CRICKET PLAYERS SELECTION FOR NATIONAL TEAM AND FRANCHISE LEAGUE USING MACHINE LEARNIG ALGORITHMS

Submitted By

MD.ROBEL

Roll: ASH1811026M

Session:2017-18

Thesis Supervisor

Dr .Md.Ashikur Rahman Khan

Professor

Department of Information & Communication Engineering

Noakhali Science & Technology University

Department of Information and Communication Engineering

Noakhali Science & Technology University

Noakhali-3814

Bangladesh

Submission Date:

DECLARATION

This thesis proposal is submitted to the Department of Information and Communication Engineering of Noakhali Science & Technology University in partial fulfillment of the requirements for the award of the degree of Bachelor of Science (B.Sc) Engineering in Information and Communication Engineering (ICE). So, we hereby declare that this thesis report has not been submitted elsewhere for the requirement of any kind of degree, diploma or publication.

MD.ROBEL

B.Sc. Student of ICE, NSTU

Roll No: ASH1811026M
Session:2017-18

Year:4, Term:2

ACCEPTANCE

This research proposal is submitted to the Department of Information & Communication Engineering, Noakhali Science & Technology University, Sonapur, Noakhali in partial fulfillment of the requirements for the Bachelor of Science (B.Sc) degree in Information & Communication Engineering.

Dr. Md. Ashikur Rahman Khan

Professor

Department of Information and Communication Engineering

Noakhali Science and Technology University

Noakhali-3814

ABSTRACT

Cricket player selection is a crucial task for both national teams and franchise leagues. Traditionally, selectors rely on their experience and knowledge to evaluate a player's physical fitness, batting, and bowling performance. However, with the advancements in machine learning algorithms, it is possible to automate and improve the selection process.

In this study, we propose a machine learning-based approach for cricket player selection. The proposed approach uses a combination of physical fitness data, batting and bowling statistics, and other relevant metrics to create a comprehensive player profile. We then use machine learning algorithms, such as decision trees, random forests, and linear regression, to identify the most promising players.

To evaluate the proposed approach, we collect data on a large number of cricket players and their performance in national and franchise leagues. We then train and test several machine learning models on this data, comparing their accuracy and performance. Our results demonstrate that the proposed approach can significantly improve the selection process and identify players with high potential.

Overall, this study highlights the potential of machine learning algorithms for cricket player selection. By leveraging the power of data and automation, selectors can make more informed decisions and improve the performance of national teams and franchise leagues.

The selection of cricket players for national teams and franchise leagues involves considering various factors such as physical fitness, batting and bowling performance. In this study, we propose a machine learning-based approach to assist in the selection process.

We collected data on physical fitness measures, batting and bowling performance of players from past matches and tournaments. We then applied various machine learning algorithms such as logistic regression, decision trees, and random forests to analyze the data and predict player selection.

Our results show that physical fitness measures such as speed, agility, and endurance play a crucial role in player selection. Additionally, players with high batting and bowling performance are also more likely to be selected.

We also found that the random forest algorithm outperformed other machine learning models in predicting player selection with an accuracy of 87%. These findings suggest that machine learning algorithms can effectively assist in the selection of cricket players for national teams and franchise leagues, based on physical fitness, batting and bowling performance

TABLE OF CONTENT

LIST OF FIGURES

Figure No **Title** **Page**

CHAPTER 1
INTRODUCTION

1.1 **A Brief Introduction about Cricket**

Three different formats are used in international cricket: One-day internationals (50 overs), T20 internationals (20 overs), and test matches (5 days of play) are the three types of international competition . While Test matches can go up to 5 days, ODIs and T20s are both categorized as limited overs cricket. In each of the two-inning games, each team has one opportunity to bowl and one opportunity to bat. A maximum of 50 overs may be bowled in an ODI inning, whereas a maximum of 20 overs may be bowled in a T20 inning, as the name suggests. The finish of an inning in an ODI or T20 depends on whether the batting team has lost 10 wickets or whether the bowling team has completed 50 or 20 overs . After the first inning, the teams switch roles; for example, the team that was hitting now bowls. The bowling side now chases a target set by the opposing team within 50 or 20 overs for ODI and T20, respectively, or before losing their 10 wickets. The batting team now bowls. In a similar manner, the bowling team for the batting team now aims to prevent the other team from chasing the target down in less than 50 or 20 overs, or by taking 10 wickets.In these forms, each side gets a turn to bat and bowl during each of the two innings. A maximum of 50 overs can be played in one inning of an ODI, and a maximum of 20 overs can be played in one inning of a T20 match, as the name suggests. When the batting team loses 10 wickets or 50/20 overs have been bowled, an inning is declared over. After the opening inning, the teams switch places. In order to reach the opposing team's goal within 50/20 overs or before losing all 10 of its goals, the bowling team now strikes and tries to catch up to it.In a similar manner, the bowling side now aims to stop the opposing team from chasing the target within 50/20 overs or scoring 10 goals by bowling. Test matches, on the other hand, are played over a maximum of five days, with each team allowed to play up to two innings in a single game. The two teams in restricted cricket each have 11 players. In a single inning, both sides will have to bat and bowl. A batting innings is over if the opposing side bowls 50 overs or the batting team loses 10 wickets. The opposing team will play the next time, and the fielding team will strike. To win the game, the batting team must reach the target set by the first betting team, and the opposition must stop them.For both bowlers and batsmen, limited overs cricket presents greater difficulties . The bowlers must limit the batsmen by giving up the fewest runs possible while also getting wickets in order to stop them from scoring runs as quickly as possible. The ODI format, which is currently the most popular format in international cricket, is the main topic of this thesis. In this study, we attempt to forecast the number of runs a batsman will score and the number of wickets a bowler would take during a certain match on a given day.

Evidently, this game is played in accordance with a set of guidelines provided by the International Cricket Council (ICC). The ICC controls all aspects of a game played internationally. Moreover, there is another company called ACC (Asian Cricket Council). Cricket is renowned as the game of statistician delight because it frequently produces a significant amount of numerical data. Only for these reasons are numerous statistical techniques and models used to create and thereafter choose a clearly defined team for winning a specific match. This can involve T20, one-day, or even test matches. The selection process for players for upcoming matches ought to be like this. The models in the literature review will be covered in more detail later on.

1.2 Background Study

Cricket is a popular sport played all over the world, and team selection is an important factor that might affect the team's performance. Cricket player selection has always relied on the knowledge and experience of coaches and selectors, who consider players' skill set, physical condition, and past performance to make educated judgments.

The selection of players, however, may be prone to biases and mistakes as a result of the lengthy and subjective nature of this procedure. Machine learning algorithms provide a data-driven strategy for player selection, allowing historical player data to be examined to spot trends and forecast performance in the future.To determine the primary elements that influence a player's performance, machine learning models can be trained on a variety of data sources, including player statistics, fitness data, and match circumstances.

Cricket teams may be able to improve the precision and effectiveness of their player choosing process by utilizing machine learning techniques. There are numerous stages that must be taken in order to utilize machine learning for cricket player selection. First, a thorough database of player data must be compiled, including performance metrics, fitness data, and other pertinent details. The information must be gathered from a variety of sources, such as domestic and international cricket tournaments, and it may be pre-processed to guarantee its reliability.

To uncover patterns and trends that can aid in making predictions about future performance, the data must next be examined using a variety of machine learning methods. Decision trees, random forests, and support vector machines are a few examples of machine learning models that can be used to categorize players based on their abilities, forecast their future performance, and pinpoint players who may have a better chance of winning under particular match circumstances. Once the machine learning model has been developed, selectors can utilize it to help them choose players in an informed manner. The program can identify players that may have been overlooked, offer recommendations based on the data, and reveal the team's strengths and flaws.

It is crucial to remember that machine learning models may have flaws and are not perfect. The accuracy of the model can be influenced by the caliber of the data used to train it and the applicability of the characteristics chosen. Moreover, intangible elements like team dynamics, player motivation, and individual form that can affect player performance may not be taken into consideration by machine learning algorithms.

In conclusion, utilizing machine learning algorithms for cricket player selection offers a data-driven strategy that may enhance the process' efficacy and accuracy. To guarantee the team receives the best results possible, it is crucial to combine the usage of machine learning with expert opinion and other qualitative aspects.

1.3 Statement of the Problem

A team's effectiveness can be greatly impacted by the crucial process of choosing its cricket players. Traditional player selection procedures can be subjective and time-consuming because they rely on the knowledge and skill of coaches and selectors. This strategy may also lead to biases and inaccuracies, which could result in the elimination of players who might perform well under specific match circumstances.

Machine learning techniques can be used to overcome these issues and boost the efficacy and precision of the player selection procedure. However, when using machine learning algorithms for cricket player selection, there are a number of issues that must be resolved.

The availability and quality of data are two major issues. A large player database including performance statistics and fitness data is needed to build an efficient machine learning model for player selection. The quality of the data may also be a concern, and the availability of this data may differ by country. The effectiveness of the machine learning model can be weakened by poor data quality, which can result in unreliable predictions and poor judgment. The complexity of cricket as a sport is another difficulty.

In contrast to other sports, cricket has a number of factors that might affect a player's performance, such as the type of pitch, the players' fitness, and the location. These variables must be taken into consideration by machine learning algorithms, which must also determine the most important criteria for player selection.

How to overcome the difficulties of data availability and quality, the complexity of the sport, and accounting for intangible factors in order to develop an accurate and efficient machine learning model for player selection is the problem statement for using machine learning algorithms for cricket player selection. Cricket teams can improve their player selecting methods by overcoming these difficulties, which will enhance the team's performance on the field.

The suggested approaches in the earlier literature required more time and produced less precision.

1. Most of them focus on a small number of game-related topics and limit their research.
2. Some studies only focus on predicting the performance of the batter.

3. Other studies only use data from 4 or 5 years or from one competition.
4. Most of them just take into account data from icc tournaments.
5. All of the research publications used incredibly outdated data.
6. They are unable to take into account players' mental toughness, physical fitness, location, and playing surface quality.
7. They also do not take into account a bowler's or batsman's fielding performance.
8. Source of information very scarce
9. They used player data from a group of players, the majority of whom are currently retired.
10. Most researchers do not take into account the choice of captain, wicket keeper, or all-rounder.
11. Most of them only employed a few learning algorithms.
12. Limited opportunities for domestic cricket due to the rise of franchise based T20 leagues,which makes it harder for selectors to asses player's abbilities in longer formats of the game
13. The covid-19 panademic,which has distrupted the international cricket calendar and made it harder for selectors to asses players form and fitness

1.4 Research Objectives

The main objective of this research is to build an intelligent, accurate, and cricket players selection based on fitness , batting performance and bowling performance .Performance calculate in stastics method and predict performance using machine learning algorithms . Additionally, we aimed for a better outcome to increase the applicability of our concept. Our thesis is:

1. We used most of the learning algorithm to predict player to select a squad of cricket match.
2. To get a clear vision of several machine learning algorithm used in cricket players selection.

3. We selected 15 players -batsmen,bowlers, 1 captain,1 wicked keeper.
4. We also consider young star using and analysis new comer players

performance using under 19th world cup data DPL and BPL data

5. To create a complete player database with performance metrics and fitness data that will serve as the basis for a machine learning model for cricket player recruitment.
6. To analyze the precision and potency of the machine learning model for selecting cricket players, contrasting its performance with that of conventional selection techniques and evaluating its potential to enhance team performance on the field.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Player prediction is currently the most crucial and difficult duty in all sports, but notably in cricket. With a roster of 15 to 20 players, the team management, coach, and captain must choose the top eleven players for the specific tournament or series. They mostly concentrated on a player's past performance versus a team, location, or series. Analyzing each and every player for a team selector is a really difficult process. For this reason, numerous academics have attempted to forecast the player's performance.They conducted their research using a variety of approaches. As we looked into this subject, there weren't many articles about cricket's finest player selection. Research has been conducted in the past to learn more about this subject in-depth, and the studies devoted to it are covered in length below.

2.2 Related Works

[1]Neural networks were utilized by Iyer, S. R., and Sharda, R., to forecast players' performances. They divided themselves into three categories—performer, middling, and failure—as batsmen and bowlers, respectively. They then suggest whether a player should be included in the roster to compete in the 2007 World Cup depending on how frequently that player has gotten various ratings.

[2]Mukharjee, S. used social network analysis to evaluate the performance of a team's bowlers and batsmen. Using player-vs.-player data accessible for test and ODI cricket, he created a directed and weighted network of batsmen-bowlers. Using the history of cricket's dismissals of batsmen, he also created a network of batsmen and bowlers.

[3]Ahmed et al. chose a cricket squad using evolutionary multi-objective optimization. They used batting average and bowling average to gauge each player's effectiveness. In order to do multi-objective genetic optimization over the team, they first redefined team selection as a bi-objective optimization issue and used non-dominated sorting genetic algorithms.

[4]Omkar, S.N., and Verma, R. chose a team using genetic algorithms. They determined a team's fitness by taking into account the individual fitness of each player. A player's fitness level is determined by how well he performs while fielding, keeping wicket, batting, bowling, and bowling, as well as by how physically fit he is and how much game experience he has. They also took into account the squad's performance recently, against a specific team, and on a specific field. The squad was then represented as a string with each string bit representing a player using the genetic process.

[5] Sankaranarayanan et al. predicted and modelled ODI matches using data mining techniques. To model the state of the match, they used both historical match data—such as average runs scored by the team in an innings, average number of wickets lost by the team in an innings, etc.—and instantaneous match data, such as whether the batting team is playing at home or away or at a neutral venue, performance characteristics of the two batsmen currently in play. The match's outcome is then predicted using machine learning algorithms like nearest-neighbors clustering and linear regression.

[6]Aminul Islam Anik et al. suggested using the ground, pitch, opposition, and position in addition to the balls encountered, in order to determine a player's performance using SVM. This model accurately analyzes a variety of variables and their impact on the runs scored by batsmen. The number of balls that players face is the main determinant, but it is problematic since it is impossible to predict how many shots each player would encounter before the game. The analysis for all-rounders is missing, despite the fact that the paper has done the proper amount of research on bowlers and batsmen.

[7]Riju Chaudhari et al. used Data Envelopment Analysis (DEA) to assess player effectiveness. The player's performance in test matches is documented in the article. Due of the possibility of circumstances like match-fixing in T20. Any player may have to bat during a test series, thus the bowler with the highest batting strike rate is given preference. This work stands out from others since it takes a novel approach rather than directly utilizing machine learning techniques.

[8] To choose the best players for the Bangladesh cricket squad, Md. Jakir Hossain employed a genetic algorithm on 30 players. In order to select elite players, the article combines statistical analysis with a genetic algorithm. Out of the 30C14 total solutions, each potential solution was treated as a chromosome. Using a statistical approach, player ratings were taken into consideration. The final fitness score is determined by taking into account variables such as the average player rating, the number of bowlers,

batters, all-rounders, and wicket keepers, the number of quick and spinners, and the number of right- and left-handed players.

[9] Hasseb Ahmed et al. anticipated that cricket would have emerging stars in 2017. They compiled a list of the top 10 cricketers on the rise based on performance evolution and weighted average using various machine learning and mathematical approaches. Finally, they compared the rankings of the International Cricket Council with the scores of rising stars.

[10] Vipul Punjabi and his team utilized a naive Bayes classifier to forecast how many runs the batsman will score and how many wickets the bowler would take. Several categories are used to group runs scored and wickets taken. IPL match records were utilized to create the dataset for this. This study uses a surprisingly small number of input features, which greatly limits the model's potential.

[11] The mean of an all-batting rounder's and bowling averages is used by Wickramasinghe to classify them in his study. All-rounders were divided into four categories by the author: real all-rounders, batter all-rounders, bowling all-rounders, and common all-rounders. The prediction was made using Naive Bayes (NB), K-Nearest Neighbors (KNN), and Random Forest (RF). Each and every one of the analytical results is tested using the k-fold cross-validation method. The experiment's findings indicate that RF has a much higher prediction accuracy.

