BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Squad Selection For Cricket Team Using Machine Learning Algorithms

AUTHORS
Meraj-Bin-Malek
Rakib Hasan Badhan
Mohaiminul Islam Shesir
Nazmul Haque Fakir

SUPERVISOR
Mr. Hossain Arif
Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

October 2018

I would like to dedicate this thesis to my loving parents ...

Declaration

It is hereby declared that this thesis /project report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Meraj-Bin-Malek
Student ID: 14101238

Rakib Hasan Badhan
Student ID: 14101212

Mohaiminul Islam Shesir
Student ID: 13201041

Md. Nazmul Haque Fakir
Student ID: 17241019

Supervisor:

Mr. Hossain Arif
Assistant Professor, Department of Computer Science and Engineering
BRAC University

December 2018

The thesis titled "Squad Selection For Cricket Team Using Machine Learning Language"
Submitted by:
Meraj-Bin-Malek Student ID: 14101238
Rakib Hasan Badhan Student ID: 14101212
Mohaiminul Islam Shesir Student ID: 13201041
Md. Nazmul Haque Fakir Student ID: 17241019
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of Computer Science and Engineering (An example is shown. It must be replaced by the appropriate Board of Examiners)

1. Mr. Hossain Arif
   Assistant Professor

2. Dipankar Chaki
   Lecturer

3. Md. Abdul Mottalib,Phd
   Head Of Department

Acknowledgements

We would like to acknowledge my supervisor Hossain Arif. We have completed our thesis with his help. We found the research area, topic, and problem with his suggestions. He guided us with our study, and supplied us many research papers and academic resources in this area. He is patient and responsible. When we had questions and needed his help, he would always find time to meet and discuss with us no matter how busy he was. Also, he arranged a lot of meetings with Kalyan Banik who had provided ideas. He also provided lots of seminars on various topics which really helped us while we were doing our process to rank players.

In addition, I would like to acknowledge Dipankar Chaki who was our co supervisor. He guided us to the right direction whenever we needed any kinds of help. He also helped us to write our thesis in a more meaningful way.

Abstract

There are mainly three renowned formats of cricket – ODI (One Day International), Test Match and T20 (Twenty Twenty). Selecting a 15 men perfect squad for a particular cricket match in a particular cricket tournament isn't an easy task. Coach and captain play a vital role to select the most perfect players for a match to win by keeping many parameters in their head such as – analyzing the past scored runs comparing with the balls faced, strike rate, total 50's , total 100's, whether the player is right handed or left handed, against which team she/ he has scored well enough (for batsman). Analyzing the past conceded runs comparing with the overs she/ he had taken, economy rate, strike rate, wickets, against which team she/ he has performed well, pitch condition, venue etc. In this paper, we have taken maximum number of parameters in consideration for selecting 1 captain (for captaincy issue), 6 batsmen (top-order, middle-order and finisher), 2 all-rounders, 1 wicket keeper and 5 bowlers (fast bowlers, spinners). Our model can be extended for the team selection in other formats of cricket too. We have used k-means clustering, Linear Regression, Naive Bayes and Page Rank algorithms for selecting batsmen and all-rounders. Support Vector Machine, Naive Bayes, 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. Our motive is to recommend 15 men squad for a cricket team to the selectors.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Introduction

In today's world, we do have a lot of games which are renowned and they are broadcasted nationally, internationally when a tournament occurs in any countries of the world. People all over the world get very excited for these kinds of tournaments. These games are a source of entertainment among various ages of people. Cricket is one of the most renowned games which have its own colors to amaze people, to attract people towards it. May be, this is the reason of having huge amount of people who are loving and are being connect directly and indirectly with this game. In south Asia, cricket is considered as a medium of more than entertainment. During this game of cricket, the people become very much interested to watch this particular game. It becomes a "festival" as mentioned before.

1.2 A Brief Introduction About Cricket

Well, cricket is both indoor and outdoor game which is being played by two teams with wooden bat and ball in a well-defined and mapped, rectangular area which is called pitch surrounded by a huge mapped circular area. Each team will have fifteen players in the squad and among these only eleven will be eligible to play on the ground. Among these players some are known as batsman, some are known as bowlers, some are known as all-rounders and a wicket keeper. There could be different types of batsmen such as – opening batsmen, middle order batsmen and finisher. There could be different types of bowlers such as – fast bowlers, spinners, medium fast bowlers in the squad. Primarily, each team has more than eleven players for the selection process (about thirty men squad used to be in listed/ being offered for camping). For indoor games there could be less than the actual number of players

(mostly governed by authorities of those particular matches). While one team is batting on the pitch, the other team must be bowling and one such session is called an "INNINGS". The batters bat on a 22 yards long pitch which differs a lot match by match. This is called "CONDITION OF PITCH". This is a huge factor to know as well as understand before the game starts. Just a half hour before the game starts, a "TOSS" happens between the two team's captains where some match officials also should be present and a spectator for the witness. Both ends of the pitch there are wickets which are made of wood and two small pieces of woods on the three sticks ("STAMPS") which are called "BAILS". Every players of bowling team should be on the ground while playing for the fielding. One specific player from the fielding side should be behind the stamps for keeping (This man is called the "WICKET KEEPER"). On the other side of the stamps, there should be a man who directs and examines the match and this man is called "UMPIRE" (Usually three umpires should be present in an international match, two umpires on the ground and the rest umpire is an additional umpire, called "THIRD-UMPIRE". He plays a very important role in a match. When the on-field umpires fail to make a decision, they refer to the third-umpire to make a decision by watching the match very carefully on television and thus send the message to the on-field umpires where the message can be seen on the big screen). One of the players throws the bowl for the batsman while the rest of all should be on the field for the fielding. The batsman who is facing the bowl is called "STRIKER". The striker takes guard on the popping crease which is four feet away in front of the wickets. The striker has to save his wicket from being hit by the ball by striking the ball hard with his bat. Striker tries his level best to hit the ball well enough to score maximum runs on each delivery. Runs can be achieved in two different ways. One way is to hit the ball hard enough for it to cross the boundary. If the batsman hits the ball into the air and the ball crosses the boundary before dropping on the ground, the batting team gets six runs, which is the maximum number of runs that can be scored on a legal delivery; otherwise the batting team gets four runs if the ball drops before crossing the boundary. Another way to score runs is by the two batsmen swapping ends running the length of the pitch in opposite directions while the fielders retrieve the ball. Who will bowl for the coming overs, is being directed by the on-field captain. The captain from the bowling side makes every single decision for the bowling side. An over consists of six balls bowled by a bowler. A different bowler comes in to bowl the next over. The number of balls may increase with illegal deliveries as wide balls or no balls which act as penalties against the bowling team in the form of an extra run and an extra ball for the batting team in that over. There are several ways in which a delivery can be declared a no ball or a wide ball. The umpires declare a delivery as a no ball or a wide ball according to those rules. Two players from the batting side are on the pitch for scoring runs. One of them bats from

one end while the other one waits at the other end where the bowler is bowling from. If a batsman gets out by any sort of rules and regulations then it's called a "WICKET". Batting team will have 10 wickets in hand to score runs as many as possible within limited overs in ODI's / T20's. For test matches, there aren't limited overs but the batters should bat very carefully to stay on the pitch as long as possible. On the other side, the bowler side must get 10 wickets as soon as possible so that the opponent cannot score a good amount on the score board. A striker can get out in various ways but should be in legal ways. When the bowler hits the wickets directly with the ball and removes the bells from the stumps, this one is called "BOWLED" by the bowler and the striker has to leave the pitch. When the batsman prevents the ball from hitting the stumps with his body, he is said to be dismissed as "Leg before Wicket" (lbw) (there are some rules and regulation which are formulated by ICC and the umpire should follow them before calling that the striker is "OUT"). If the striker leaves the popping crease and misses the ball and the wicket keeper removes the bells by hitting wickets with the ball, the batsman is dismissed as "STUMPED". If the batsman hits the ball into the air and the ball is caught by a fielder without dropping on the ground, the type of dismissal is called "CAUGHT". If a fielder retrieves the ball and removes bells from the stumps by hitting them with the ball, before the batsman reaches the crease while swapping ends to get a run, the batsman running towards the end where the bells have been removed, is said to be dismissed by "RUN OUT". Any type of wicket except run out is said to be taken by the bowler who bowled the ball. Finally, the question is which team is the winning team among the two teams? Well, if team A scores more run than team B then team A wins that match and regarded as a winning team where the best player (known as "MAN OF THE MATCH") comes out from the winning team most of the time.

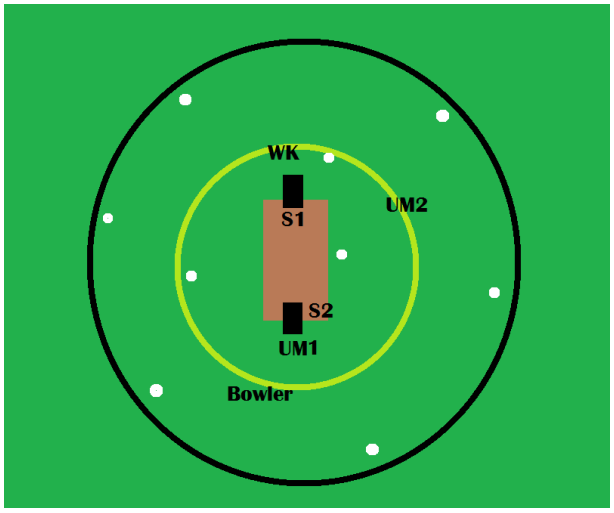

Fig. 1.1 A simple structure of a cricket match

Explanation of above picture:

White dots are representing the bowler's side fielding position.

1. Bowler: Starting point of a bowler to run.
2. M1 and UM2: Umpire 1 and 2 respectively.
3. WK: Wicket-keeper.
4. S1 and S2: Striker 1 and 2 respectively.
5. Black circle representing the boundary of the field.
6. Light green circle representing the inner circle.
7. Black rectangular box is representing STAMPS.
8. Light brown box is representing the PITCH.

Obviously, this game is played under some rules and regulations which are given by international cricket council (ICC). An international game is totally governed by the ICC. There is also another organization which is called ACC (Asian Cricket Council). Cricket usually generates a huge amount of numeric data that's why it's known as the game of statistician delight. Only because of these reason many statistical tools and model are used to generate and thus, select a well-defined team for winning a particular game. This can be T20 or One day international or even test matches. This should be the way of the selectors to select an individual for the matches to be played. Later on, in the literature review we will discuss the models in a broad way.

