

Predicting Rising Stars in Indian Cricket

MINI PROJECT

By

Dhruv Talati	60004180022
Naitik Rathod	60004180054
Nishit Mistry	60004180066
Manan Parikh	60004180049

Guide:

Prof. Aniket Kore
Assistant Professor



Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)



University of Mumbai
2020-2021

CERTIFICATE

This is to certify that the mini project entitled **“Predicting Rising Stars in Indian Cricket”** is a bonafide work of **“Dhruv Talati(60004180022), Naitik Rathod(60004180054), Nishit Mistry(60004180066) and Manan Parikh(60004180049)”** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering.

A handwritten signature in black ink, appearing to read 'Aniket Kore', is written over a light gray rectangular background.

Prof. Aniket Kore
Guide

Dr. Meera Narvekar
Head of Department

Dr. Hari Vasudevan
Principal

Mini Project Report Approval

This mini project report entitled *Predicting Rising Stars in Indian Cricket* by *Dhruv Talati, Naitik Rathod, Nishit Mistry and Manan Parikh* is approved for the partial fulfillment of the degree of *B.E. in Computer Engineering*.

Examiners

1.-----



Prof. Aniket Kore

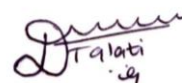
2.-----

Date: 3rd May 2021

Place: Mumbai

Declaration

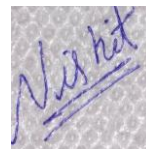
I/We declare that this written submission represents my/our ideas in my/our own words and where others' ideas or words have been included, I/We have adequately cited and referenced the original sources. I/We also declare that I/We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my/our submission. I/We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.



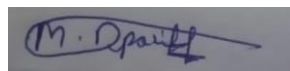
(Dhruv Talati – 60004180022)



(Naitik Rathod - 60004180054)



(Nishit Mistry - 60004180066)



(Manan Parikh - 60004180049)

Date: 3rd May 2021

Abstract

There are immense number of predictions and speculations about cricket ongoing on the internet all the time. These are usually based on how good the player plays in the previous matches. A system which predicts the future scope of a player is exigent. Authors propose a system based on machine learning techniques for predicting the rising stars in cricket for the bowling, batting, and all-rounders domains in Indian Cricket. Data will be collected from various sources and an indigenous database will be created for the project. The preprocessing and analysis based on considerations of different features will be done. Generative machine learning algorithms will be used, and various factors will be taken for making the predictions about the rising stars. Multiple factors will be considered for the predictions as in, number of runs, wickets, 50s, 100s, etc.

The system will predict top 10 Rising Stars in each domain i.e., batting, bowling, and all-rounders. The system will display which current player will be replaced by the future star. There will be a cap on players age and threshold will be set for the rising stars by considering the lowest score from analysis of the current players.

For proof of applicability, the older data will be fed into the system and the predictions made will be compared with the current player rankings. The older data will be taken from espncriinfo datasets. The system will comprise of data from different trophies going on in India, IPL data, and then considering the parameters for the predictions.

Multiple machine learning algorithms like SVM and CART will be employed to check for better accuracies among all. This system will be very useful for selectors who need to predict the player performance and pick the players for a team. Selection process will become easier, and all the data and stats will be available at one place.

Contents

Chapter	Contents	Page No.
1	INTRODUCTION	1-3
	1.1 Description	1
	1.2 Problem Formulation	1
	1.3 Motivation	2
	1.4 Proposed Solution	2
	1.5 Scope of the project	3
2	REVIEW OF LITERATURE	4-11
	2.1 Previous work	4
	2.2 Research Gap	11
3	SYSTEM ANALYSIS	12-14
	3.1 Functional Requirements	12
	3.2 Non-Functional Requirements	12
	3.3 Specific Requirements	13
	3.4 Use-Case Diagrams and description	13
4	ANALYSIS MODELING	15-21
	4.1 Data Modeling	15
	4.2 Activity Diagrams / Class Diagram /Sequence /Collaboration /State	17
5	DESIGN	22-23
	5.1 Architectural Design for proposed system	23
6	IMPLEMENTATION	24-27
	6.1 Algorithms / Methods Used	24
	6.2 Working of the project	24
7	RESULTS AND DISCUSSIONS	28
8	CONCLUSIONS & FUTURE SCOPE	29

List of Figures

Fig. No.	Figure Caption	Page No.
3.1	Use-case diagram of the proposed system	14
4.1	Data Flow Diagram of the proposed system	16
4.2	Class Diagram of the proposed system	17
4.3	State Diagram of the proposed system	18
4.4	Activity Diagram of the proposed system	19
4.5	Sequence Diagram of the proposed system	20
4.6	Collaboration Diagram of the proposed system	21
5.1	Architectural Design of the proposed system	23
6.1	Home Page - i	24
6.2	Home Page - ii	25
6.3	Home Page - iii	25
6.4	Home Page - iv	26
6.5	Login Page	26
6.6	Batting Stats	27
6.7	Bowling Stats	27

List of Abbreviations

Sr. No.	Abbreviation	Expanded form
i	U-19	Under 19
ii	RSs	Rising Stars
iii	BCCI	Board of Control For Cricket In India
iv	BN	Bayesian Network
v	NB	Naïve Bayesian
vi	SVM	Support Vector Machines
vii	CART	Classification and Regression Tree
viii	RSP	Rising Star Prediction
ix	OWA	Ordered Weighted Averaging
x	MEMM	Maximum-Entropy Markov Model
xi	ICC	International Cricket Council
xii	SR	Strike Rate
xiii	HS	Highest Score
xiv	AVG	Average
xv	SNA	Social Network Analysis
xvi	PIB	Performance Index of A Batsman

xvii	ABa	Batting Average of The Batsman Against The Bowler He Faced
xviii	CBo	Career Bowling Average of The Bowler
xix	AHP	Analytic Hierarchy Process
xix	AHP	Analytic Hierarchy Process
xx	ODI	One Day International
xxi	IPL	Indian Premier League
xxii	CSV	Comma Separated Values
xxiii	API	Application programming interface

Chapter 1

Introduction

1.1 Description

Cricket is no longer a sport only played by the elites or the maharajas in India as it was a century ago. Today, in India the sport is played by the masses and is loved by all. No other sport has enjoyed the love and affection that cricket has in India. The complexity of cricket has also increased many folds. Earlier it was just the test matches today there are Test matches, ODIs, Twenty20s and the matches are played internationally, domestically and by various age groups like U-19 etc. In India currently there are many tournaments being played like Indian Premier League, Ranji Trophies, Syed Mushtaq Ali and many more. The data and analysis that is available is mainly on the capped players with very little data analysis done on the uncapped players. This makes selectors task of different teams a cumbersome one. The authors therefore propose a solution to this predicament. The proposed system will be using machine learning algorithms and classification algorithms for prediction of rising stars of cricket. There will also be a classification of the rising stars based on their role in the team as a batsman, bowler, all-rounder or a wicket keeper. This will enable teams to have a data on the players who have performed consistently using the rising star prediction algorithm. The teams can make better informed decisions during the selection process and make valuable addition to their teams accordingly. The players will also be benefitted as they can be assured that a one bad game during which the selectors have come to watch their performance will not be a roadblock in their career because selectors have data on their consistency, their performance with respect to co-players etc.