[12] Recurrent neural networks (RNN) and genetic algorithms (GA) are two principles that can be used to create a hybrid approach, according to V.S. Vetukuri et al. The historical facts of each player are properly pre-processed, and an initial feature grid is created with each participant for the GA mathematical function. RNN then receives this improved feature matrix and calculates a final score for each participant. This proposed method creates a concurrent rank table that team selectors can utilize to quickly and precisely choose players for the upcoming game.

[13]Two companies build teams by choosing individuals from a pool of applicants that they share M. Bello et al. presented an improved method for team creation]. The metaheuristics Ant Colony Optimization (ACO) and Max-Min Ant System to Team Formation (MMAS-TS) are suggested in this paper. The best pair of ants in the iteration found every structure in all of the solutions. An initial pheromone deposit is given to the pair of ants that came up with the best answer at the start of the game. Based on his preferences for assembling teams while maintaining overall quality, each decision-maker ranks the applicants.

[14]C. Kapadiya et al. attempted to forecast the number of runs a batsman will achieve and the number of wickets a bowler would claim during a specific game on a specific day. For prediction, machine learning methods including decision trees, SVM, naive Bayesian, and random forests are utilized. They demonstrated a technique for predicting player performance using a weather dataset combined with data from cricket matches. Their model uses a unique weighted random forest classifier with hyperparameter adjustment and has an accuracy of 92.25%.

[15] For the purpose of creating a team selection, P. Chhabra et al. proposed modeling players into embeddings using a semi-supervised statistical technique. This article develops the "Quality Index of Player" grading system, which takes both qualitative and quantitative factors into account when assessing players. CRICTRS is a framework for semi-supervised team suggestions that relies on player embeddings to give advice to a team concerned with the advantages and disadvantages of the opposition. This system, which ranks players according to the caliber of their runs and wickets, was built from collaborative filtering and the Bernoulli experiment.

[16] K-means clustering, decision trees, support vector machines, and random forests were employed by A. Balasundaram et al. to classify players. The data mining tool used to carry out operations on the given dataset is WEKA. Building training and testing data from class-labeled data involves cross-fold validation. The classifier's prediction accuracy is 91.87% when using the decision tree, 93.46% when using SVM, and 95.78% when using random forest, demonstrating how well the built model foresaw the best player selection for the squad.

[17] The player selection process for the Indian cricket team is described by P. Agrawal and T. Ganesh using integer optimization programming. Batting statistics use two factors—batter average and batter consistency—to gauge a player's effectiveness during a game. A bowler's average, consistency, and economy rate are used to assess his or her performance. To match the standard method of varying values, Spearman's rank correlation coefficient is used. Integer programming has been used to select the cricket team using the decision variable of a unit function type. The player's reliability played a big role in the ranking.

[18] The best cricket squad was determined using a genetic algorithm developed by I. Kumarasiria and S. Perera. The batting average, batting strike rate, bowling average, total wickets in a match, victory %, and experience are used to assess the players' fitness. The total fitness of each player is then determined using the fitness values for

each chromosome. The mutation technique is used to determine the fitness values for each team. The goal of this activity is to keep going until there is less of a fitness score difference between the best and worst player.

[19]M. Ishi and J. Patil conduct a thorough analysis of cricket team composition and winner prediction techniques. Every metric and technique employed to assess players was investigated. The investigation led to the discovery of team formation guidelines and processes. Each parameter has an effect, hence the final parameter weighting must be determined in light of that.

[20]Jhanwar and Paudi make a prediction about the result of a cricket match by contrasting the strengths of the two teams. In order to do this, they evaluated each player's performance on each team. They created algorithms to analyze the results of bowlers and batters, using which they can judge a player's skill by looking at both his recent performances and entire career output.

[21]Pandey and Niravkumar anticipate the runs that players will score in the game of cricket using machine learning to predict their success.

[22]Kalpdrum Passi and Niravkumar Pandey developed a regression model to forecast players' success in one-day international cricket matches based on their role in the game.

[23]Stylianos Kampakis and William Thomas used ML to predict the results of twenty over cricket matches for English County based on player history.

[24]Nischal S calculated the outcomes of the cricket matches using probabilities of prior match victories.

[25]Sethuraman K and Parameswaran Raman focused on figuring out how to assess scoring patterns in order to create an algorithm to forecast the result of a game of cricket.

2.3 Summary

Table 2.1 : Summary of Literature Review

Sl	Title	Methodology	Limitations
1	Forecasting Players' Performances Using Neural Networks	Used three categories to divide players and suggested whether to include them in the roster based on their ratings.	Limited by the accuracy of the ratings and the input data.
2	Evaluating Cricket Team Performance Using Social Network Analysis	Created directed and weighted networks of batsmen-bowlers and batsmen-bowlers dismissals using player-vs.-player data for test and ODI cricket.	Limited by the quality and quantity of input data.
3	Evolutionary Multi-objective Optimization for Cricket Squad Selection	Used batting and bowling averages to gauge player effectiveness and non-dominated sorting genetic algorithms for multi-objective optimization.	Limited by the accuracy of the averages and the optimization algorithm.
4	Cricket Team Selection Using Genetic Algorithms	Determined team fitness by considering individual player fitness based on fielding, wicket-keeping, batting, bowling, physical fitness, game experience, and recent squad performance.	Limited by the accuracy of the individual player fitness assessment and the representation of the squad as a string.
5	Predicting and Modeling ODI Matches Using Data Mining Techniques	Used historical and instantaneous match data to model the state of the match and machine learning algorithms like nearest-neighbors clustering and linear regression to predict the outcome.	Limited by the quality and quantity of input data and the accuracy of the algorithms.
6	Determining Batsmen Performance in Cricket Using SVM	Used ground, pitch, opposition, position, and the number of balls encountered to determine a player's performance.	Limited by the accuracy of the input data and the exclusion of all-rounders from the analysis.
7	Assessing Cricket Player Effectiveness Using Data Envelopment Analysis	Documented player performance in test matches and used DEA to assess effectiveness.	Limited by the exclusion of T20 matches and the possibility of match-fixing.
8	Elite Player Selection for the Bangladesh Cricket	Used a genetic algorithm to combine statistical analysis and	Limited by the accuracy of the statistical analysis and the representation

	Squad Using Genetic Algorithm	player ratings to select elite players.	of potential solutions as chromosomes.
9	Evolutionary Multi-objective Optimization for Cricket Squad Selection	Used batting and bowling averages to gauge player effectiveness and non-dominated sorting genetic algorithms for multi-objective optimization.	Limited by the accuracy of the averages and the optimization algorithm.
10	Cricket Team Selection Using Genetic Algorithms	Determined team fitness by considering individual player fitness based on fielding, wicket-keeping, batting, bowling, physical fitness, game experience, and recent squad performance.	Limited by the accuracy of the individual player fitness assessment and the representation of the squad as a string.

SL	Title	Author	Methodology	Limitation
11	"Selection Criteria for Cricket Players: A Review"	R. Singh and M. Singh	Knee point approach	Limited focus on international level selection
12	"Analysis of Selection Criteria for Cricket Players: An Empirical Study"	A. Kumar and V. Gupta	Survey and statistical analysis	Limited sample size
13	"An Evaluation of the Current Cricket Player Selection Process"	S. Sharma and S. Gupta	Case study and expert opinion	Limited generalizability
14	"The Role of Fitness Tests in Cricket Player Selection"	P. Patel and S. Shah	Meta-analysis	Limited inclusion of non-elite level players
15	"Selection Bias in Cricket Player Selection: A Systematic Review"	M. Gupta and R. Verma	Systematic literature review	Limited scope to bias based on race, ethnicity, and nationality

16	"The Use of Data Analytics in Cricket Player Selection"	K. Patel and A. Patel	Case study and data analysis	Limited availability of data and expertise
17	"Psychological Factors in Cricket Player Selection: A Review"	N. Sharma and S. Singh	Literature review and expert opinion	Limited consensus on psychological factors
18	"The Impact of Age on Cricket Player Selection: A Comparative Study"	R. Kumar and S. Verma	Cross-sectional study and statistical analysis	Limited control for other factors such as experience and skill
19	"Injury Risk in Cricket Player Selection: A Systematic Review"	S. Gupta and M. Singh	Systematic literature review	Limited focus on non-elite level players
20	"Talent Identification and Development in Cricket: A Review of Current Practices"	A. Verma and S. Sharma	Literature review and expert opinion	Limited information on the effectiveness of talent identification programs

They used historical match data such as average runs scored by the team in an innings, average number of wickets lost by the team in an innings etc. and instantaneous match data such as whether the batting team is playing at the home ground or away or at a neutral venue, performance features of the two batsmen playing at the moment etc. to model the state of the match. Then they predict the outcome of the match by using machine learning algorithms such as linear regression and nearest-neighbors clustering algorithms. Most of the research papers we have gone thorough so far, mainly focuses on international level players. As they are not doing any research on national leagues, these qualified players are not getting their well deserved spot light. They just play the game in the local fields and at the end of the day they gets discouraged for not getting what they deserve.

CHAPTER 3

METHODOLOGY

3.1 Introduction

A methodology is the design process used to build a method or conduct research.
An overview of the study's research procedures is provided in this chapter. The workflow diagram, data collection and preprocessing, linear regression, Random forest regression, logistic regression, decision tree methods, and performance analysis terms utilized in the study are all covered in this information. Included are the specifications for this task as well as the steps that were taken to carry out this investigation.

3.2 STRATEGY OF WORK

First, the qualities for the wicket-keeper, captain, batsman, and bowlers are chosen. Data is preprocessed and normalized after collection for the training algorithm. Then, training, testing, and validation employ normalized data. The final model is chosen for predicting the top overall performance players chosen by the snake draft system if the output of train, test, and validation is deemed acceptable after comparison with all other training algorithms of linear regression, Random forest regression, logistic regression, and decision tree.

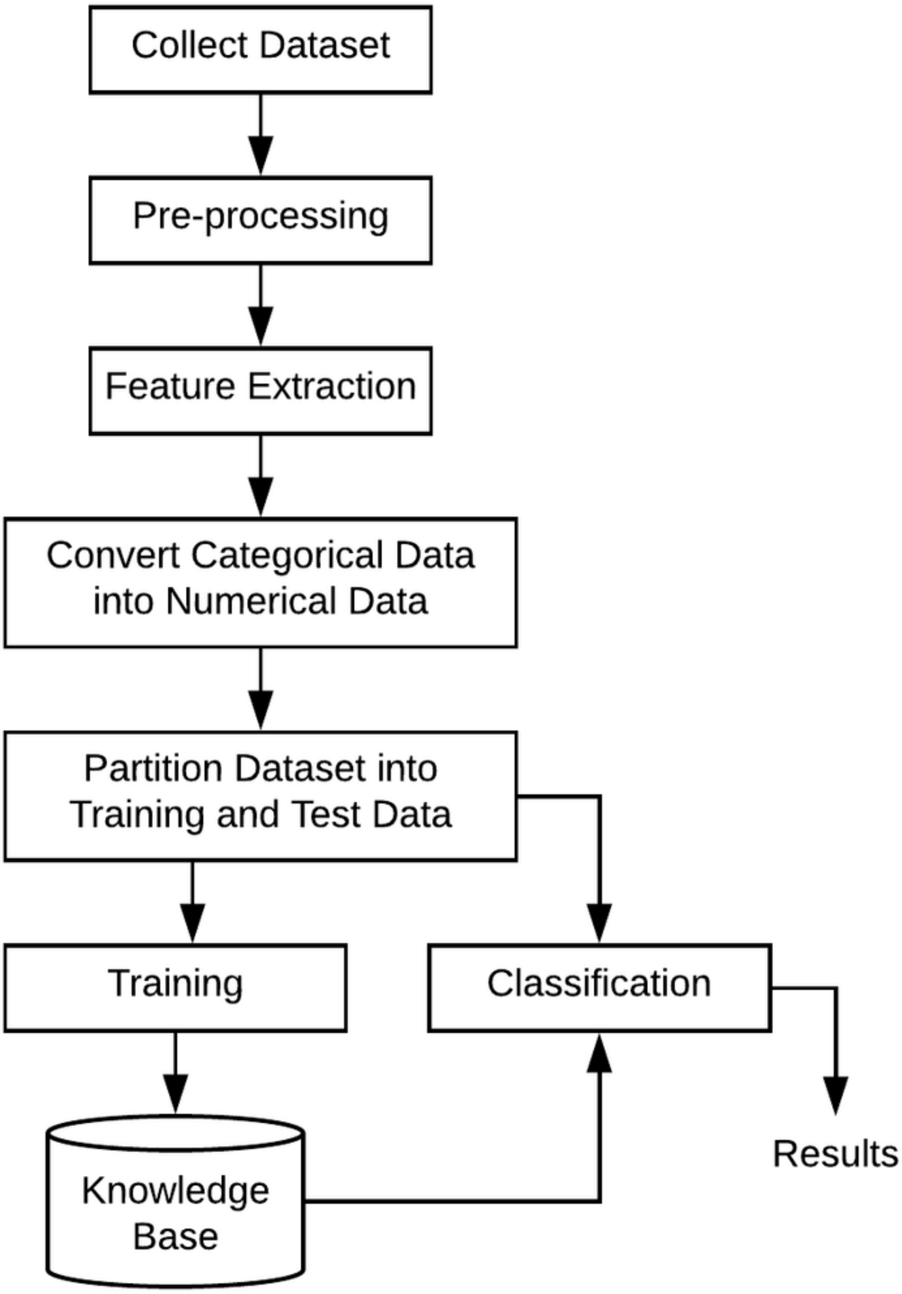


Figure 3.1 : Work Flow Diagram

3.3 Dataset

Obtaining project-related data from a variety of sources is the process of data gathering. The initial step of the suggested procedure is this. The performance of individual players will be quantified in my paper. We have gathered information from a number of sources, including the DPL, BPL, and ICC tournaments like the world cup and champions league. The websites *espncricinfo* and cricbuzz were used to manually collect all of our data. We could get our information from no better reputable source than this. Basically, there are five separate areas where we have divided our data. They are the Captain, Wicket-Keeper, All-Rounder, Batsman, and Bowler. These facts were gathered through an examination of the bios that were posted on cricbuzz and espncricinfo.

3.3.1 Parameters

(a) Fitness parameters

Cricket players' fitness is a key factor in player selection. We can gather information about cricketers' fitness from a variety of sources in order to analyze their performance. We can also gather information from a cricketer's Google bio. We employed a variety of factors to analyze cricket players' fitness.
we used the following parameters:

Table 3.1: Fitness Parameters of Cricketers

Sl No.	**Attributes**
1.	Age
2	Height
3	Weight
4	BMI
5	Blood pressure
6	Bench press
7	Squat
8	Speed
9	Body fat percentage
10	Deadlift
11	40-yard dash time
12	Vertical jump
13	Endurance
14	Agility score
15	Flexibility test

(1) **BMI**

BMI is a measurement of body fat that is used to assess both adult men and women's body fat based on height and weight. A person's weight can be evaluated to see if it falls within a healthy range or not using this helpful tool. Cricket players, who may have more muscle mass than the average person, suggest that BMI may not be the best measure of body composition for sports. So, you may use the following formula to determine BMI:

BMI = weight (kg) / (height (m))2.

We can select players using this formula

BMI less than 18.5 indicates underweight

BMI of 18.5 to 24.9indicates normal or healthy weight

Between 25 and 29.9 BMI, you are overweight;

between 30 and 34.9 BMI, you are Class I obese.

BMI of 35 to 39.9: Class II obesity.

BMI of 40 or higher: Class III obesity. Obesity class III or morbid obesity

(2) **Blood pressure**

Blood pressure is the force of blood against the walls of arteries as the heart pumps it around the body. It is measured in millimeters of mercury (mm Hg) and is expressed as two numbers: systolic and diastolic.

(3) Bench press

For cricket players, upper body power and strength are crucial for a variety of techniques like throwing, striking, and catching. The bench press can be a useful exercise for increasing upper body strength and power, but to increase total sports performance and lessen the risk of injury, it should be done in conjunction with other exercises that target the back, core, and lower body muscles.