1.3 Understanding Formats of Cricket

Usually, there are three formats of cricket such as - one-day matches, T20 matches and test matches. One-day matches, also known as ODIs (One Day International) and T20 (twenty-twenty) matches are also known as limited over's cricket as over is less than ODI's (each team will face 20 over of bowling). In these formats there are two innings, so each team gets one chance to bat and later on, one chance to bowl. In ODIs, a maximum of 50 over can be bowled in one innings and in T20s, as the name defines/ describes this game, a maximum of 20 over can be bowled in one innings. So, an innings ends if 50/20 over have been bowled or the batting team has lost 10 wickets. At the end of the first innings the teams change roles. The bowling team now bats and tries to chase the target set by the other team within 50/20 over or before losing their 10 wickets. According to the previous process, the batting team now bowls and tries to prevent the other team from chasing the target down within 50/20 over or by taking 10 wickets. Test matches on the other hand are played over a maximum of five days and each team can play up to two innings in a match. Limited over cricket is more challenging for both batsmen and bowlers. Batsmen need to

score runs as fast as possible and the bowlers need to restrict them by conceding the least runs and taking wickets. The focus of this thesis is to define the best model for selecting the best batsmen (in three categories such as – opener, middle order and finisher), bowlers (fast bowler, spinner), all-rounder's, captain (for the captaincy issues), wicket-keeper. First of all, we will categories them/ expands them in their section. Then we will find the best of them by considering maximum attributes/ operators/ issues. Thus, in this study, we are trying to measure the performance of each selected player, and then we will announce the best fitted players (the best 15 men squad) for a cricket match.

1.4 Contributions

The main contributions of this thesis are:

1. We have read the data of DPL (Dhaka Premiere League) players separately. We have created a file for only registered batsmen, another file for only registered bowlers, another file for registered all rounder's, another file for the wicket keeper, another file for the captains. So, before jumping to the algorithms we will create a database for each category.

2. Then, we have analyzed them – classified them by measuring their performance (this performance will be based on different parameters and obviously will differ from each database such as – the measurement of the performance will differ from the bowlers as well as all-rounder's, captain, wicket-keeper)

3. Our proposed model is the best fitted model for "any sorts of cricket format" just because we have treated each registered player differently (Primarily we have collected the data of the "registered" DPL players by seeing their "Playing Role" from the ESPN database).

4. We have compared different parameters for each database.

5. We have compared the accuracies of different multiclass classification algorithms on our datasets.

1.5 Outline

The rest of the thesis is organized as the following table -

Chapter 2 highlights some work related to the game of cricket.

Chapter 3 describes the data and the preprocessing that we did on the data. We describe the statistics and the attributes that are used to measure the players' performance. We also introduce some new attributes that we used in this study.

Chapter 4 gives a brief description of the machine learning algorithms that we used to create the prediction models.

Chapter 5 reveals the results of the experiments that we carried out on our data. We also discuss and compare the results and performances of different machine learning algorithms on our data.

Chapter 6 concludes the thesis and gives some directions for future work in the field.

Chapter 2

Literature Review

2.1 Related Works

Now-a- days player prediction is the most important and challenging task for any sports especially in cricket. The team management, the coach and captain have to select best eleven players for the particular tournament or series from a squad of 15 to 20 players. They mainly focused on player's previous performance against the team or venue or series. It's a very tough task to analyses each and every player for a team selector. For that reason many researchers have tried to predict the performance of the player. They used various methodologies for their research. As far as we researched about this topic, we found very few articles related to best player selection in the game of cricket.

[11] In Industrial Engineering Research Conference, 2008 Muthuswamy, S. and Lam, S. submitted an article "Blower Performance Prediction for One-day International Cricket Using Neural Network". In their research they mainly predicted the performance of Indian bowler against some international cricket team whom are played against India frequently. They used Backpropagation network and radial basis function network to predict how many runs a bowler is likely to concede and how many wickets a bowler is likely to take in a given ODI match.

Wikramasinghe, I. [18] predicted the performance of batsmen in a test series using a hierarchical linear model.

Iyer, S. R. and Sharda, R. [1] used neural networks to predict the performance of players. They classified themselves into batsmen and bowlers separately in three categories – performer, moderate and failure. Then based on the number of times a player has received different ratings, they recommend if the player should be included in the team to play World Cup 2007.

Saikia, H. and Bhattacharjee, D. [7] classified all-rounders in four categories using

Naïve Bayes classification: Performer, Batting All-rounder, BowlingAll-rounder and Underperformer. They used the data of 35 all-rounders who played in first three seasons of IPL to generate the classification model and used the model to predict the expected classes of six new all-rounders.

Saikia et al. [8] predicted the performance of bowlers in IPL IV using artificial neural networks. They used the bowlers' performance measures from ODI and T20I (T20 international) matches. A lot of work has been done to measure players' performance and rank them. These rankings can then be used to select players for matches and tournaments.

Lemmer, H.H. [9] defined a new measure called Combined Bowling Rate to measure the performance of bowlers. The Combined Bowling Rate is a combination of three traditional bowling measures: bowling average, strike rate and economy.

Mukharjee, S. [10] applied Social Network Analysis to rate batsmen and bowlers in a team performance. He generated a directed and weighted network of batsmen-bowlers using player-vs.-player information available for test and ODI cricket. He also generated a network of batsmen and bowlers using the dismissal record of batsmen in the history of cricket.

Parker, D., Burns, P. and Natarajan, H. [13] defined a model for valuation of players for IPL auction. Their model considered factors like previous bidding price of the player, experience of the player; strike rate etc.

Barr, G. D. I. and Kantor, B. S. [5] produced a new range to compare and selecting batsman for limited over cricket team. They mainly focused on provability of getting out and introduced a new formula for the strike rate. The used two dimensional graph and put the strike rate in x-axis and put out in y-axis. Than they decide a selection criteria for selection the player.

Bhattacharjee, D. Pahinkar,D. [2]used bowling rate to find out the best bowler for Indian premier league (IPL). They also considered the other factor of any bowler like wicket taking tendency, fielding efficiency, and catches of his career.

Shah, P. [17] additionally characterized new measures to players' execution. The new measure for batsmen considers the nature of every bowler he is confronting and the new measure for bowlers considers the nature of every batsman he is rocking the bowling alley to. The total of individual execution of a batsman against every bowler is the aggregate execution record of the batsman. Likewise, the total of individual execution of a bowler against every batsman is the aggregate execution record of the bowler.

Ahmed et al. [4] used evolutionary multi-objective optimization for cricket team selection. They used batting average and bowling average as a measure of performance for batsmen and bowlers. They redefined team selection as a bi-objective optimization problem and then

used non-dominated sorting genetic algorithm for multi-objective genetic optimization over the team.

Omkar, S.N. and Verma, R. [12] used genetic algorithms for selecting a team. They defined the fitness of a team by considering the individual fitness of each player on the team. The fitness of a player is calculated based on his performance in batting, bowling, wicket-keeping, fielding, his physical fitness and his experience in the game. They also considered the team's performance against a particular team, on a particular pitch and the recent performance of the team. Then they used the genetic algorithm by representing the team as a string where each string bit represented a player.

Sankaranarayanan et al. [16] used data mining techniques to model and predict ODI matches. 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. Out of all those papers, none of them were nationally focused. And they mainly tell us about the batting average or strike rate of a batsman. And for the bowlers they are focusing on the economy or strike rate. We find this as a big issue. If we really think about it, we can find many other possibilities to measure the capability of a player and we also think we have some really qualified players in our national team. 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. Thus most of these talents are not getting recognized. But if we shift our focus on the national level players, observe them close enough, we can surly increase our opportunity to find another player like Sakib-al-Hasan or Tamim Iqbal among them. For the above reason we are doing our research on local players to find the next talents. If we can create an opportunity for them, they will surly get encouraged and put their passion into work. Keeping this in mind we decided to find out the next best players for our country based on their performance. So in our research we collect the data of all the national level players of Bangladesh and research on them. In our paper we are going to focus on the players in a more efficient way, we are going to categorize the batsmen based on how many runs they take per ball.

Chapter 3

Data and Processing

3.1 The Data

We obtained all our data manually from a website named espncricinfo. This is the best renowned source from where we could collect our data. We basically have collected our data in five different sections. Those are - Batsman, Bowler, All-Rounder, Wicket-Keeper and Captain. We have collected these data by looking their bio which is given in espncricinfo. Like - for batsman, we have collected only those batsman whose playing role is "Batsman", for bowler, we have collected only those bowlers whose playing role is "Bowler", for all-rounders, we have collected only those all-rounders whose playing role is "All-Rounder", for wicket keepers, we have collected only those wicket keepers whose playing role is "Wicket keeper", for captain, we have collected only those captains who acted as Captain in these tournament. For batsman, bowler and all-rounder, we have considered matches played in the 2017-2018 secession in Dhaka Premier Division Cricket League. We mainly go through all the matches that played in that secession and collected all the data by manually input in excel file by their playing role.

3.1.1 Batting Measures

Innings: The number of innings in which the batsman has batted. The more innings the player has played, the more experienced the player is.

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.

$$BattingAverage = \frac{RunsScored}{TimesOut} \tag{3.1}$$

Strike Rate (SR): Strike rate is the average number of runs scored per 100 balls faced. In limited overs cricket, it is important to score runs at a fast pace. More runs scored at a slow pace is rather harmful to the team as they have a limited number of overs. This attribute indicates how quickly the batsman can score runs.

$$BattingStrikeRate = \frac{RunesScore * 100}{BallsFaced} \tag{3.2}$$

Centuries: Number of innings in which the batsman scored more than 100 runs. This attribute indicates the capability of the player to play longer innings and score more runs.

Fifties: Number of innings in which the batsman scored more than 50 runs (but less than 100). This attribute indicates the capability of the player to play longer innings and score more runs.

3.1.2 Bowling Measures

Innings: The number of innings in which the bowler bowled. The more innings the player has played, the more experienced the player is.

Overs: The number of overs bowled by a bowler. This attribute also indicates the experience of the bowler. The more overs the bowler has bowled, the more experienced the bowler is.

