Cricket Data Analytics and Prediction

A Project Report of Capstone Project - 2

Submitted by

SAURABH KUMAR

(1613101647/16SCSE101159)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

DR. SANSAR SINGH CHAUHAN (PROFESSOR)

APRIL / MAY-2020

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"CRICKET DATA ANALYTICS AND PREDICTION"** is the bonafide work of **"SAURABH KUMAR(1613101647)"** who carried out the project work under my supervision.

SIGNATURE OF HEAD	**SIGNATURE OF SUPERVISOR**
Dr. MUNISH SHABARWAL,	Dr. SANSAR SINGH CHAUHAN,
PhD (Management), PhD (CS)	PhD(CS), **Professor**
Professor & Dean,	**School of Computing Science &**
School of Computing Science &	**Engineering**
Engineering	

Abstract

Cricket is the one of most watched sport now a days. Cricket winning depends on the various Circumstances like home crowd advantages, performances in past, experience in match, performance at specific venue, Performance against the specific team and the current form of the team and the player. In this work a model has been proposed that has two methods, first predicts the score of first inning not only on the basis of current run rate but also considers number of wicket fallen, venue of match and batting team.

The second method predicts the outcome of the match in the second inning considering the same attributes as of former method along with the target given to the batting team. The methods have been implemented using Linear regression Classifiers or Q Learning base decision tree approach or Naïve Bayes classifiers for first inning and second inning respectively. In the above both methods, the five overs intervals have been made from 50 overs of the match and at each interval. The above attributes have been recorded for all nonCurtailed matches played between 2002 and 2014 of every team independently.

Table of Contents

LIST OF FIGURES

CHAPTER 1

INTRODUCTION TO PROJECT

1.1   OVERVIEW

Cricket is being played in many countries all around the world. There are lots of domestic and international tournaments being held in many countries which play cricket. Cricket is a game played between two team comprising of 11 players in each team. The result is either a win, loss or tie. However, sometime due to bad weather conditions the game is also washed out as cricket is a game which cannot be played in a rain. Moreover, the game is also extremely unpredictable because at every stage of the game the momentum shifts to one of the team between the two. A lot of times the results gets decided on the last ball of the match where the game gets really close. Considering all these unpredictable scenarios of this unpredictable game, there is a huge interest among the spectators to do some predictions either at the start of the game or during the game. Many spectators also play betting games to win money. So, keeping in minds all these possibilities, this report aims at studying the problem of predicting the game results before the game has started based on the statistics and data available from the data set. These are the different way to do prediction. The prediction can be done taking into consideration the players performance as well as team performance. These are many unpredictable things that happen in a cricket game like matches being washed out due to rain, a key plyer getting injured before the game, players changing their teams, etc. sometimes a key player also gets injured during the game and is not able to take further part in the game. All these factors do affect the prediction to some extent. The report discusses a methodology that I followed for the game result prediction. The methodology consists of first the attribute selection algorithms which trim down the list of attributes to only important ones and the data mining algorithms which can be applied on those attributes. The game prediction problem that I studying does not take into consideration the

players performance but it does not take into considerations the team past performance at high level extent along with the other factors like toss winner, toss decision, home support etc.

1.2 FACTORS

Cricket winning can be predicted like all other games. We need to find the best attributes or factors that influences the match outcome. The result of a cricket match depends upon more of in-game and more of pre-game attributes. Pre-game attributes like pitch, Team strength, weather, venue etc. and in-game attributes like run rate, total run, strike rate, wickets in hand etc. influences a match result predominantly. Below are the attributes that decides outcome of the cricket match

1.2.1 Pitch:

Unlike other sports, cricket stadiums shape and size is not fixed except the dimensions of the inner circle and pitch which are 30 yards and 22 yards respectively. Outfield variations and pitch can have a substantiate effect on bowling and batting. The spin of the ball, seam movement and the bounce depend upon the nature of the pitch. It depends on the how the wet is the pitch. The more wet the pitch, the slower it will play. On the off chance that it is drying out, those balls will change significantly, yet all it will get less difficult those drier it gets.

Green pitches tend to get easier to bat on. Wickets can get significantly more dry or wet. They might start to break up if they are soft.

1.2.2 TOSS:

According to cricket analysts, there is sure measure of advantage for a team if it wins the toss. This might not be deciding factor in a match but it would give the team the opportunity of choosing "what they want".

1.2.3 TEAM STRENGTH:

The team strength should be balanced for winning a match. Captainship in a team is also a deciding factor. Past records: The past team performances can be considered to predict the outcomes of match. History of games at that venues how did the team performs, performances at that specific venue, performance against the specific opposition and experience at the specific venue.

1.2.4 HOME GROUND ADVANTAGE:

This is another attribute which determines the winner in the match. If you are playing in the home ground condition everything would be in your hands like climatic factor, pitch nature and major role is played by the home crowd. Home team gets better motivation.

1.3PROBLEM STATEMENT:

In this to design a system that can be provide the score and winning prediction in cricket match, the system can analyze multiple parameters like winning toss, batting side, DL approach, home ground advantage, player wise performance etc. while declaring a time for particular championship it is very important to select the best team so that the chances of the team winning the championship become easy. This problem had to be solved to generate the best players from both the team for the best battle. To solve this Problem we have collected historical data of all some team like (India, Australia, New Zealand, South Africa etc.), and using prediction algorithm like Naïve Bayes algorithm we are predicting the best starting players for both the team can be used in fantasy league for winning the maximum points.

1.4 RELATED WORK:

There has been a lot of related study to this problem in various different sports. The paper I have used for references are all related work that had been done on

this problem. The paper by Trawinski described the prediction of results using a fuzzy classification system. This paper was predicting the results for basketball games. I had used the attribute selection technique mentioned in this paper for my project. The attribute selection technique proposed in this paper was done using WEKA so it was good reference point for me too. The wrapper method algorithm and the ranker method algorithm implemented in this paper was also used in my project. But the prediction part was done by using the fuzzy classification system and I did not use that system for my prediction part.

The paper by Haghighat also described the prediction of results in sports using the data mining techniques. But this paper was not specific to any particular sport rather it was in for all general sports. The attribute selection algorithm that it used was more of an elimination approach where the attributes were eliminated one by one and the classification accuracy is computed. Once a good subset of attributes is achieved, then the eliminated attributes were again added one by one to see the accuracy improves, But, in my approach I did not used this elimination approach for the attribute selection part. Then the paper used various data mining algorithms to perform classification. I used the naïve Bayes and the decision tree algorithms from the paper in my project and compared my accuracy with that of the paper.

The paper by Zdravevski takes into consideration some of the external factors like the number of players injured before the particular game, the winning streak of a team before a particular game, the fatigue factors of the player etc. The approach I used in the project does not take into account the external features like player injury, fatigue or the winning streak. My data set contains more data about the matches and the events happening in the match like toss and player of match rather than the data about external factors.