1.2 Problem Formulation

The current models and systems available are not designed to analyze the data of different tournaments and make accurate predictions of the Rising Stars. Authors here propose to create a system that is capable of handling the live data from different formats so that the predictions made can be close to perfect.

The proposed methodology will not only be beneficial for the selectors who need to pick a player for a team, but it will also be helpful for people to analyze the player's performance and also have their stats along with it.

1.3 Motivation

In India, one could ask anyone what they dreamt of becoming in childhood and most often than not the answer would be “in my childhood I always dreamt of becoming a cricketer”. With the 2nd largest population in the world and sheer amount of people playing cricket it is always a tough task for one to predict the rising stars that would become one of the legends of the game in the future.

Authors aim to build a system for predicting the rising stars of India's favorite sport cricket. With thousands of cricket matches being played every day in a country of India's size it is necessary for franchises and BCCI to have access to data which shows them the rising stars on whom they could keep an eye. There are large number of tournaments like Vijay Hazare, Ranji Trophy, Syed Mushtaq Ali, under-19 matches etc., being played and this data is residing in different locations making it difficult for teams to identify players suited for the slots they are looking to fill in their teams.

1.4 Proposed Solution

The authors aim to provide a system using a machine learning based models for predicting the rising stars of the game with high accuracy to help the selectors build a better team. The system will be based on the live data being fetched and analysis being done on the spot. There is ever increasing data as the player's keep on playing the matches. The results might vary every time the data changes and this can be very difficult to handle if the data is being fed manually. Every time our system is called; data will be checked for any updates and then the further processing will begin. This will ensure the latest live data is being considered and accurate predictions are being made for the RSs. The RSs of all domains will be displayed, along with a module where the current player which can be replaced by the upcoming star will be mentioned too. Stats of all the players will also be visible along with the RSs list so that all the information could be fetched in one place.

1.5 Scope of the project

With the current pace at which the cricket matches are being played, there are numerous amounts of changes that are being done in the statistics and the data available on the player. Having a system which is based on live data is necessary and holds credibility over the models which are based on relatively older data. Due to the ease of use which it will provide having the RSs and the stats in one place, it could be a very informative project that can be of great use to the people involved over the selection process of the players. Similar systems can be used by people from other countries for their tournaments to achieve the RSs of their country.

Chapter 2

Review of literature

2.1 Previous work

Authors in [1] suggest that finding the rising stars (RSs) within the domains is of great importance as the organizations can put efforts in betterment and expertise of the RS. Rising star is an emerging player who could become a star in future based on the consistent performance.

The authors have put forward the concept of co-players who play with the RS as it is an essential factor for the rising star prediction in cricket. Three features are defined namely, co-batsmen, team, opposite team for batsmen and similarly co-bowlers, team, opposite teams for bowlers. Total 9 features for the RS in batsmen and 11 features for RS in bowling domain have been considered. Data used is taken from espncricinfo and processed into two datasets.

Multiple Machine learning algorithms have been used and four most appropriate have been used for binary classification. Generative classifiers outperform the others. The RSP is made with high accuracies and rankings are compared with the ICC rankings of 2013-2016 players. Basic terminologies and concepts of the game of crickets are also elaborated. Among the generative models, Bayesian Network (BN) and Naïve Bayesian (NB) are used. Support Vector Machines (SVM) and Classification and Regression Tree (CART) are used in the discriminative models. For the feature evaluation, authors have used the state-of-the-art evaluators information gain, gain ratio and chi-squared statistics. The weighted average of the batsman and bowler is calculated and then their performance is evaluated. Statistical analysis of each feature is shown for their data. The authors used 10-fold cross validation for training and to validate the classifiers for using their datasets of each domain.

In batting domain, highest accuracy achieved was 87.5%, 87.3%, 84%, 78% using BN, NB, CART, SVM on the first dataset. Even the second dataset shows the highest accuracies of 89%, 88%, 80%, 73% for CART, BN, NB, SVM for the same feature.

In bowling domain, the highest accuracies are 80%, 78.8%, 78.5% and 78.5% by applying SVM, NB, CART and BN using their first dataset and 77.8%, 75%, 74% and 72% accuracies

for BN, CART, SVM and NB classifiers using the other dataset. Category wise and model wise analysis is done, and the rankings of ICC are compared with the predictions.

In [2], two types of classification models are considered to learn the desired predictive function $F^{\text{RS}}(.)$ and two algorithms are chosen for each model category.

It is the first attempt that uses supervised machine learning methods for prediction of rising stars. Four famous algorithms are chosen for binary classification of rising stars, although other ML algorithms may also be used. In this work, famous algorithms are selected from wide collection, based on efficient performance and classification accuracy. A set of eleven features are designed on the basis of content and graph information. This feature combination was not considered for prediction of rising stars in previous research. The performance of recommended algorithms is critically analysed in terms of evaluation metrics and MEMM classifier demonstrates best performance. This novel idea is implemented for rising stars prediction in database domain. It can be implemented for other domains and may be utilized for rising paper prediction.

In this work, the performance of applied classifiers is analysed by Precision, Recall and F1 evaluation metrics. They have mainly used F1 score to examine the effects of different features for rising star classification accuracy and prediction.

The authors in [3] suggest a solution to the problem of choosing a good batsman in twenty20 using a two-stage method for measuring and ranking batting parameters in cricket using ordered weighted averaging (OWA) operator and regression.

Authors have collected the raw data from the cricinfo website.

The first stage measures the performance of players taking five various capabilities such as Highest Score (HS), Average (AVG), Strike Rate (SR), 4s, 6s using OWA operator and then establishes the ranking of parameters using a regression model.

The authors of this paper try to provide a scientific basis to prove that strike rate is an important measure for good batsmen in shorter version of cricket using OWA and regression methods.

It is shown that for different generated OWA weights, corresponding to various uncertainty levels, the ranking obtained for batting parameters is not sensitive to the change of these weights. In OWA operator there are two important measures, the dispersion (or entropy) and the orness.

In this section, authors propose a two-stage OWA-regression method to prioritize batting parameters including, HS, Avg, S/R, 4s, and 6s for 40 players chosen from IPL 4. In the first stage they use the OWA method for finding the performance of batsmen.

The second stage they compute the estimated parameters of the model. The corresponding result shows that the most important batting parameters are S/R, HS, Avg, 4s and 5s in the decreasing order of importance.

For different OWA weights, the ranking of batting parameters does not change. This indicates that the changes of OWA weights, obtained from different orness levels, does not affect the ranking of batting parameters.

On the basis of the mathematical results, they have shown that S/R is the most important parameter to choose a batsman for Twenty20 format.

This research paper can help selectors better identify the good batsman further increasing their chances of winning in the T20 format.