Bowling Average: Bowling average is the number of runs conceded by a bowler per wicket taken. This attribute indicates the capabilities of the bowler to restrict the batsmen from scoring runs and taking wickets at the same time. Lower values of bowling average indicate more capabilities.

$$BowlingAverage = \frac{NumberofRunConceded}{WicketTaken} \tag{3.3}$$

Bowling Strike Rate: Bowling strike rate is the number of balls bowled per wicket taken. This attribute indicates the wicket taking capability of the bowler. Lower values mean that the bowler is capable of taking wickets quickly.

$$BowlingStrikeRate = \frac{BallsBowled}{WicketTaked} \tag{3.4}$$

Bowling Economy: Economy rate is the average number of runs conceded for each over bowled. A lower economy rate is seen as preferable – it means that the bowler is able to get more batsmen out with fewer balls. The statistic is considered to be more important in shorter games than longer test matches.

$$BowlingEconomy = \frac{RunsConceded}{OversBalled} \tag{3.5}$$

Three/Five Wicket Haul: Number of innings in which the bowler has taken more than four wickets. This attribute indicates the capability of the bowler to take more wickets in an innings. Higher the value, more capable the player.

3.1.3 Captain

This is a binary attribute indicating whether a player is captain of the team. This attribute tries to indicate the control and responsibilities the player has. Some players perform well as captains while some perform worse.

3.1.4 Wicket Keeper

This is a binary attribute indicating whether a player is a wicket keeper. Wicket keepers are primarily batsmen. They are expected to score more runs as they specialize in batting and are less fatigued than other players as they are physically less active during fielding compared to other fielders.

3.1.5 All-Rounder

3.2 Data Cleaning

While collecting the data of batsman, all-rounder, bowler, captain, wicket-keeper, we have faced one problem like - there were some players who were in the best 11 squad but didn't play because his team already declared as a winning team before his time to perform or sometimes he didn't play for a particular match (was in the bench as an extra player). In these cases we have collected the data from previous session of Dhaka Premiere League's.

3.3 Oversampling

TWe observed that a majority of the records fall within class 1 in both batting and bowling. This created a major imbalance in the distribution of values and affected the performance of the learning algorithms. To solve this problem, we applied an oversampling technique Supervised Minority Oversampling Technique (SMOTE) [36] on minority classes to make all the classes equally distributed. SMOTE over-samples minority classes by creating synthetic

example tuples. To create synthetic tuples of minority class, SMOTE takes each minority class sample and creates synthetic examples along the line segment joining any or all of its nearest neighbors. To generate a synthetic sample, the difference between the feature vector under consideration and its nearest neighbor is taken. This difference is then multiplied by a random number between zero and one and the product is added to the feature vector under consideration. This way, a random point along the line segment joining two specific features is selected. Neighbors from the k nearest neighbors are selected based on the amount of oversampling required. e.g. to oversample a minority class by 300(in percentage), three neighbors from a tuple's nearest neighbors are selected and one sample in the direction of each is generated.

3.4 Batsman Parameters

1. Data Set collected from WWW.ESPNCRICINFO.COM

2. This data set collected from the "Dhaka Premere League" matches manually.

3. Total Team = 12.

4. Team Names :

a. AL = Abahani Limited.

b. LR = Legends of Rupgang.

c. KSKS = Khelaghar Samaj Kallyan Samity.

d. PDSC = Prime Doleshwar Sporting Club.

e. SJDC = Sheikh Jalal Dhanmondi Club.

f. GGC = Gazi Group Cricketers.

g. MSC = Mohammedan Sporting Club.

h. PBCC = Prime Bank Sporting Club.

i. BU = Brothers Union.

j. SSC = Shinepukur Cricket Club.

k. KKc = Kalabagan Krira Chakra.

l. ABCC =Agrani Bank Cricket Club.

5. Total Batsman = 61 (person)

6. Attributes :

"x is defining the number of match. In dataset you will see M1, M2, M3..... There M1 stands for Match no 1 and so on....."

a. X(Run) = Runs gained.

b. x(Bowl) = Bowl faced to gain runs.

c. x(SR) = Strike rate for that particular match.

d. x(4's) = Total 4's for that particular match.
e. x(6's) = Total 6's for that particular match.
f. x(OP) = Opponent.
g. x(VNU) = Venue.
h. x(W/L) = Win or Lose?
i. T50's = Total 50's.
k. T100's = Total 100's.
l. TRuns = In Total Runs.
m. TBF = Total Ball Faced.
n. Avg.Run = Averege Run.
o. Avg.SR = Averege Strike Rate.
p. RF(50's) = Relative Frequency of 50's.
q. RF(100's) = Relative Frequency of 100's.
p. L = Match lost.
q. W = Match won.
r. D = Match Drawn.

3.5 All-Rounder Parameters

1. Data Set collected from WWW.ESPNCRICINFO.COM
2. This data set collected from the "Dhaka Premere League" matches manually.
3. Total Team = 12.
4. Team Names :
a. AL = Abahani Limited.
b. LR = Legends of Rupgang.
c. KSKS = Khelaghar Samaj Kallyan Samity.
d. PDSC = Prime Doleshwar Sporting Club.
e. SJDC = Sheikh Jalal Dhanmondi Club.
f. GGC = Gazi Group Cricketers.
g. MSC = Mohammedan Sporting Club.
h. PBCC = Prime Bank Sporting Club.
i. BU = Brothers Union.
j. SSC = Shinepukur Cricket Club.
k. KKc = Kalabagan Krira Chakra.
l. ABCC =Agrani Bank Cricket Club.
5. Total Registered All Rounder = 12 (person)

6. Attributes :

"x is defining the number of match. In dataset you will see M1, M2, M3..... There M1 stands for Match no 1 and so on....."

T stands for total.

1. X(RAB) = Runs gained As Batsman.
2. x(Bowl) = Bowl faced to gain runs.
3. x(SR) = Strike rate for that particular match.
4. x(4) = Total 4's for that particular match.
5. x(6) = Total 6's for that particular match.
6. x(OP) = Opponent.
7. x(VNU) = Venue.
8. x(W/L) = Win or Lose?
9. T(50) = Total 50's.
10. T(100) = Total 100's.
11. T.RAB = In Total Runs as batsman.
12. TBF = Total Ball Faced.
13. Avg.Run = Averege Run.
14. Avg.SR = Averege Strike Rate.
15. RF(50's) = Relative Frequency of 50's.
16. RF(100's) = Relative Frequency of 100's.
17. L = Match lost.
18. W = Match won.
19. D = Match Drawn.
20. x(Over) = Over taken as a bowler.
21. x(Run) = Runs given.
22. X(Wic) = Wickets Taken.
23. x(Mdn) = Maiden Over.
24. x(ECN) = Economy.
25. T(4's) = Total 4's.
26. T(6's) = Total 6's.
27. T.over = Total Over taken as a bowler.
28. T.Run.Given = Total Run Given.
29. T.Wic = Total Wickets.
30. T.Mdn = Total Maiden.
31. T.ECN = Total Economy.
32. T.Win = Total Win.

33. T.Loss = Total Loss.
34. Avg.RAB = Average run as batsman.
35. Avg.BF = Average Ball Faced.
36. Avg.SR = Average Strike Rate.
37. Avg.over = Average Over taken as a bowler.
38. Avg.Run.Given = Average Run Given as a bowler.
39. Avg.Wic = Average Wickets taken.
40. Avg.Mdn = Average Maiden.
41. Avg.ECN = Average Economy.
42. Avg.Win Average win.
43. Avg.loss = Average Loss.
44. Win(in percent) = In percentage win.
45. Loss(in percent) = In percentage Loss.

3.6 Captain Parameters

1. Win(in percent) = In percentage win.
2. Loss(in percent) = In percentage Loss.
3. T(4's) = Total 4's.
4. T(6's) = Total 6's.
5. T(50) = Total 50's.
6. T(100) = Total 100's.

3.7 Wicket-Keeper Parameters

1. T.match = Total Match.
2. T.Stamp = Total Stumping.
3. T.Catch = Total Catch.
4. Run Out = Total Run Outs.

3.8 Bowler Parameters

1. Data Set collected from WWW.ESPNCRICINFO.COM
2. This data set collected from the "Dhaka Premere League" matches manually.
3. Total Team = 9.

4. Team Names :

a. AL = Abahani Limited.

b. LR = Legends of Rupgang.

c. KSKS = Khelaghar Samaj Kallyan Samity.

d. PDSC = Prime Doleshwar Sporting Club.

e. SJDC = Sheikh Jalal Dhanmondi Club.

f. GGC = Gazi Group Cricketers.

g. MSC = Mohammedan Sporting Club.

h. PBCC = Prime Bank Sporting Club.

i. SSC = Shinepukur Cricket Club.

5. Total Batsman = 61 (person)

6. Attributes :

"x is defining the number of match. In dataset you will see M1, M2, M3..... There M1 stands for Match no 1 and so on....."

W/L=Win or loose the match

OP = Opponent

VNU = Venue of Particuler Match

TO = Total Over

RG = Total Run Given

WT = Total Wicket Take

Avg = Average

Sr = Strick Rate

Eco= Economy

5W = Five Wicket Take

3W = Three Wicket Take

Chapter 4

Learning Algorithms

4.1 Supervised Algorithm

Supervised learning is a machine learning technique of deriving a function from a labeled training sample. A training sample is a set of training tuples. A training tuple consists of a set of input attributes and an associated output value. A supervised learning algorithm generates an inferred function by analyzing the training data. This function is then used to classify an unseen data. In predictive analytics, the generated function is called a predictive model. For our study, we used Naïve Bayes, Decision Tree, linear regression, Google Page Rank Algorithm, KNN Classification and multiclass SVM to generate the prediction models.

4.2 Naïve Bayes Classifier

Bayesian classifiers are statistical classifiers that predict the probability with which a given tuple belongs to a particular class [6]. Naïve Bayes classifier assumes that each attribute has its own individual effect on the class label, independent of the values of other attributes. This is called class-conditional independence. Bayesian classifiers are based on Bayes' theorem.

Bayes Theorem:

Let X be a data tuple described by measurements made on a set of n attributes. Let H be a hypothesis such that X belongs to a specified class C. Bayesian classifiers calculate P(H|X), the probability with which the hypothesis H holds true for the observed attribute values of the data tuple X. P(H|X) is called the posterior probability or posteriori probability of H conditioned on X. Similarly, P(X|H) is the posterior probability or posteriori probability of X given H i.e. the probability with which the data tuple X exists, given the hypothesis H is true. P(H) is the prior probability or a priori probability of H which means that H holds

true for a data tuple regardless of the values of its attributes. P(X) is the prior probability or a priori probability of X, which is the probability with which the data tuple X with given attribute values exists. Now, Bayes Theorem is defined as,

$$P(H/X) = \frac{P(X/H).P(H))}{P(X)} \tag{4.1}$$

Naive Bayes Classifier:

1. Let D be a training set of data tuples and their associated class labels, where each tuple is represented by an n-dimensional attribute vector, X=(x1, x2, x3,. . . ,xn).