CHAPTER 2

2.1 Methodology

I have followed the following methodology in the course of my project. The methodology consists of 5 different phases as shown in fig. 1i.e. Data set generation, Data cleaning, Attribute selection Data mining and analysis of results. Each of this phase was part of my project milestones submitted and I had created a gantt chart to keep track of timeline. I will be discussing each phase in the following section in detail.
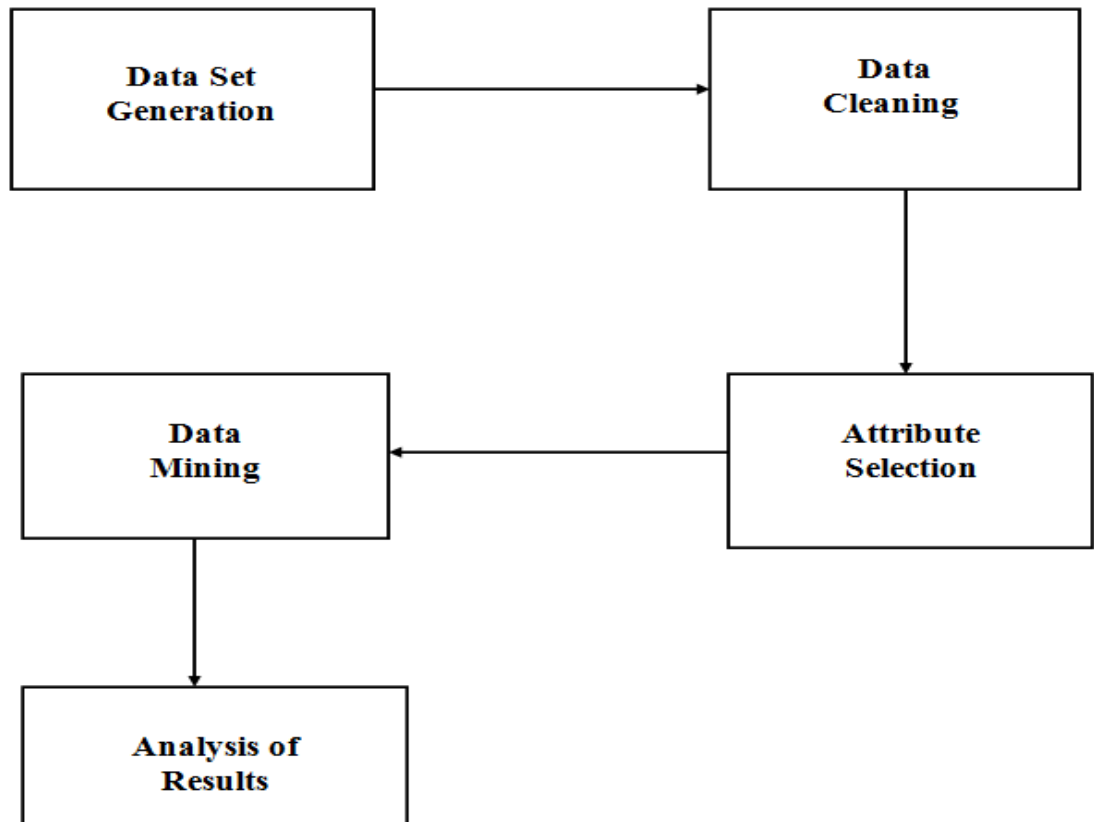

Fig:- Methodology diagram

2.1.1 Data Set Generation:

The data was collected from the http://cricsheet.org website. The website has data about all matches (from 2008 to 2016) of international tournaments held. It is a ODI format of tournaments. It means that each team bats or bowls for maximum 50 overs each. The data set downloaded from this website was in the .CSV format. I used the java classes file and the file writer to read the .CSV file and write the content to new file. The .CSV file had data about 5219 different entries and the java code basically copied the important data from all the 5219 files had then combined multiple files into a single file. The data from the .CSV file contains more irrelevant information like gender, date, umpire information etc. which were all discarded and the java code copied only data about the team, venue, toss, toss _winner, player _of_ match and thee ball by ball data. The .CSV file had data shown in the figure.
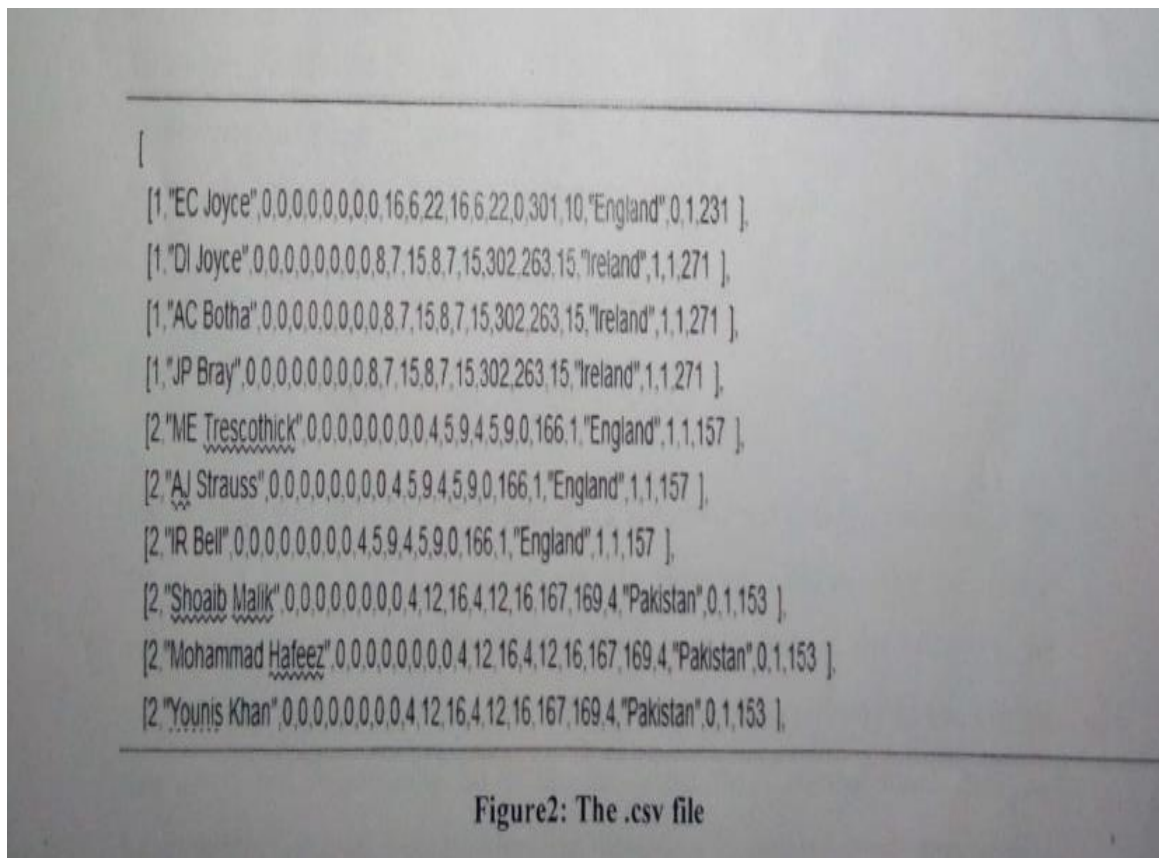
```
[
[1,"EC Joyce",0,0,0,0,0,0,0,0,16,6,22,16,6,22,0,301,10,"England",0,1,231 ],
[1,"DI Joyce",0,0,0,0,0,0,0,0,8,7,15,8,7,15,302,263,15,"Ireland",1,1,271 ],
[1,"AC Botha",0,0,0,0,0,0,0,0,8,7,15,8,7,15,302,263,15,"Ireland",1,1,271 ],
[1,"JP Bray",0,0,0,0,0,0,0,0,8,7,15,8,7,15,302,263,15,"Ireland",1,1,271 ],
[2,"ME Trescothick",0,0,0,0,0,0,0,0,4,5,9,4,5,9,0,166,1,"England",1,1,157 ],
[2,"AJ Strauss",0,0,0,0,0,0,0,0,4,5,9,4,5,9,0,166,1,"England",1,1,157 ],
[2,"IR Bell",0,0,0,0,0,0,0,0,4,5,9,4,5,9,0,166,1,"England",1,1,157 ],
[2,"Shoaib Malik",0,0,0,0,0,0,0,0,4,12,16,4,12,16,167,169,4,"Pakistan",0,1,153 ],
[2,"Mohammad Hafeez",0,0,0,0,0,0,0,0,4,12,16,4,12,16,167,169,4,"Pakistan",0,1,153 ],
[2,"Younis Khan",0,0,0,0,0,0,0,0,4,12,16,4,12,16,167,169,4,"Pakistan",0,1,153 ],
```

Figure2: The .csv file

In Computing, Commas separated values (CSV) file is a delimited text file that uses a comma to separate values. A CSV files stores tabular data (numbers and text) in a plaintext. Each line of the file is a data record. Each record consists of one or more fields separated by commas. The use of commas as a field separator is the source of the name for this file format.

