565
S Dhawan	RG Sharma	1	193	31.0	6.22	-	1/193	1	v Australia	Mohali	10 Mar 2019	156.264574	24.862108	0.487758	-0.461883	322.075565
S Dhawan	Shubman Gill	1	192	30.5	6.22	-	0/192	2	v Zimbabwe	Harare	18 Aug 2022	155.264574	24.362108	0.487758	0.538117	320.375565
KL Rahul	RG Sharma	1	189	30.1	6.26	-	1/189	2	v Sri Lanka	Leeds	6 Jul 2019	152.264574	23.962108	0.527758	0.538117	314.327565

In the above screenshot, it has been observed that after the player’s ranking in the player-partnership section, a query about who has played 30 overs partnership among most top 10 partnership players compared to others was asked for verification of the ranking dataset.

b. Who has played for top partnership against Australia

```
In [187]: data_select = Playerquery_partnership_copy.nlargest(10,['Runs'])
data_select=data_select.loc[(data_select['Opposition'] == 'v Australia' )]
```

```
Out[187]:
```

Player1	Player2	Wkt	Runs	Overs	RR	In	Out	Inns	Opposition	Ground	Start Date	Rank_Runs	Rank_Overs	Rank_RR	Rank_Inns	Total_Weight
S Dhawan	RG Sharma	1	193	31.0	6.22	-	1/193	1	v Australia	Mohali	10 Mar 2019	156.264574	24.862108	0.487758	-0.461883	322.075565
RA Jadeja	HH Pandya	6	150	18.0	8.33	5/152	5/302	1	v Australia	Canberra	2 Dec 2020	113.264574	11.862108	2.597758	-0.461883	230.118565
MS Dhoni	KM Jadhav	5	141	24.5	5.67	4/99	4/240	2	v Australia	Hyderabad (Deccan)	2 Mar 2019	104.264574	18.362108	-0.062242	0.538117	216.760565

In the above screenshot, it has been observed that after the player’s ranking in the player-partnership section, a query about who has played for top partnership against Australia compared to others was asked for verification of the ranking dataset.

6.2 Win/loss ratio analysis from the datasets extracted



Table 2: India batting in 1st innings (Records from last 4 years)


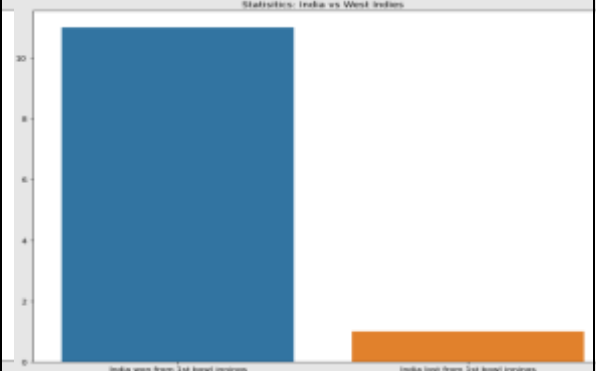
Against Australia	Against West Indies
 <p>India won from 1st bowl innings = 8 India lost from 1st bowl innings = 7</p>	 <p>India won from 1st bowl innings = 11 India lost from 1st bowl innings = 1</p>

Table 3: India bowling in 1st innings (Records from last 4 years)

In table 2 it has been observed that when India took the first batting in the innings and had a face against Australia and West Indies then the ratio against win to loss is measured. Similarly, from taking bowling in the innings the ratio is also measured for observing winning and losing against Australia and West Indies. It is secondary important to select a team for doing up and down ranked selection. It is also preferred for selecting a new player and a rising player from the high-ranked players. The result observed from the above graphs is that while considering teams such as Australia and West Indies, the win ratio is huge when team India is bowling in the first innings.