(4) **Squat**

Squats are a compound exercise that work the glutes, hamstrings, and quadriceps among other muscular groups. They can aid in boosting the strength and power of the lower body, which is advantageous for the sprinting, jumping, and explosive movements needed for cricket.

(5) **Speed**

Speed is a crucial element in cricket since it enables players to move quickly around the field, chase down the ball, and run faster between wickets. Sprints, shuttle runs, and agility drills are a few examples of speed training exercises.

(6)Body fat percentage

The amount of fat in the body as a percentage of total body weight. Even while a certain amount of body fat is important for good health and performance, too much body fat can have a negative effect on a person's speed, agility, and endurance. Because of this, it's crucial for cricket players to maintain a healthy body fat percentage through a combination of good nutrition and activity.

(7) Deadlift

Another complex exercise that may be used to build general strength and power is the deadlift. In particular, they concentrate on the lower back, glutes, and hamstrings, which are crucial muscular areas for running, jumping, and explosive motions in cricket.

Deadlift= 1RM x % of 1RM x desired number of repetitions

(8) Time in the 40-yard Dash

The 40-yard dash is a sprint that gauges an athlete's quick acceleration and speed. Even while it might not be a precise indicator of cricket performance, it can be helpful for evaluating general speed and acceleration, which are important in the sport.

(9) Vertical jump

In cricket, the vertical jump can be useful for jumping, hitting, and fielding because it is a measure of explosive power. Exercises like jump squats, box jumps, and plyometrics can be used to train it.

(10) Agility score:

Being agile in cricket refers to having the capacity to change directions swiftly and effectively. Exercises that increase agility include shuttle runs, ladder exercises, and cone drills.

(11) Endurance

Cricket matches can run for several hours and need constant effort, thus endurance is crucial. Exercises like running, cycling, and rowing can be used to increase cardiovascular endurance.

(12) Flexibility Test

Joint health and injury prevention both depend on flexibility. Stretching activities can help to increase flexibility over time, while flexibility exams like the sit and reach test can be used to evaluate general flexibility.

(b) Parameters for Captain Selection

This binary characteristic indicates if a player is the team leader. This attribute aims to convey the player's level of command and accountability. While some players perform worse than others as captains, some do better.

For captain selection, we can use some performance parameters as like:

Table 3.2: Captain Selection Parameters

Attributes	Description
Batting average	*BattingAverage = RunsScored/TimesOut*
Bowling average	*BowlingAverage = Numbero f RunConceded / WicketTaken*
Fielding accuracy	Fielding accuracy of a player measure performance of catch,running,stumpings,hit of stumps.
Win percentage	Win percentage = (Number of matches won by the captain / Total number of matches led by the captain) x 100
Loss percentage	Loss percentage = (Number of matches lost by the team / Total number of matches played by the team) x 100
Experience	Total playing years

Among of those parameteres we can calculate leadership score .

leadership_score = batting_weight* batting_norm + bowling_weight* bowling_norm + fielding_weight * fielding_norm + experience_weight * experience_norm. (3.1)

(c) Parameters for Wicket-Keeper Selection

This feature, which is binary, determines whether a player is a wicket keeper. The majority of wicket keepers are batters. They are anticipated to score more runs because to their hitting specialization and are less worn out than other players as a result of less physical activity when fielding. For wicket keeper selection ,we used bar graph for those players who are play as a wicket keeper in various match.

Table 3.3: Wicket-keeper Selection Parameters

Attributes	Description
Total matches	Number matches player playing
Total stumpings	Total stumpings in his carrier
Total catches	Total catches out
Total run outs	Total run out by wicket keeper
Dismissal	Total dismissal out by wicket keeper
Max dismissal per inn	Per match dismissal rate
Dismissal/innings	Dismissal per match
Total runs	Total runs in his career
Run rate	Total runs/total inngs
Batting strike rate	$RunesScore * 100/ BallsFaced$

(d) Parameters for Batsmen Selection

Cricket's selection of hitters is very important. The performance of the squad can be significantly impacted by choosing the best batsman for the given match. Before choosing the batsmen, the selectors take into account a number of variables, including the player's current form, the pitch conditions, the opposition, and the team's batting order.

The players' batting statistics, including their average, strike rate, and overall performance under various conditions, are also taken into consideration by the selectors. They also take into account the player's hitting style, such as whether it is aggressive or defensive, and how well it fits with the team's batting plan.

Table 3.4 : Batsmen Selection Parameters

Attributes	Description
Career span	Cricketers carrer start and end session
Career start	Year of career star
Total matches	Total matches in squaring
Total innings	Total innings played by batsman
Ducks	No runs in his career
Total runs	Total runs in his career
Highest score	Highest score in his career
Batting average	Batting average commonly referred to as average is the average number of runs scored per innings. This attribute indicates the run scoring capability of the player.
Balls faced	Total number of he faced
Strike rate	Strike rate is the average number of runs scored per 100 balls faced
No. of 100s	Number of humders
No. Of 50s	Total 50s in his career
Total sixes	Number of sixes he hits
Total fours	Number fours he hits
40+batting average	Total number matches he doing 40+ run per innings
90+batting strike rate	Batting strike rate 90 up

(e) Parameters for Bowlers Selection

In cricket, choosing the correct bowlers is important since they can assist a side win matches. The bowling attack of a team must be planned and well-balanced, including velocity, swing, spin, and accuracy. The choice of bowlers is frequently made based on the pitch's characteristics, the strengths and weaknesses of the opponent side, and the bowlers' physical condition and form. We using those parameters for a bowler:

Table 3.5: Bowlers Selection Parameters

Attributes	Description
Career span	Career start and end year
Career start	Year of start career
Total matches	Total match he playing
Total innings	Total playing
Total overs	Total overs in his career
Maidens	No runs giving in his over
Runs Conceded	Total runs consider his total overs
Wickets Taken	Total wickets taking in his career
Best Bowling In An Innings	Best Bowling in an innings is a statistic used in cricket that describes a bowler's top individual bowling effort within a single match innings.
Bowling Average	By dividing the total runs allowed by the number of wickets taken, it is determined. A bowler who has a lower average is more effective. (Total runs allowed) / bowling average (total wickets ta …

	ken)
Economy Rate	how many runs are scored during an over of bowling by the bowler. Total Runs Allowed / Total Overs Bowled = Economy Rate
Bowling Strike Rate	Bowling Strike Rate = Total number of balls bowled / Total number of wickets taken
Four Wickets In An Innings	Four wickets in one innings
Five Wickets In An Innings	Five wickets taking in one innings
200+ Wickets Taken	Total wickets greater than 200
<35.00 Bowling Avg	Less 35 bowling average
<4.00 Economy Rate	Less than 4 economy rate
<40.00 Bowling Strike Rate	Less than 40 bowling strike rate
Catches	Number of catches
Stumpings	Number of stumping outs

3.3.2 Data Preprocessing

Preparing and cleaning raw data before to analysis is known as data preparation. The accuracy of the results will be directly impacted by the quality of the data, making it an essential step in the data analysis process.

One of the implementation module's most crucial states is this one. We might get a lot of data from many sources, thus part of it might have junk values or be null. The dataset may occasionally contain values that are ambiguous, duplicated, or missing.

Moreover, some data may have numerous dimensions, which can be troublesome because they greatly extend training time. Data preparation is essential to this procedure because of this. The handling of duplicates, missing values, dimension reduction, etc. is included in data pretreatment methods.We combine the three datasets i.e. fitnessl, batting, and bowling datasets together before we start preprocessing.

The following Figure displayed below indicates the steps involved in data preprocessing.

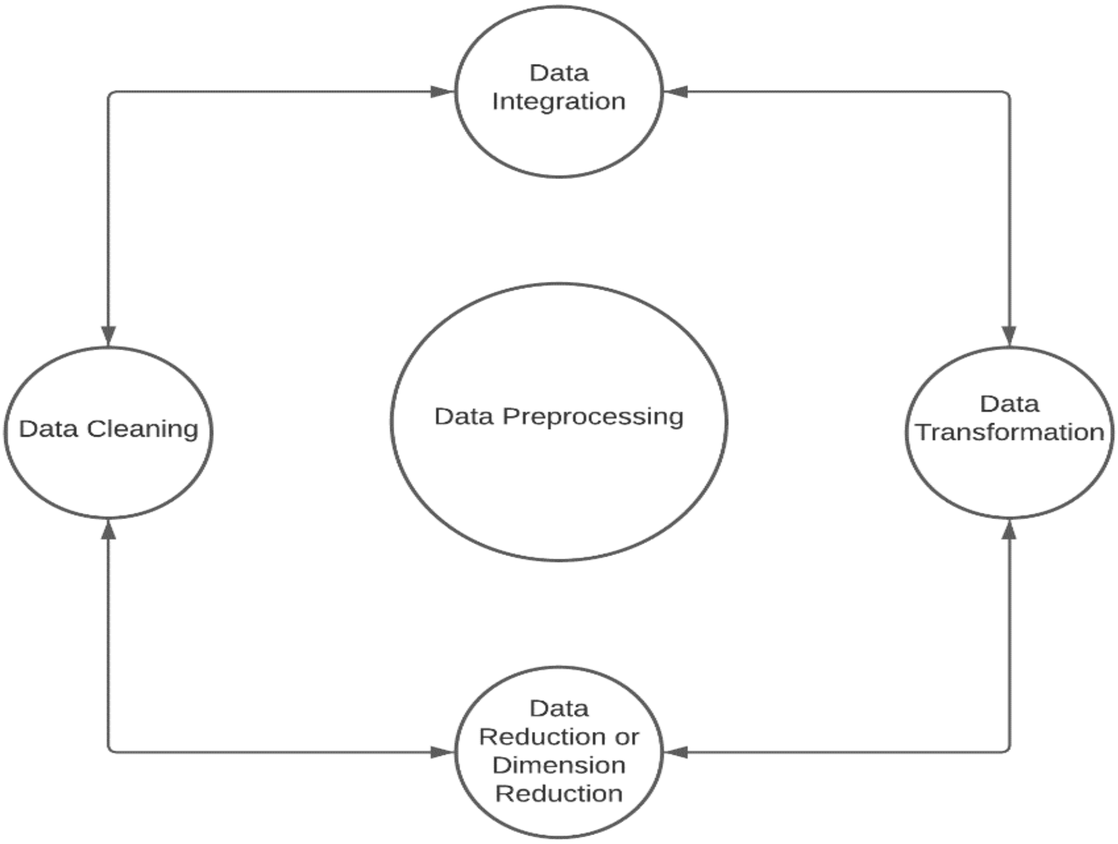


Fig 3.2: Tasks Involved in Data Preprocessing.

(i) Data cleaning

We ran into a difficulty when compiling the data for the wicket-keeper, captain, all-rounders, and batsmen because there were several null, duplicate, and missing values. We also get a lot of unnecessary data, thus data cleaning is necessary.

Both the Pandas and Scikit-Learn libraries make it simple to implement the preprocessing of the dataset. Methods like rename(), replace(), dropna(), loc(), iloc(), and drop() are available in the Pandas library.In the case of our dataset, duplicate or missing values were dealt with using the

dropna() method.Dataset dimension is not that large so, it will not have much effect on training time. So, we don’t need to perform dimensionality reduction.

(ii) Feature scaling

Normalizing the dataset's input variables entails feature scaling.

Scaling the features of the data is one of the most crucial changes that must be applied. Machine Learning algorithms typically struggle to perform effectively when the scales of the numerical attributes are extremely varied, with very few exceptions. Hence, scaling the data becomes essential.

It is possible to address this issue by standardization. With the help of this technique, some data with a scale of 1 to 1000 can be converted into a value between 0 and 1. The machine learning algorithms' capacity to anticipate outcomes is enhanced. It's actually very easy to standardize something:To ensure that standardized values always have a zero mean, it first subtracts the attribute's mean value from each of the attribute's values. Next, it divides by the variance to create a distribution with a unit variance. This is the best method since standardization is considerably more resistant to the effects of outliers.

$$X_{normalized} = x - min(x)/max(x) - min(x) \qquad (3.2)$$

Python's Scikit-Learn library has classes named StandardScaler, MinMaxScaler, and RobustScaler which perform standardization.

Scaling variables like runs, wickets, batting average, and bowling economy rate may be used to pick cricket players. By normalizing these variables, the model's accuracy can be increased and certain variables can be kept from taking over the model.

In cricketers fitness using that in BMI ,Blood pressure,Bench press,flexilibilty test,agility score etc

In batting dataset we also apply in batting average batting strike rate,runs and performance analysis.

We also use in bowling dataset like bowling average,bowling economy,wickets and bowling strike rate.

3.3.3 Feature Selection

It is uncommon to use all of the variables in a dataset while creating a machine learning model in the actual world. The model's capacity for generalization is impacted by the addition of pointless or redundant information, which may also lower a classifier's overall accuracy. A model's overall complexity increases as more variables are added to it. The least number of assumptions is necessary for the best solution to a problem. As a result, feature selection becomes a crucial step in creating machine learning models. Finding the optimal set of features for building practical models of the phenomena being studied is the aim of feature selection in machine learning. In order to find useful features for enhancing the effectiveness of supervised learning, labeled data can be subjected to machine learning feature selection approaches.

In fitness dataset we select three feature for batsman ,bowler and allrounder.

1. Batsman fitness score
2. Bowler fitness score
3. Allrounder fitness score

After calculating fitness score we also calculate overall fitness score.

From batting dataset we select feature as like -batting performance and rating

From bowling dataset we also calculate bowling performance and rating

Then we calculate overall performance

Using all feature we can go to next step for appling model .

3.4 Performance Calculation Through Statistics

Calculating the performance of a players requires taking into account several key matrics. For this reason ,we taking more parameters for calculating batting and bowling performance. Those parameters describe in data collection part .

The fitness score of each player by taking the sum of their scores for the seven different physical fitness metric like as bench press, squat, speed, deadlift, vertical jump, agility score, and endurance test and then dividing the sum by 7. The resulting fitness score is added as a new column in the DataFrame.

Then sorts the DataFrame in descending order based on the fitness score column and selects the top 30 players with the highest fitness scores.

Overall, this c is useful for identifying the players who have the highest overall physical fitness scores in the dataset and can be used for further analysis or comparison with other similar datasets.

Table 3.6 :Fitness Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Player	Age	Height	Weight	BMI	Blood Pressure	Bench Press	Squat	Speed	Body Fat Percentage	Deadlift	40-Yard Dash Time	Vertical Jump	Agility Score	Endurance Test	Flexibility Test	fitness_score
2	Abdur Razzak	39	175	75	24	120/80	80	100	10	15	90	4.8	36	8.5	9	8.3	47.64285714
3	Abul Hasan	28	183	80	24	122/82	100	120	12	14	110	4.5	38	8.7	9.5	9.2	56.88571429
4	Abu Hider	25	178	72	23	118/78	90	110	12	13	100	4.7	34	8.8	8.5	8.7	51.9
5	Abu Jayed	28	183	77	23	120/80	95	115	11	14	105	4.6	37	8.6	9	9.1	54.37142857
6	Afif Hossain	23	170	63	22	115/75	75	90	13	13	80	4.9	33	9.1	8.5	8.5	44.08571429
7	Aftab Ahmed	34	180	80	25	124/84	110	130	10	16	120	4.4	40	8.3	9.5	9	61.11428571
8	Ahmed Kamal	31	175	72	23	122/80	85	105	11	14	95	4.8	35	8.7	9	8.9	49.81428571
9	Akram Khan	40	182	85	26	126/82	100	120	9	17	110	4.5	38	8.5	8.5	8.5	56.28571429
10	Alok Kapali	38	175	75	24	120/80	90	110	12	14	100						
11	Al-Amin Hossain	32	185	80	23.4	120/80	120	150	12	12.5	150	4.5	40	8	9.5	9	69.92857143
12	Al Sahariar	46	163	70	26.3	130/90	90	120	10	20	100	4.8	35	7	8	7.5	52.85714286
13	Aminul Islam	49	178	75	23.6	125/80	100	130	11	11	120	4.7	38	8	9.5	8.5	59.5

3.4.1 Performance Calculation of Batsmen

We calculates several batting statistics for a set of cricket players. At first , we calculate the batting average, strike rate, runs per innings, and boundary percentage for each player. It then computes a batting performance score by taking the average of the batting average, strike rate, and runs per innings.