2.1.2 Data Set Description:

The new combined data set that I generated had the data from the .CSV files. The data set consist of 5 different attributes and 5219 different instances. The data set spans across all the seasons of the one day international from 2008 to 2016 season. The winning team is the classifier in the data set and this project is to predict the winning team in the match. The attributes in the data set are as follows:

1). **Venue**: Playing venue.

2). **Team1**: The Playing team 1.

3). **Team 2**: The Playing team 2.

4). **Player**: Add player 1st team.

5). **Player**: Add player 2nd team.

6). **Winning team**: This is the class attribute i.e. winning team.

2.1.3 Data Cleaning:

The data obtained from the http://cricsheet.org[1] website was already cleaned. So, I did not have to do any sort of cleaning on the data. , I had to tackle the missing values data and the data for the matches which were washed out due to rain. Those matches data were present in the .json file. after combining the two data into a single data set, I had to manually fill in the data about the 7 missing teams from the http://espncricinfo.com/ website. Moreover there were 10 matches which were washed out so those instance were filled with null values and they were discarded from the data set. So, the final dataset had 5219 instances with 21 attributes.

2.1.4 Attribute Selection

The data set obtained after handling the missing values had to be filtered with the help of the attribute selection algorithm. Since, there were 21 attributes it was necessary to identify all important attributes which would be useful for the data mining tasks. The paper by Haghighit [3] explained the attribute elimination process. Here, first a data set with some number of attributes is selected and then each attribute is eliminated one by one from the set of attributes. The elimination is based upon the result of running the classification algorithm on the set of attributes. An attribute is completely eliminated if the accuracy improves after its removal or else the attribute is kept in the data set. So, at the end of this elimination process we get a set of attributes using which we get the highest accuracy of prediction for the classification algorithms. The paper by Trawinski [2] describes the two types of selection algorithms i.e. the wrapper method and the ranker method that were used by the authors in their study. I used this paper as my reference for the attribute selection phase. I wanted to evaluate the worth of each attribute and rank each attribute to identify their importance before using them for making the prediction during the data mining phase. So I decided to choose these two types of attributes selection algorithms from the papers by Trawinski [2]. These two types of attribute selection methods are available to use in WEKA. After the attribute selection phase, The list of attributes gets minimized to 10 i.e. Team 1, Team 2, Venue, Home _ Team, Toss _ winner, Toss _ decision, Player _ of _ match, Team _ batting _first, Team _batting _ second, and winning _ team. Here the winning _ team is the classifier.

CHAPTER 3

3.1 Class Diagram

The class diagram for the proposed system describes the system in term of classes attributes, operations, and their associations. In UML, classes and objects are shown by boxes composed of three compartments. Top compartment displays the name of the class or object. The center compartment displays its attributes; the bottom compartment displays its operations.
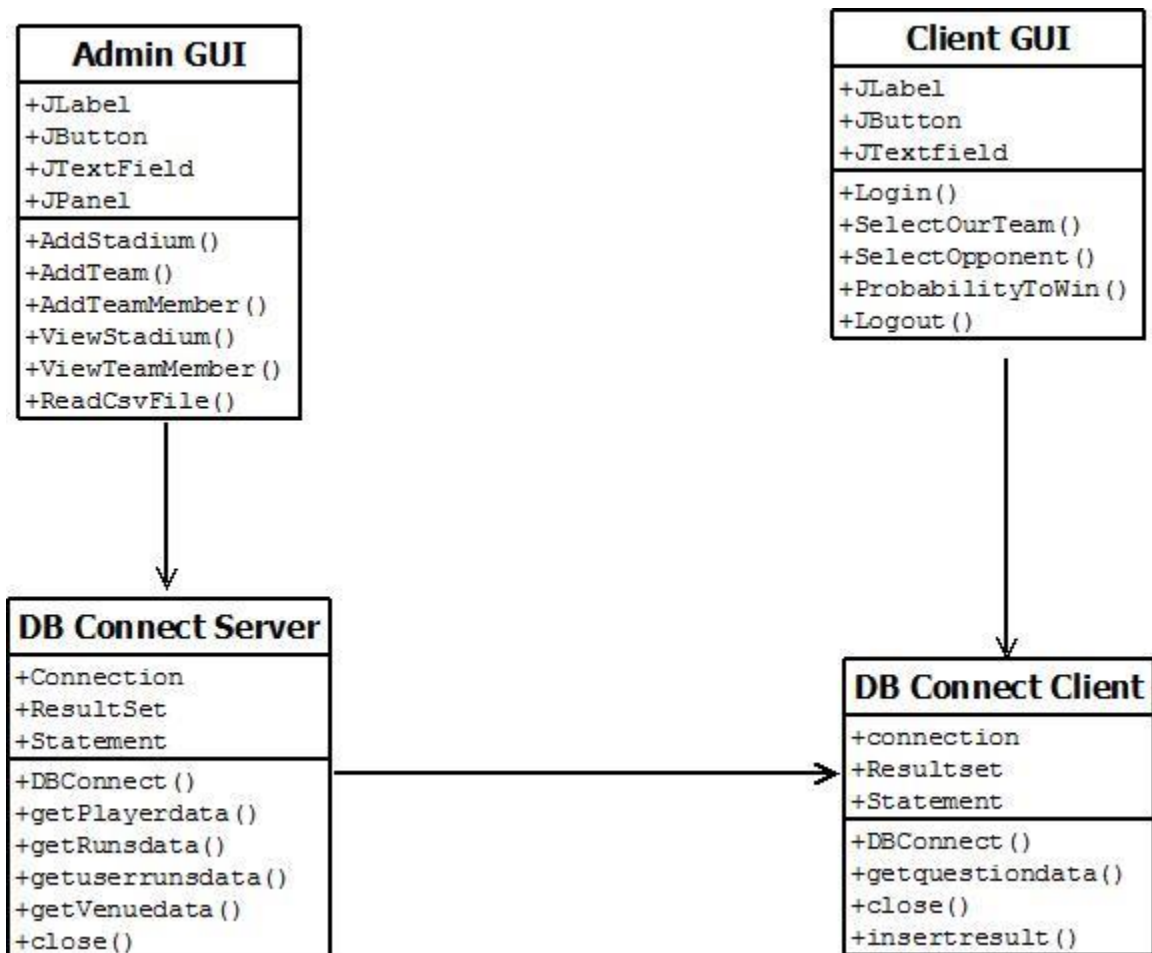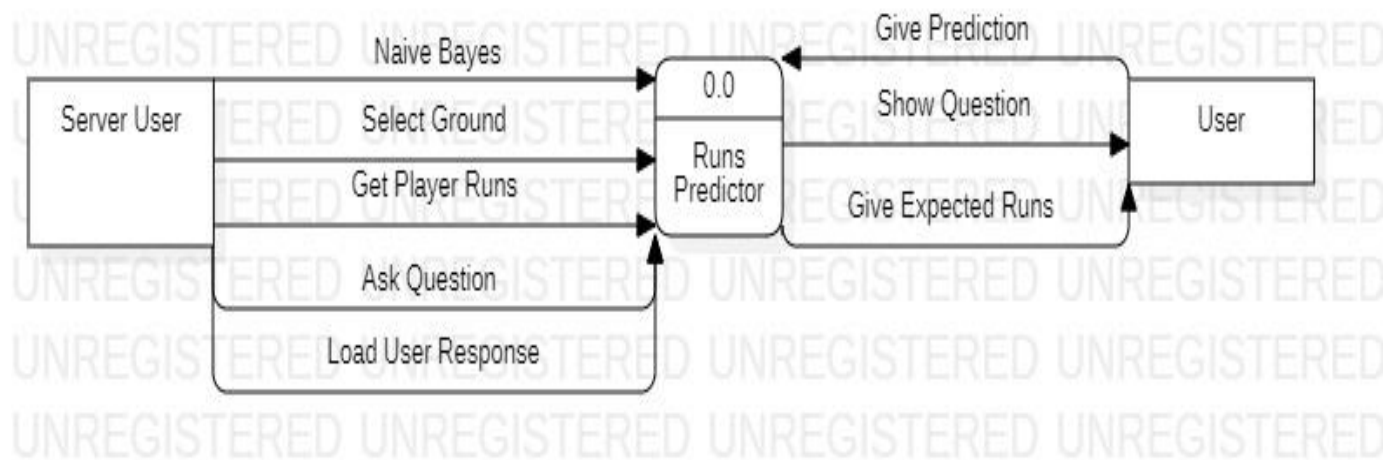
3.2 Data Flow Diagram

