PREDICTIVE MODELING FOR CRICKET DATA ANALYSIS

A Mini Project report submitted to

Jawaharlal Nehru Technological University, Kakinada,

in the partial fulfillment for the award of the Degree in

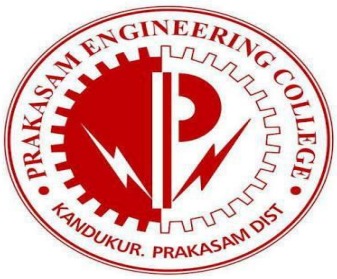
MASTER OF COMPUTER APPLICATIONS

Submitted by

BOGALA MOUNIKA
(20F91F0006)

Under the noble guidance of

Mr.M M RAYUDU , M.Tech(Ph.D)
AssociateProfessor



PRAKASAM ENGINEERING COLLEGE

(An ISO 9001-2008 & NAAC Accredited Institution)

(Affiliated to Jawaharlal Nehru Technological University, Kakinada)

O.V.ROAD, KANDUKUR-523105, A.P.

2021-2022

PRAKASAM ENGINEERING COLLEGE

(An ISO 9001-2008 & NAAC Accredited Institution)

(Affiliated to Jawaharlal Nehru Technological University, Kakinada)

O.V.ROAD, KANDUKUR-523105, A.P.

DEPARTMENT OF

MASTER OF COMPUTER APPLICATIONS

BONAFIDE CERTIFICATE

This is to certify that the mini project entitled **"PREDICTIVE MODELING FOR CRICKET DATA ANALYSIS"** *is a bonafide work of* **BOGALA MOUNIKA (20F91F0006)** *in the partial fulfillment of the requirement for the award of the degree in* **MASTER OF COMPUTER APPLICATIONS** *for the academic year 2021-2022. This work is done under my supervision and guidance.*

Signature of the GUIDE
Mr. M. M. RAYUDU
M.Tech(Ph.D)

Signature of the HOD
Mr. M. M. RAYUDU
M.Tech, (Ph.D)

Signature of the External Examiner

DECLARATION

I do here by declare that the project work entitled **"PREDICTIVE MODELING FOR CRICKET DATA ANALYSIS"** is a genuine work carried out by me under the guidance of **Mr.M.M.RAYUDU M.Tech (Ph.D)** in partial fulfillment for the award of the degree of **"Master of Computer Applications"** of **Jawaharlal Nehru Technological University, Kakinada.**

BOGALA MOUNIKA

(20F91F0006)

ACKNOWLEDGEMENT

I feel to render my thankful acknowledgement to the following distinguished personalities, who stretched their helping hand to me, in completing my mini project work.

I am very grateful and my sincere thanks to our secretary & correspondent **Dr.K.RAMAIAH** of **PRAKASAM ENGINEERING COLLEGE** for giving this opportunity.

I hereby, express my regards and extend my gratitude to our PRINCIPAL, **Dr.K.RAVI KUMAR** , for giving this opportunity to do the thesis as a part of our course.

I express my deep sense of gratitude to **Mr.M.M.RAYUDU, M.Tech(Ph.d),** Head of the Department, **Department of MCA** for having shown keen interest at every stage of development of our thesis and guiding us in every aspect.

And I am thankful to my Internal Guide **Mr.M.M.RAYUDU, M.Tech(Ph.d)** who has channeled my thoughts and timely suggestions.

I would also like to thank all my Faculties in Prakasam Engineering College for their constant encouragement and for being a great group of knowledgeable and cooperative people to work with.

BOGALA MOUNIKA

(20F91F0006)

TABLE OF CONTENTS

CHAPTER **PAGE NO.**

ABSTRACT

1. INTRODUCTION

1.1 Introduction 1

1.2 Existing System 3

1.2.1 Drawbacks 3

1.3 Proposed System 3

1.3.1 Features or Objectives of Proposed 4

1.3.2Advantages of Proposed 4

2. **LITERATURE SURVEY**

2.1Reference paper1 5

2.2 Reference paper2 5

2.3 Reference paper3 6

2.4 Reference paper4 6

2.5Reference paper5 7

3. SYSYEM ANALYSIS

3.1 Introduction 8

3.1.1Methodology 8

3.1.2.CRISP-DM 8

3.1.3.Data Collection 9

3.1.4.Data Preprocessing 9

3.1.5.Exploratory Data Analysis 10

3.2 Feasibility Study 11

3.2.1Operational Feasibility 12

3.2.2 Economic Feasibility 13

3.3 RequirementSpecifications 14

3.3.1Hardware requirements 14

3.3.2 Software requirements 15

4. SYSTEM DESIGN

4.1Introduction 16

4.1.1Module Description 16

4.1.2 Classification by Decision Tree Induction 16
4.1.3 .Classification by Random Forest 17
4.1.4.Classification by Logistic Regression 18
4.2. Data Design 19
4.3. UML DIAGRAMS 20
4.3.1.Class Diagram 21
4.3.2.UsecaseDiagram 22
4.3.3 Sequence Diagram 23
4.3.4 Deployment Diagram 24

CONCLUSIONSANDFUTUREENHANCEMENT 25
REFFERENCES 26

ABSTRACT

Data Science / Analytics is all about finding valuable insights from the given dataset. In short, Finding answers that could help business. We will see how to get started with Data Analysis in Python. Since usually such are based on inbuilt datasets like iris, It becomes harder for the learner to connect with the analysis and hence learning becomes difficult. IPL is one of the most popular cricket tournaments in the world, thus the problems.

Cricket is like Religion in our Country. Every fan tries to predict the score and also they want the playing 11 according to their choice. Cricket is increasingly popular among the statistical science community, but the unpredictable and inconsistent natures of this game make it challenging to apply in common probability models.

utilized to determine the runs that a given team will score. It does not take into consideration other important factors which are of equal, if not more, importance such as match venue(home or away), number of wickets fallen as well as the rating of a given player(player's past performance against a given team). This number of important parameters along with their interdependence creates a challenging scenario in developing an accurate model of the game.