2. For a multiclass classification, suppose that there are m classes, C1, C2, C3,. . . ,Cm. The Naïve Bayes classifier predicts that a given tuple X belongs to the class with the highest posterior probability conditioned on X. That is X belongs to class Ci if and only if

$$P(Ci/X) > P(Cj/X) \tag{4.2}$$

Thus, we need to maximize P(Ci|X). The class with maximum P(Ci|X) is called the maximum posteriori hypothesis.

From Bayes Theorem,

$$P(Ci/X) = \frac{P(X/Ci).P(Ci))}{P(X)} \tag{4.3}$$

As P(X) is constant for all classes, we need to maximize P(X/Ci)P(Ci). If the class prior probabilities are unknown, all the classes are assumed to be equally probable i.e. P(C1) = P(C2) = P(C3) =. . . =P(Cm) and in that case, all we need to do is to maximize P(X/Ci) . Otherwise, we maximize P(X/Ci)P(Ci).

3. For high dimension data, it would be very expensive computationally to calculate P(X/Ci) . For this, the naïve assumption of class-conditional independence is made which assumes that the attribute values are conditionally independent of each other. Thus,

$$P(X/Ci) = P(X1/Ci) * P(X2/Ci) * P(X3/Ci) * .....P(Xn/Ci) \tag{4.4}$$

4.Now it is easy to estimate the probabilities P(X1/Ci) * P(X2/Ci) * P(X3/Ci) * ..... P(Xn/Ci) from the training tuples. Here, xk refers to the value of the corresponding attribute Ak of tuple X. xk is calculated based on the type of attribute i.e. whether the attribute is categorical or continuous valued. For different types xk is calculated differently as follows:

a. If Ak is categorical, then P(xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci,D|, the number of tuples of class Ci in D. b. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean and standard deviation .

5.P(X/Ci)P(Ci) are calculated for each class Ci. The Naïve Bayes classifier predicts that the tuple X belongs to class Ci if and only if P(Ci|X) > P(Cj|X) for 1 j m, j i

$$P(Ci/X) > P(Cj/X) \tag{4.5}$$

4.3 Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

y = c + b*x

Where, y = estimated dependent variable score,

c = constant,

b = regression coefficient, and

x = score on the independent variable.

4.4 Support Vector Machine

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. Vladimir Vapnik, Bernhard Boser and Isabell Guyon introduced the concept of support vector machine in their paper. SVMs are highly accurate and less prone to overfitting. SVM transforms the original data into a higher dimension using a nonlinear mapping. It then searches for a linear optimal hyperplane in this new dimension separating the tuples of one class from another. With an appropriate mapping to a sufficiently high dimension, tuples from two classes can always be separated by a hyperplane. The algorithm finds this hyperplane using support vectors and margins defined by the support vectors. The support vectors found by the algorithm provide a compact description of the learned prediction model. SVM takes different approaches to classify linearly separable and linearly non-separable data.

Classification of SVM In this SVM training involves the minimization of the error function:

subject to the constraints:

Where C is the capacity constant, w is the vector of coefficients, b is a constant, and represents parameters for handling non separable data (inputs). The index i labels the N training cases. Note that represents the class labels and xi represents the independent variables. The kernel is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

Regression SVM: the error function:

which we minimize subject to:

4.5 Decision Tree Classification Algorithm

J. Ross Quinlan introduced a decision tree algorithm called ID3 in his paper [14]. Later he introduced a successor of ID3 called C4.5 in [15] to overcome some shortcomings such as over-fitting. Later on, L. Breiman, J. Friedman, R. Olshen and C. Stone described the generation of binary decision trees in their book Classification and Regression trees (CART) [3]. ID3 and CART follow a similar approach to learn decision trees from training data. ID3, C4.5 and CART are greedy algorithms which construct decision trees from top to down in a recursive divide-and-conquer manner. They start with a training set with tuples and their associated class labels. The training set is then recursively partitioned into smaller subsets as the tree is being built. The general strategy of the decision tree algorithms is described as follows:

1. The algorithm starts with a data-partition D, an attribute list and an attribute selection method. The data partition D is the entire training set at the beginning. Attribute list is the list of attributes describing the data tuples. Attribute selection method is a procedure that determines the best attribute that discriminates the data tuples according to their class. This procedure uses an attribute selection measure such as information gain or Gini index. The attribute selection measure determines if the decision tree is binary or non-binary.

2. The tree starts at a single node N which contains all the tuples from D. If all the tuples in D belong to the same class, N becomes a leaf and the algorithm stops. Otherwise, the attribute selection method is called which determines the splitting criterion. The splitting criterion determines the best way to partition the training tuples into individual classes and returns the attribute that should be tested at node N. The splitting criterion also tells us which branches to grow from node N with respect to the outcomes of the chosen test. Ideally, the

splitting criterion is determined so that the tuples in the same partition belong to the same class.

3. The node N is labeled with the splitting criterion. Each branch from the node N represents the outcome of the splitting criterion. The tuples in D are then partitioned according to the test determined by the splitting criterion. There are three possible scenarios as shown in figure 4. Let A be the attribute determined by the splitting criterion, having v different values a1, a2, . . . , av.

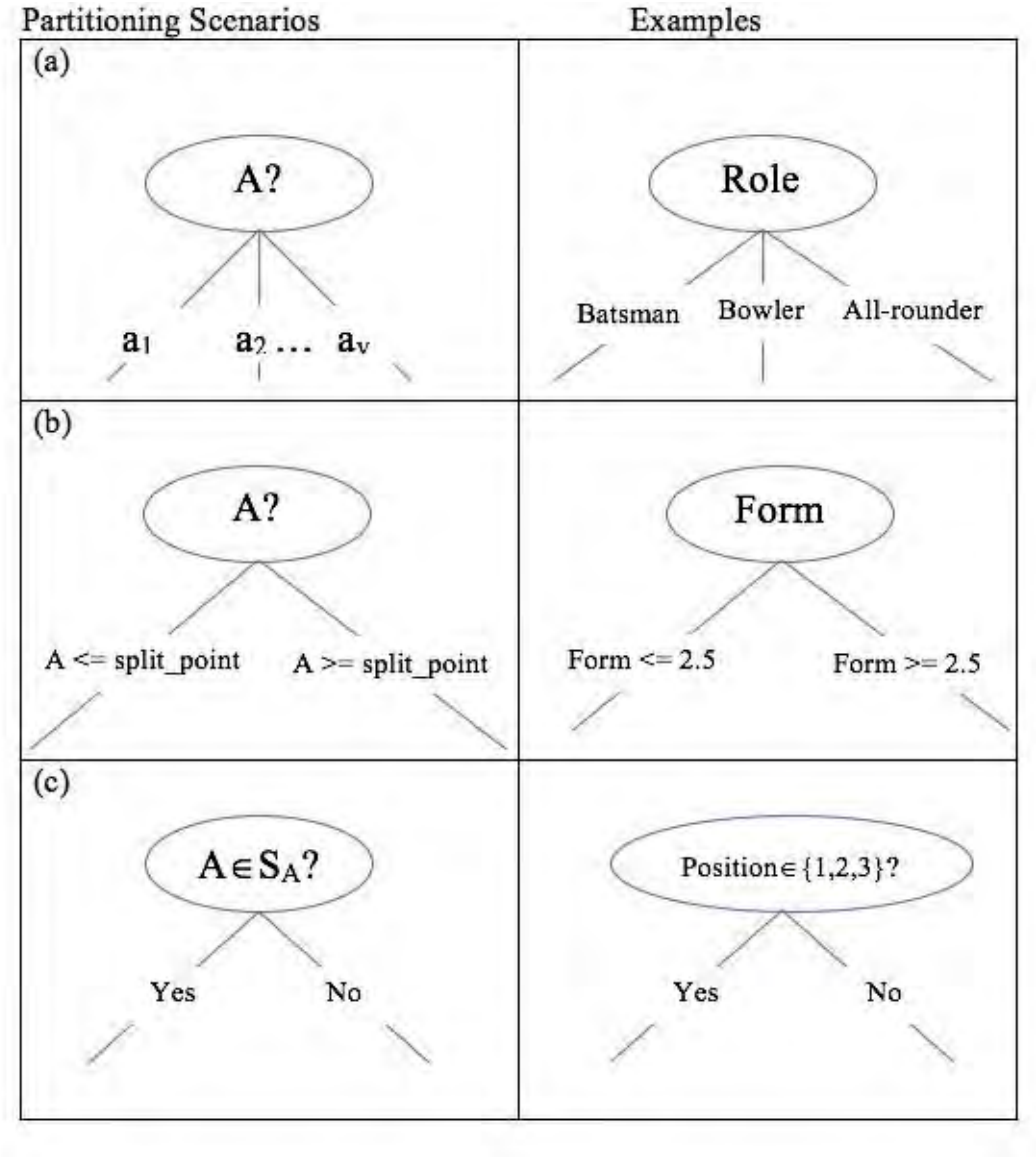

Fig. 4.1 Possibilities Of partitioning tuples

a. A is discrete valued: In this case, the outcomes of the test at node N are simply the known discrete values of A. A branch is created for each value and is labeled with that

value. Partition Dj is the subset of class-labeled tuples in D having value aj of A. As all the tuples in a given partition have the same value of A, A need not be considered in any future partitioning of the tuples.

b. A is continuous values: In this case, there are two possible outcomes of the test at node N based on the spilt point determined by the splitting criterion. Let a be the split point. The two possible outcomes are . Two branches are created from N corresponding to the two outcomes and the tuples are partitioned accordingly.

c. A is discrete valued and a binary tree must be produced: The test at node N is of the form "A SA?," where SA is the splitting subset for A, returned by Attribute selection method as part of the splitting criterion. It is a subset of the known values of A. If a given tuple has value aj of A and if aj SA, then the test at node N is

satisfied. Two branches are grown from N. By convention, the left branch out of N is labeled yes so that D1 corresponds to the subset of class-labeled tuples in D that satisfy the test. The right branch out of N is labeled no so that D2 corresponds to the subset of class-labeled tuples from D that do not satisfy the test.

• The above procedure is called recursively to form a decision tree. The recursive partitioning stops when one of the following conditions is met:

a. All the tuples in the partition D belong to the same class.

b. There are no remaining attributes on which the tuples can be partitioned. In this case, node N is converted to a leaf and labeled with the most common class in D.

c. There are no more tuples to be partitioned. In this case, a leaf node is created with majority class in D.