Authors in [4] explore the application of Social Network Analysis (SNA) to rate the players in a team performance. They generate a directed and weighted network of batsmen–bowlers using the player-vs-player information available for Test cricket and ODI cricket. Additionally, they generate a network of batsmen and bowlers based on the dismissal record of batsmen in the history of cricket—Test (1877–2011) and ODI (1971–2011).

They obtain data from the cricinfo website. The website contains the information of proceedings of all Test matches played since 1877 and all ODI matches from 1971 onwards. These include the runs scored by batsmen, wickets taken by bowlers, outcome of a game and also the information of the mode of dismissal of a batsman. They collect the data of player-vs-player for Test cricket (2001–2011), and ODI cricket (1999–2011) from the cricinfo website. They have also collected the batting and bowling average of the player

The performance of a batsman is judged by the ‘quality’ of runs scored and not the number of runs scored. Hence, runs scored against a bowler with a lower bowling average carry more credit than runs scored against a bowler of less importance. Authors introduce a performance index of a batsman (PIB) against a bowler given by the following equation

$$PIB = ABa/CBo$$

where ABa is the batting average of the batsman against the bowler he faced and CBo refers to the career bowling average of the bowler.

In this paper [5] authors have used a concept of co-players is introduced for predicting the rising stars for better team selection. Analysis for selecting the best features is done using machine learning techniques, the actual prediction of top 10 rising stars is done.

In this work, authors have defined a concept of co-players, i.e., a player is evaluated considering not only his own performance but also of other players he has played in a common time span. Team performances and opposite team performances are also considered. The authors have also introduced the concept of allrounders which will have its own set of features.

The data is taken from Espncricinfo for years 2006–2018 for all three domains, i.e., batsman, bowler, and all-rounder.

Authors have used the Support vector machine (SVM) which is a discriminative classifier trained for classification and analysis of data based on standard data sets and it is said to be a supervised learning model, where data is plotted in space as a point.

The authors have used features like co-batsmen runs, co-batsmen strike rate and other.

Features for the batting domain. Similarly, they have defined the features for bowling domain and allrounder domain.

To evaluate a player as a rising star or not a rising star, the player who had played at least 20 matches is considered. SVM model is trained, and further applicable rising stars are noted. And the final list of rising stars is derived, according to their RS score. The result obtained based on computation on features states player ability depends not only on his performance but also on performance of other players and team which ensures a good evaluation system. The list was compared by the authors with ICC and accuracy obtained as: 60% for batting domain, bowling accuracy: 70%, all-rounder accuracy:40%.

The authors in [6] have discussed the existing methods to find the rising stars and the pros and cons of the methods are discussed. The datasets and the evaluation of the performances are described too. Open challenges and the future scope is discussed towards the end.

Methods for finding Rising stars in Bayesian Networks are sub categorized into four: Ranking methods, Prediction Methods, Clustering Methods, Analysis Methods. Earlier published papers using these methods are discussed and their improvements done later as discussed. Findings and

Limitations of multiple publications and their dataset are stated and the scope of improvement among those is described. Basically, the authors state that collaborations with renowned researchers can also lead to a rising future. Similarly, all the methods in Bayesian networks are discussed, their earlier works, best works and future scopes are mentioned. Applications of finding rising stars in other domains like community question answering networks, sports networks, telecommunication networks are elaborated with relevant works in those fields. Datasets of academic networks and miscellaneous networks are specified with the main features of all the datasets.

Challenges like falsely predicted rising stars, using multiple data sources, applications in multiple domains, long term prediction impacts and many other challenges and future directions are discussed in brief.

The authors in [7] obtained all the data from www.cricinfo.com using scraping tools, parsehub and import.io. For batting, they considered matches played from January 14, 2005 to July 10, 2017. The senior most player during this span was SR Tendulkar, so they collected innings by innings list of the performance of all the batsmen from December 18, 1989 when he played his first ODI match. For bowling, they considered matches played from January 2, 2000 to July 10, 2017. The senior most player during this span was PA de Silva, so they collected innings by innings list of the performance of all the batsmen from March 31, 1984 when he played his first ODI match. Since the past stats of the players such as average, strike rate etc. are not available directly online for each match they played, they calculated from the innings by innings list for each match. They imported all the data in MySQL tables and used php to manipulate them.

For predictive analytics, they used Weka and Dataiku. Both these tools are a collection of machine learning algorithms for data mining and also provide some pre-processing functionalities.

They have made various attributes i.e., for Batting attributes:

- No of innings: The number of innings the player has played till the day of that match, which shows the experience of the player
- Batting Avg: This attribute depicts the run scoring capability of that player
- Strike Rate: This attribute depicts how fast the player can score runs.
- Other attributes included the centuries, fifties, zeros and the highest score of that player

Bowling attributes:

- No of innings: the total number of matches the player has played till the day of the match
- Overs: Total number of overs bowled by the bowler, for experience
- Bowling Avg: Bowling average is the number of runs conceded by a bowler per wicket taken.
- Bowling Strike Rate: Bowling strike rate is the number of balls bowled per wicket taken.
- Four/Five Wicket Haul: Number of innings in which the bowler has taken more than four wickets.

Derived attributes:

- Consistency: This attribute describes how experienced the player is and how consistent he has been throughout his career.
- Form: Form of a player describes his performance over last one year. All the traditional attributes used in this formula are calculated over the matches played by the player in last 12 months from the day of the match.
- Opposition: Opposition describes a player's performance against a particular team. All the traditional attributes used in this formula are calculated over all the matches played by the player against the opposition team in his entire career till the day of the match.
- Venue: Venue describes a player's performance at a particular venue. All the traditional attributes used in this formula are calculated over all the matches played by the player at the venue in his entire career till the day of the match.

This paper [8] is centred on the implementation of machine learning to foretell the winner of an IPL match. The historical dataset was obtained from various sources like Espncricinfo and IPLt20. Feature engineering techniques were applied by authors to derive more insights about the current dataset. Most of the Machine Learning algorithms work better with numerical values than the string values. Hence all the string formats in the dataset were converted to the numerical formats utilizing the Label Encoding by the authors.

To produce accurate results, all the unnecessary features from the dataset were eliminated eg- Umpire Name, StadiumName etc.

To rule out the class imbalance authors have designed the model to predict the winner based on the essential features instead of the Team names, declaring either Team 1 or Team 2 as a winner. Authors have considered the data of only 11 players for a team based on the highest number of matches they have played during the IPL was considered.

Authors have referred to the DREAM 11 points table like total score of bowlers etc to derive the formula.

Authors have weighted the features according to their relative importance over other measures eg-For Batting Analytic Hierarchy Process (AHP) the attributes were arranged by the authors in their decreasing order of importance: -

Batting Average > Innings > Strike Rate > 50's > 100's > 0's

Similarly Bowling AHP and the yearly ranks of each team based on the win ratios was noted by the authors and the ranks were derived using AHP.

Teams' ranking was done according to the teams' points, and past performance features were fed to the model for predicting the results.

The consistency of a team adds more weightage to its current performance than the overall performance and hence authors have assigned 80 percent weightage to the current performance of a team and 20 percent weightage to their overall performance.