CHAPTER1

INTRODUCTION

1.1.INTRODUCTION

Cricket, especially the twenty20 format, has maximum uncertainty, where a single over can completely change the momentum of the game. With millions of people following the Indian Premier League(IPL), therefore developing a model for predicting the outcome of its matches beforehand is a real-world A cricket match depends upon various factors, and in this work, various feature selection methods were used to reduce the number of features to 5 from 15. Player's performance in the field is considered to find out the overall weightage (relative strength) of the team. Predicting the outcome of a game using players strength and weakness against the players of the opponent team by considering the statistics of a set of matches played by players helps captain and coaches to select the team and order the players.

Indian Premier League (IPL) is a professional cricket league based on twenty20 format and is governed by Board of Control for Cricket in India. The league happens every year with participating teams representing various cities of India. There are many countries active in organizing t20 cricket leagues, and when most of them are being over hyped; team franchises routinely losing money, IPL has stood out as an exception. As reported by espncricinfo, with Star Sports spending $2.5 billion for exclusive broadcasting rights, the latest season of IPL (2018, 11th) saw the increment in number of viewers including both the digital streaming media & television. The 10th season had million people streaming the league through their digital devices and watching directly on television. So, with millions of people eying the league, it would be an interesting problem to make use of statistics and machine learning to predict the outcome of IPL matches.

As of now, there are eight teams that compete with one another in a double round-robin fashion during the league stage. After the league stage, the top four teams in the league points table qualify to the playoffs.

In playoffs: The winner between first & second team qualify for the final, while the loser gets an extra chance to qualify by playing against the winner between third & fourth team

that qualified for the playoffs. Finally, the two qualifying teams go against each other for the league title.

Among all formats of cricket, T20 format sees a lot of turnaround in the momentum of the game. An over can completely change a game. So, predicting an outcome for a t20 game is quite a challenging task. On top of that, developing a prediction model for a league which is completely based on auction is another hurdle. In this work, we have analyzed various factors that might affect the outcome of a cricket match, and found out that home team, venue, team strength, toss decision i.e. first batting or first fielding play respective vital role in determining the winning side.

Some works have been published in this area of predicting outcome in sports. In our literature review, we found out that the published works were for test or one-day-international (ODI) cricket format.

The person has analyzed the factors like home field advantage, winning the toss, game plan (first batting or first fielding) and the effect of D/L (Duckworth Lewis) method for one day cricket format. Similarly, Bailey and Clarke mention in their work that in one day cricket format, home ground advantage, past performances, venue, performance against the specific opposition, current form; are statistically significant in predicting total runs and predicting the match. Discusses about modelling home-runs and non-home runs prediction algorithms and considers taking runs, wickets, frequency of being all-out as historical features into their prediction model. But, they have not leveraged bowler's features and have given more emphasis to batsmen have proposed a tool that predicts match a-priori, but player's performance has not been considered into their model.

The IPL works on a franchise-system based on the American style of hiring players and transfers. These franchises were put for auction, where the highest bidder won the rights to own the team, representing each city.Over 200 million Indian viewers,10 million international viewers,4 million live spectators : the Indian premier league(IPL) is a sports and entertainment revolution in the making, surpassing all records of viewership on ground and on media. Advertising revenue and ticket sales have exceeded all expectations ,making IPL highly profitable for the organizers ,broadcasters and successful team owners. Zealous fan following-even hostility for visiting teams-shows local loyalties are building up faster than anyone expected.

HISTORY AND BACKGROUND OF THE IPL:

Kerry Francis Bull more Packer, AC(17 December 1937-26December 2005),was an Australian media tycoon whose family company owned controlling interests in both the nine television network and leading Australian publishing company Australian consolidated press. packer was best known for founding world series Cricket. In 1997 the nine network cricket rights deal led to a confrontation with the cricket authorities, as top players from several countries rushed to join him at the expense of their international sides. packer's aim was to secure broadcasting rights for Australian Cricket, and he was large successful. Many of the well-known cricketers of that period left their national team to play in Kerry packer's world series cricket .Some of our legendry Cricketers also contacted to play in that series. But due to some controversies, mainly with Australian board due to television rights, this league could not be successful

1.2 EXISTING SYSTEM:

As Cricket Fever goes every time with big leagues coming up ,Franchise owners of IPL in Bidding of players is done based on the previous team, players performance data and the analysis is done in some basis. Only on the detailed analysis on a particular attribute.

1.2.1 DRAWBACKS :

In existing system the franchise owners have bid the players in the auction based on Previous data analysis. The analysis done is not done with classification and regression algorithms as Now the are present in proposed system by applying this algorithms. We get results based on our required field root cause problem Impacted stakeholders/product users Impacts of the issues. Effects a successful solution must include the proper analysis data of the teams, players, player statistics.

1.3.PROPOSED SYSTEM:

In this proposed system, we are going to predict the winning team based on some key factors by taking IPL Dataset from the previous years and will come to prediction based on the statistical data provided.

1.3.1 FEATURES & OBJECTIVES OF PROPOSED SYSTEM

The proposed system will have the following features.

- The extracted data from Kaggle will tell us what are the features or attributes present used and how the statistics differ accordingly.
- It provides flexibility to the user to transfer the data through the network very easily by compressing the large amount of file only if the user have the account as he need to download the matches data because manually it takes very much time to create a dataset.

It should also identify the user and provide the communication according to the prescribed level of security with transfer of the file requested and run the required process at the server if necessary.

1.3.2 ADVANTAGES OF PROPOSED SYSTEM

Apart from normal analysis we can have the clear predictive outcome once the complete analysis is performed. Thus the predictive analysis helps for the Franchises, Players and Team Management to come to a conclusion regarding their team profile and players.