A data flow diagram is a graphical representation of the flow of data of CSPS. It helps us to understand the flow of system. A data flow diagram is a graphical representation of the "flow" of data through an information system, Modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system without going into a great detail, which can later be elaborated. Data flow diagram is one of the three essential perspectives of the structured system analysis and design method SSADM.

3.2.1 DFD Level 0

DFD level 0 is also called a context diagram. It's a basic overview of the whole system or processes being analyzed or modeled. It's designed to be an at-a-glance view showing the system as a single high-level process with its relationship to external entities. It should be easily understood by a wide audience, including Stakeholders, business analyst, data analyst and developers.

From the above data flow diagram we get a detailed understanding of how the flow of each function will take place. DFD level 0 gives the detailed of CSPS.

3.2.2 DFD Level 1

The DFD level 0 is broken down into more specific, Level 1 DFD. Level 1 DFD depicts basic modules in the system and flow of data among various modules. Level 1 DFD also mentions basics process and source of information.

- It provides a detailed view of context level diagram.
- Here, the main function carried out by the system are highlighted as we break into sub-processes.

Whereas DFD level 1 of the different functions give a more in depths knowledge of the methods and database usage. The DFD gives us a good visualization of the steps involved in functions of the user and the database that is the connected to the various methods.
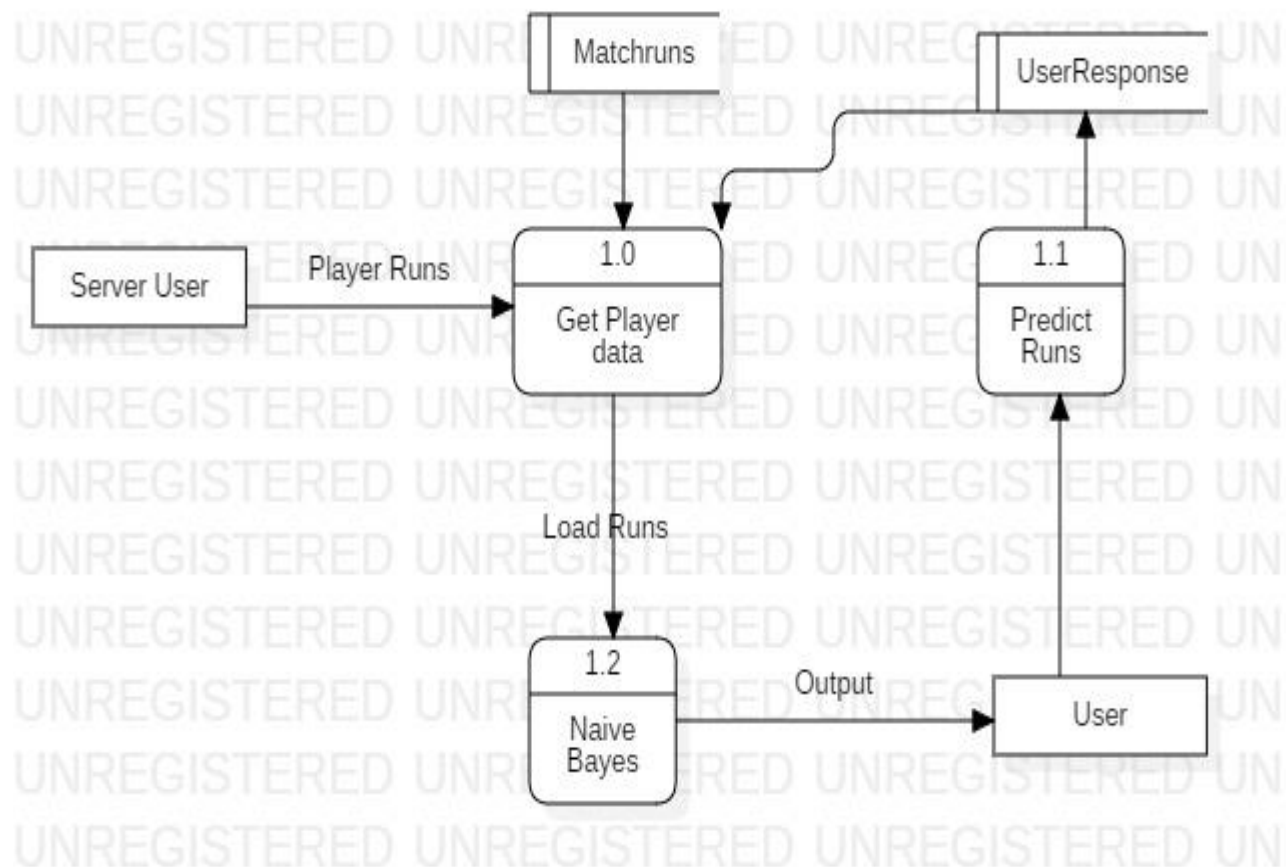
CHAPTER 3

3.1 Motivation:

There are several areas in our day to day life when we need such analysis of data and some of them are listed below that arises the urgent need for such software.

- For instances the are users who bet on the application DREAM 11 and there this software can be useful as it will give information about every all players so that the individual can choose his/her team efficiently and it will be productive.
- The commentators or analyst can also use it to demonstrate the status and the records of players on a specific field. It is also useful for some youtubers to analyze the play of some players in their videos.
- It can be also used on a small scale like inter school tournaments, District level tournaments to analyze the games of various teams and so to improve the game

3.2 Scope:

- It is basically used for the analysis that at which the particular ground the winning percentage is higher of particular team.
- It can be used in predicting the teams on the application like DREAM 11 where people choose their teams and compete with each other to win the prizes. It will help the person to form the team of such players that are compatible with the playing conditions.
- It can be used to predict the overall winning chances of team in a particular match on particular ground.
- It can be used by team management to analyze and made such team that they can give higher percentage of winning.