A machine learning model is asymmetric in nature and is neither capable of identifying the symmetry of features nor has a way to input the information about the symmetry of features. Hence, this information was entered to the model by generating a symmetric duplicate for every row in the training set by the authors.

Several machine learning models were applied by the authors to the selected features to predict the IPL match results. The best results were concluded using the tree-based classifiers. The highest accuracy of 60.043% with Random Forest with a standard deviation of 6.3% and an ambiguity of 1.4% was observed by them.

The research by authors focused on predicting the winner for an IPL match using machine learning and utilized the available historical data of IPL from season 2008-2019.

The authors in [9] made initial step involved in developing the system to pre-process the raw data. The raw data pertaining to players of countries are fetched from espnricinfo portal. Also, the statistical information pertaining to the players are fetched from espnricinfo, statsguru repository. In this work, the data mining algorithms used are k-means clustering, decision trees, random forest and support vector machines. The software used to perform data mining operations over the selected dataset is WEKA

The data fetched from espnricinfo portal is populated into an excel file and the CSV loader is used to feed the data into the knowledge system. The dataset is then passed on to the class assigner where the typical classes are provided to the data. The class labelled data is then passed

to the cross-fold validation to create training and testing data. The test set and training set is then fed to the classifier algorithms and the results are observed. The classifier block is changed according to the type of classification algorithm that is to be used.

The precision values for batsman, bowler and all-rounder are 0.94, 0.95 and 0.81 respectively. The recall values for batsman, bowler and all-rounder are 0.93, 0.92 and 0.81 respectively. The prediction accuracy of the classifier is 91.87% when decision tree classification is used, 93.46% when SVM is used and 95.78% when random forest used.

The authors in [10] have used the cricket players dataset collected over the past 10 years from a website and the main outcome of the system is to predict if the team will win the match or not. Random forest model is used as it gave the best results for the authors as compared to the other algorithms. Various metrics were used for performance testing of the players like weather conditions, previous scores, maiden overs, etc.

2.2 Research Gap

All the previous work done in this domain has been done for the rising stars in the international players only. There is no study where the upcoming players/ uncapped players of any country are analyzed for the RSs. Authors in [5] have considered the current players and compared the RSs with the current ICC rankings, while the relatively newer players are not considered. The authors here propose to predict the RSs considering the players performances in the Vijay Hazare trophy, IPL, Ranji Games. Upcoming players capable of being a rising star will be predicted and can be used for the selection purposes.

Chapter 3

System analysis

3.1 Functional Requirements

Get Cricket Data - download, filter, and store the required data in the local database. Structured data from apis like cricapi and espnricinfo will be used. System should be able to save the records of players and store it in the local database for analysis.

Analysis Strategy - specify new equations that will consider the players performance in all different formats played. The equation will calculate the feature score for all the players. The minimum score among the capped players will be set as the threshold for predicting the rising stars among the upcoming players.

Displaying RSs and player stats: For each player, obtain data from various reliable sources. Analyze the data and display the RSs and their stats. This will make it easier for selectors as all the data will be available at one place.

3.2 Non-Functional Requirements

The non-functional requirements of the system are explained below as performance requirements and design constraints.

1. Performance requirements:

- a) Accuracy - Since we will give the priority to the accuracy of the software, the performance of the Prediction System will be better and accurate results will be obtained
- b) Openness - The system should be extensible to guarantee that it is useful for a reasonable period. The live data being fetched will ensure that the system is working until major changes occur in the formats of cricket being played.
- c) Reliability - Data is being fetched from reliable sources such as Espnricinfo and cricapi. This will have lesser errors and accurate predictions will be made for the RSs.

2. Design constraints:

- a) Hardware Constraints - The system will be integrated with a web application. To use the recommendation system, the user should enter from a personal computer or access website from mobile where the predictions and the stats will be displayed.
- b) Software System Attributes
 - Usability - The software will be embedded in the backend of an application. It should be scalable designed to be easily adopted by a system.
 - Reliability - The system should have accurate results and fast responses when user checks for the RSs.

3.3 Specific Requirements

Player data in local database used for training the models should be secured and no manipulation should be done to the data after the model is trained as it will lead to faulty results.

Data obtained from various sources should be stored in similar manner for faster training and prediction purposes.

3.4 Use-Case Diagrams and description

A UML use case diagram is the primary form of system/software requirements for a new software program underdeveloped. Use cases specify the expected behavior (what), and not the exact method of making it happen (how). Use cases once specified can be denoted both textual and visual representation (i.e., use case diagram). A key concept of use case modeling is that it helps us design a system from the end user's perspective. It is an effective technique for communicating system behavior in the user's terms by specifying all externally visible system behavior.

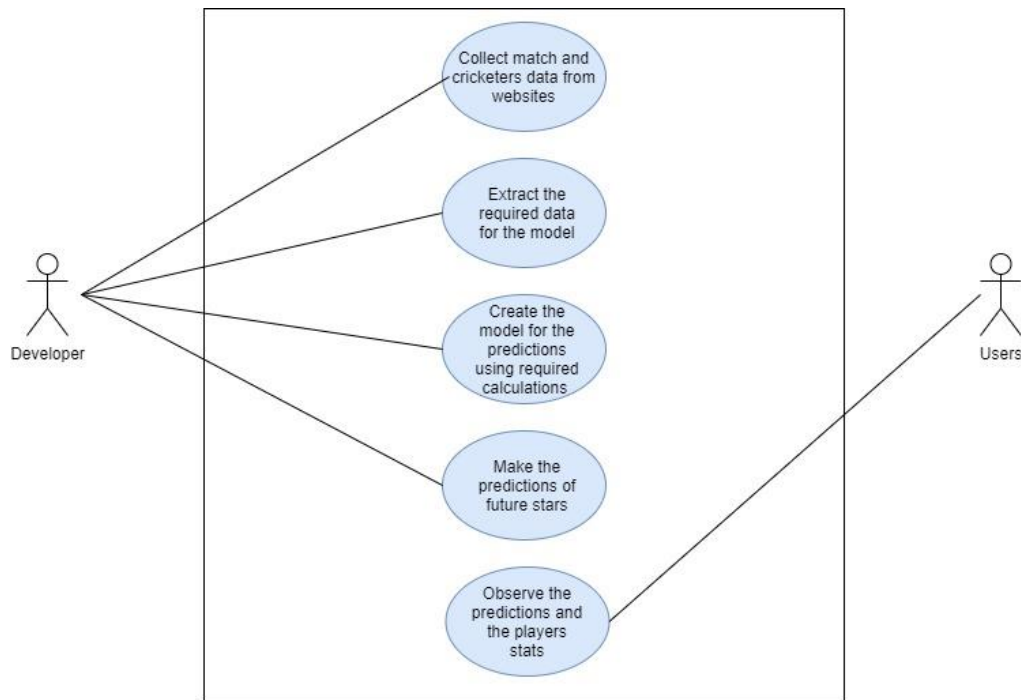


Fig 3.1 Use-case diagram of the proposed system

1. The initial players data is fetched from the reliable sites.
2. The model analyzes the data and calculates the feature scores of the players.
3. The scores are compared with the threshold set from the current players' scores.
4. Model makes its predictions.
5. The user gets the RSs list displayed with the stats of the players too.