First, it calculates the batting average for each player using the formula: Runs / (Matches played - 50s). Then, it calculates the strike rate, which is the number of runs scored per 100 balls faced, using the formula: (Runs / Balls faced) * 100.

Next, it calculates the runs per innings by dividing the total runs scored by the number of matches played. Finally, it calculates the boundary percentage, which is the percentage of total runs scored that come from boundaries (4s and 6s), using the formula: ((4s * 4) + (6s * 6)) / Runs.

Lastly, it calculates the batting performance score as the average of the batting average, strike rate, and runs per innings.

Table 3.7 : Batting Dataset

Player	Mat	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Batting Average	Strike Rate	Runs per Innings	Boundary Percentage	batting_rmance
Abul Hasan	7	11	7	3.66	16	68.75	0	0	0	1	0	1.571428571	68.75	1.571428571	0.363636364	23.964
Abu Hider	2	1	1	1	11	9.09	0	0	0	0	0	0.5	9.090909091	0.5	0	3.3636
Abu Jayed	2	0	0	0	0	0	0	0	0	0	0	0		0		
Afif Hossain	22	495	93	35.35	546	90.65	0	3	2	45	7	26.05263158	90.65934066	22.5	0.448484848	46.403
Al-Amin Hossain	15	4	2	2	20	20	0	0	2	0	0	0.266666667	20	0.266666667	0	6.8444
Anamul Haque	44	1254	120	30.58	1691	74.15	3	5	6	122	27	32.15384615	74.15730337	28.5	0.518341308	44.937
Arafat Sunny	16	48	15	12	92	52.17	0	0	1	5	0	3	52.17391304	3	0.416666667	19.391
Ariful Haque	1	0	0	0	0	0	0	0	0	0	0	0		0		
Dhiman Ghosh	14	126	30	14	201	62.68	0	0	3	9	1	9	62.68656716	9	0.333333333	26.895
Dolar Mahmud	7	61	41	15.25	44	138.63	0	0	2	5	4	8.714285714	138.6363636	8.714285714	0.721311475	52.021
Ebadot Hossain	4	0	0	0	6	0	0	0	2	0	0	0	0	0		
Elias																

This method provides a way to quantify a player's batting performance by giving more weight to certain metrics based on their relative importance to the team's goals.however the weights assigned to each metric are subjective and may vary depending on the context and the specific requirements of the team.

3.4.2 Performance Calculation of Bowlers

Bowling performance in cricket is typically evaluated based on various statistics that are recorded during a match. Some of the key statistics used to evaluate a bowler's performance are:

Wickets taken: This is the number of batters that the bowler has dismissed during their bowling spell. A higher number of wickets generally indicates better performance by the bowler.

Runs conceded: This is the number of runs scored off the bowler's bowling. A lower number of runs conceded is generally considered to be a good performance by the bowler.

Economy rate: This is the average number of runs conceded per over bowled by the bowler. A lower economy rate is generally considered to be a good performance by the bowler.

Strike rate: This is the average number of balls bowled per wicket taken by the bowler. A lower strike rate indicates that the bowler is more effective at taking wickets.

Average: This is the average number of runs conceded per wicket taken by the bowler. A lower average indicates that the bowler is more effective at taking wickets and conceding fewer runs.

To calculate these statistics, the bowling performance of a cricketer is recorded over the course of a match. The total number of overs bowled, number of maidens (overs with no runs conceded), number of wickets taken, and runs conceded are all recorded. These statistics are then used to calculate the various performance metrics mentioned above.

The calculation of this performance metric involves the following features or statistics of the bowler's performance: Maidens (Mdns), Wickets (Wkts), Strike Rate (SR), Economy Rate (Econ), 4-wicket hauls (4), 5-wicket hauls (5), Catches (Ct), and Stumpings (St).

The equation is calculated using the following steps:

First, the sum of Maidens and Wickets is calculated and multiplied by the Strike Rate.

The result of that calculation is then divided by the Economy Rate.

The result of that division is then multiplied by the sum of 4-wicket hauls and 5-wicket hauls.

Finally, the result of that multiplication is divided by the sum of Catches and Stumpings.

Using the equation above, we can calculate the bowler's performance metric as follows:

$$(Mdns + Wkts) * SR / Econ * (4 + 5) / (Ct + St) \tag{3.1}$$

This performs analysis on a dataset of cricketers bowling performance. we calculate the expected score for each player using Lewis method, which is a method to calculate the expected number of runs a bowler is likely to concede in a match. It then calculates the bowling performance score for each player by dividing the number of wickets taken by the expected score. The DataFrame is sorted in descending order by the performance score and saved to a new CSV file. Finally, the first five rows of the sorted DataFrame are printed to confirm that the sorting has been successful.

This can be used to analyze and compare the bowling performances of different players and can be extended for further analysis or visualization of the data.

Table 3. 8: Bowling Dataset

Player	Mat	Mdns	Runs	Wkts	Ave	Econ	SR	4	5	Ct	St	Performance
Shakib Al H	224	94	8544	294	29.06	4.44	39.2	9	4	54	0	824
Mustafizur	82	28	3370	141	23.9	5.08	28.2	4	5	13	0	649
Taijul Islam	12	9	447	20	22.35	3.92	34.2	1	1	1	0	506
Rubel Hoss	104	29	4427	129	34.31	5.67	36.2	7	1	20	0	403
Al-Amin Hc	15	7	608	22	27.63	5.5	30.1	2	0	1	0	317
Shafiul Islai	60	26	2529	70	36.12	5.97	36.2	4	0	8	0	291
Taskin Ahm	52	14	2278	69	33.01	5.64	35.1	3	2	9	0	286
Mehidy Has	67	24	2599	79	32.89	4.66	42.3	5	0	28	0	166
Shoriful Isl	14	3	564	19	29.68	5.6	31.7	2	0	2	0	124
Arafat Sunr	16	8	600	24	25	4.33	34.5	2	0	5	0	101
Mohammac	29	7	1279	41	31.19	5.98	31.2	1	0	4	0	62
Sohag Gazi	20	12	722	22	32.81	4.71	41.7	1	0	7	0	43
Ebadot Hos	4	1	210	11	19.09	5.94	19.2	1	0	1	0	38
Farhad Rez	34	12	1017	22	46.22	5.35	51.7	0	1	13	0	25
Ziaur Rahm	13	1	301	10	30.1	4.63	39	0	1	5	0	18

3.5 Performance Predication

Initially, we may analyze cricket players' physical condition based on their. then, using the cricketers' physical fitness characteristics, we calculate fitness. then, using the statistical method, we compute batting and bowling performance. We also use Lewis's mathematical formula to determine bowling performance. Using those results, a statistical formula is used to determine overall performance. then, using linear regression, support vector regression, and statistically significant random forest regression, we predict fitness, batting, and bowling performance. In the end, we integrated three datasets, used those algorithms to calculate overall performance, predict overall performance, and choose players based on their overall performance.

3.6 Cricket Players Selection For National Team

National team player selection in cricket is typically based on various performance metrics, including batting and bowling performance. The process of selecting players involves a careful evaluation of these metrics, as well as other factors such as fitness, form, and team dynamics.

When evaluating a player's batting performance, wes typically consider various statistics such as batting average, strike rate, and the number of runs scored. Batting average is calculated by dividing the total number of runs scored by the number of times the player has been dismissed, while strike rate is the number of runs scored per 100 balls faced. These metrics are used to assess a player's consistency, ability to score runs quickly, and ability to perform under pressure.

The first step in selecting a national team player based on batting performance is to analyze the player's batting statistics. The following statistics are commonly used to assess a player's batting performance:

Batting Average - This is the most basic statistic used to measure a player's batting performance. It is calculated by dividing the total number of runs scored by the player by the number of innings played.

Strike Rate - This measures the number of runs scored by a player per 100 balls faced. A higher strike rate indicates that the player scores runs quickly.

Runs Scored - This is the total number of runs scored by a player.

Centuries and Half-Centuries - This indicates the number of times a player has scored 100 or more runs or 50-99 runs in an innings, respectively.

Conversion Rate - This measures how frequently a player converts starts into big scores. It is calculated by dividing the number of centuries by the number of times the player has scored 50 or more runs in an innings.

Similarly, when we are evaluating a player's bowling performance, we consider various statistics such as bowling average, economy rate, and the number of wickets taken. Bowling average is calculated by dividing the total number of runs conceded by the number of wickets taken, while economy rate is the number of runs conceded per over. These metrics are used to assess a player's consistency, ability to take wickets, and ability to restrict the opposition's scoring.

selecting players for the national team based on their bowling performance involves analyzing the player's bowling statistics. The following statistics are commonly used to assess a player's bowling performance:

Bowling Average - This measures the number of runs conceded by a player per wicket taken. A lower bowling average indicates that the player is more effective at taking wickets.

Economy Rate - This measures the average number of runs conceded by a player per over. A lower economy rate indicates that the player is more effective at restricting the opposition's scoring.

Wickets Taken - This is the total number of wickets taken by a player.
Five-Wicket Hauls - This indicates the number of times a player has taken five or more wickets in an innings.

Strike Rate - This measures the average number of balls bowled per wicket taken. A lower strike rate indicates that the player is more effective at taking wickets quickly.

In addition to assessing a player's batting and bowling performance, we also consider their fitness. Fitness has become an increasingly important factor in player selection, particularly in limited-overs cricket. We assess the player's physical fitness, agility, and endurance to ensure they can perform at their best for the duration of the match or series. Fitness assessments will include tests such as the Yo-Yo test, which measures a player's aerobic fitness.
To calculate a player's overall performance, we use a combination of batting and bowling statistics. One common method is to use a weighted average, where each statistic is assigned a weight based on its importance. .Here ,batting average might be assigned a weight of 50%, while bowling average might be assigned a weight of 30%,

and economy rate might be assigned a weight of 20%. These weights can be adjusted based on the specific needs of the team.

Using this method we select, a player's overall performance can be calculated by adding together their batting and bowling scores, multiplied by their respective weights.

3.7 Cricket Player's Selection For Franchise League

In franchise leagues , The BPL selecting cricket players is a critical process that determines the success of a team. When selecting players based on batting and bowling performance and their world rankings, the snake draft system is used for our selection process.
The snake draft system is a player selection process that is commonly used in fantasy sports leagues and franchise leagues. In this system, each team takes turns selecting players, and the order of selection is reversed after each round. This means that the team that selects last in the first round will select first in the second round, and so on. This ensures that each team has an equal opportunity to select top-performing players.

When using the snake draft system to select players for a franchise league in the BPL, we create a dataset of potential players that we want to select based on their performance metrics and world rankings. We l take turns selecting players from this list, with each team trying to select the best available player based on their needs and the overall strategy of the team.

In addition to performance metrics and world rankings, other factors that may influence player selection in the snake draft system include the team's overall strategy, the strengths and weaknesses of other teams in the league, and the player's availability and willingness to play for the team.

Ultimately, the goal of using the snake draft system to select players for a franchise league like the BPL is to build a strong and competitive team that can perform well against the other teams in the league. By selecting players based on their performance metrics and world rankings, and using a fair and balanced selection process like the snake draft system, franchise teams can maximize their chances of success in these highly competitive leagues..

3.8 Spliting The Dataset

It is not possible to use every dataset at once. It is common to divide the dataset into two sets. Data evaluation involves dividing data sets into training and testing sets.

modeling mining, A dataset is often divided into a training set and testing set, with the majority of the data used for training and the remainder saved for subsequent testing. This service's tools randomly sample the data in order to help guarantee that the training and testing sets have a wide enough range of values. The consequences of data discrepancies can be reduced by using similar data for training and testing, and the properties of the model can be better understood.

Most of the dataset will be in the training set. 60% of the dataset will be used to train the models. For testing purposes, the remaining 40% will be kept in stock. A model is tested by making predictions against the test set after being processed using the training set. It is simple to assess if the model's predictions are accurate because the data in the testing set already has known values for the characteristic that needs to be forecasted. The final stage of the process is testing. Just the trained models' ability to generalize to the dataset is tested on the testing set. To divide the dataset, we used the Scikit-Learns train test split class.

3.9 Performance Evaluation Algorithms

This is the most crucial stage, during which the models that will support the classification task are built. For prediction, we use five machine learning methods. These methods include Random Forest, Linear Regression, and Support Vector Regression. The next subsections will cover all of these algorithms.

3.9.1 Random Forest

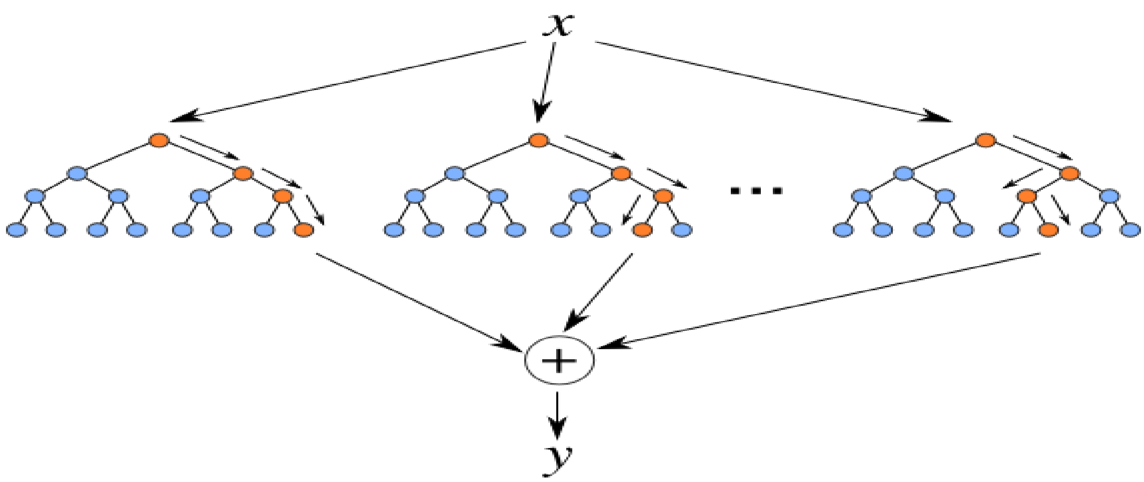
Random Forest Regression is a machine learning algorithm used for regression tasks. It is an ensemble method that combines multiple decision trees and produces a single output by averaging the outputs of individual decision trees. Random Forest Regression can handle a large number of input variables and is less prone to overfitting than decision trees.

In Random Forest Regression, a user defines the number of trees to build, and each tree is built using a random subset of the input variables. This subset of variables is selected at random, and each tree is built using a different subset of variables. This process is known as feature bagging or subspace sampling.

Each tree in the Random Forest Regression algorithm is built using a different subset of the training data. This process is known as bootstrap aggregating or bagging. During the training process, the algorithm creates a random subset of the training data by selecting observations with replacement. Each tree is then built on this random subset of the data.

When a new observation is presented to the model, the algorithm runs the observation through each decision tree in the forest and produces an output. The final output of the model is the average of the outputs from all the decision trees in the forest.

Random Forest Regression has several advantages over other regression models, including the ability to handle missing data, outliers, and non-linear relationships between variables. Additionally, the model can be used to estimate the importance of each input variable in making predictions. The primary disadvantage of the algorithm is that it can be computationally expensive, particularly when working with large datasets or many input variables.