Decision Tree Induction:

Decision tree induction is the process of creating decision trees for class-labeled training tuples [14]. A decision tree is basically a tree structure like a flowchart [6]. Each internal node of the tree represents a test on an attribute and each branch is the outcome of the test. Each leaf node is a class label. The first node at the top of the tree is the root node. Figure 3 shows a typical decision tree. It is a sample tree describing prediction rules for predicting runs based on the four derived attributes explained on section 3. Internal nodes of the tree are denoted by rectangles and leaf nodes are represented by ovals.

To classify a given tuple X, the attributes of the tuple are tested against the decision tree starting from the root node to the leaf node which holds the class prediction of the tuple. The construction of decision trees is easy as it does not require any domain knowledge or parameter setting. Decision trees can easily handle multidimensional data. The representation of the classification rules in a tree form is intuitive and easy to understand by humans.

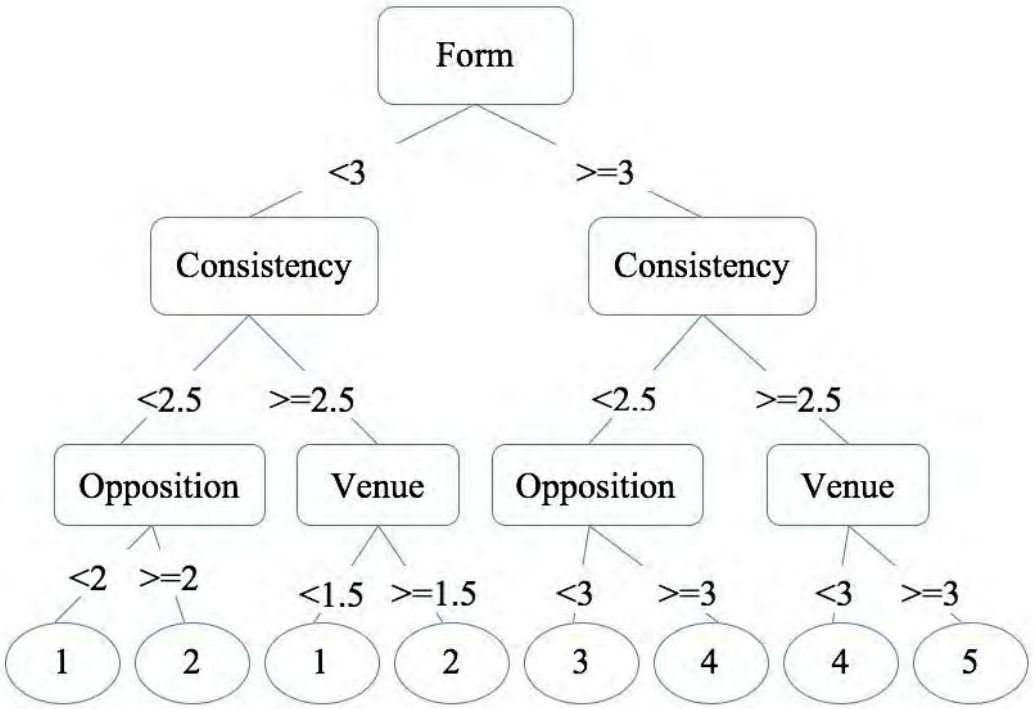

Fig. 4.2 A decision tree describing prediction rules

Decision tree classifiers are fast at learning and classification and have a good accuracy in general.

4.6 Page Rank Algorithm and Implementation

Page Rank (PR) is an algorithm used by Google Search to rank websites in their search engine results. Page Rank was named after Larry Page, one of the founders of Google. Page Rank is a medium of calculating the importance of website pages. According to Google:

"Page Rank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known.

Algorithm:

The Page Rank algorithm gives a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. Page Rank can be calculated for collections of documents of any size. It is assumed in several research papers

that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The Page Rank computations require several passes, called "iterations", through the collection to adjust approximate Page Rank values to more closely reflect the theoretical true value.

Simplified algorithm:

Assume a small universe of four web pages: A, B, C and D. Links from a page to itself, or multiple outbound links from one single page to another single page, are ignored. Page Rank is initialized to the same value for all pages. In the original form of Page Rank, the sum of Page Rank over all pages was the total number of pages on the web at that time, so each page in this example would have an initial value of 1. However, later versions of Page Rank, and the remainder of this section, assume a probability distribution between 0 and 1. Hence the initial value for each page in this example is 0.25.

The Page Rank transferred from a given page to the targets of its outbound links upon the next iteration is divided equally among all outbound links.

If the only links in the system were from pages B, C, and D to A, each link would transfer 0.25 Page Rank to A upon the next iteration, for a total of 0.75.

$$PR(A) = PR(B) + PR(C) + PR(D) \tag{4.6}$$

Suppose instead that page B had a link to pages C and A, page C had a link to page A, and page D had links to all three pages. Thus, upon the first iteration, page B would transfer half of its existing value, or 0.125, to page A and the other half, or 0.125, to page C. Page C would transfer all of its existing value, 0.25, to the only page it links to, A. Since D had three outbound links, it would transfer one third of its existing value, or approximately 0.083, to A. At the completion of this iteration, page A will have a Page Rank of approximately 0.458.

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3} \tag{4.7}$$

In other words, the Page Rank conferred by an outbound link is equal to the document's own Page Rank score divided by the number of outbound links L( ).

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} \tag{4.8}$$

4.7 K-NN Classification Algorithm

K-NN is a non parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

1. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

2. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

Algorithm:

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the overlap metric (or Hamming distance). In the context of gene expression microarray data, for example, k-NN has also been employed with correlation coefficients such as Pearson and Spearman. Often, the classification accuracy of k-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighbourhood components analysis.

A drawback of the basic "majority voting" classification occurs when the class distribution

is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the k nearest neighbors due to their large number.[4] One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its k nearest neighbors. The class (or value, in regression problems) of each of the k nearest points is multiplied by a weight proportional to the inverse of the distance from that point to the test point. Another way to overcome skew is by abstraction in data representation. For example, in a self-organizing map (SOM), each node is a representative (a center) of a cluster of similar points, regardless of their density in the original training data. K-NN can then be applied to the SOM.

Chapter 5

Results and Discussion

5.1 Experiment Setup

We divided our data into training and test sets to find the best combination that gives the most accuracy. We experimented our data on every algorithms, then tried to find the best batsman, all-rounder, captain, wicket-keeper and bowler. (Batsman, all-rounder, captain, wicket-keeper and bowler are being selected if their name pops up in every results. like - if player A pops up first in the result of linear regression, naive bayes, SVM, PG ranking algorithms then player A will be finally selected).

We analyzed and compared the performance of the algorithms in terms of several performance measures, which are described below in short:

Accuracy: The prediction accuracy of an algorithm is the ratio of the number of test instances correctly classified by the algorithm to the total number of test instances. The higher the accuracy, the better the performance.

$$Accuracy = \frac{NumberOfCorrectlyClassifiedInstances}{TotalNumberOfInstances} \tag{5.1}$$

Precision: Precision of a class is the ratio of the number of instances which were correctly predicted to be in that class(true-positive) to the total number of instances which were predicted to be in that class. Precision indicates how useful the model is, as it shows the how many instances were classified correctly from the ones that were classified. Let x be a class,

$$Precision = \frac{NumberOfInstancesOfClassXPredictedCorrectly}{TotalNumberOfInstancesWhichWereToBePredictedInClassX} \tag{5.2}$$

Recall: Recall of a class is the ratio of the number of instances which were correctly

predicted to be in that class(true-positive) to the total number of instances of that class. Recall indicates how complete the model is, as it shows how many instances was the model able to find out correctly out of the total number of instances of a class. Let x be a class,

$$Recall = \frac{NumberOfInstancesOfClassXPredictedCorrectly}{TotalNumberOfInstancesInClassX} \tag{5.3}$$

F1 Score: F1 score of a class is the harmonic mean of precision and recall of the class. It captures the meaning of both precision and recall.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.4}$$

Area under the ROC curve (AUROC): A Receiver Operating Characteristic curve is a graphical plot of true positive rate also known as sensitivity against false positive rate also known as specificity of a binary classifier. True positive rate of a classifier is the ratio of the number of instances that were classified correctly as positives to the total number of positive instances. False positive rate is the ratio of the number of instances that were incorrectly classified as positives to the total number of negative instances. For multiclass problems, ROC curves are generated for each class by using one-vs-all approach. The area under the ROC curves is a measure of accuracy. Its value ranges from 0.5 to 1, with 0.5 meaning least accurate and 1 meaning most accurate. The values of precision, recall, F1 score and AUROC in the tables in the following sections are weighted averages of the values of these measures for each class. We used four machine learning algorithms: Naïve Bayes, Decision Trees, Random Forest and Support Vector Machine in our experiments. We simulated these algorithms in Weka [32] and Dataiku [33]. Following is a brief discussion on the performance of these algorithms and then we compare their performance based on prediction accuracy. All the results in this study have been obtained from Weka [32] 3-9-1-oracle-jvm and Dataiku Data Science Studio [33] on Mac OS 10.11.6 and Windows 10.

5.2 Selection Process Of Selected Captain

Well, we have found that there were twelve captains in the tournament. So we picked them up for the process and plotted different parameters in the "BAR DIAGRAM" to show the performance of the captain.