3.3 Proposed Model:

USERS

LOGIN

DATASET

SYSTEM A

SYSTEM B

NAÏVE BAYES CLASSIFIER

PREDICTING WINNING TEAM
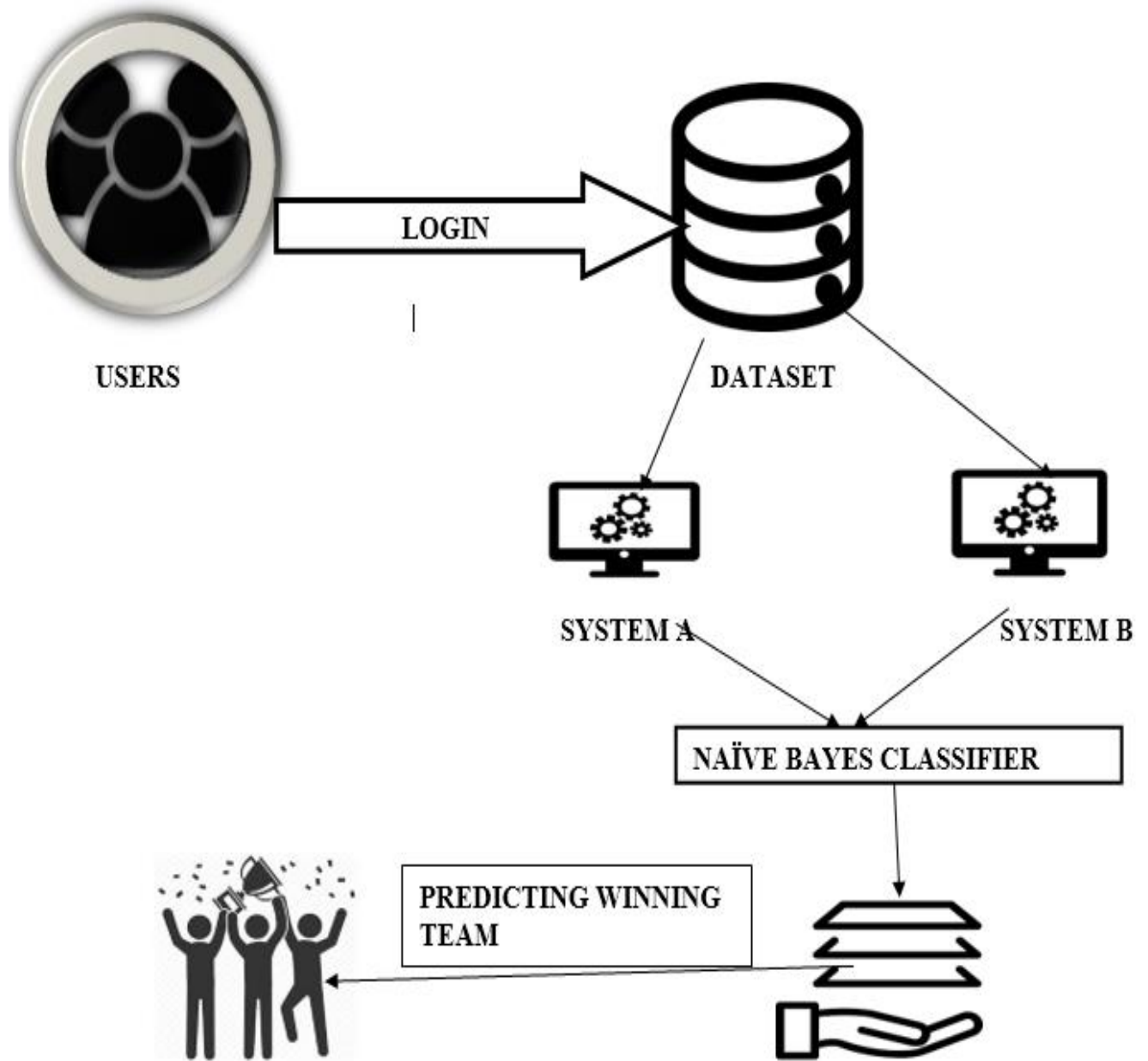
3.4 Technologies used:

3.4.1. JAVA

Java Programming language is concise which makes it easy to use and learn. Java virtual machine enables (JVM) enables java to be executed in any environment and platform making it portable. Web applications and applets can be accessed in a secure way using java. It is objects oriented support multithreading. Java supports cross platform optimized code called bytecode which are faster to execute Hence, it has high performance.

3.4.2. My SQL

My SQL is an open source database management system suitable for relational data. My SQL is popular choice of database for use in web applications. It has high availability and can also run on cloud computing platforms. Administration of My SQL is handled with the use of php My Admin which is free and open source tool for web browser.

3.5 Algorithms

Naïve Bayes:

The naïve Bayes is a probabilistic classification algorithm which is based upon the Bayes theorem of conditional probability. The naïve Bayes algorithm assumes that all the attributes in the data set are independent of each other and then compute the probabilities of each attribute. The naïve Bayes algorithm with a percentage split method divide the data set into 70% training data and 30% test data and gives an accuracy rate of 57.06%. Here, this accuracy improves only slightly when we use the 10-Fold cross validation method.

Input: user input file data record which contains (C score, overs, WF) segment of average score from train database of k to n over's.

Output: Projected score Sc

Step1: Read R (run, overs, WF) from current parameters.

Step2: map with train features with each sample.

Step3: calculate average score of train DB with same evidences.

$$AvgTScore = \sum_{k=0}^{n}(Sc)$$

Step4: pre score= (C Score + Avg T Score)

Step5: return pre score

The accuracy rate of the model on the test data set was used as the evaluation criteria. The table below gives the accuracy rate of each of the algorithm along with how it performed using the percentage split method as well as the K – Fold cross validation method.

Algorithm	Percentage Split	K-Fold Cross Validation
Decision Tree (J48)	31.07%	63.24%
Random Forest	48.02%	71.08%
Naïve Bayes	57.06%	60.97%
KNN	52.54%	51.39%

Figure 7: Algorithms Comparison

This model gives an accuracy rate of about 60% to 70% when using the K - Fold cross validation whereas the accuracy rate in the case of the percentage split is not very good. This is mainly because of the fact that in percentage split method, WEKA splits the data manually on the basis of 70% training and 30% test split that I mentioned. The K _ cross validation method overcomes that by dividing the data set into k different subsets. and thereby computing the results for k different subsets. We can notice that the accuracy rate almost doubles in some algorithm as shown above. The random forest algorithm

gives the best accuracy rate for the model with 71.08 accuracy rate.

3.6 Hardware and software requirements:

3.6.1 Hardware requirements

Processor- Dual Core

Speed -1.1 GHZ RAM-

512 MB(MIN4)

Hard – 20GB

Keyboard-Standard windows keyboard

Mouse- Two or three button Mouse

3.6.2 Software requirements

Operating system- Windows XP,7,8,10

Front end- java 7

Technology- My SQL connector java IDE-

Net beans

CHAPTER 4

4.1 Module Split-up:

Whole Project is divided into two Module:
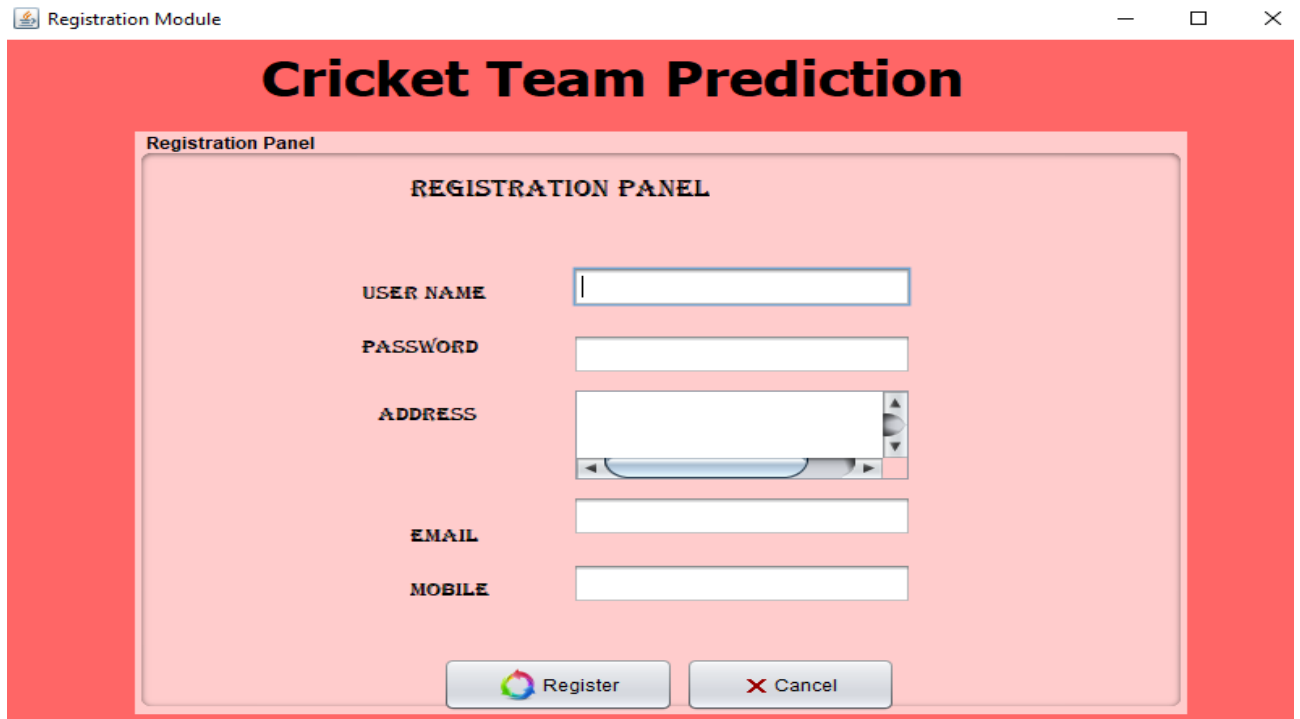
- Admin Module