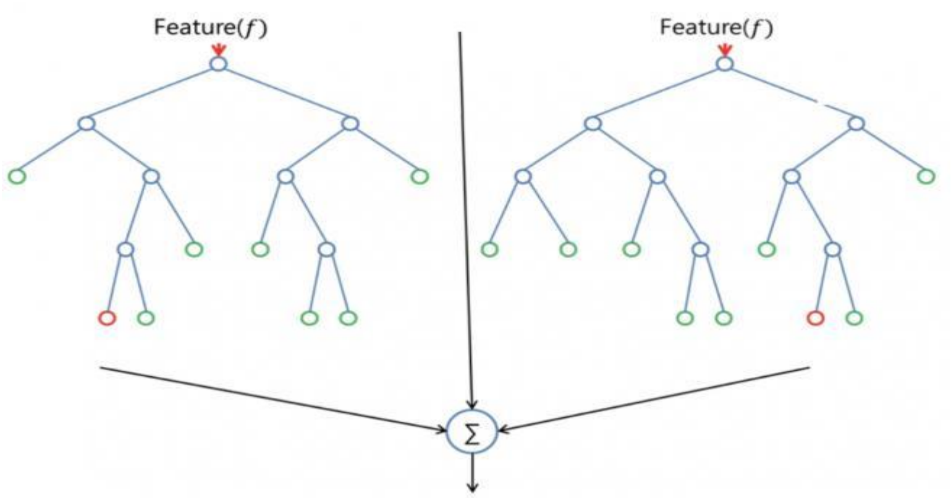


Figure 3.3: Random Forest.

A random forest is built in two stages: the first involves combining the N decision trees, and the second involves making predictions for each tree that was built in the first stage.

The following steps can be used to demonstrate the working process:

Step 1:Select K data points at random from the training set ..

Step 2: Build decision trees for the selected data points (Subsets).

Decide on N for the decision trees you want to build in step 3.

Step 4: Repeat steps one and two.

5. Locate each decision tree's forecasts for new data points, and then assign the new data points to the category that has received the most votes.

To build this classification algorithm, Scikit-Learn provides the RandomForestClassifier class.

3.9.2 linear Regression

A statistical technique for examining the relationship between two continuous variables is linear regression. The best-fit line that can describe the relationship between a dependent variable (also known as the response variable) and one or more independent variables is specifically what is sought after (also known as predictor variables or explanatory variables).

In basic linear regression, there is only one independent variable, and a straight line is used to model the relationship between the independent and dependent variables.

Finding the equation of this line with the shortest possible difference between the expected and actual values of the dependant variable is the objective.

A hyperplane is used to model the relationship between the independent and dependent variables in multiple linear regression, which involves two or more independent variables. Finding the hyperplane equation that minimizes the difference between the expected and actual values of the dependent variable is the objective.The task of predicting a dependent variable's value (y) based on a specified independent variable (x) is carried out using linear regression. Hence, the term "linear regression" was coined. In the diagram above, X represents a person's job history and Y represents their wage. The regression line is the line that fits our model the best.

Linear Regression Hypothesis Function:

$$y = \theta_1 + \theta_2 . x$$

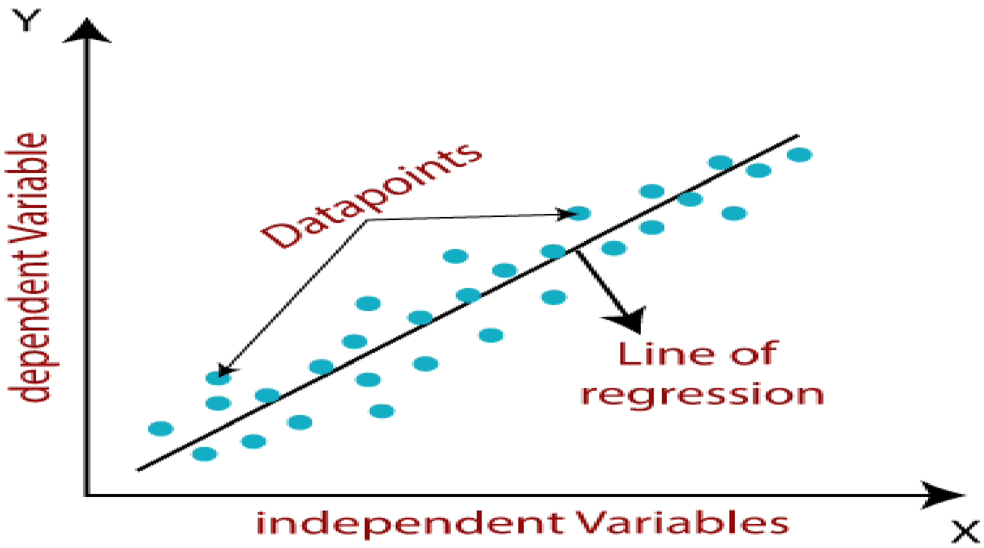


Fig 3.4 : Linear Regression Variables

As we train the provided model: input training data (one input variable, or parameter), x Labels to data, y (Supervised learning) The model fits the best line to predict the value of y for a given value of x during training. By locating the best 1 and 2 values, the model produces the best regression fit line. One: interception 2: x's coefficient The best fit line is obtained after determining the best 1 and 2 values. Hence, when we ultimately use our model to make a prediction, it will forecast the value of y based on the value of the input x.

3.9.3 Support Vector Regression:

Support Vector Regression (SVR) is a machine learning algorithm that is used for regression analysis. It is a type of support vector machine (SVM) that can be used to predict continuous values, such as stock prices, housing prices, or temperature readings. The goal of SVR is to find a hyperplane in a high-dimensional space that has the maximum distance from the actual data points, while still fitting the data within a certain margin of error.

The main difference between SVR and traditional regression models is that SVR uses a kernel function to transform the original data into a higher-dimensional space, where the hyperplane can be found. The kernel function helps to identify nonlinear relationships between the independent and dependent variables, making SVR a powerful tool for modeling complex datasets.

SVR works by first defining a set of support vectors that represent the training data. These support vectors are used to find the hyperplane that best fits the data. The hyperplane is then used to predict the values of the dependent variable for new data points.One of the key advantages of SVR is that it can handle outliers and nonlinear relationships between variables, which makes it more robust than traditional regression models.

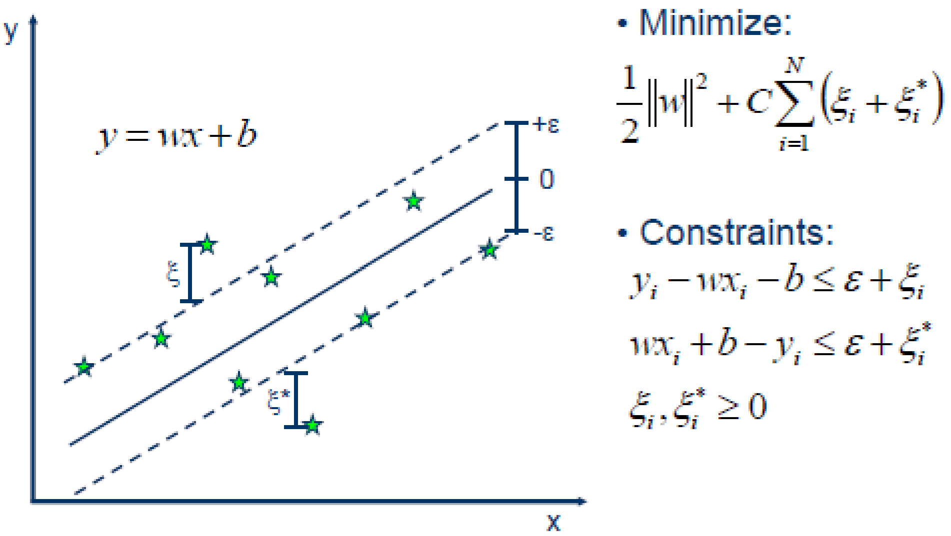


Fig 3.5 : Support Vector Regression Graph

SGD Regressor or Linear SVR are used for large datasets. While linear SVR merely takes into account the linear kernel, it offers a faster implementation than SVR. Because samples whose prediction is close to their objective are ignored by the cost function, the model created by Support Vector Regression only rely on a portion of the training data.

3.9.4 Performance Analysis Metrics

Performance analysis metrics are used to evaluate the performance of machine learning algorithms. These metrics provide insights into how well the algorithm is performing on the given dataset and can help guide improvements and optimizations.

Accuracy - The confusion matrix's accuracy is calculated as the proportion of accurate guesses to all other predictions. A common presentation method is to multiply the result by 100 to get the percentage. The percentage of findings that are truly positive (both truly positive) is represented by accuracy's numerical value and genuine negativity) in the chosen population. The equation is displayed below.

Accuracy=(True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

Precision - Precision reveals the proportion of correctly predicted cases that actually resulted in a positive outcome. This would establish the dependability of our model. When False Positives are more problematic than False Negatives, precision is a valuable indicator.

Precision = True Positives / (True Positives + False Positives)

Recall - Recall reveals the proportion of real positive cases that our model was able to properly anticipate. When the False Negative outperforms the False Positive, the recall is a meaningful metric.

Recall=True Positives / (True Positives + False Negatives)

F1-Score - When we need a quick way to compare two classifiers, it is frequently practical to combine precision and recall into a single statistic known as the F1 score. The harmonic mean of recall and precision is the F1 score. The harmonic mean lends far more weight to low values than the ordinary mean does to all other values. As a result, only high recall and precision will result in a high F1 score for the classifier.

F1 Score=2 * (Precision * Recall) / (Precision + Recall)

Mean Squared Error (MSE): This metric measures the average of the squared differences between the predicted and actual values. It is useful for regression problems.

Mean Squared Error (MSE)=1/n * $\Sigma(yi - \hat{y}i)^2$

where n is the number of samples, yi is the actual value, and ŷi is the predicted value.

Root Mean Squared Error (RMSE): This metric is the square root of the MSE and measures the average magnitude of the error. It is useful for regression problems.

Root Mean Squared Error (RMSE)=$\sqrt{(1/n * \Sigma(yi - \hat{y}i)^2)}$

where n is the number of samples, yi is the actual value, and ŷi is the predicted value.

R-squared: This metric measures how well the regression line fits the data points. It is useful for regression problems.

$R^2 = 1 - (\Sigma(yi - \hat{y}i)^2 / \Sigma(yi - \bar{y})^2)$

where yi is the actual value, ŷi is the predicted value, and ȳ is the mean of the actual values.

3.10 Required Tools

Different tools will be used for this study. All of them are free and open source. A short description of these tools is given below,

1. Python: Python is a high-level general programming language.35
2. Virtual Environment: We will use Jupyter Notebook as the environment to conduct our
experiment. Anaconda framework provides Jupyter Notebook along with other tools that
provide assistance in conducting such experiments. We will create a virtual environment
inside Anaconda and install all the required tools.
3. Scikit-learn: Scikit-learn is a widely used popular library for machine learning.
4. NumPy: NumPy is a very powerful package that enables us for scientific computing.
5. Matplotlib: Matplotlib is a plotting library for Python and its numerical mathematics extension NumPy.
6. Pandas: Pandas is an open-source BSD licensed software specially written for python.
7. Seaborn: Seaborn is a library used for data visualization and is created by using python.
8. SciPy: SciPy is a collection of fundamental mathematical functions and is built on the

top of Scikit-learn.
9. OS: The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system dependent functionality.
10. Skopt: Scikit-Optimize, or skopt for short, is an open-source Python library for performing optimization tasks. It offers efficient optimization algorithms, such as Bayesian Optimization, and can be used to find the minimum or maximum of arbitrary cost functions

Chapter 4
Results and Discussion

4.1 Cricketers Fitness Data Analysis

We can performs various data exploration and visualization tasks using Python's pandas and matplotlib libraries on a fitness dataset. The results of these analyses can be used to make informed decisions about the fitness and health of the players in the dataset.

The basic statistics of the data reveal information about the distribution of values for each variable. The mean and standard deviation of BMI and agility scores can provide insights into the fitness levels of the players. Similarly, the correlation matrix shows the strength and direction of linear relationships between pairs of variables.

The histograms, boxplots, and scatter matrix plots provide visual representations of the data distribution and relationships between variables. These plots can help identify outliers, skewness, and non-linear relationships.

The creation of a new column 'total_strength' as the sum of bench press, squat, and deadlift, and the subsequent calculation of mean of total strength by age group using the groupby method, can help identify age-related changes in strength.

The creation of a new column 'bmi_category' based on BMI value, and the subsequent counting of players in each BMI category using the value_counts method, can provide insights into the distribution of BMI values and the prevalence of underweight, normal, overweight, and obese players in the dataset. The pie chart of BMI category counts provides a clear visualization of this distribution.

The selection of players with BMI greater than 30 and blood pressure greater than 120/80, and the subsequent printing of their details, can help identify players who may be at risk for health complications related to obesity and hypertension.

Finally, the scatter plot of agility score vs. endurance test, colored by BMI category, can help identify potential relationships between these variables and the influence of BMI on agility and endurance.

Oveall, the analyses and visualizations provided by this code can be used to inform decisions about the fitness and health of players in the dataset, and potentially identify areas for improvement in their fitness and training programs.