Used Parameters:

1. Win(in percent) = In percentage win.
2. Loss(in percent) = In percentage Loss.
3. T(4's) = Total 4's.

4. T(6's) = Total 6's.
5. T(50) = Total 50's.
6. T(100) = Total 100's.

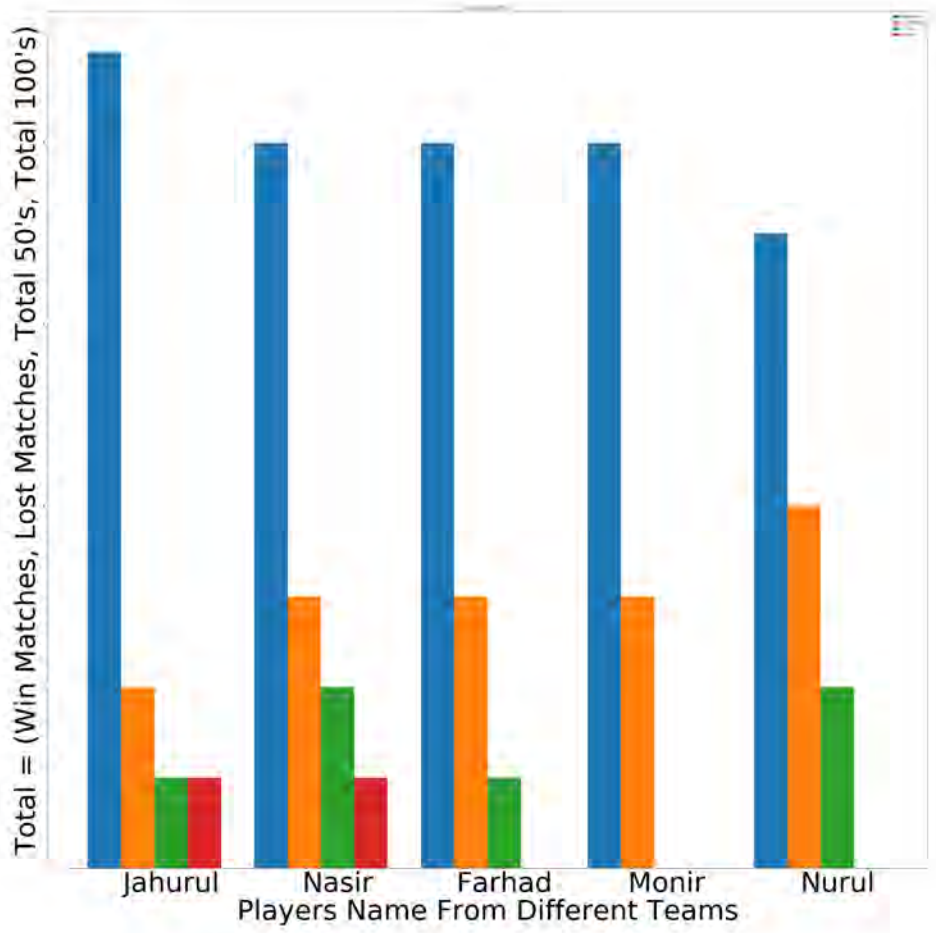

Fig. 5.1 Captain Bar Diagram

As, we can see from the above bar diagram (5.1) that Jahurul Islam's winning percentage is higher than the other captain's so Jahurul Islam is going to be selected as a captain for the best squad.

We didn't measure only the winning 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 "Jahurul Islam"

5.3 Selection Process Of Selected Wicket Keeper

Well, we have found that there were twelve wicket keepers in the tournament. So we picked them up for the process and plotted different parameters in the "BAR DIAGRAM" to show the performance of the wicket keeper.

Used Parameters:

1. T.match = Total Match.
2. T.Stamp = Total Stumping.
3. T.Catch = Total Catch.
4. Run Out = Total Run Outs.

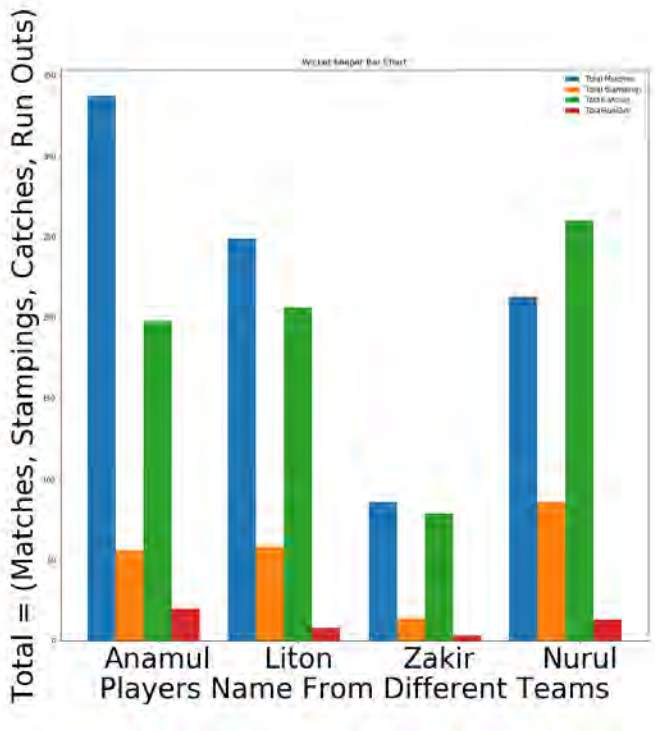

Fig. 5.2 Wicket Keeper Bar Diagram

As, we can see from the above bar diagram (5.2) that Nurul Hasan's performance is higher than the other wicket keeper's so Nurul hasan is going to be selected as a wicket keeper for the best squad.

We didn't measure only one parameter's percentage for the wicket keeper but also we also have measured his performance in other areas only because of examining one area we can't take a player in the team. He also should have create an impact on other parameters also like here anamul haque played highest number of innings as compare to nurul hasan and liton das. When we look at other bars we can see that nurul hasan is the best for building up a perfect team where he can play the role of a wicket keeper.

So. Our Wicket keeper is "Nurul Hasan".

5.4 Selection Process Of Selected Batsmen

5.4.1 Initial Listed Players In Batsmen

We have collected only those players whose playing role is a batsman. In our data file there isn't a single player who is a bowler. The names of initial batsmen are given below :

Initial List :

Table 5.1 Initial Batsmen List

Number	Player Name	Number	Player Name
1	Saif Hassan	31	Imrul Kayes
2	Anamul Haque	32	Jony Talukder
3	Nazmul Hossain Shanto	33	Enamul Haque
4	Nasir Hossain	34	Rony Talukder
5	Mosaddek Hossain	35	Shamshur Rahman
6	Abdul Mazid	36	Raqibul Hasan
7	Mohammad Naim	37	Mehedi Maruf
8	Abhishek Mitra	38	Mehrab Hossain JNR
9	Naeem Islam	39	Zakir Hasan
10	Nazmul Hossain Milon	40	Al-Amin
11	Robiul Islam Robi	41	Nahidul Islam
12	Mahidul Islam Ankon	42	Mizanur Rahman
13	Amit Majumder	43	Junaid Siddique
14	Nazimuddin	44	Myshukur Rahman
15	Rafsan Al Mahmud	45	Alok Kapali
16	Imtiaj Hossain	46	Nazmus Sadat
17	Fazle Mahmud	47	Fardeen Hasan Ony
18	Marshal Ayub	48	Shadman Islam
19	Farhad Hossain	49	Towhid Hridoy
20	Farhad Reza	50	Afif Hossain

Above table is showing the initial listed batsmen from different teams. The other players name has been given below:

Table 5.2 Initial Batsmen List 2

Number	Player Name	Number	Player Name
21	Shykat Ali	51	Minhaz khan
22	Tanbir Hyder	52	Waliul Karim
23	Nurul Hasan	53	Taibur Rahman
24	Ziaur Rahman	54	Mohammad Ashraful
25	Shohag Gaji	55	Munim Shahriar
26	Mahedi Hasan	56	Faruque Hossain
27	Shafiul Hayat	57	Shahriar Nafees
28	Mominul Haque	58	Somya Sarkar
29	Jahurul Islam	59	Salman Hossain
30	Jaker Ali	60	Shamshul Islam

Above are the listed players (batsmen) on which we are going to perform various calculations (machine learning algorithms).

5.4.2 Clustering

Clustering: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology (from Greek "grape") and typological analysis. The subtle differences are often in the use of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest.

Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939[1][2] and famously used by Cattell beginning in 1943[3] for trait theory classification in personality psychology.

Connectivity-based clustering (hierarchical clustering):

Connectivity-based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

Connectivity-based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances), and UPGMA or WPGMA ("Unweighted or Weighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

Centroid-based clustering:

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach

is to search only for approximate solutions. A particularly well known approximate method is Lloyd's algorithm, often just referred to as "k-means algorithm" (although another algorithm introduced this name). It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (k-means++) or allowing a fuzzy cluster assignment (fuzzy c-means).

Most k-means-type algorithms require the number of clusters – k – to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders of clusters (which is not surprising since the algorithm optimizes cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. First, it partitions the data space into a structure known as a Voronoi diagram. Second, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based clustering, and Lloyd's algorithm as a variation of the Expectation-maximization algorithm for this model discussed below.

Clustering batsman:

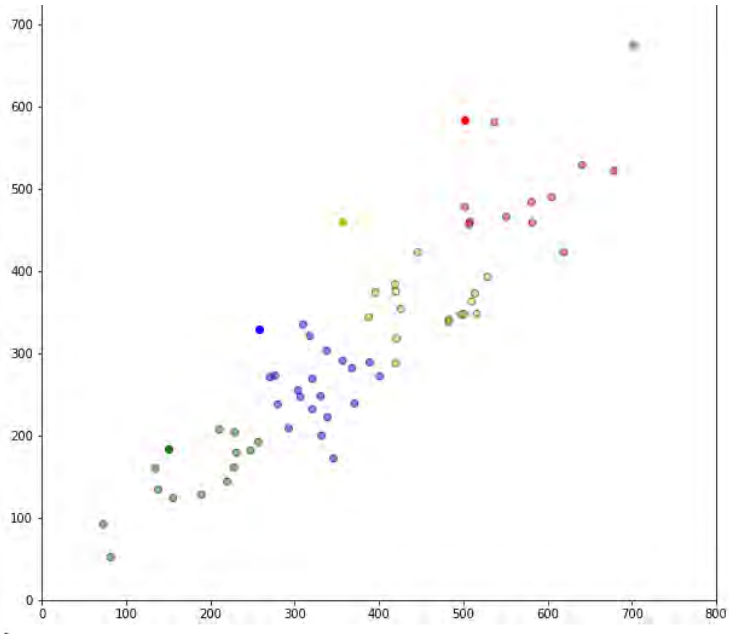

Fig. 5.3 Clustering Of Batsmen

Above diagram (Fig. 5.3) we have plotted "Total Ball Faced" in the direction of X

axis and "Total Runs" in the direction of Y axis. Red coloured and yellow coloured dots representing the best clustered batsmen because they have faced more balls and also scored more runs in this league. So we have picked them up for the further processes.

After Clustering Players In Group A1:

Table 5.3 Clustered A1 Batsmen

Number	Player Name
1.	Nazmul Hossain Shanto.
2.	Mohammad Naim.
3.	Naeem Islam.
4.	Fazle Mahmud.
5.	Shykat Ali.
6.	Al-amin.
7.	Mizanur rahman.
8.	Juniad Siddique.
9.	Shadman Islam.
10.	Towhid Hridoy.
11.	Mohammad Ashraful.
12.	Shahriar Nafees.