Chapter 4

Analysis Modeling

4.1 Data Modeling

Data modeling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures. The goal is to illustrate the types of data used and stored within the system, the relationships among these data types, the ways the data can be grouped and organized and its formats and attributes. Data can be modeled at various levels of abstraction.

The DFD diagram is displayed below for the proposed system.

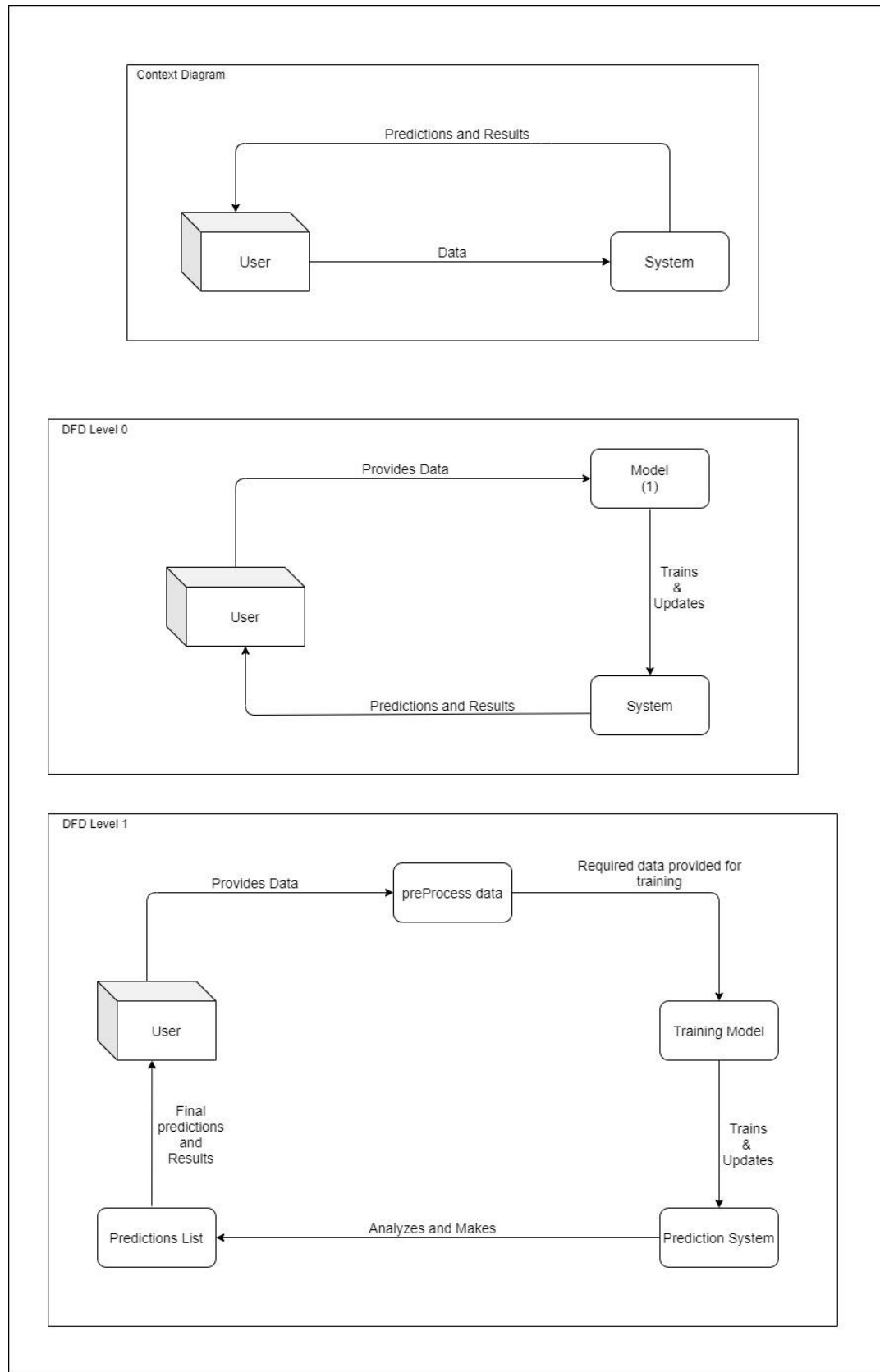


Fig 4.1 Data Flow Diagram of the proposed system

4.2 Activity Diagrams / Class Diagram / Sequence / Collaboration / State

4.2.1 Class Diagram

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application.

Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modelling of object-oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram.

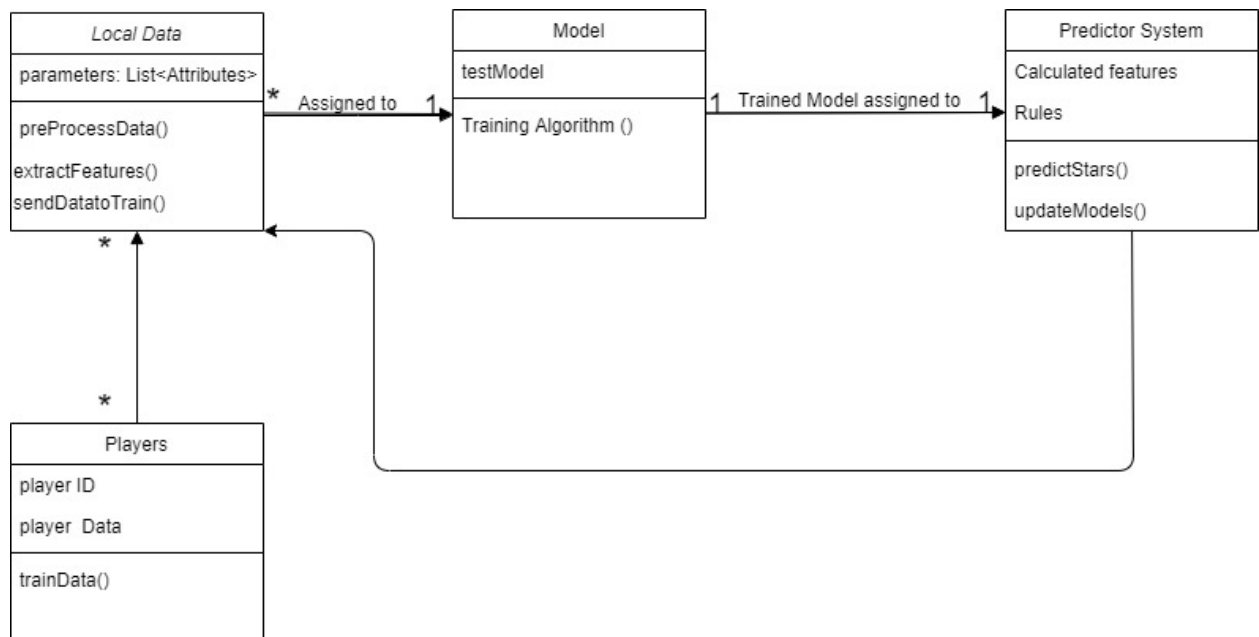


Fig 4.2 Class diagram of the proposed system

4.2.2 State Diagram

A state diagram is used to represent the condition of the system or part of the system at finite instances of time. It's a behavioral diagram and it represents the behavior using finite state transitions. State diagrams are also referred to as State machines and State-chart Diagrams. These terms are often used interchangeably. So simply, a state diagram is used to model the dynamic behavior of a class in response to time and changing external stimuli. We can say that each and every class has a state but we don't model every class using State diagrams. We prefer to model the states with three or more states.