CHAPTER 2

LITERATURE SURVAY

2.1. http://en.wikipedia.org/wiki/Cricket :

The purpose of this article is to develop models that can help team selectors build talented teams with minimum possible spending. In this study, we build several predictive models for predicting the selection of a player in the Indian Premier League, a cricket league, based on each player's past performance. The models are developed using SAS® Enterprise MinerTM 7.1. The best-performing model in the study is selected based on the validation data misclassification rate. The selected model provides us with the probability measure of the selection of each player,which can be used as a valuation factor in the bidding equation. The models that are developed can help decision-makers during auction set salaries for the players.

2.2 . http://www.duckworth-lewis.com/mags/dlmethod/ :

Cricket prediction is comparatively difficult as there are many factors that can influence the result or outcome of the cricket match. Earlier basic prediction systems for cricket match consider only the venue and disregard the factors like weather, stadium size, captaincy etc. The factors like venue of the match, pitch ,weather conditions first batting or fielding all play a vital rolein predicting the winner of the match. Suitable models are necessary to predict and data mining makes it possible to extract required information from the data files. This paper presents the usage of Google Prediction API to analyse the data of previous cricket matches and predict outcome of a given cricket match.

2.3. http://www.espncricinfo.com/ :

Cricket, especially the twenty20 format, has maximum uncertainty, where a single over can completely change the momentum of the game. With millions of people following the Indian Premier League, therefore developing a model for predicting the outcome of its matches beforehand is a real-world problem. A cricket match depends upon various factors, and in this work various feature selection methods were used to reduce the number of features to 5 from 15. Player's performance in the field is considered to find out the relative strength of the team. A Linear Regression based solution is proposed to calculate the weightage of a team based on the past performance of its players who have appeared most for the team. Finally, a dataset with the features: home team, away team, stadium, toss winner, toss decision, home team weightage and away team weightage, is fed to a Random Forest Classifier to train the model and make prediction on unseen matches. Classification results are satisfactory. Problem in the dataset and how the accuracy of the classifier can be improved is discussed.

2.4. Predicting the Winner in One Day International Cricket

Consider the decision problem of when to declare during the third innings of at least cricket match. There are various factors that affect the decision of the declaring team including the target score, the number of over s remaining, the relative desire to win versus draw, and the scoring characteristics of the particular match. Decision rules are developed and these are assessed against historical matches. We observe that there are discrepancies between the optimal time to declare and what takes place in practice. consider the determination of optimal team line ups in Twenty20 cricket where a line up consists of three components: team selection, batting order and bowling order. Via match simulation, we estimate the expected runs scored minus the expected runs allowed for a given line up.

The line up is then optimized over a vast combinatorial space via simulated annealing. We observe that the composition of an optimal Twenty20 line up sometimes results in non traditional roles for players. As a by-product of the methodology, we obtain an “all-star” line up selected from international Twenty20 cricket. is a first attempt to investigate the importance of fielding in cricket. We introduce the metric of expected runs saved due to fielding which is both interpretable and is directly relevant to winning matches. The metric is assigned to individual players and is based on a textual analysis of match commentaries using random forest methodology.

2.5. Score and Winning Prediction in Cricket through Data Mining

Cricket is one of those sports where a large amount of numerical information is generated in every game. The game of cricket got a new dimension in April 2008, when Board of Control for Cricket in India (BCCI) initiated the Indian Premier League(IPL). It is a franchise based Twenty20 cricket tournament where teams are formed by competitive bidding from a collection of Indian and International players. Since, valuations of the players are determined through auction, so performance of individual player is always under scanner. The objective of this study is to analyze and predict the performance of bowlers in IPL, using artificial neural network. Based on the performance of bowlers in the first three seasons of IPL, the paper tries to predict the performances of those bowlers who entered in the league in its fourth season as their maiden IPL venture. The performances of these bowlers in IPL-IV are predicted, and the external validity of the model is tested using their actual performance in IPL-IV. This prediction can help the franchises to decide which bowler they should target for their team.

CHAPTER3

SYSTEM ANALYSIS

3.1. INTRODUCTION

Cricket is an outdoor team sport. A cricket match is played involving two teams of eleven players each. The game is mainly played at domestic and international. The sport is played mainly in the following ten countries, who are full time members of the International cricket council and are also referred to as the test nations; Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka,West Indies and Zimbabwe. Cricket is followed by a billion plus audience worldwide and growing in popularity rapidly and humongous amount of money being spent on the game, be it in telecast rights or sports betting In comparison to other team sports such as football and baseball, the amount of work that has been done on cricket in the end of analytics is less. The economic and social perspectives such as fan following, media coverage, etc. provides strong incentive to analyse the game.

3.1.1. METHODOLOGY

This section discusses the methodology followed for the research and the valuation of the outcome of the research conducted.

3.1.2. CRISP-DM

CRISP-DM stands for Cross Industry Standard Process for Data Mining. It is a comprehensive process model for data mining projects. This project uses CRISP-DM due to its independence from technology and industry sector(Wirth and Hipp; 2000).

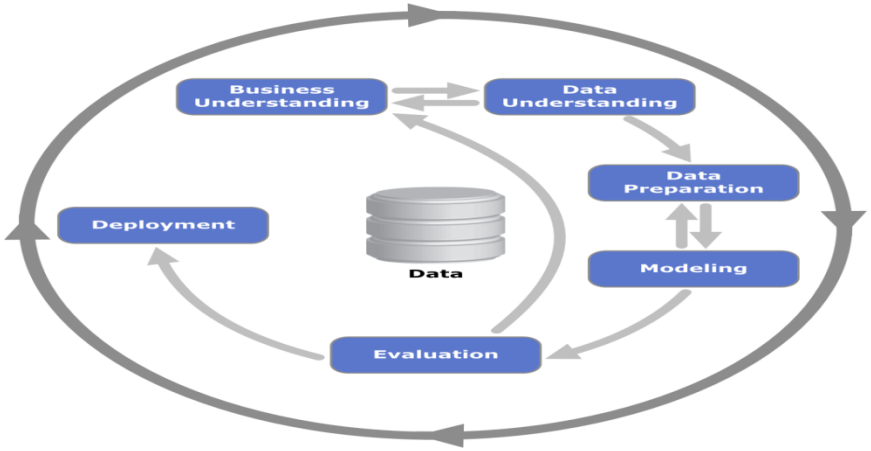


Fig 1: CRISP-DM model phases.