In the above table, we can see there are the 12 batsmen in Cluster A1. (Best performance 1)

These twelve batsmen are the best so far we can say that because they have faced too many balls in the tournament of Dhaka Premiere League session 17/18. They also have score the top most runs in this league. As we can see that they are representing red dots in the cluster. It means that they have faced many balls and scored many runs.

This includes that they have more potential to stay on the crease and also can perform well by scoring more runs. Besides the players who are close to the point 0 they didn't face too many balls as a result couldn't score more runs in this league. It indicates that they can't stay on the crease for a longer period of time. So we didn't pick them up.

After Clustering Players In Group A:

In the above table, we can see there are the 12 batsmen in Cluster A. (Second Best performance)

These 16 batsmen are the second best so far we can say that because they have faced too many balls in the tournament of Dhaka Premiere League session 17/18. They also have score

Table 5.4 Clustered A Batsmen

Number	Player Name
1.	Saif Hasan
2.	Anamul Haque
3.	Nasir Hossain
4.	Robiul Islam robi
5.	Mahidul Islam Ankon
6.	Amit Majumder
7.	Rafsan Al Mahmud
8.	Marshall Ayub
9.	Farhad Hossain
10.	Tanbir Hayder
11.	Shamsur Rahman
12.	Raqibul Hasan
13.	Myshukur Rahman
14.	Waliul Karim
15.	Taibur Rahman
16.	Fardeen Hasan Ony

the second top most runs in this league. As we can see that they are representing yellow dots in the cluster. It means that they have faced many balls and scored many runs.

This includes that they have more potential to stay on the crease and also can perform well by scoring more runs. Besides the players who are close to the point 0 they didn't face too many balls as a result couldn't score more runs in this league. It indicates that they can't stay on the crease for a longer period of time. So we didn't pick them up.

So far we have got two clusters of batsmen. As we know that, clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). So, the players in cluster A1 are not similar to the cluster A. They are different in many ways. To find out that difference we have to do some tests. Those are given below, actually now comes the algorithms parts.

5.4.3 Linear Regression Result

Player Listed In (A1)

Table 5.5 Accuracy and Predicted Runs By Linear Regression (A1)

Number	Player Name	Accuracy	Predicted Runs
1.	Shadman Islam.	.95	121.49
2.	Towhid Hridoy.	.96	114.37
3.	Nazmul Islam Shanto.	.98	104.68
4.	Shahriar Nafees.	.96	100
5.	Mizanur Rahman.	.90	97.8
6.	Naeem Islam.	.67	131.07
7.	Sykat Ali.	.87	100
8.	Al-Amin.	.97	82.76
9.	Mohammad Ashraful.	.94	83.21
10.	Fazle Mahmud.	.72	108.82
11.	Junaid Siddique.	.86	83.97
12.	Mohammad Naim.	.79	86.75

In the above table we can see that the accuracy (R2 factor) of the algorithm on each players data set and their predicted runs. On based on this predicted runs we can easily ranked them as linear regression reads the behavior of each batsman's scoring level then draw a base line then gives the out put (Predicted run).

Player Listed In (A)

In the table given below, we can see that the accuracy (R2 factor) of the algorithm on each players data set and their predicted runs. On based on this predicted runs we can easily ranked them as linear regression reads the behavior of each batsman's scoring level then draw a base line then gives the out put (Predicted run).

Table 5.6 Accuracy and Predicted Runs By Linear Regression (A)

Number	Player Name	Accuracy	Predicted Runs
1.	Marshal Ayub.	.95	111.12
2.	Farhad.	.84	122.27
3.	Raqibul Islam.	.882	114.67
4.	Shamsur.	.914	105.08
5.	Anamul.	.94	99.61
6.	Ankon.	.96	96.01
7.	Robi.	.979	91.75
8.	Saif.	.88	92.52
9.	Rafsan.	.633	128.83
10.	Amit.	.95	83.62
11.	Myshukur.	.91	85.82
12.	Fardeen.	.92	79.83
13.	Waliul.	.86	84.04
14.	Tanbir	.83	82.18
15.	Nasir.	.713	90.67
16.	Taibur.	.919	69.017

5.4.4 Nive Bayes Result

As we know, Bayes Theorem:

Let X be a data tuple described by measurements made on a set of n attributes. Let H be a hypothesis such that X belongs to a specified class C. Bayesian classifiers calculate P(H|X), the probability with which the hypothesis H holds true for the observed attribute values of the data tuple X. P(H|X) is called the posterior probability or posteriori probability of H conditioned on X. Similarly, P(X|H) is the posterior probability or posteriori probability of X given H i.e. the probability with which the data tuple X exists, given the hypothesis H is true. P(H) is the prior probability or a priori probability of H which means that H holds true for a data tuple regardless of the values of its attributes. P(X) is the prior probability or a priori probability of X, which is the probability with which the data tuple X with given attribute values exists. Now, Bayes Theorem is defined as,

$$P(H/X) = \frac{P(X/H).P(H))}{P(X)} \tag{5.5}$$

So we can use this on A1 clustered batsmen.

Player Listed In (A1)

Table 5.7 Predicted Runs By Naive Bayes (A1)

Number	Player Name	Predicted Runs
1.	Shadman Islam.	144
2.	Towhid Hridoy.	83
3.	Nazmul Islam Shanto.	150
4.	Shahriar Nafees.	121
5.	Mizanur Rahman.	102
6.	Naeem Islam.	76
7.	Sykat Ali.	55
8.	Al-Amin.	94
9.	Mohammad Ashraful.	102
10.	Fazle Mahmud.	65
11.	Junaid Siddique.	83
12.	Mohammad Naim.	88

Player Listed In (A)

As we know, Bayes Theorem:

Let X be a data tuple described by measurements made on a set of n attributes. Let H be a hypothesis such that X belongs to a specified class C. Bayesian classifiers calculate P(H|X), the probability with which the hypothesis H holds true for the observed attribute values of the data tuple X. P(H|X) is called the posterior probability or posteriori probability of H conditioned on X. Similarly, P(X|H) is the posterior probability or posteriori probability of X given H i.e. the probability with which the data tuple X exists, given the hypothesis H is true. P(H) is the prior probability or a priori probability of H which means that H holds true for a data tuple regardless of the values of its attributes. P(X) is the prior probability or a priori probability of X, which is the probability with which the data tuple X with given attribute values exists. Now, Bayes Theorem is defined as,

$$P(H/X) = \frac{P(X/H).P(H))}{P(X)} \tag{5.6}$$

Table 5.8 Predicted Runs By Naive Bayes (A)

Number	Player Name	Predicted Runs
1.	Marshal Ayub.	28
2.	Farhad.	63
3.	Raqibul Islam.	28
4.	Shamsur.	38
5.	Anamul.	57
6.	Ankon.	21
7.	Robi.	8
8.	Saif.	94
9.	Rafsan.	46
10.	Amit.	71
11.	Myshukur.	82
12.	Fardeen.	104
13.	Waliul.	79
14.	Tanbir	61
15.	Nasir.	129
16.	Taibur.	23

5.5 Selection Process Of Selected All Rounders

5.5.1 Initial Listed Players In All Rounders

We have collected only those players whose playing role is a all rounder. In our data file there isn't a single player who is a bowler(who only bowls). The names of initial all rounders are given below :

Initial List :

Table 5.9 Initial All Rounders

Number	Player Name
1.	Nasir Hossain
2.	Mehedy Hasan Miraj
3.	Naeem Islam
4.	Mosarraf Hossain
5.	Fazle Mahmud
6.	Farhad Reza
7.	Nurul Hasan
8.	Ziaur Rahman
9.	Afif Hossain
10.	Shuvagata Hom
11.	Mohammad Saifuddin
12.	Jaimul Hasan

5.5.2 Clustering

Clustering: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Centroid-based clustering:

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

Most k-means-type algorithms require the number of clusters – k – to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders of clusters (which is not surprising since the algorithm optimizes cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. First, it partitions the data space into a structure known as a Voronoi diagram. Second, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based clustering, and Lloyd's algorithm as a variation of the Expectation-maximization algorithm for this model discussed below.

Clustering Diagram Of All Rounders:

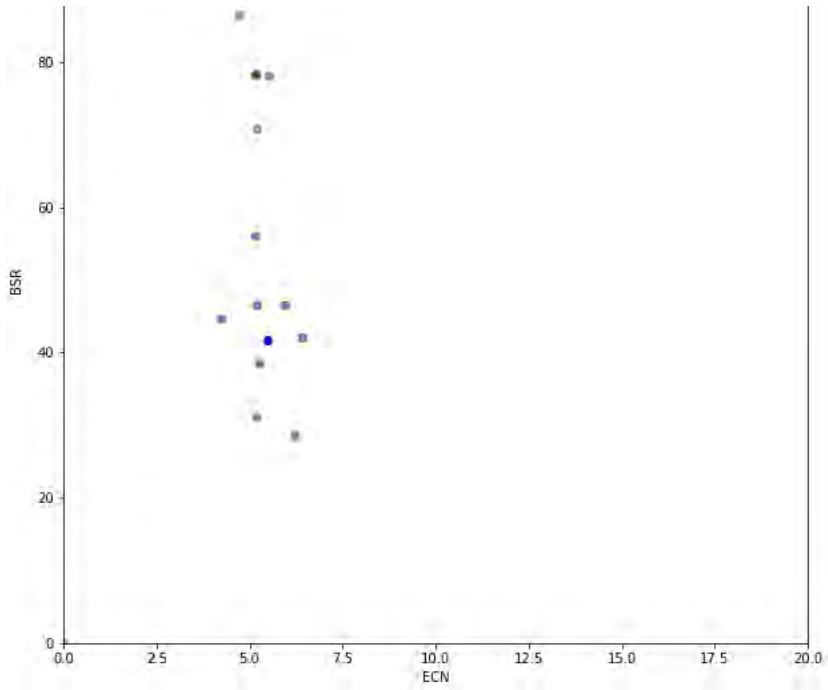

Fig. 5.4 Cluster of all rounder

After Clustering All Rounders List A1

As we do have too many all rounders so we need to cluster first to identify the best all rounders. And we already know that cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many

fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Table 5.10 After Clustering All Rounder List

Number	Player Name
1.	Nasir Hossain.
2.	Mahedi Hasan Miraj.
3.	Naeem Islam.
4.	Fazle mahmud.
5.	Farhad Reza.
6.	Shubhagata Hom.
7.	Mohammad Shaifuddin.
8.	Jaimul Islam.