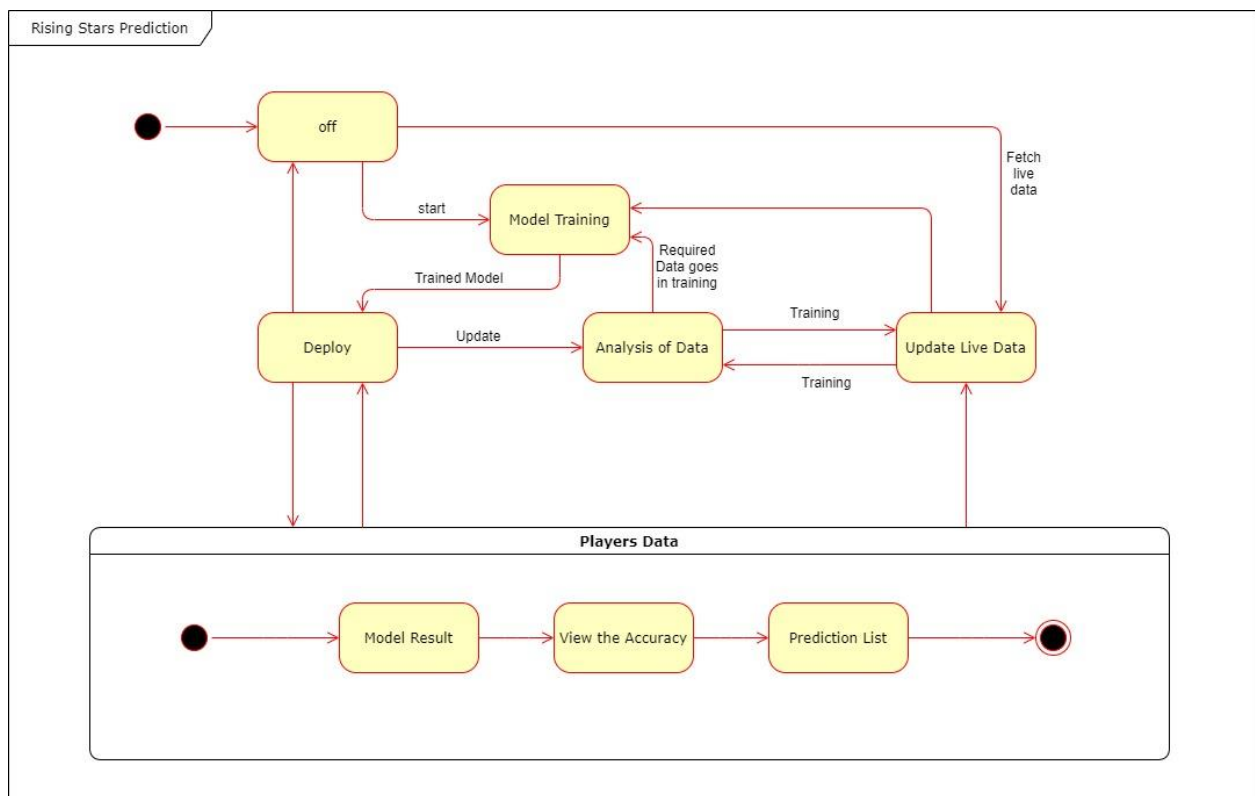


Fig 4.2 State diagram of the proposed system

4.2.3 Activity Diagram

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent.

Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc.