3.1.3. Data collection

Data was web scrapped from www.kaggledataset This source contained match information such as date, teams (countries), ground, result, day and night match or not, toss winner, team batting first, etc. From this, only the factors that are reported before theplay starts were considered for model building keeping in mind the objective of the re-search, rest of them were used for exploratory data analysis as applicable. However, it must be noted that a common problem faced while scrapping large amount of data from web, is with Timeouts. The same problem was faced with and overcome by making the system sleep for few seconds after collecting data. To establish the veracity of the data, data instances were picked at random and information was compared by performing a manual search on the web.

3.1.4. DATA PRE-PROCESSING

This section discusses the steps taken to tidy the data and convert it to a format suitable for analysis. All the column names were checked for inconsistency and suitably handled to maintain consistency. All the observations with unwanted results, such as abandoned matches, conceded matches, cancelled matches, walkovers etc. were dropped. Unnecessary columns such as the one indicating row number and the one providing link text were dropped. The links were used during data collection to fetch game play details and were not required any further Data types of the columns were converted to the required data types, since the information was retrieved by web scrapping hence all the columns were stored as object types. There were outliers in the data set, but were left untransformed because they are natural outliers.

Further it was made sure data was represented in a format that is suitable for analysis. Each row represents unique individual observations. Each of the columns in the data represents separate variables. It was made sure data table was stored in melted format rather than a pivoted format.

3.1.5 Exploratory Data Analysis

The approach to exploratory data analysis was to first generate hypothesis from the data and then try to either prove or disprove the hypotheses generated. The thought was that in doing so it will help in the later stages with feature engineering without being influenced by the data available in the corpus.

The effect of location on match outcome:

The below figure visually analyses the difference in performance of teams according to the location.

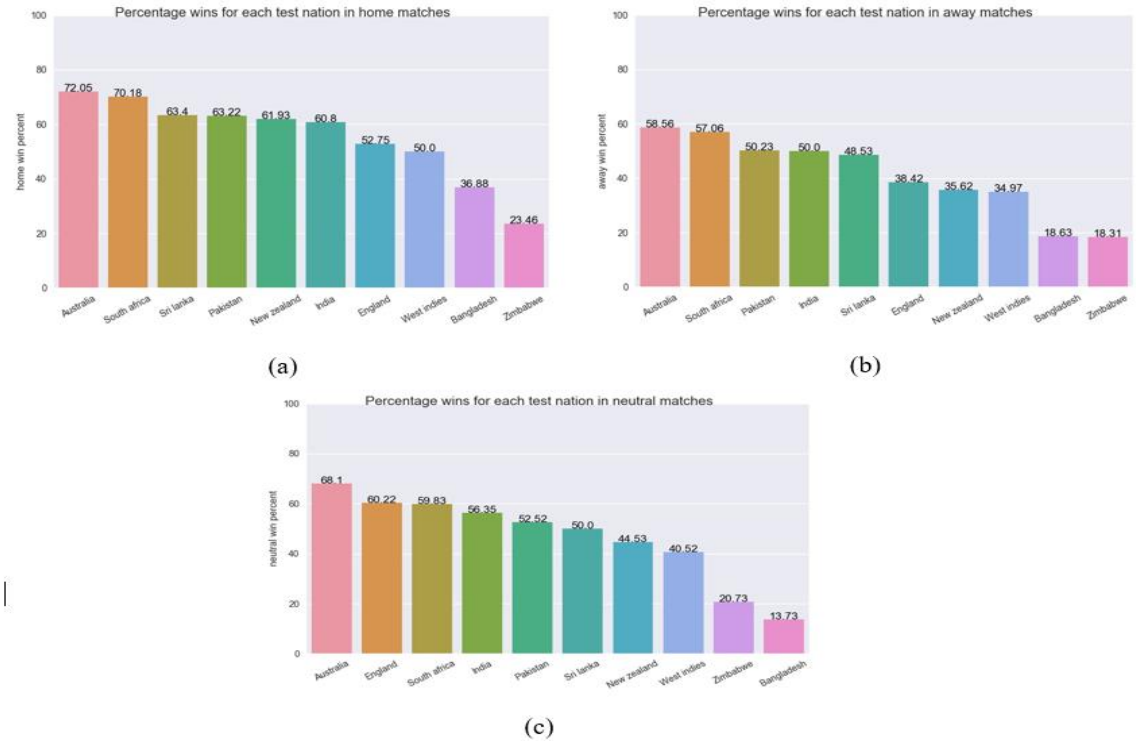


Fig 2: Location wise win percentage of each team; (a) home, (b) away and (c)neutral.

We can observe from Figure 2 that, the percentage of wins for each team variety according to the location. Not only does the win percentage differ for the teams, but also the positional order of the teams changes according to location. Thus, it can be assumed that location is important. The toss plays an important factor in deciding the winner of a match

There is some ambiguity observed about the importance of toss from the works of other scholars as discussed in the related works section 2. Hence, a Chi-square test of independence was conducted between toss and winner attributes to establish the importance of toss.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	1.328	.000
	Cramer's V	.443	.000
N of Valid Cases		2475	

Table 2 : Chi-square test of independence between toss and winner

Referring to Cramer's V, value of .443 which is less than 0.5 and hence does not signify strong association between the two variables. Thus, it is safe to assume that toss is not a significant attribute.