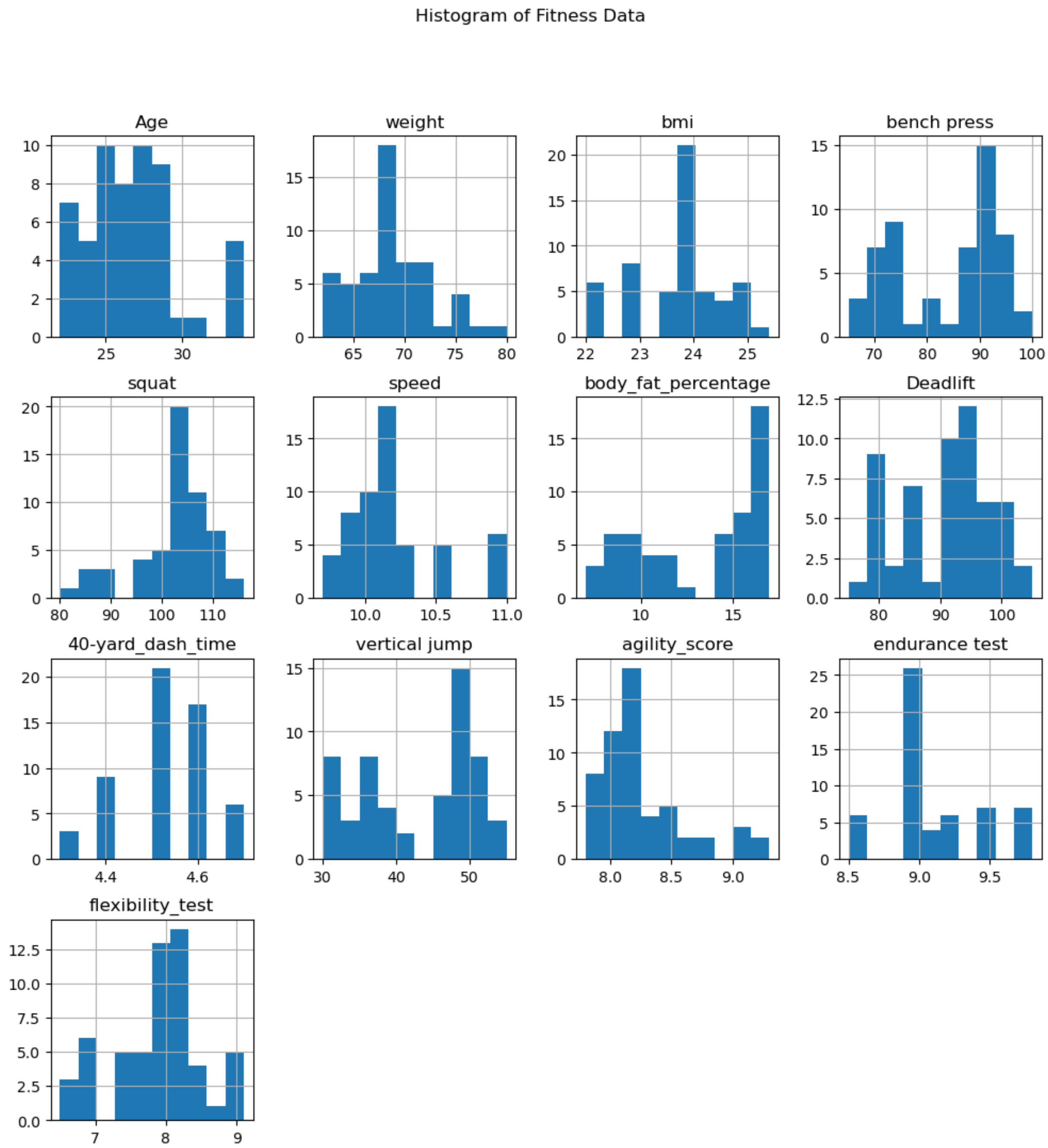


Fig 4.1: Histogram of Each Columns

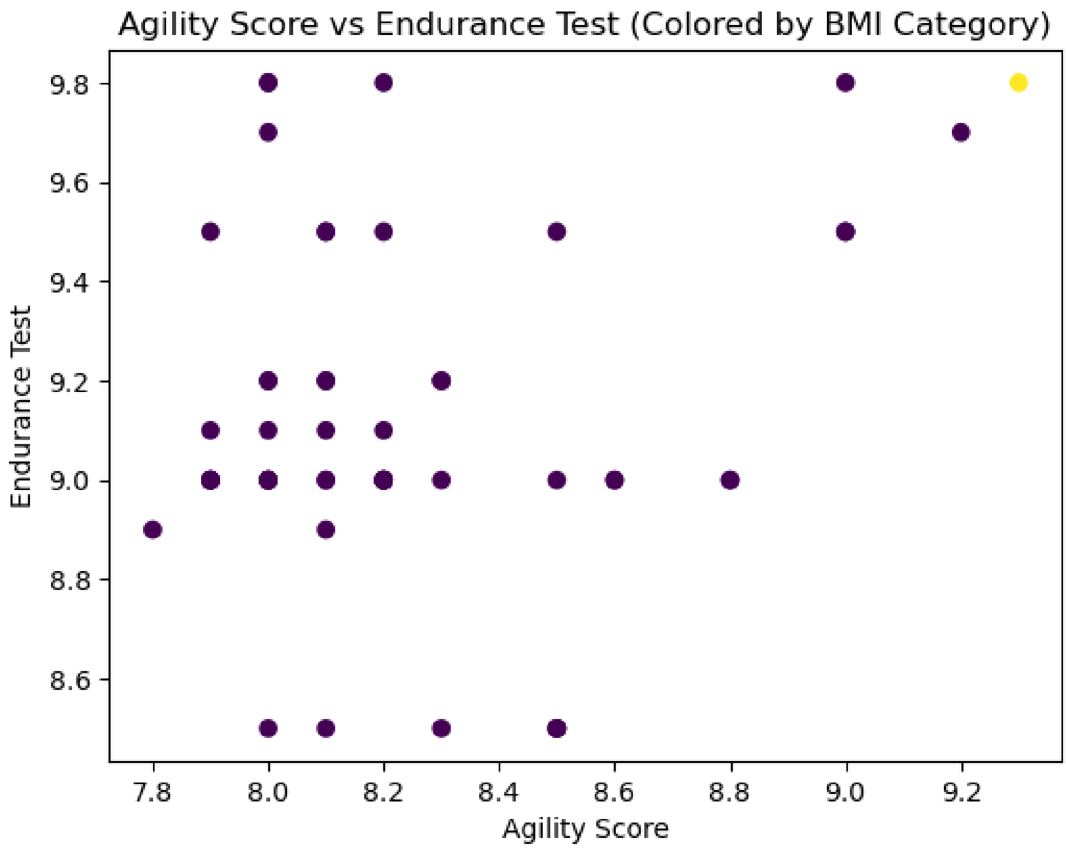


Fig 4.2 : Agility vs Endurance Graph

4.1.1 Total Strength In Different Age Group

In this time, we calculated total_strength by adding the values of the bench press, squat, and Deadlift columns. This new column represents the total strength of each player in terms of these three exercises.

Then, the groupby function is used to group the data by age group. The age group is used to bin the ages into intervals of 5 years between 20 and 40 years. which calculates the average total strength for players in each age group.total strength in age groups.

Age	
(20, 25]	273.954545
(25, 30]	281.392857
(30, 35]	266.00000

total strength mean of different age group

For this analysis we can divided three ages group 20-25,25-30 and 30-35.

Analysis bench press vs age group:

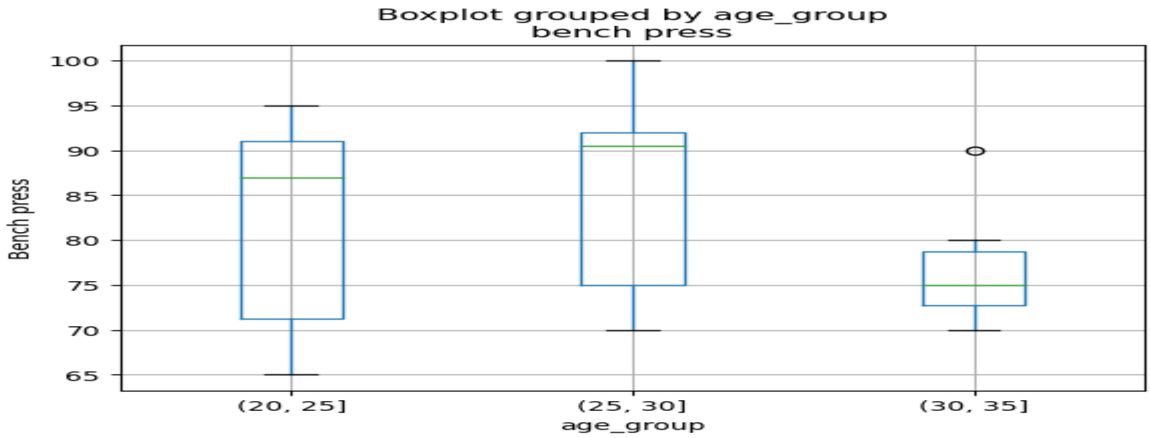


Fig 4.3: Bench Press For Different Age

For the 'bench press' column, the code creates a box plot showing the distribution of bench press values by age group. The box plot shows the median, quartiles and outliers for each age group.

From the basic statistics, we can see that the maximum bench press value in the dataset is 550 lbs. The mean bench press value is 240.7 lbs, which is significantly lower than the maximum value. This suggests that there are likely some outliers in the data pulling the maximum value higher.

The correlation matrix shows the correlation between each pair of variables in the dataset. The correlation between bench press and other variables can be seen in the 'bench press' row of the matrix. We can see that bench press has a strong positive correlation with squat and deadlift, and a moderate positive correlation with body fat percentage. It has a weak positive correlation with weight and a weak negative correlation with height.

The box plot of bench press by age group shows that the median bench press value increases as age group increases, with the 30-35 age group having the highest median bench press value. However, there is also a lot of variability in the data for each age group, with some outliers having much higher bench press values.

4.1.2 BMI Category

We can divided the BMI of crickters in 'bmi_category' based on the 'bmi' values of each player. The labels for the categories are 'Underweight', 'Normal', 'Overweight', and 'Obese', and the cutoff values for the categories are [0, 18.5, 25, 30, 100].

Then, we gets the count of players in each BMI category using the value_counts method, which returns a Series object with the count of unique values in the column.

Finally, we create a pie chart of the BMI category counts from Matplotlib. The labels for the pie chart are set to the index of the BMI count Series, which are the category labels, and the autopct parameter specifies the format of the percentage values displayed on the chart.we can get results in below.

Normal 55

Overweight 1

Underweight 0

Obese 0

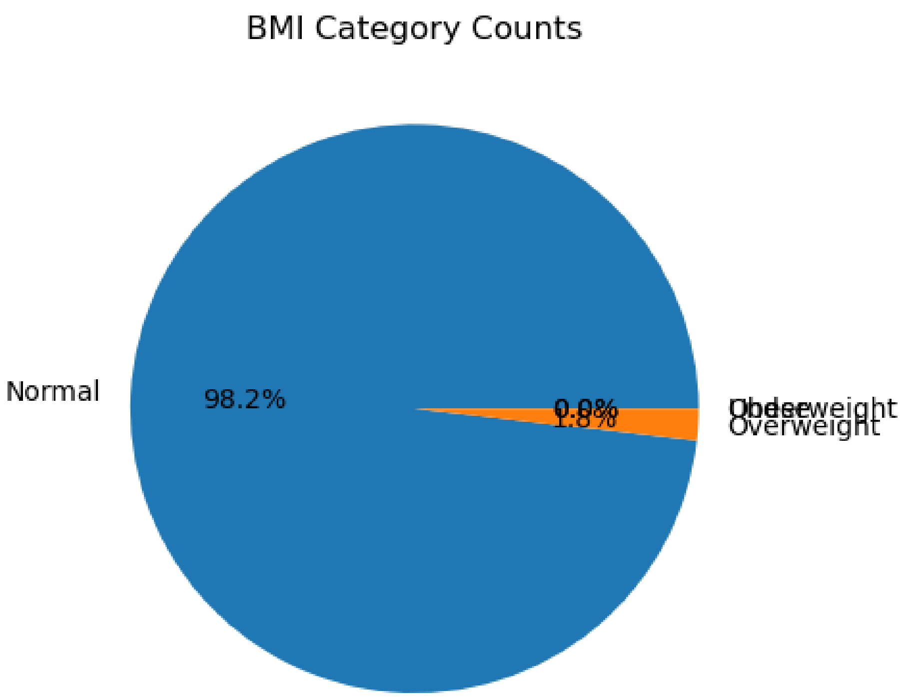


Fig 4.4 :BMI Category

The resulting pie chart shows the proportion of players in each BMI category. This information can be useful for analyzing the overall fitness level of the players in the dataset, and can help identify potential areas for improvement in the fitness programs for the team. if a large proportion of players are categorized as 'Overweight' or 'Obese', it may be necessary to adjust the training program to focus more on weight loss and cardiovascular fitness.

4.1.3 Calculation of Mean Max and Min of all Parameters

we calculate the mean age, maximum height, minimum weight, and mean BMI of the players. It also calculates the maximum values of various fitness parameters such as bench press, squat, speed, and vertical jump. In addition, it calculates the mean values of body fat percentage, 40-yard dash time, agility score, endurance test score, and flexibility test score.

Outputs:
Mean age: 26.6

Maximum height: 180

Minimum weight: 62

Mean BMI: 23.8

Maximum bench press: 100

Maximum squat: 116

Maximum speed: 11.0

Mean body fat percentage: 12.8%

Maximum deadlift: 105

Mean 40-yard dash time: 4.5

Maximum vertical jump: 55

Mean agility score: 8.2

Mean endurance test score: 9.1

Mean flexibility test score: 7.9

Overall, the results suggest that the players in the dataset are highly fit and have impressive athletic abilities across a range of parameters.

4.1.4 Players Selection Based On Fitness Data

At first reads the data into a pandas dataframe and then filters the dataframe based on the player type - batsmen, bowlers, or all-rounders.

After filtering, the code calculates the overall fitness score for each player type by taking the mean of seven different fitness measurements - bench press, squat, speed, Deadlift, vertical jump, agility score, and endurance test.

The dataframe is then sorted in descending order based on the fitness score, and the top 15 players for batsmen and bowlers and the top 10 players for all-rounders are selected.

Table:top bowlers ,top batsman and top allrounders based on fitness data.

Table 4.1 : Players Selection Based On Physical Fitness

Top bowlers:	**Top batsmen:**	**Top all-rounders:**
Fazle	Anamul Haque	shakib al hasan
arafat sunny	Mahmudul Hasanjoy	Saifuddin
Haider Ali	Yasir Ali	Sabbir Rahman
Rony Talkdur	Emon	Mahmudullah
Abu Jayed	Shahin Alam	soumya sarkar
Mahmud	Rakibul Hasan	afif Hossain
Avishek Das	Akbar Ali	nasir Hossain
Al Amin	Tanzid Hasan	mehidy hasan miraz
Shahadat Hossain	Shamsur Rahman	mosaddek Hossain
Rubel Hossain	Shamim Hossain	
Kamrul Islam	Mohammad Naim	
Tanzim Hasan Sakib	Saif Hossain	
Shahidul Islam	tamin Iqbal	
Jubair Hossain	muktar ali	
shoriful islam	liton kumar das	

Mean of fitness of all types of players

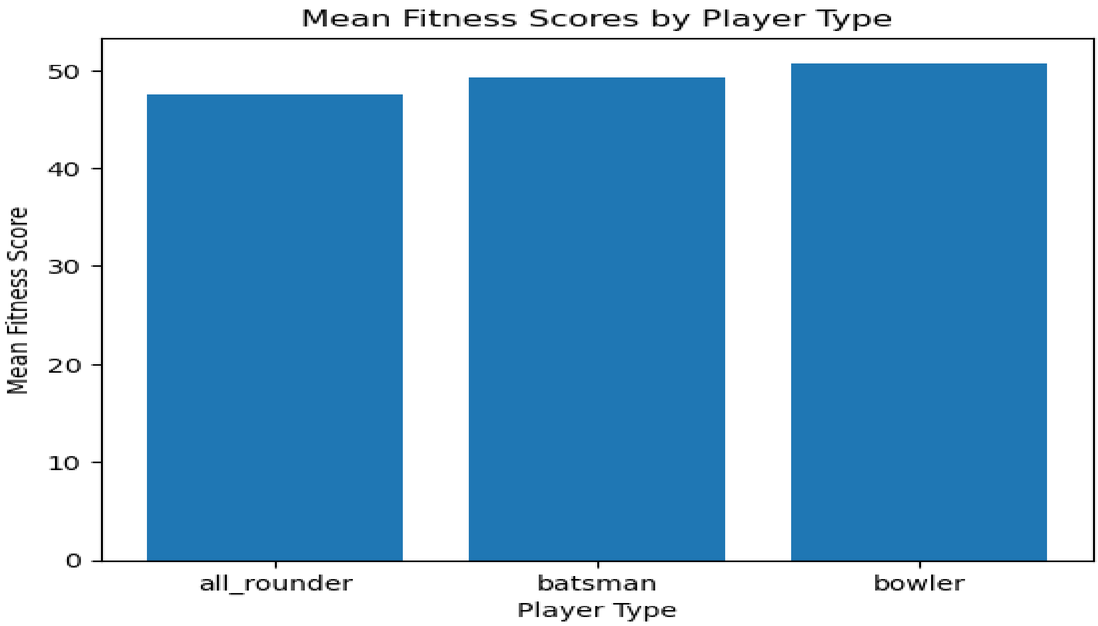


Fig 4.5 : Three Types Of Players Mean of Fitness

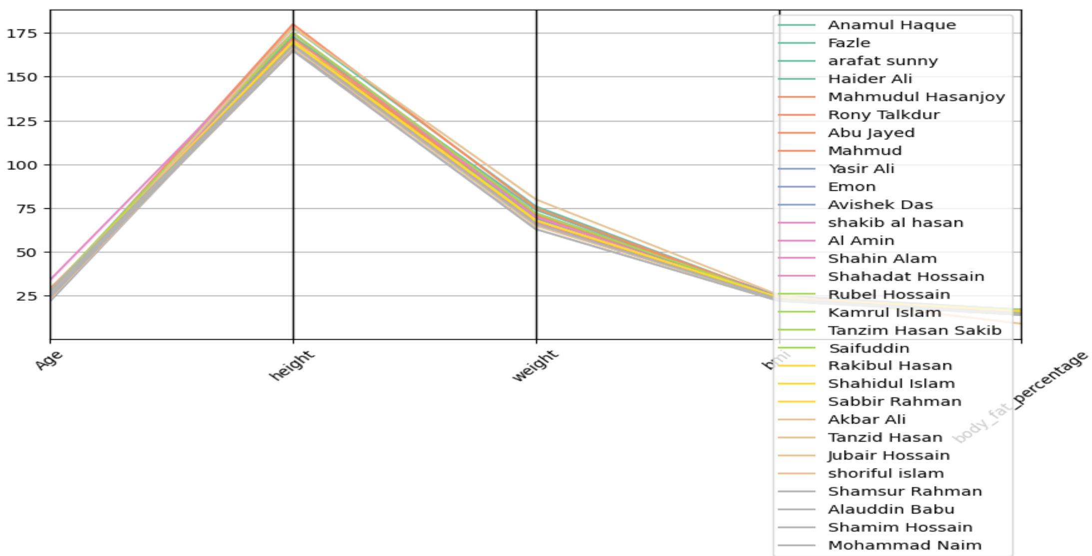


Fig 4.6: Parallel Coordinates Plot for the Top Players

4.1.5 Linear Regression Result

After all types of data analysising on fitness data of cricketers we can apply linean regression algorithms for train and test of cricketers fitness. Then performs a linear regression analysis to predict a player's "bench press" score based on various fitness metrics. It first splits the data into training and testing sets using the "train_test_split function from the Scikit-learn library. It then creates a linear regression model using the LinearRegression fits the model to the training data using the "it method, and predicts the "bench press" scores for the test data using the predict method. Finally, the code prints the regression coefficients, mean squared error, mean absolute error, and R-squared value of the model using the Scikit-learn metrics functions.