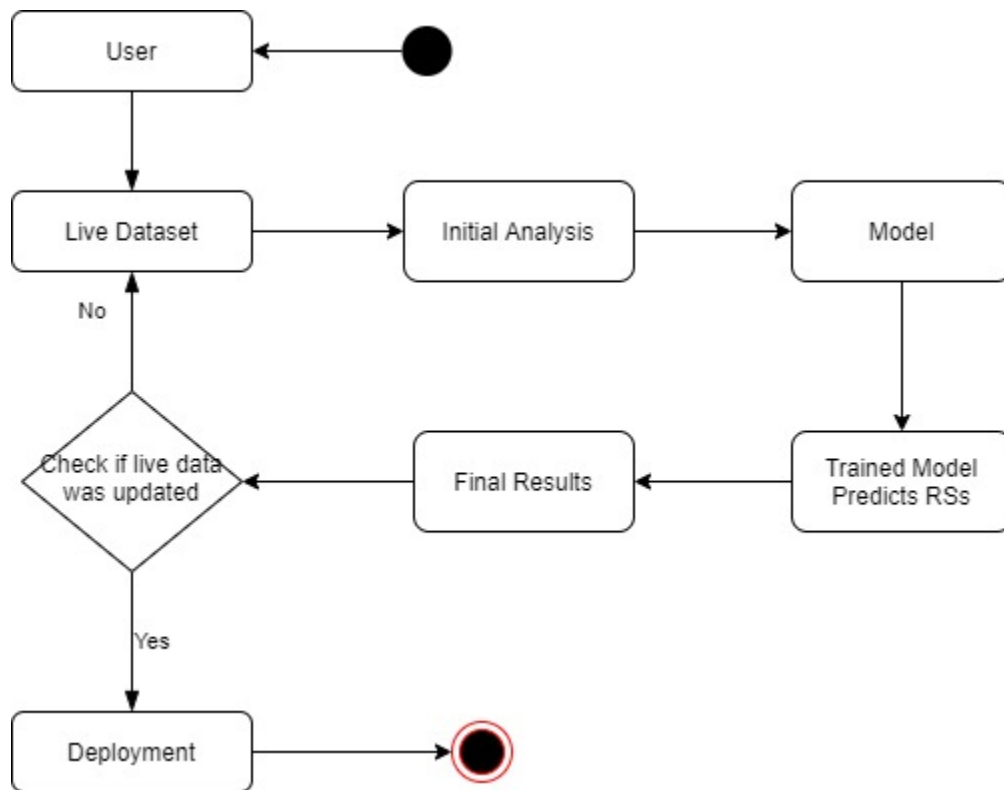


Fig 4.3 Activity diagram of the proposed system

4.2.4 Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

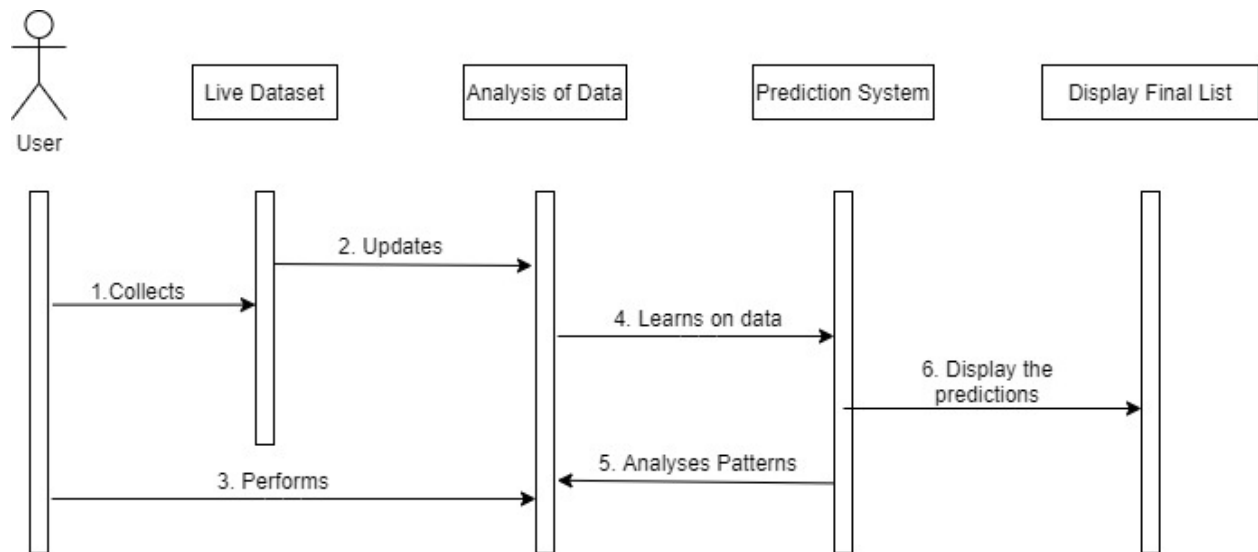


Fig 4.4 Sequence diagram of the proposed system

4.2.5 Collaboration Diagram

The collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system.

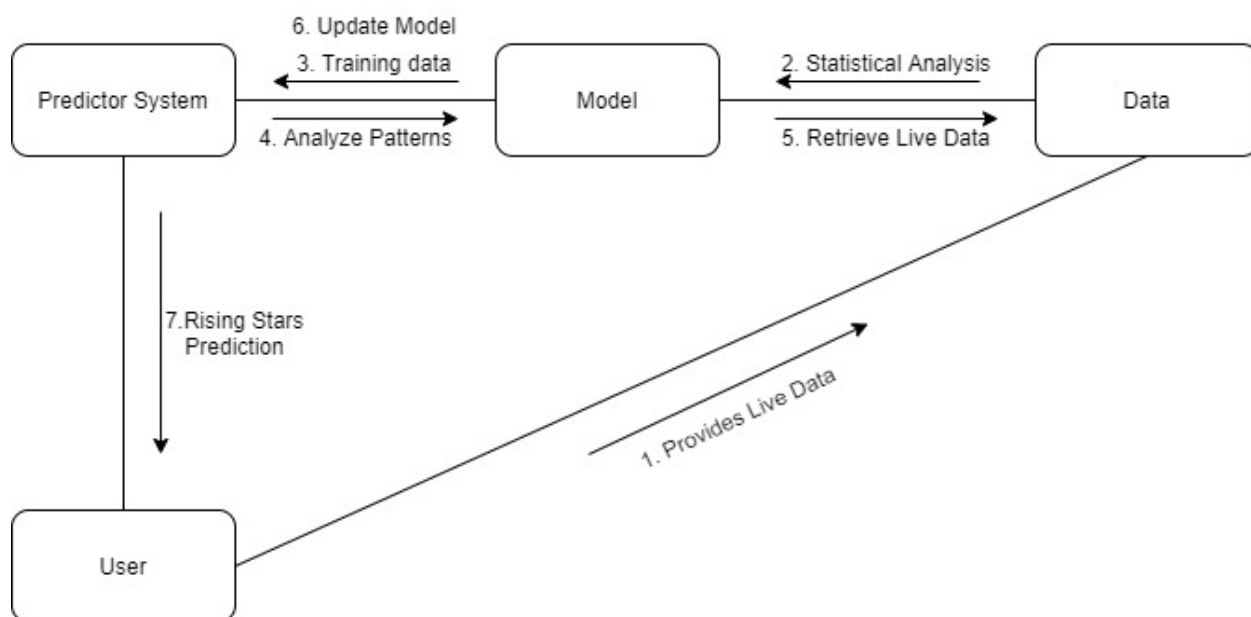


Fig 4.5 Collaboration diagram of the proposed system

Chapter 5

Design

5.1 Architectural Design for proposed system

The architectural design of our proposed system would represent the software needs and design of the system. IEEE defines architectural design as “the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system.” The software that is built for computer-based systems can exhibit one of these many architectural styles.

Each style will describe a system category that consists of:

1. A set of components (for example: a database, computational modules) that will perform a function required by the system.
2. The set of connectors will help in coordination, communication, and cooperation between the components.
3. Conditions that how components can be integrated to form the system.
4. Semantic models that help the designer to understand the overall properties of the system.