5.5.3 Linear Regression Result

Table 5.11 Accuracy and Predicted Runs By Linear Regression (On Batting performance)

Number	Player Name	Accuracy	Predicted Runs
1.	Nasir Hossain.	.95	121.49
2.	Mahedi Hasan Miraj.	.96	102.37
3.	Naeem Islam.	.98	125.68
4.	Fazle mahmud.	.96	100
5.	Farhad Reza.	.90	78.8
6.	Shubhagata Hom.	.67	120.07
7.	Mohammad Shaifuddin.	.87	100
8.	Jaimul Islam.	.97	70.76

Table 5.12 Accuracy and Predicted Wickets By Linear Regression (On Bowling performance)

Number	Player Name	Accuracy	Wickets
1.	Nasir Hossain.	.95	5
2.	Mahedi Hasan Miraj.	.96	6
3.	Naeem Islam.	.98	3
4.	Fazle mahmud.	.96	10
5.	Farhad Reza.	.90	8
6.	Shubhagata Hom.	.67	1
7.	Mohammad Shaifuddin.	.87	16
8.	Jaimul Islam.	.97	2

5.6 Selection Process Of Selected Bowlers

5.6.1 Initial Listed Players In Bowlers

We have collected only those players whose playing role is a bowler. In our data file there isn't a single player who is a batsman(who only bats). We also have taken this parameter while taking the names of the bowlers : 1. Data Set collected from WWW.ESPNCRICINFO.COM, 2. This data set collected from the "Dhaka Premere League" matches manually, 3. Total Team = 9, 4. Team Names : a. AL = Abahani Limited, b. LR = Legends of Rupgang, c. KSKS = Khelaghar Samaj Kallyan Samity, d. PDSC = Prime Doleshwar Sporting Club, e. SJDC = Sheikh Jalal Dhanmondi Club, f. GGC = Gazi Group Cricketers, g. MSC = Mohammedan Sporting Club, h. PBCC = Prime Bank Sporting Club, i. SSC = Shinepukur Cricket Club, 5. Total Batsman = 61 (person) 6. Attributes : "x is defining the number of match. In dataset you will see M1, M2, M3..... There M1 stands for Match no 1 and so on....." W/L=Win or loose the match, OP = Opponent, VNU = Venue of Particuler Match, TO = Total Over, RG = Total Run Given, WT = Total Wicket Take, Avg = Average, Sr = Strick Rate, Eco= Economy, 5W = Five Wicket Take, 3W = Three Wicket Take.

Initial List:

Table 5.13 Initial Bowler List

Number	Player Name	Number	Player Name
1	Sanjamul Islam	19	Nazmul Islam
2	Taskin Ahmed	20	Shohag Gaji
3	Mehedi Hasan Miraj	21	Ellias Sunny
4	Sandip Roy	22	Robiul Haque
5	Mohammad Shahid	23	Abu Haider
6	Asif Hasan	24	Mahedi Islam
7	Mosarraf Hossain	25	Nayeem Islam
8	Syed Rasel	26	Tipu Sultan
9	Tanvir Islam	27	Qazi Anik
10	Mohammad Shadman	28	Taijul Islam
11	Abdul Halim	29	Enamul Haque
12	Masum Khan	30	Mohammad Azim
13	Hasan Mahmud	31	Mohammad Saifuddin
14	Arafat Sunny	32	Subhagata Hom
15	Sarifullah	33	Naeem Islam Jnr
16	Mamun Hossain	34	Raihan Uddin
17	Salauddin Shakil	35	Nahidul Islam
18	Nazmul Islam	36	Shoriful Islam

5.7 Overall Selection Process Of Players

Our overall process in one diagram is given below:

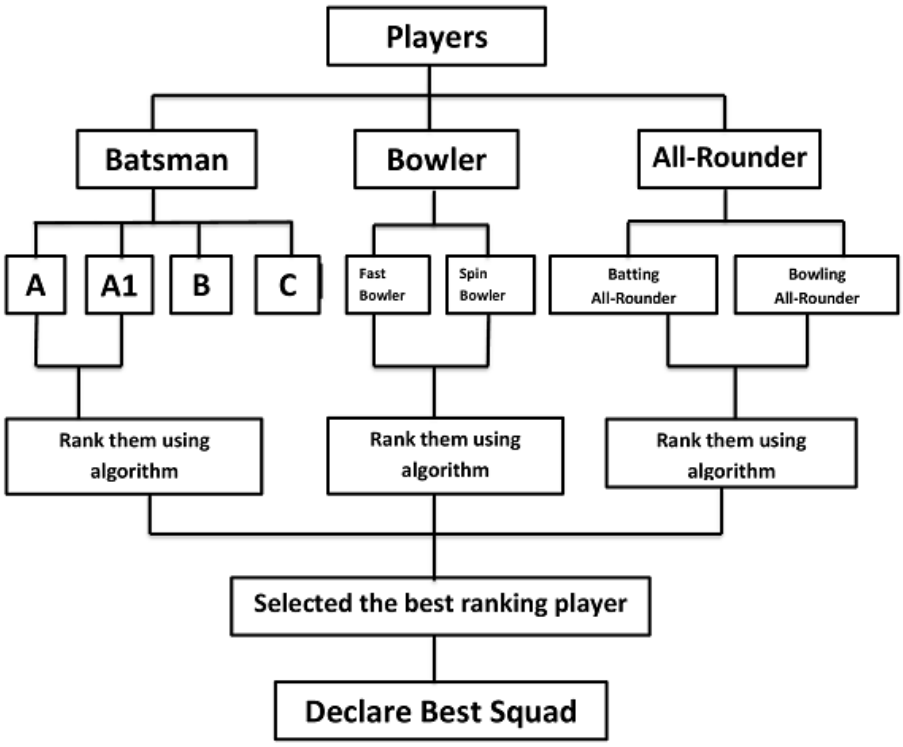

Fig. 5.5 Overall Selection Diagram

Above diagram is showing the full model/ process that we have followed or build up to search for the best batsmen, bowlers, all-rounders, captain and a wicket-keeper.

At the very beginning we have classified them into three sections - Batsman, bowler, all-rounder. Then batsmen into 4 groups, bowlers into 2 groups, All-rounders into 2 groups. Gradually we have reached at the final destination which is "Making A Perfect Squad For A Team"

5.8 Algorithms Result

Algorithms Result:

For regression of bowler, Logistic regression had the highest accuracy of 66.666 percent , sensitivity of 58.33 percent and specificity of 75.0 percent. Here the regression accuracy increases as we increase the size of the training sample and decrease the test sample.

Table 5.14 Performance of logistic regression on bowlers data set

Accuracy	Sensitivity	Specificity
66.66666	58.3333	75.0

In Support Vector Machine had the highest accuracy of 70.666 percent , sensitivity of 66.33 percent and specificity of 75.0 percent. Here the regression accuracy increases as we increase the size of the training sample and decrease the test sample.

Table 5.15 Performance of SVM on bowlers data set

Accuracy	Sensitivity	Specificity
70.66666	66.3333	75.0

Decision tree showed accuracy of 58.666 percent , sensitivity of 50.00 percent and specificity of 66.6 percent. Here the regression accuracy increases as we increase the size of the training sample and decrease the test sample.

Table 5.16 Performance of decision tree on bowlers data set

Accuracy	Sensitivity	Specificity
58.66666	50.0	66.6

From the above discussion, we can say that SVM algorithm has the best result for the prediction. So we have taken the output of SVM.

Table 5.17 Recommended 15 Men Squad

No	Name	Designation
1	Jahurul Islam	Captain
2	Nurul Hasan	Wicket-Keeper
3	Nazmul Hossain	Batsman-1
4	Mizanur Rahman	Batsman-2
5	Shahriar Nafees	Batsman-3
6	Marshal Ayub	Batsman-4
7	Raqibul Hasan	Batsman-5
8	Farhad Hossain	Batsman-6
9	Rabiul Haque	Bowler-1
10	Abu Haider	Bowler-2
11	Nahidul Islam	Bowler-3
12	Nayeem Hasan	Bowler-4
13	Mohammad Azim	Bowler-5
14	Akash Majumder	Batting All-Rounder
15	Farhad Reza	Bowling All-Rounder

Chapter 6

Conclusion and Future Work

6.1 Conclusion and Future work

Selecting the right players for a team plays a significant role in a team's victory in a particular match. An accurate prediction of how many runs a batsman is likely to score , how many wickets a bowler is likely to take and how efficient an all-rounder will be in the upcoming match will help the team management to select. In this paper, we have collected the data of Dhaka Premiere League 2017/2018 session from ESPN. 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 k-means clustering, Linear Regression, Naive Bayes and PageRank algorithms for selecting batsmen and all-rounders. Support Vector Machine, NaiveBayes, 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), 6 batsmen(top-order, middle-order and finisher), 2 all-rounders, 1 wicket keeper and 5 bowlers (fast bowlers, spinners). 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. We are working on a project called "quickCric" that will help organizers to arrange Cricket tournaments in all formats. All types of domestic and other sponsored tournaments can be arranged with live scoring and data storage facilities. The team managers can track the performance of each player with the help of our model and can easily select the 15 man squad for the next available tournament.

References

[1] 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.

[2] 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.

[3] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[4] Bukiet, B. and Ovens, M. (2006). A mathematical modelling approach to one-day cricket batting orders. *Journal of sports science & medicine*, 5(4):495.

[5] 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.

[6] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[7] 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.

[8] 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)*.

[9] Lemmer, H. H. (2002). The combined bowling rate as a measure of bowling performance in cricket. *South African Journal for Research in Sport, Physical Education and Recreation*, 24(2):37–44.

[10] 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.

[11] 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).

[12] Omkar, S. and Verma, R. (2003). Cricket team selection using genetic algorithm. In *International congress on sports dynamics (ICSD2003)*, pages 1–3. Citeseer.

[13] Parker, D., Burns, P., and Natarajan, H. (2008). Player valuations in the indian premier league. *Frontier Economics*, 116:1–17.

[14] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[15] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[16] Sankaranarayanan, V. V., Sattar, J., and Lakshmanan, L. V. (2014). Auto-play: A data mining approach to odi cricket simulation and prediction. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 1064–1072. SIAM.

[17] Shah, P. (2017). New performance measure in cricket. *ISOR Journal of Sports and Physical Education*, 4(3):28–30.

[18] Wickramasinghe, I. P. (2014). Predicting the performance of batsmen in test cricket.