- Users Module

- **Home Page:**

  After opening the first page this page will display by clicking only continue button we can access more part of application.
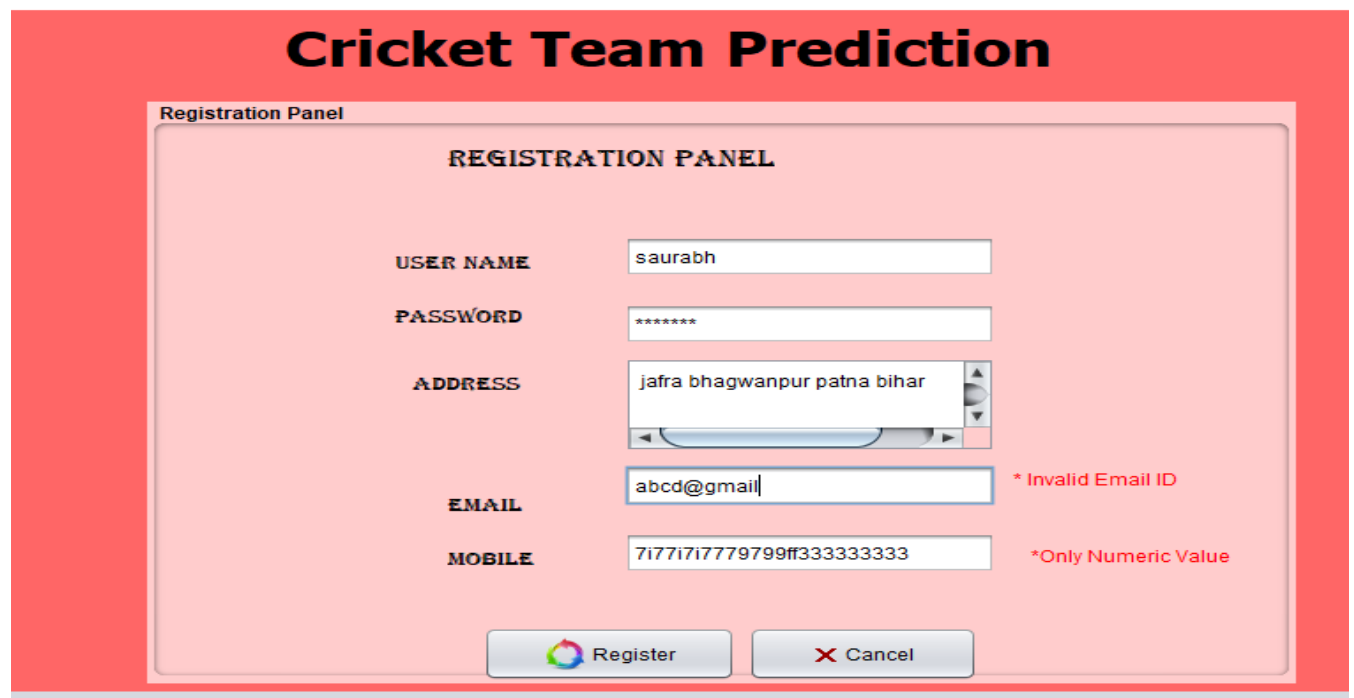
Registration module:-

- In this module new users can do registration to get access to user clients page.

- In this registration module each field is mandatory.

- Figure to show when we enter invalid email and mobile number:

- Figure to show when we enter valid character to register.

4.1.1 Admin Module:

Admin Module has several sub Module

- First sub module is admin login panel. In which admin enter their username and their password to control different function of a system.

- After entering correct username and password, admin gets access to the admin account module. In this Module admin get various option like Add Stadium, Add Team, Add Team Members, View Stadium, View Team members, Read CSV file. All these option help admin to run system properly for their users.
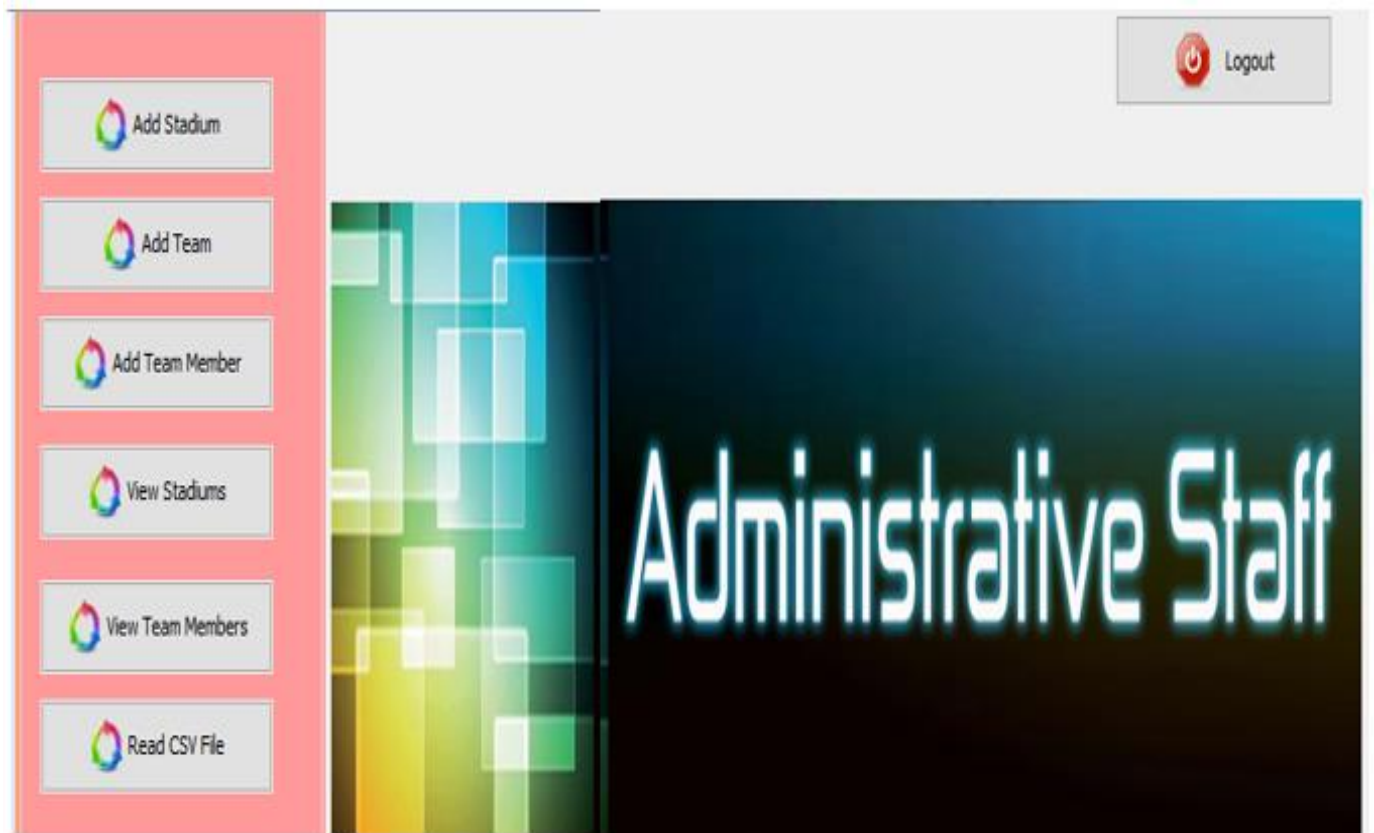
- The add stadium button allows admin to add a new stadium name for the user.