The important point in history of data model

The thought behind this hypothesis is that if there was a change in the nature of the game due to any reason such as introduction of twenty-twenty format, etc. that could render data previous to that point in time meaning less. To prove the above stated point an independent sample t-test was conducted on the scores, by splitting the corpus into sets at the year 2008. It was intuitively decided upon 2008 by observing the increase in number of twenty-twenty matches in 2007. The data is

produced a result of p=0.01, signifying a significance difference between the scores from before 2008 and that of after 2008.(Pallant; 2013) However on checking for the veracity of the test, it was found that similar results were obtained if the corpus was split at any annual value. Hence it was decided to include whole corpus for analysis.

3.2 FEASIBILITY STUDY

A project feasibility study is a comprehensive report that examines in detail the five frames of analysis of a given project. It also takes into consideration its four Ps, its risks and POVs, and its constraints (calendar, costs, and norms of quality). The goal is to determine whether the project should go ahead, be redesigned, or else abandoned altogether.

The five frames of analysis are: The frame of definition; the frame of contextual risks; the frame of potentiality; the parametric frame; the frame of dominant and contingency stThe

four Ps are traditionally defined as Plan, Processes, People, and Power. The risks are considered to be external to the project (e.g., weather conditions) and are divided in eight categories: (Plan) financial and organizational (e.g., government structure for a private project); (Processes) environmental and technological; (People) marketing and socio cultural; and (Power) legal and political. POVs are Points of Vulnerability: they differ from risks in the sense that they are internal to the project and can be controlled or else eliminated.

The constraints are the standard constraints of calendar, costs and norms of quality that can each be objectively determined and measured along the entire project lifecycle. Depending on projects, portions of the study may suffice to produce a feasibility study; smaller projects, for example, may not require an exhaustive environmental analysis. This assessment is based on an outline design of system requirements, to determine whether the company has the technical expertise to handle completion of the project. When writing a feasibility report, the following should be taken to consideration: A brief description of the business to assess more possible factors which could affect the study. The part of the business being examined. The human and economic factor. The possible solutions to the problem. At this level, the concern is whether the proposal is both technically and legally feasible (assuming moderate cost). The technical feasibility assessment is focused on gaining an understanding of the present technical resources of the organization and their applicability to the expected needs of the proposed system. It is an evaluation of the hardware and software and how it meets the need of the proposed system.

3.2.1 OPERATIONAL FEASIBILITY

Operational feasibility is the measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

The operational feasibility assessment focuses on the degree to which the proposed development project fits in with the existing business environment and objectives with regard to development schedule, delivery date, corporate culture and existing business processe.To ensure success, desired operational outcomes must be imparted during design and development. These include such design-dependent parameters as reliability, maintainability, supportability, usability, reducibility, disposability, sustainability,

affordability and others. These parameters are required to be considered at the early stages of design if desired operational behaviours are to be realised. A system design and development requires appropriate and timely application of engineering and management efforts to meet the previously mentioned parameters. A system may serve its intended purpose most effectively when its technical and operating characteristics are engineered into the design. Therefore, operational feasibility is a critical aspect of systems engineering that needs to be an integral part of the early design phases.

3.2.2 ECONOMIC FEASIBILITY

In case of a new project, economic viability can be judged on the following parameters:

- Total estimated cost of the project
- Financing of the project in terms of its capital structure, debt to equity ratio and promoter's share of total cost
- Existing investment by the promoter in any other business
- Projected cash flow and profitability

The financial viability of a project should provide the following information:

- Full details of the assets to be financed and how liquid those assets are.
- Rate of conversion to cash-liquidity (i.e., how easily the various assets can be converted to cash).
- Project's funding potential and repayment terms.
- Sensitivity in the repayments capability to the following factors:
- Mild slowing of sales.
- Acute reduction/slowing of sales.
- Small increase in cost.
- Large increase in cost.

In 1983 the first generation of the Computer Model for Feasibility Analysis and Reporting (COMFAR), a computation tool for financial analysis of investments, was released. Since then, this United Nations Industrial Development Organization (UNIDO) software has been developed to also support the economic appraisal of projects. The COMFAR III Expert is intended as an aid in the analysis of investment projects. The main module of the program accepts financial and economic data, produces financial and

economic statement. Graphical displays and calculates measures of performance. Supplementary modules assist in the analytical process. Cost-benefit and value-added methods of economic analysis developed by UNIDO.

3.3 REQUIREMENT SPECIFICATIONS:

FUNCTIONAL REQUIREMENTS:

Functional requirements are associated with specific functions, tasks or behaviours the system must support. The functional requirements address the quality characteristic of functionality while the other quality characteristics are concerned with various kinds of non-functional requirements. Because non-functional requirements tend to be stated in terms of constraints on the results of tasks which are given as functional requirements (e.g., constraints on the speed or efficiency of a given task), a task-based functional requirements statement is a useful skeleton upon which to construct a complete requirements statement.

NON-FUNCTIONAL REQUIREMENTS:

Non-functional requirements are requirements that specify criteria that can be used to judge the operation of a system, rather than specific behaviours. This should be contrasted with functional requirements that specify specific behavior or functions. In general, functional requirements define what a system is supposed to do whereas non-functional requirements define how a system is supposed to be. Non-functional requirements are often called qualities of a system. Other terms for non-functional requirements are "constraints", "quality attributes", "quality goals" and "quality of service requirements". Qualities. Non-functional requirements can be divided into two main categories.

1. Execution qualities, such as security and usability, are observable at run time.
2. Evolution qualities, such as testability, maintainability, extensibility and scalability are embodied in the static structure of the software system.