6.3 Add/remove player's condition

```
In [293]: select_batting_player
```

	Player	Batting_Rank	Total_Weight_Batting	Highest_Rank_partnership
0	RG Sharma	1.0	229.74216	1.0
1	S Dhawan	2.0	216.73816	2.0
2	V Kohli	3.0	207.27816	12.0
3	SS Iyer	4.0	135.24916	6.0
4	KL Rahul	5.0	132.14016	1.0
5	RR Pant	6.0	98.57516	16.0
6	HH Pandya	7.0	89.60016	8.0
7	Shubman Gill	8.0	75.44716	3.0
8	SV Samson	10.0	49.42316	43.0

So Again the batting partnership of SV Samson is not acceptable

In this figure above, SV Samson has a very low value of weight for rank partnership so this player cannot be selected. So looking for a new player under acceptable and marginal conditions like that good SR, average partnership, and good fielding rank also in the batting side. After searching the players, the results have been taken out.

```
array(['KL Rahul', 'S Dhawan', 'Ishan Kishan', 'RA Jadeja', 'MS Dhoni',
      'RG Sharma', 'Shubman Gill', 'SS Iyer', 'HH Pandya', 'KM Jadhav'],
      dtype=object)
```

Three players (RA Jadeja, MS Dhoni, and KM Jadhav) are reviewed for team selection. Now the question is:

- Which player is having best partnership amongst RA Jadeja, MS Dhoni, and KM Jadhav?

Player1	RA Jadeja	Player1	MS Dhoni
Player2	HH Pandya	Player2	KM Jadhav
Rank	8.0	Rank	9.0
Name: 7, dtype: object		Name: 8, dtype: object	

From the above table, it is seen that RA Jadeja has a better partnership than MS Dhoni and KM Jadhav. Hence, RA Jadeja is selected.

6.4 Final Rank Estimation

Player	Rank	Batting_Rank	Highest_Rank_partnership	Rank_for_allrounder	Rank_for_fielding	Rank_for_WicketKeeper
V Kohli	1.0	3.0	12.0	19	1	0
S Dhawan	2.0	2.0	2.0	25	2	0
RG Sharma	3.0	1.0	1.0	21	5	0
Kuldeep Yadav	4.0	34.0	200.0	13	19	0
RA Jadeja	5.0	15.0	8.0	4	7	0
HH Pandya	6.0	7.0	8.0	5	14	0
YS Chahal	7.0	37.0	206.0	7	10	0
Mohammed Shami	8.0	33.0	157.0	10	9	0
SN Thakur	9.0	14.0	52.0	9	25	0
KL Rahul	10.0	5.0	1.0	22	4	2
JJ Bumrah	11.0	35.0	106.0	17	34	0
SS Iyer	12.0	4.0	6.0	23	8	0
B Kumar	13.0	32.0	54.0	16	18	0
RR Pant	14.0	6.0	16.0	27	3	1
Shubman Gill	15.0	8.0	3.0	20	17	0

From the above table, it is observed that the final ranks of the selected 15 players are most essential to select the 11 players instead of calculation of Combination ${}^{15}C_{11} = \frac{15!}{(15-11)! \times 4!}$.

For instance, player V Kohli has the greatest batting as well as fielding rank. So, this player is considered a batsman. After performing the calculation, and cumulating all the weights of the respective players, the final squad has been selected.

6.5 Discussion

From the valuation part, it has been observed how the best players are selected through the linear and weightage-based rank system. It has also been verified how the rank selection is predicted from the match (India v Australia 2020 3rd ODI Sun 19 January, 13:30 Local

(13:30 IST) M.Chinnaswamy Stadium, Bangalore, India). In this match, India had batted at first with a run rate of 6.08 and scored 289 with a fall of wickets 3. And Australia lost this match they had scored 286 runs at a run rate of 5.72 and fell off nine wickets. Below is the player's selection for this match.

Actual Player (Who played the match)	Predicted Ranked Players	Player Matched
Rohit Sharma	V Kohli	V Kohli
KL Rahul	S Dhawan	S Dhawan
Virat Kohli	RG Sharma	RG Sharma
Shreyas Iyer	Kuldeep Yadav	Kuldeep Yadav
Manish Pandey	RA Jadeja	Mohammed Shami
Shikhar Dhawan	HH Pandya	JJ Bumrah
Ravindra Jadeja	YS Chahal	SS Iyer
Mohammad Shami	Mohammed Shami	RA Jadeja
Kuldeep Yadav	SN Thakur	KL Rahul
Navdeep Saini	KL Rahul	
Jasprit Bumrah	JJ Bumrah	
	SS Iyer	
	B Kumar	
	RR Pant	
	Shubman Gill	

Table 4: Comparison between Prediction and Matched Player

From this above table, it can be found that the total player matched is 9 out of 11 players who have been selected to play the match. The accuracy rate is about 81.8%. The rest of the work which is included in the report has been added to the configuration manual.