- The add team button allows admin to insert a new team with a stadium name which is already added by the admin.

- The team member can be added with stadium and team name which is already added by the admin.

- The remaining button like view team Member will display the added member where as view stadium button will display the stadium name which is already inserted by the admin.


4.1.2 Users module:

Users module is also divided into several sub module.

- First sub module is user login panel. In which user enter their registered their username and password to get the benefits of our system.

- After entering correct username and password, User gets access to user Account module. In this module users get various options like Select Team.
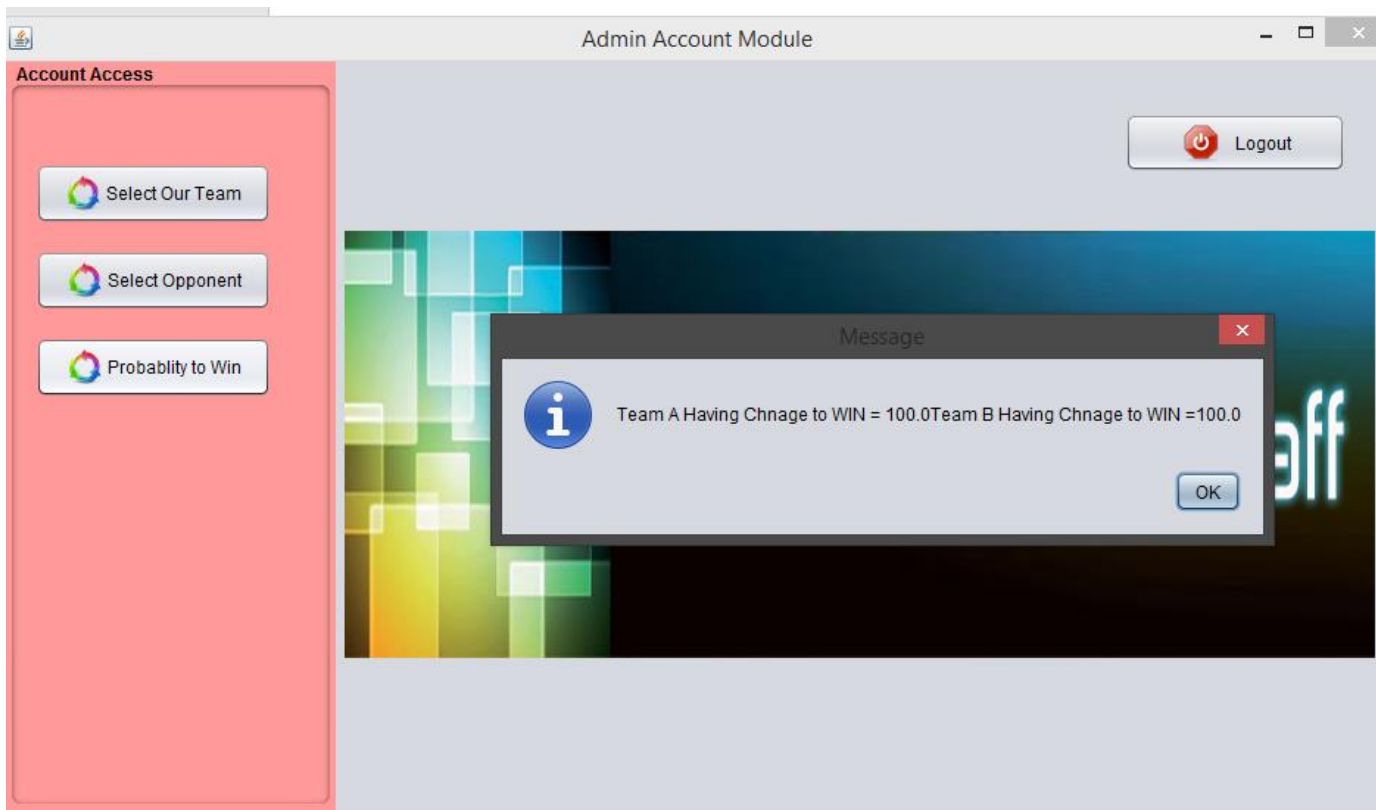
- After selecting two team User click on probability to win option to see the upcoming match probability i.e. which is going to win.
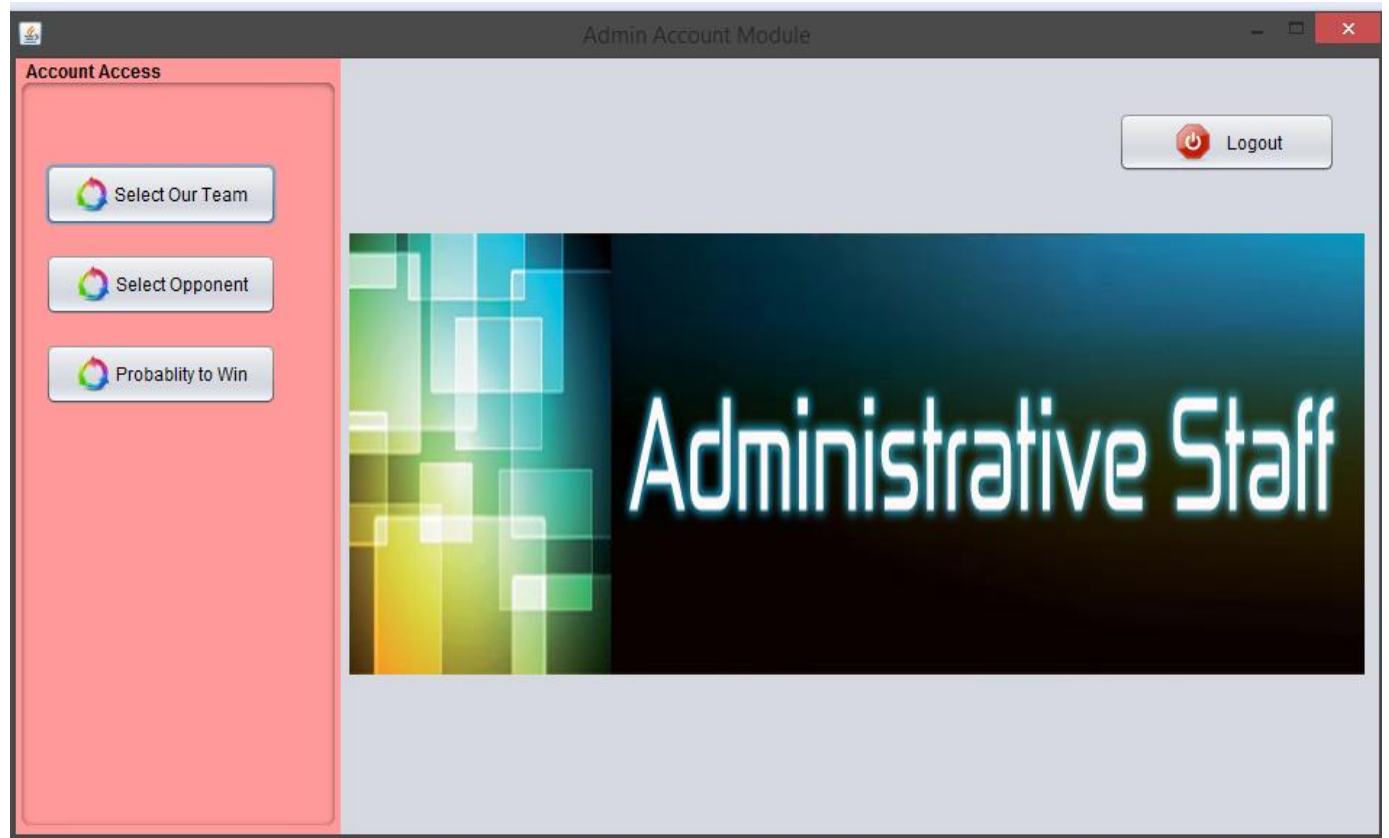
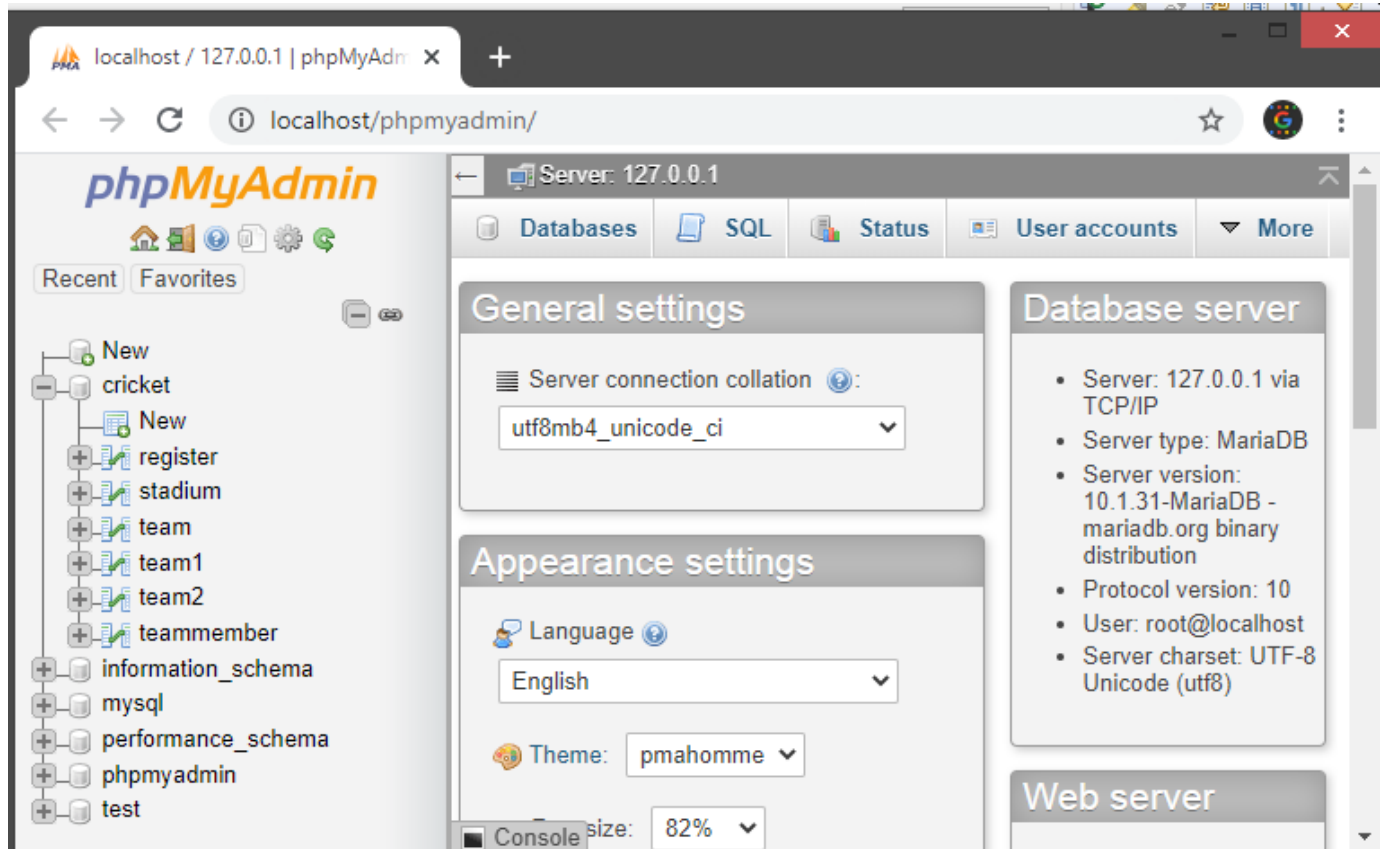- We can select our team members and opponent by clicking on given button.