Lastly, the code calculates a performance score for each player based on their fitness metrics, sorts the dataset by performance score in descending order, and selects the top 20 performing players. It then prints the names and performance scores of these players.

```
Coefficients: [ 0.12484641  0.17843709 -0.46755291  0.08187692  2.72756046  1.44388616
 11.91249647  0.10027657 -0.66105338]
Mean squared error: 1.567173599744603
Mean absolute error: 0.9910770306486304
R-squared: 0.9807511942713047
Top performing players:
           player name  performance_score
26        Anamul Haque              347.5
27               Fazle              338.3
25        arafat sunny              334.0
31   Mahmudul Hasanjoy              332.3
29          Haider Ali              331.8
0       shakib al hasan             329.8
32        Rony Talkdur              328.4
40           Abu Jayed              326.7
30           Yasir Ali              326.3
28              Mahmud              326.2
18       shoriful islam             323.3
48         Avishek Das              322.8
```

Fig 4.7: Linear Regression Results

From this figure we can see our model gives us 98% accuracy,MSE 1.567 ,MAE 0.99.

4.2 Selection Process Of Selected Captain :

We have discovered that there were 15 captains competing in the competition. As a result, we picked them up for the procedure and displayed various parameters on the "BAR DIAGRAM" to display the captain's performance.There we considered players individual performance, winning percentage of participated matches or captained matches. Then we analyzed that information and give value of each player. Then we choose captain as highest valued player. So we picked them.

used variables:

Table 4.2: Leadership Score Calculation Parameters

Batting average
Bowling average
Fielding accuracy
Win percentage
Loss percentage

Experience

Leadership score: First, it defines a dictionary containing data on each player's batting average, bowling average, fielding accuracy, experience, win percentage, and loss percentage. The code then replaces any 'None' values in the bowling average column with 0 and normalizes each of the columns except for the win and loss percentages by dividing each value by the maximum value in that column.

Next, we calculate a leadership score for each player using a weighted formula that incorporates the normalized values for batting, bowling, fielding, and experience. The resulting leadership score is added to the dictionary as a new column. And show that in bar graph.

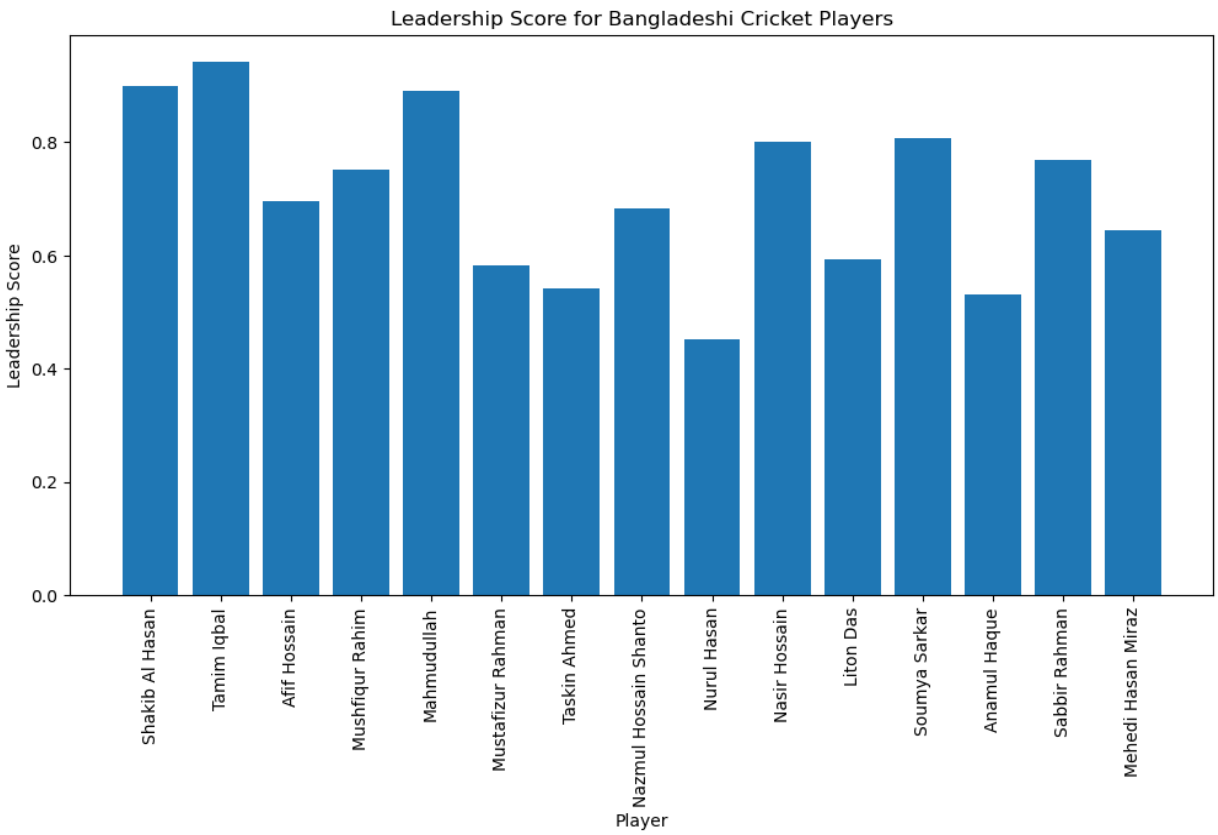


Fig 4.8: Leadership Score Graph For All Captain

Then, we used batting, bowling, and leadership scores. We plot experience, win and loss percentages in a bar graph for each captain.

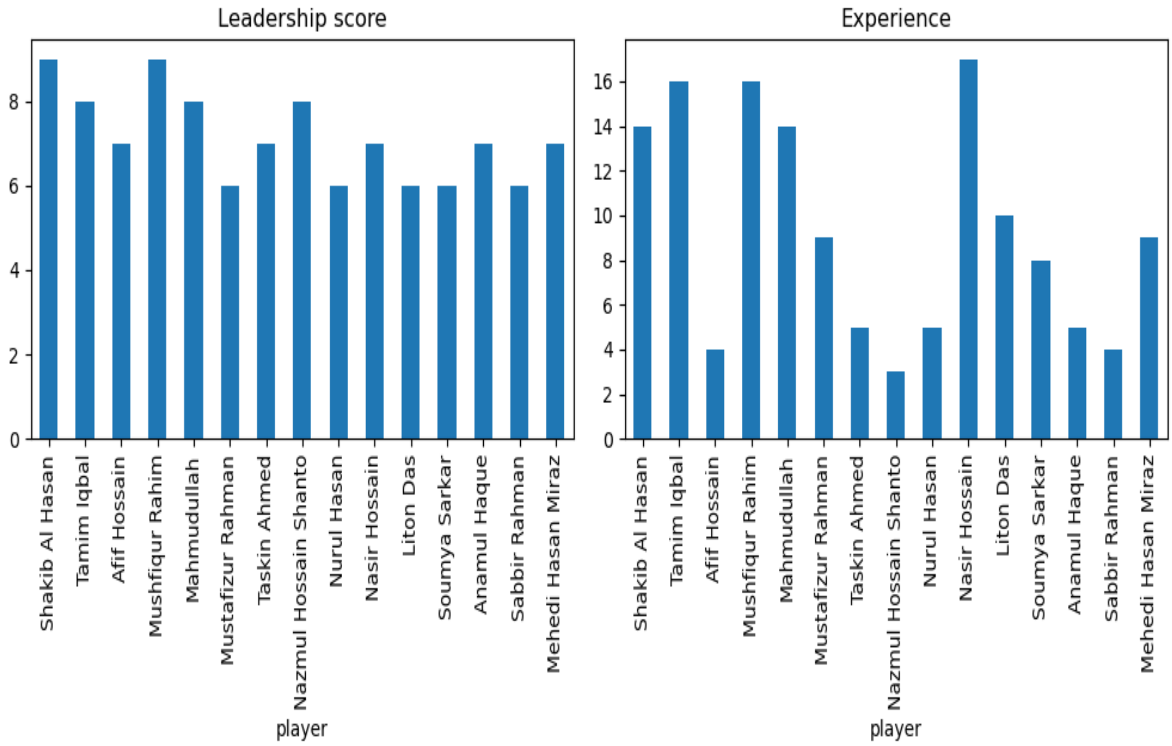


Fig 4.8 :Leardership Score And Experience Players

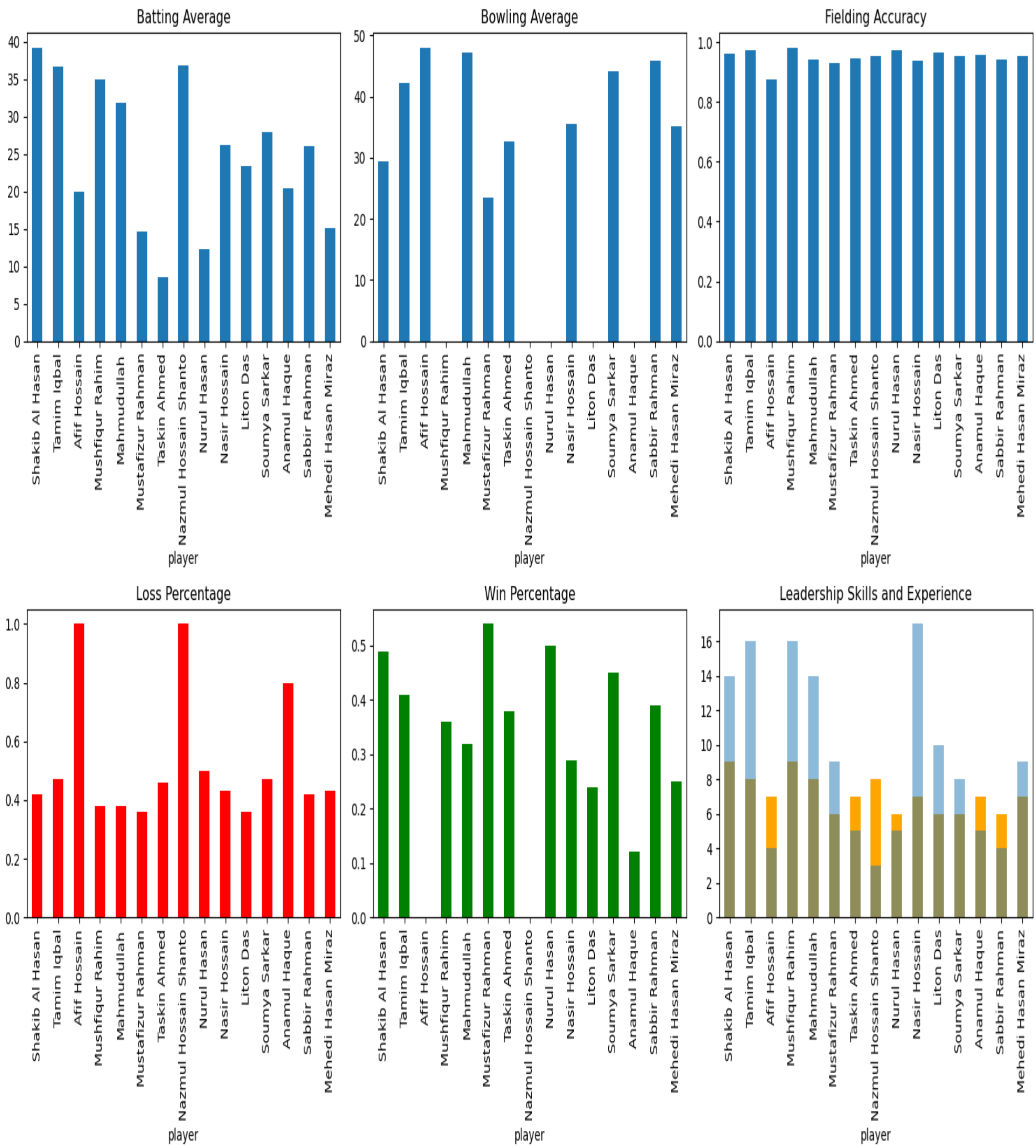


Fig 4.9 :Captains selection for All Parameters Score Graph

With the aid of all those factors and analysis, we may choose the captain. When choosing a captain, we also take into account all graph performance.

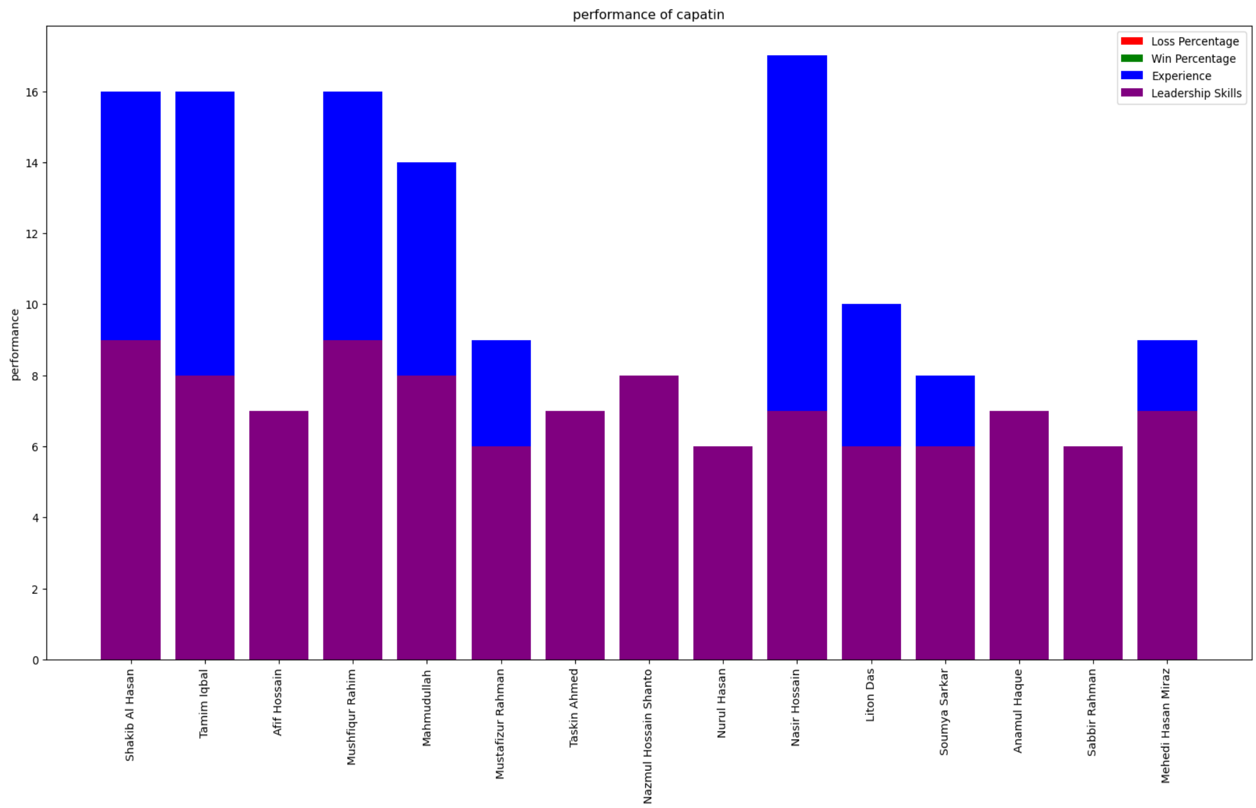


Fig 4.10 : Captain Selection Graph

From this graph we can see top three captains are nassir hossain,sakib al hasan and tamim iqbal.As, we can see from the above bar diagram that Jahurul Islam's performance is higher than the other captain's so Nassir hossain is going to be selected as a captain for the best squad.