7 Conclusion and Future Work

In this report, to achieve the outcome of the research objective KDD methodology has been applied. Initially, the data is extracted from the sports website called 'espnricinfo' with the required query from the last 4 years to evaluate the model. The portal is based on selection features like batting, bowling, all-rounder, fielding, wicket-keeping, player-partnership, and team first-innings win ratio for both batting and bowling are web-scraped using python coding. Then the data quality has been assessed, all the null values are checked and replaced with zeroes. Later, the data was transformed into proper datatypes for further steps. After getting the data ready, a ranking-based indexing strategy based on linear programming is applied to the attributes of each player on every selective feature. Each player's features and other aspects are thoroughly analysed using the linear discriminate approach to determine their ranking. After achieving the ranks for the attributes, the total weightage is found and on the criteria of 9 batsmen, (including both all-rounders and wicket-fielders) and 6 full-time bowlers are selected by considering all the aspects. After the final 15 are selected, for checking the accuracy it has been cross validated with one of the previous matches and it is found that nine players have been matched which says the experiment has an accuracy of 81.8%. In future work the classifiers can be used for finding the rank estimation when new youngsters emerge onto the ground and based on the importance of that

match, players can be substituted and given rest for upcoming big tournaments. The accuracy of the findings can be improved without compromising the factors that help in the selection.

8 Acknowledgement

I would like to thank my supervisor Dr. Cristina Hava Muntean for her constant support and feedback without whom this project would be accomplished. I also want to thank the NCI library mentor (Shannon Mallon) for the support without hesitation. And all the staff NCI staff directly or indirectly help me reach this stage with all knowledge and guidance. And finally, my lovable parents for standing by me at all odds and encouraging me.

References

- Agrawal, P. and Ganesh, T., 2020, April. Selection of Indian Cricket Team in ODI using Integer Optimization. In *Journal of Physics: Conference Series* (Vol. 1478, No. 1, p. 012001). IOP Publishing.
- Bhattacharjee, D. and Saikia, H., 2016. An objective approach of balanced cricket team selection using binary integer programming method. *Opsearch*, 53(2), pp.225-247.
- Chakraborty, S., Kumar, V. and Ramakrishnan, K., 2018. Selection of the all-time best World XI Test cricket team using the TOPSIS method. *Decision Science Letters*, 8(1), pp.95-108.
- Hussain, A., Qiang, Y., Bilal, A.Q.M., Ullah, U. and Ullah, N., 2019. Region-based teams ranking in the game of cricket using PageRank algorithm. *Int. J. Comput. Appl*, 177(16), pp.10-15.
- Hussain, A., Yan, Q., Bilal, M.A.Q., Wu, K., Zhao, Z. and Ahmed, B., 2019. Region-wise ranking for one-day international (ODI) cricket teams. *International Journal of Advanced Computer Science and Applications*, 10(10).
- Jayanth, S.B., Anthony, A., Abhilasha, G., Shaik, N. and Srinivasa, G., 2018. A team recommendation system and outcome prediction for the game of cricket. *Journal of Sports Analytics*, 4(4), pp.263-273.
- Jha, A., Kar, A.K. and Gupta, A., 2022. Optimization of team selection in fantasy cricket: a hybrid approach using recursive feature elimination and genetic algorithm. *Annals of Operations Research*, pp.1-29.
- Kissell, R. and Poserina, J., 2017. *Optimal sports math, statistics, and fantasy*. Academic Press.
- Mahbub, M.K., Miah, M.A.M., Islam, S.M.S., Sorna, S., Hossain, S. and Biswas, M., 2021, November. Best Eleven Forecast for Bangladesh Cricket Team with Machine Learning Techniques. In *2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)* (pp. 1-6). IEEE.
- Perera, W.A.S.C., 2021. *Predicting an Optimal Sri Lankan Cricket Team for ODI Matches According to the Nature of the Game* (Doctoral dissertation).

Prakash, C.D., Patvardhan, C. and Singh, S., 2016. A new machine learning based deep performance index for ranking IPL T20 cricketers. *International Journal of Computer Applications*, 137(10), pp.42-49.