3.2.1 HARDWARE REQUIREMENTS

Operating System	:	Microsoft Windows 10/8/7/Vista/2003/XP
RAM Size	:	4GB
Processor	:	Minimum Core i5

Disk Space : 3GB (download and install)

3.3.2 SOFTWARE REQUIREMENTS :

Software for Python : Anaconda

Tool for Anaconda : Spyder

Programming Functionality : Python

Packages : Matplotlib,Seaborn,Pandas,Numpy

CHAPTER 4

SYSTEM DESIGN

4.1 INTRODUCTION

Data Mining is the process of discovering large patterns in large data sets involving methods at the intersection of machine learning, statistics and the database systems. Data mining refers to extracting or "mining" knowledge from large amounts of data. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Others view data mining as simply an essential step in the process of knowledge discovery.

4.1. 1 MODULE DESCRIPTION

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data.

CLASSIFICATION TECHNIQUES :

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are

known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The various types of classification algorithms are:

- Decision Tree Induction
- Random Forest
- Logistic Regression

4.1.2 CLASSIFICATION BY DECISION TREE INDUCTION

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non- leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce binary.

During the late1970s and early1980s, J.Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). This work expanded on earlier work on concept learning systems, described by E. B. Hunt, J. Marin, and P. T. Stone. Quinlan later presented C4.5 (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared. In1984, a group of statisticians (L. Breiman, J. Friedman, R. Olshen, and C. Stone) published the book Classification and Regression Trees (CART), which described the generation of binary decision trees. ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples. These two cornerstone algorithms spawned a flurry of work on decision tree induction.

4.1.3.CLASSIFICATION BY RANDOM FOREST:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.Random decision forests correct for decision trees' habit of over fitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Hoand later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

4.1.4. CLASSIFICATION BY LOGISTIC REGRESSION

In statistics, the logistic model (or logit model) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick;

These are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a log it, from logistic unit, hence the alternative names.

Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model, the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales

the odds of the given outcome at a constant rate, with each dependent variable having its own parameter; for a binary independent variable this generalizes the odds ratio.

Logistic regression was developed by statistician David Cox in 1958. The binary logistic regression model has extensions to more than two levels of the dependent variable: categorical outputs with more than two values are modeled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a classifier.

4.2. DATA DESIGN

Data design is the first design activity, which results in less complex, modular and efficient program structure. The information domain model developed during analysis phase is transformed into data structures needed for implementing the software. The data objects, attributes, and relationships depicted in entity relationship diagrams and the information stored in data dictionary provide a base for data design activity. During the data design process, datatypes are specified along with the integrity rules required for the data. For specifying and designing efficient data structures, some principles should be followed. These principles are listed below.

1. The data structures needed for implementing the software as well-as the operations that can be applied on them should be identified.
2. The data dictionary should be developed to depict how different data
Objects interact with each other and what constraints are to be imposed on the elements of data structure.
3.Stepwise refinement should be used in data design process and detailed design decision should be made late in the process.
4. Only those modules that to access date stored in a data structure directly should be aware of the representation of the data structure.
5. A library containing a set of useful data structures along with the operations that can be performed on them should be maintained.

4.3.UML DIAGRAMS:

UNIFIED MODELLING LANGUAGE:

The unified modeling is a standard language for specifying, visualizing, constructing and documenting the system and its components is a graphical language which provides a vocabulary and set of semantics and rules. The UML focuses on the conceptual and physical representation of the system. It captures the decisions and understandings about systems that must be constructed. It is used to understand, design, configure and control information about the systems. Depending on the development culture, some of these artifacts are treated more or less formally than others. Such artifacts are not only the deliverables of a project; they are also critical in controlling, measuring, and communicating about a system during its development and after its deployment. The UML addresses the documentation of a system's architecture and all of its details. The UML also provides a language for expressing requirements and for tests. Finally, the UML provides a language for modeling the activities of project planning and release management.

BUILDING BLOCKS OF UML:

The vocabulary of the UML encompasses three kinds of building blocks:

- Things
- Relationships
- Diagrams

4.3.1 CLASS DIAGRAM:

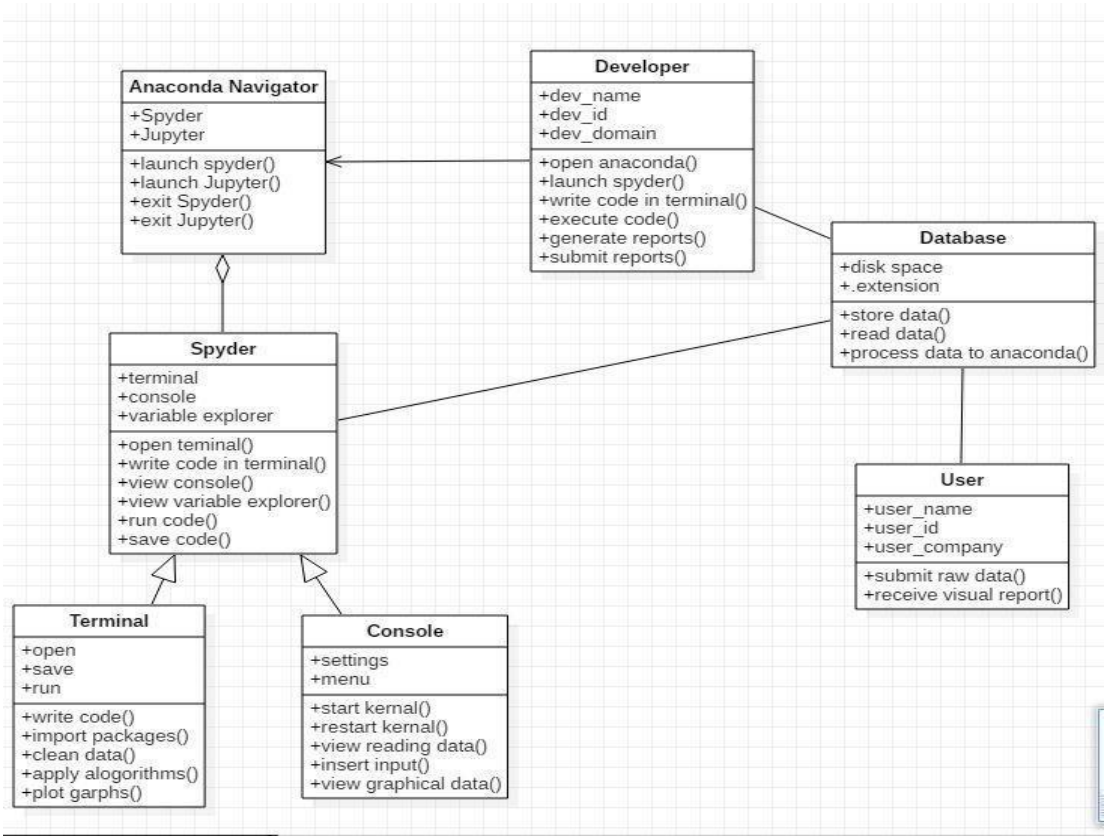


Fig 4.3.1: Class Diagram

4.3.2. USE CASE DIAGRAMS

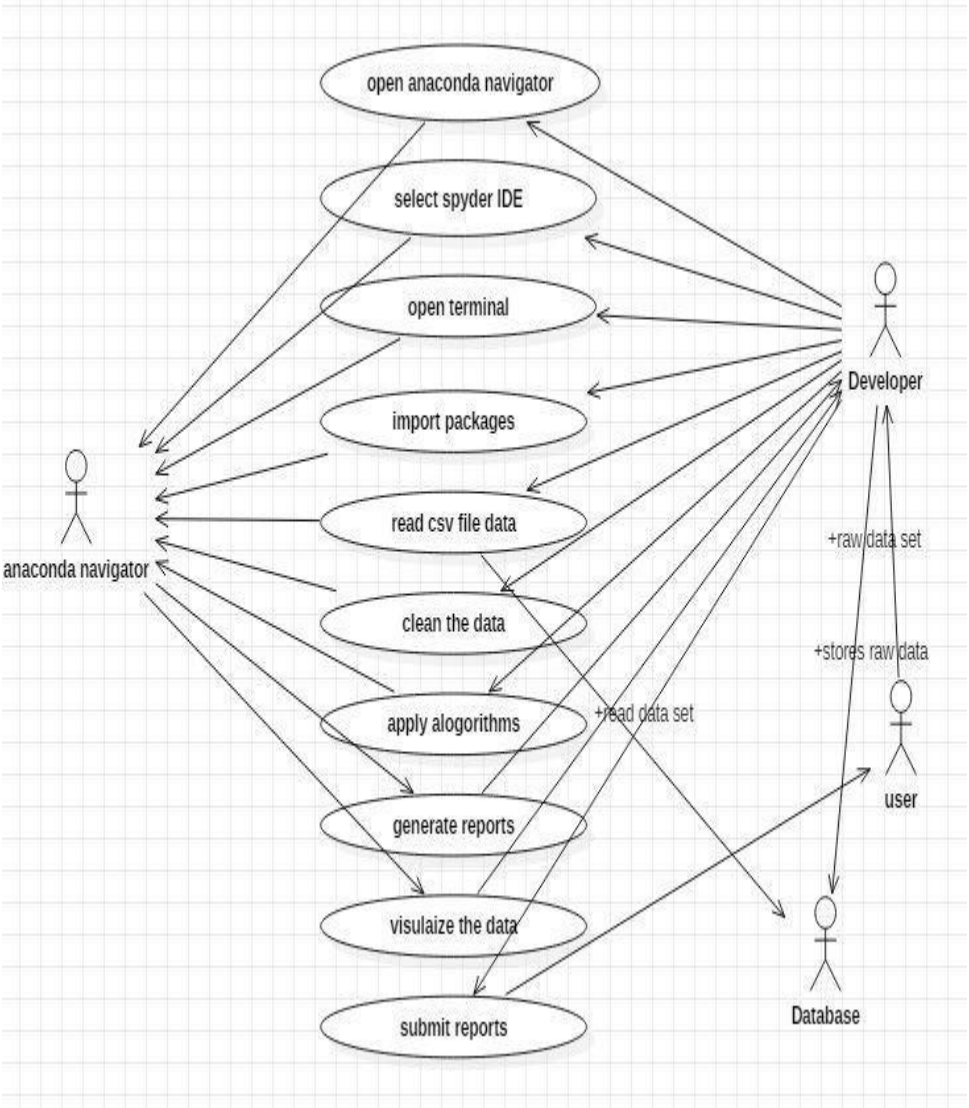


Fig 4.3.2 Use case Diagram

4.3.3. SEQUENCE DIAGRAMS :

UML sequence diagrams are used to represent the flow of messages, events and actions between the objects or components of a system. Time is represented in the vertical direction showing the sequence of interactions of the header elements, which are displayed horizontally at the top of the diagram. Sequence Diagrams are used primarily to design, document and validate the architecture, interfaces and logic of the system by describing the sequence of actions that need to be performed to complete a task or scenario. UML sequence diagrams are useful design tools because they provide a dynamic view of the system behavior which can be difficult to extract from static diagrams or specifications.