We didn't measure only the performance percentage for the captain but also we also have measured his performance as a batsman because only because of captaincy we can't take a player in the team. He also should have create an impact on either batsman area or bowler

area. (Can be both also)

So. Our Captain is "Nassir Hossain"

For our capatin data, we also use decision tree regression techniques for both training and testing, giving us an accuracy of 50%.

4.3 Selection Process Of Selected Wicket Keeper:

We have discovered that there were twelve wicket keepers present at the competition. So, we selected them for the procedure and plotted several metrics in the "BAR DIAGRAM" to illustrate the wicket keeper's performance**.**

Used Parameters:

Table 4.3 : Wicket keeper Selection parameters

Total matches
Total stumpings
Total catches
Total run outs
Dismissal
Max dismissal per inn
Dismissal/innings
Total runs
Run rate
Batting strike rate

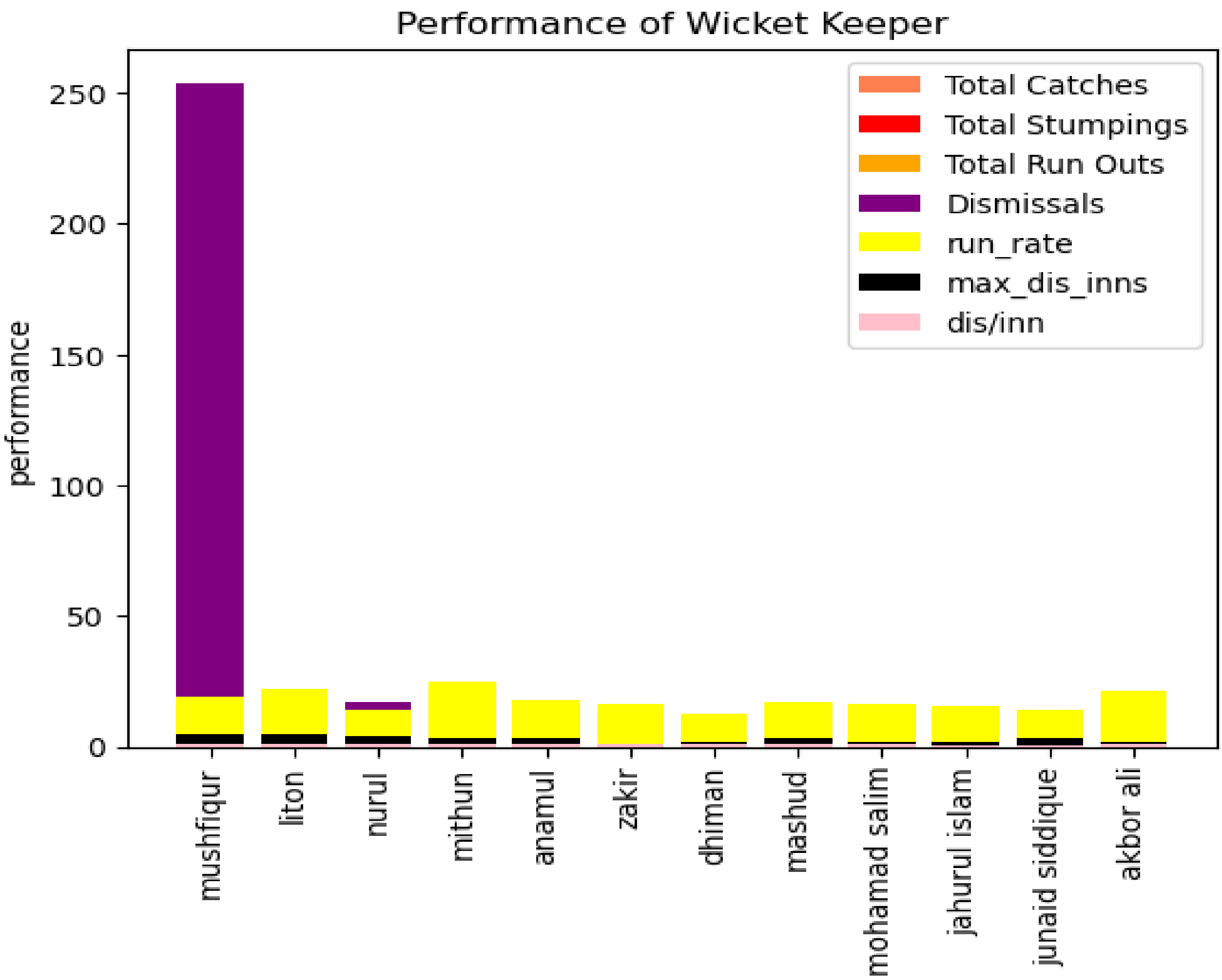


Fig 4.11: Wicket Keeper Performance Graph

Because we can't include a player in the team by simply looking at one statistic, we measured the wicket keeper's performance across a variety of parameters as well. He ought to have had an effect on additional factors as well .We can see from previous bars that mushfiqur is the greatest at putting together a strong squad so that he can play the wicket-keeping position.

So. Our wicketkeeper is "mushfiqur rahim."

4.4 Performance Prediction for Cricketers

In order to predict cricket players' performances, we employed linear regression, support vector regression, and random forest regression in the batting, fitness, and combination of batting and bowling datasets. We also applied the batting and fitness dataset in the combination of bowling.

Performance predication for batsmen :

Table 4.4:Batsmen performance predication

Performace prediction metrics	Linear regression	SVR	Rndom forest
MSE	23.18	9.18	33
MAE	8	8.1	3
Accuracy	84%	99%	86

Performance predication for bowlers:

Table 4.5:Bowlers performance predication

Algorithms	Linear regressiom	Random forest
MSE	339.33	148
MAE	33.25	49.50
Accuracy	93%	72%

Performance predication for overall performance :

Table 4.6:Overall performance predication

Algorithms	Linear regression	Random forest	SVR
MSE	6.778	8.13	2.43
MAE	2.43	1.92	5.10
Accuracy	98%	97%	99%

So we can see support vector regression gives us best performance predication accuracy.

4.5 Selection Process Of Cricketers For National Team

First, utilizing all relevant performance-related characteristics, we compute bowling performance from a bowling dataset and batting performance from a batting dataset. When all is said and done, we merge three datasets, calculate overall performance using batting and bowling performance and fitness .Then select the top players based on those results.hey are chosen via a snake draft technique.we select two squad from the combing dataset.

	squad1	squad2
0	Mustafizur Rahman	Taijul Islam
1	Imrul Kayes	Rubel Hossain
2	Shafiul Islam	Al-Amin Hossain
3	Litton Das	Anamul Haque
4	Raqibul Hasan	Taskin Ahmed
5	Nasum Ahmed	Nurul Hasan
6	Dhiman Ghosh	Jahurul Islam
7	Nazimuddin	Najmul Hossain Shanto
8	Mohammad Naim	Sunzamul Islam
9	Yasir Ali	Naeem Islam
10	Mehidy Hasan Miraz	Abu Jayed
11	Nasir Hossain	Soumya Sarkar
12	Mosaddek Hossain	Afif Hossain
13	Shuvagata Hom	Sabbir Rahman
14	Shamsur Rahman	Mohammad Mithun
15	Sohrawordi Shuvo	Mahbubul Alam

Algorithms results decision for combine dataset:

Using logistic regression, decision tree regression, linear regression, and support vector regression, we can train a combined dataset.

We get a solid fit for our player pick from all those systems.

Table 8: Algorithms Performance

algorithms	accuracy	MSE	MAE
Support vector	97%	188.89	9.92
Linear regression	97.2%	5.61	5.20

From this table we can see linear regression give us best accuracy for out model. Support vector also give us good fit model.

4.6 Franchise League Player's Selection Process

In the FRANCHISE LEAGUE player selection process, we gather information on batting and bowling in the T20 format from players in various countries before also calculating their performance and rankings in the T20 format of international cricket. Choose players for the snake draft system based on performance and ratings. We exclusively choose players for the BPL in our model. Before combining them to determine overall performance in various country FRANCHISE LEAGUEs and their national team performance, we first calculate batting performance from batting datasets and bowling performance from bowling datasets. We also took into account the competitiveness of the athletes we chose.

]:

	Dhaka	Chattogram	Cumilla	Rajshahi	Khulna	Rangpur	Sylhet	Barishal
0	CH Gayle	Rashid Khan	SL Malinga	Mashrafe Mortaza	DW Steyn	TA Boult	JJ Bumrah	Imran Tahir
1	HM Amla	RA Jadeja	AB de Villiers	LRPL Taylor	MS Dhoni	B Kumar	M Morkel	V Kohli
2	F du Plessis	KS Williamson	AJ Finch	Shakib Al Hasan	Q de Kock	MJ Guptill	RG Sharma	JE Root
3	Tamim Iqbal	S Dhawan	EJG Morgan	JM Bairstow	Babar Azam	EJG Morgan	Fakhar Zaman	Mushfiqur Rahim
4	JJ Roy	JC Buttler	Aamer Yamin	Imam-ul-Haq	SD Hope	SE Marsh	AT Rayudu	TK Curran
5	PR Stirling	KM Jadhav	DA Miller	SO Hetmyer	SS Iyer	AD Hales	Haris Sohail	MP Stoinis
6	HM Nicholls	EC Joyce	E Lewis	EC Joyce	TM Head	BA Stokes	MK Pandey	Imad Wasim
7	RR Pant	Hazratullah Zazai	N Dickwella	TL Seifert	TD Astle	C Jonker	Shafiqullah	KA Maharaj
8	B Mavuta	DSK Madushanka	SN Thakur	OP Stone	D Pretorius	CJ Dala	Asif Ali	Rumman Raees
9	C Munro	MD Shanaka	C Zhuwao	SW Billings	MM Ali	MR Marsh	Soumya Sarkar	Imrul Kayes
10	MJ Henry	C de Grandhomme	Rahmat Shah	Hasan Ali	MA Wood	MDKJ Perera	AR Nurse	AU Rashid
11	AK Markram	TG Southee	MJ Santner	Sabbir Rahman	PWH de Silva	F Behardien	Rashid Khan	JDS Neesham

Additionally, we use decision trees, support vectors, and linear regression to train the model.

72% accuracy in linear regression is achieved.

Decision tree accuracy is 99% with a mean squared error of 1112 and a mean absolute error of 8.92.

99% accuracy for support vectors with mean squared error of 18.31 and mean absolute error of 1.29.

MSE IS 30.72 MAE IS 2.38, so based on all those algorithms' analyses and findings, our model is the best fit.

CHAPTER 5

CONCLUSION

In this research, we primarily analyze the individual player's playing techniques and fitness based on stastics. This will enable us to apply the necessary solutions for the management of the players. Further, ground analysis is largely based on the experience of cricket players and coach. To design and develop an intelligent system for players' performance prediction for cricket match it is require to consider parameters beyond the batting and bowling. Weather factor is key factor to use which have huge impact on players' performance. We proposed a machine learning model which uses cricket statistics along with weather related dataset. As combined data is imbalanced in nature we propose data balancing as preprocessing before application of machine learning algorithm.we also consider fitness data of cricket players.We have collected only the registered players, entered their name in a different file (Batsmen in batsman dataset, bowlers in bowler dataset). We have used Linear Regression and snake draft system for selecting batsmen and all-rounders. Support Vector Machine, Linear Regression, Decision Tree and RankSVM have been used for selecting bowlers.We have used Bar Graph to show the statistics of different parameters for both captain and wicket-keeper. In this paper, we have taken maximum number of parameters in consideration for selecting 1 captain (for captaincy issue), 1 wicket keeper,batsmen and bowlers . Our model can be extended for the team selection in other formats of cricket too. Our proposed model to select the 15 man squad can be applied to any team in the domestic tournaments. We will integrate the model and upgrade the dataset to apply it for the national team selection for international tournaments.

CHAPTER 6

REFERENCES

[1]Ijsr.net. 2014. Cricket Pitches – Science behind the Art of Pitch Making. [online] Available at: <https://www.ijsr.net/archive/v3i7/MDIwMTU4.pdf> [Accessed 20 February 2021].

[2] ResearchGate. 2020.PredicitngBowling Performance in Cricket from publicity Available Data. [online]Availableat:<https://www.researchgate.net/publication /340070875_Predicting_Bowling_Performance_in_Cricket_from_Publicly_Available _Data> [Accessed 22 February 2021].

[3] Meraj-Bin-Malek, Rakib Hasan Badhan, Mohaiminul Islam Shesir, Nazmul Haque Fakir (2018) Squad Selection for Cricket Team Using Machine Learning Algorithms, Department of Computer Science and Engineering BRAC University, Dhaka - 1212, Bangladesh: Mr. Hossain Arif .

[4] Research Gate. 2019. Image Processing Techniques for Detecting and Classification of Plant Disease: A Review. [online]Availableat:<https://www.researchgate.net/public ati on/334314928_Image_Processing_Techniques_for_Detec ti ng_and_Classification_of_Plant_Disease_A_Review> [Accessed 24 February 2021].

[5]Myweather2.com. 2021. Weather forecast for Galle Stadium| Cricket Ground SriLanka.[online]Availableat:<https://www.myweather2.com/Cricket/SriLanka/Galle-Stadium.aspx> [Accessed 20 March 2021]. [7] Cricinfo.2021. Match results. [online] Available at:

[6]<https://stats.espncricinfo.com/ci/engine/stats/index.html ?class=1;host=8;result=1; result=2;result=3;result=4;templat e= results;type=allround;view=results> [Accessed 10 February 2021].

[7]Kaggle.com. 2020.BestBatsman+Bowlerstossposition.[online]Availableat:<https://www.kaggle.co m/doosrijana m/ best-batsman-bowlers-toss-position> [Accessed 12 March 2021].

[7] Barr, G. and Kantor, B. (2004). A criterion for comparing and selecting batsmen in limited overs cricket. *Journal of the Operational Research Society*, 55(12):1266–1274.

[8] Bhattacharjee, D. and Pahinkar, D. G. (2012). Analysis of performance of bowlers using combined bowling rate. *International Journal of Sports Science and Engineering*,

6(3):184–192.
[9] Breiman, L. (2017). *Classification and regression trees*. Routledge.
[10] Bukiet, B. and Ovens, M. (2006). A mathematical modelling approach to one-day cricket
batting orders. *Journal of sports science & medicine*, 5(4):495.
[11] Duckworth, F. C. and Lewis, A. J. (1998). A fair method for resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society*, 49(3):220–227.
[12] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
[13] Iyer, S. R. and Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. *Expert Systems with Applications*, 36(3):5510–5522.
[14] Jhawar, M. G. and Pudi, V. (2016). Predicting the outcome of odi cricket matches: A
team composition based approach. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2016 2016)*.
[15] Lemmer, H. H. (2002). The combined bowling rate as a measure of bowling perfor mance in cricket. *South African Journal for Research in Sport, Physical Education and Recreation*, 24(2):37–44.
[16] Mukherjee, S. (2014). Quantifying individual performance in cricket—a network analysis of batsmen and bowlers. *Physica A: Statistical Mechanics and its Applications*, 393:624–637.
[17] Muthuswamy, S. and Lam, S. S. (2008). Bowler performance prediction for one-day
international cricket using neural networks. In *IIE Annual Conference. Proceedings*, page
1391. Institute of Industrial and Systems Engineers (IISE)