Premkumar, P., Chakrabarty, J.B. and Chowdhury, S., 2020. Key performance indicators for factor score-based ranking in One Day International cricket. *IIMB Management Review*, 32(1), pp.85-95.

Shah, P. and Patel, M.N., 2018. Ranking the cricket captains using principal component analysis. *Int. J. Physiol. Nutr. Phys. Educ*, 3(2), pp.477-483.

Shah, P., 2017. New performance measure in Cricket. *ISOR Journal of Sports and Physical Education*, 4(3), pp.28-30.

Singla, S. and Shukla, S.S., 2020. Integer optimisation for Dream 11 cricket team selection. *International Journal of Computer Sciences and Engineering*.

Vetukuri, V.S., Sethi, N. and Rajender, R., 2021. Generic model for automated player selection for cricket teams using recurrent neural networks. *Evolutionary Intelligence*, 14(2), pp.971-978.

Viva Question and answers

Q1: How many players were there in total in the dataset? Was there a way to automate the ranking/team creation with machine/deep learning solutions?

Ans1: In this research project, there are total 5 datasets based on the feature selection such as players based on their batting performance, bowling performance, all-rounder performance, fielding impact, and player-partnership built. In here, all datasets are stored in solution artifact folder which has already been submitted in the moodle. The batting data has 50 players named as 'bat.csv', bowling dataset has 50 players named as 'bowl.csv', all-rounders dataset has 50 players named as 'allround.csv', fielding dataset has 50 players named as 'fielding.csv', and player-partnership dataset has 446 rows with each player batting with other players in that time frame named as 'partnership.csv'.

There can be many machine/deep learning methods to arrive at the solution of ranking or team creation. In this scenario, rank method was used and performed by linear discrimination method (which is one of the machine learning approach) and weightage calculation based.

Q2: Are there other queries that could have been used in the ranking process?

Ans2: Other than the feature selection that has been mentioned in the introduction of the report, wicket-keeping is also one of the important factors that need to be considered while ranking the players. Apart from these in the historical data used in the project, there can be filtered as players who can play best on different types of pitches/grounds. One more query that can be used is the weather condition on the day they are playing, it can be a hot climate, a cold place, a day match, or a night match. These could have been used if a proper dataset is found. If the queries have to be within the dataset used, then the query can be the dependency of the bowling pair that can be used to break the partnership of the opposite team or the bowling pair that can be used in powerplay or death over bowling. For making rank-based selection there are queries for best feature selection.

Q3: Does the presence of a player in the team influence the performance of another player? (e.g. when player A plays, player B plays better, so it is good to have player A and B together)

Ans: Yes, the presence of one player will definitely influence the performance of another player. In this case, one of the feature selections which has been included is player-partnership. Here, there are situations where the batsman has been performing better when another batsman whom he is comfortable with is playing on the other end of the batting. While ranking the player partnership, queries have been used to check if there were any influencing players that need to be added to the final list.

Q4: Would you be able to predict a player's performance based on the historical data? How would you do that?

Ans: Yes, it is possible to predict the player's performance based on historical records. In this situation, the first datasets were gathered through web scrapping from www.espn.com/cricket, which are completely based on their statistics in past years. For instance, the performances/statistics of the players based on feature selections were taken from the last four years. And then take the data of the players and by linear discriminatory method and weightage calculations of each attribute the players' statistics are checked and the rank of each player as well as the player's performance can be achieved. Finally, the queries are used to check the accuracy of the values.

Q5: Briefly present the limitations of your research work.

Ans: When it comes to limitations, there are a few aspects that are not considered while predicting the players. In this work, only players who have played for India have been considered. Youngsters who have played well but didn't get the chance to play for the Indian team are not included. Another limitation is that this is the method applied on Indian one-day international, but when it comes to the 20-over format the players need to be more aggressive, runs should be scored quickly, and the bowlers should focus more on taking wickets than reducing the runs. Some of the conditions are the player's health condition before, during, and after the match is not considered here. So, one cannot predict the player's fitness without these. Also, in this project players' emotion is not considered, and their mental being was not included. Hence in the project, further research can be carried out through these parameters to obtain well suit players for match specific.