- By clicking on probability to win system can predict winner.

4.2 Admin Page:

PHP My Admin is open source free software, designed to handle the administration and management of MySQL database through a graphic user interface. Written in PHP, PHP My Admin has become one of the most popular Web-Based MYSQL management tools. PHP My Admin comes with detailed documentation and is being supported by a large multi-language community. PHP My Admin's ever growing list of features supports all commonly used operations such as browsing, dropping, creating, altering MySQL database tables fields and indexes. Also, PHP My Admin enables you to manage MySQL users and user privileges. Another commonly used PHP My Admin features is its import function. With PHP My Admin, MySQL database imports from backup is made easy and you can import an SQL or CSV dump with a few mouse clicks. Also, you can easily export your database CSV, SQL, XML and other popular formats.

- Admin can also see all registered users from PHP MyAdmin page. Figure shows all registered users.

CONCLUSION

The model which is used to predict the result of the matches was built successfully with an accuracy rate of about 60% to 70%. The list of attributes was cut down to 10 important one out of the 21 attributes available in the dataset by using the attribute selection algorithm. The 4 data mining algorithms that were performed on the model was naïve Bayes. The prediction results were better when K-Fold cross validation was used as compared to the percentage split. The accuracy of the random forest algorithm was the best with 71.08%. Although the accuracy was between 60% and 70% but it was still low because of the fact that the total number of instances in the data set was 574 and the total number of classes were 11. We need at least 100 instances per class to identify pattern in the data set and perform a prediction with a high accuracy rate. So, with 11 classes in the dataset we need at least 1100 instances to perform prediction with a high accuracy rate. Since, the data set consisted of 574 instances, in future it may improve the accuracy with more number of instances in the data set because with a larger number of instances the model will have the flexibility to deduce better rules and identify more patterns in the data set as compared to with a lesser number of instances.

FUTURE WORK

There are some future works that can be done in order to improve this project.

- The data set can include some of the external factors like player injury, player fatigue, winning streak with a particular team ,overall winning streak, average run scored by a team against a particular team in previous matches etc. and on the basis of these data we can try to do the prediction and to check to see the accuracy improves.
- The prediction can also be done taking into consideration the performance of the players in the team like the total number of runs scored by a player in the tournament, the form guide of the player, the number of man of the match awards earned etc. rather than only using a high level data about the different matches like toss winner, toss decision, home team etc.
- There is no web/mobile applications or UI that my project contains. So, a web/mobile applications can be made which would take in the entire data set as input and display the prediction result for each of the instances to a pdf or text file.

REFERENCES

1. Dataset: http://cricsheet.org [online accessed 02-september -2016]

2. http://espncricinfo.com/ [online accessed 20-september-2016]

3. Pramila M. Chawan. "Prediction of live cricket score and winning", International journal of trend in research and development volume 5(4), ISSN:2394-9333, www.ijtrd.com.

4. Haghighat, Maral, Hamid Rastegari, and Nasim Nourafza. "A review of data mining techniques for result prediction in sports." Advances in computer science : an International Journal2.5(2013)