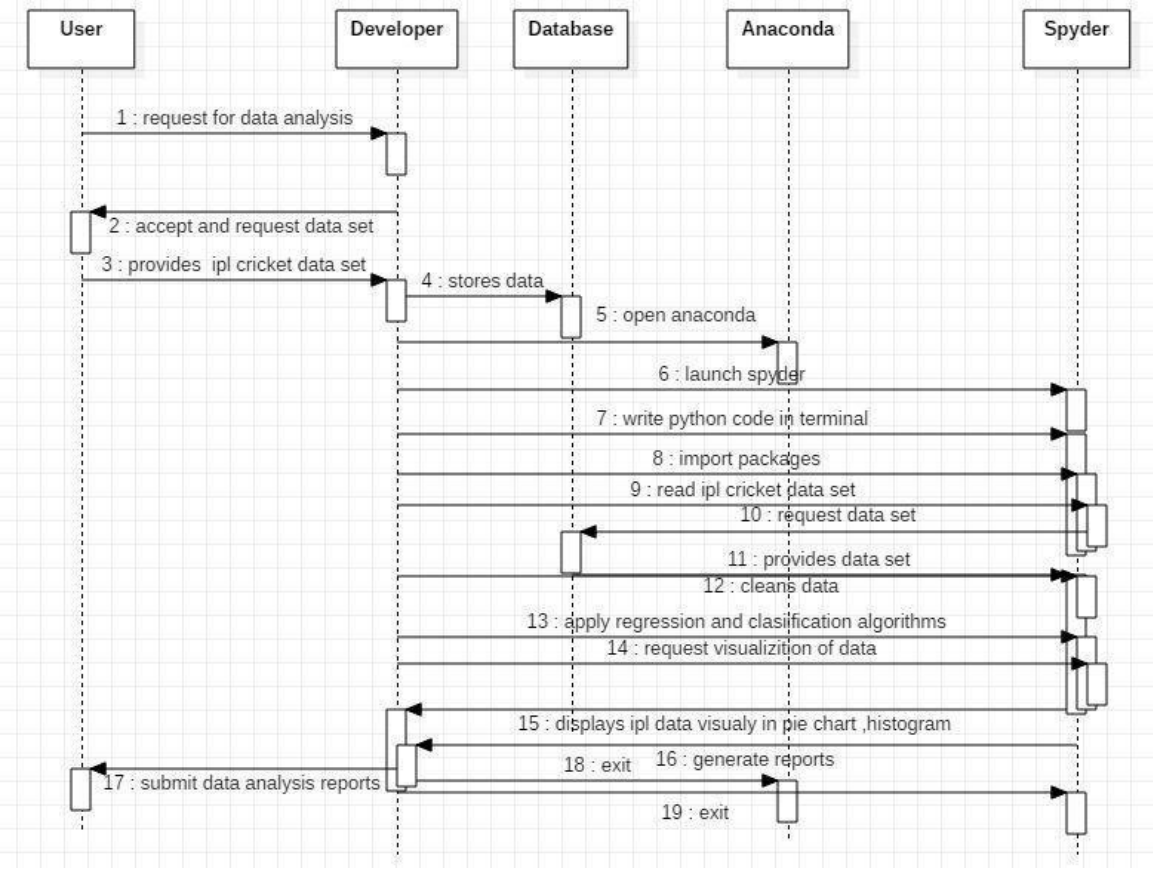


Fig 4.3.3 Sequence diagram

4.3.4DEPLOYMENT DIAGRAM:

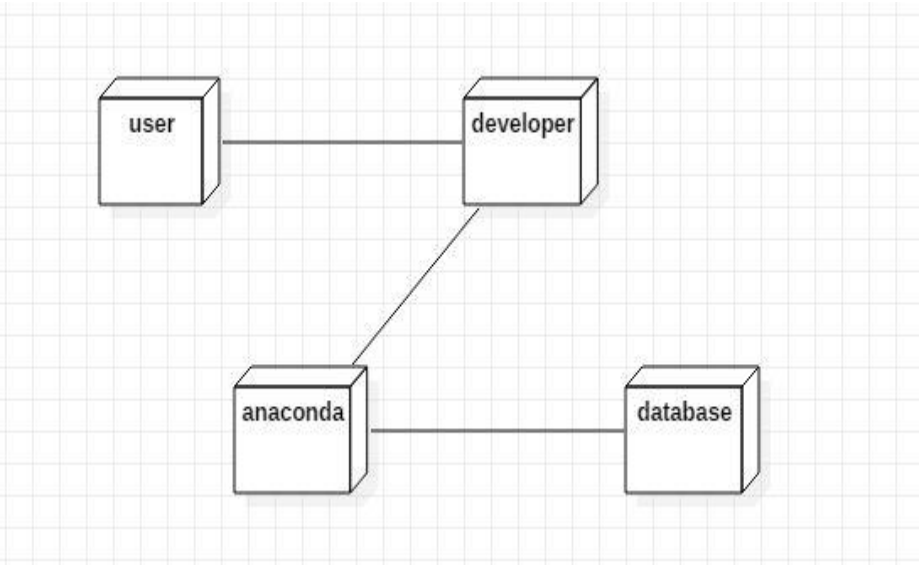


Fig 4.3.4: Deployment Diagram

CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

Data-driven predictive models could be a way forward in IPL team management. Data-driven recommendations could also be developed for player selection. Predictive analytics could seek to pick probable winners and help manage risk better. Analytics bridges the gap between team managers and team coaches. These data insights and quantifications provide precise and timely answers. These compelling charts, reports, and predictive models can be automated for continuous updates by streaming input data.

FUTURE ENHANCEMENTS:

Thus with the help of IPL data we can ensemble the module and by taking player statistics like his average, strike rate and team run rate we can improve the prediction accuracy.To implement a ensemble module with the help of multiple learning algorithm which can be applicable for all type of cricket matches like 20-20, ODI and test cricket also.

REFERENCES

[1] http://en.wikipedia.org/wiki/Cricket

[2] http://www.duckworth-lewis.com/mags/dlmethod/

[3] http://www.espncricinfo.com/

[4] Ananda Bandula Siri, “Predicting the Winner in One Day International Cricket” Journal of Mathematical Sciences & MathematicsEducation.

[5] Tejinder Singh, Vishal Singla and Parteek Bhatia, “Score and Winning Prediction in Cricket through Data Mining” 8 October2015.

[6] Amal Kaluarachchi and Aparna S Varde, ”CricAI: A classification based tool to predict the outcome in ODIcricket.