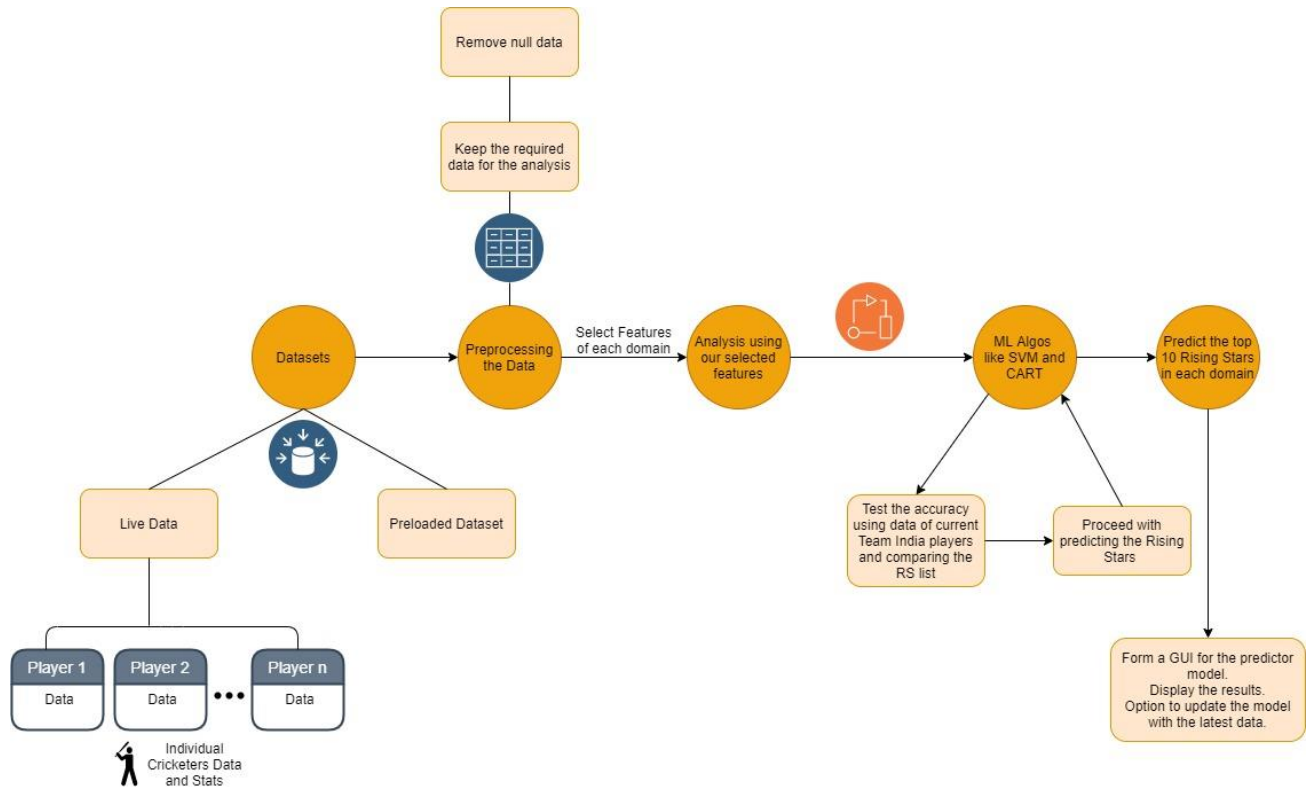


Fig 5.1 Architectural design for the proposed system

Chapter 6

Implementation

6.1 Algorithms / Methods Used

Currently, no algorithms were used in the implementation. The Player's data is being fetched from the cricapi and being displayed in the webpage.

The player's stats including the batting and bowling in all the formats is fetched and displayed currently. The data analysis and the analysis based on different features will be done further, with the model training and testing and the prediction part.

6.2 Working of the project

The system displays some of the players pictures and their famous quotes on the home page. There is a login page for the user to start with. On clicking the player, the complete batting and bowling stats are visible which are being fetched from cricapi. The live data is being fetched and displayed. The further analysis and predictor models will be developed in the upcoming time.

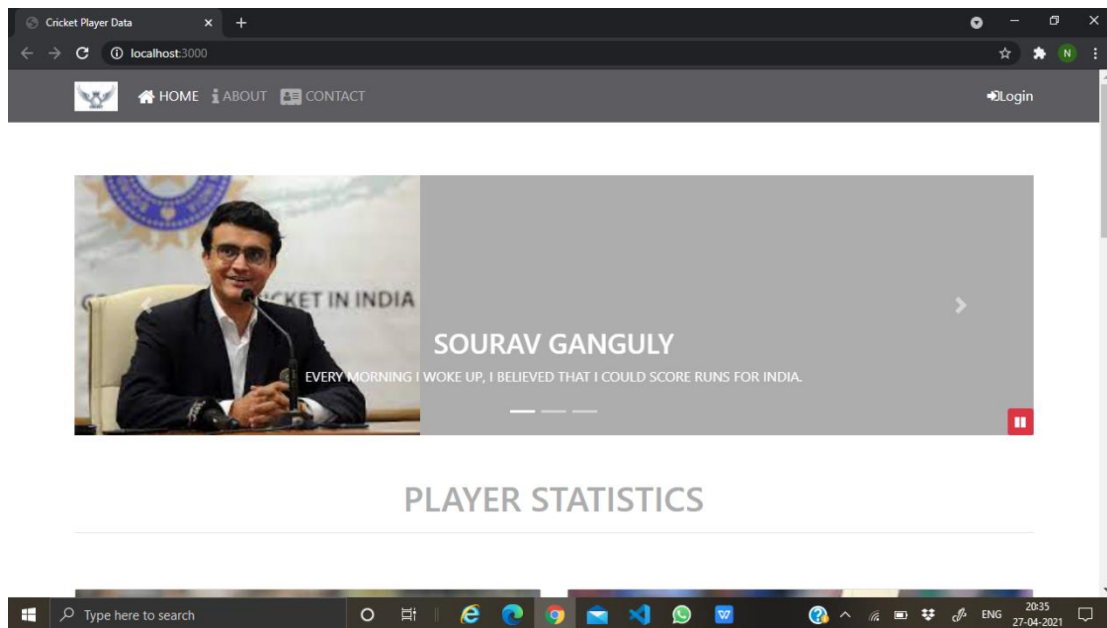


Fig 6.1 Home Page-i

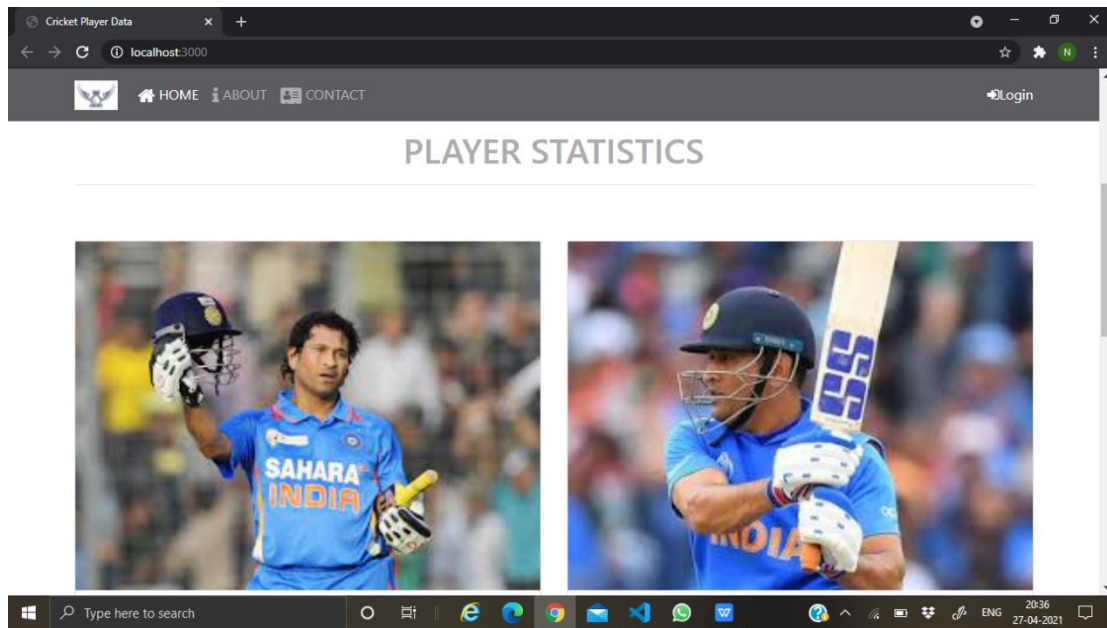


Fig 6.2 Home Page-ii

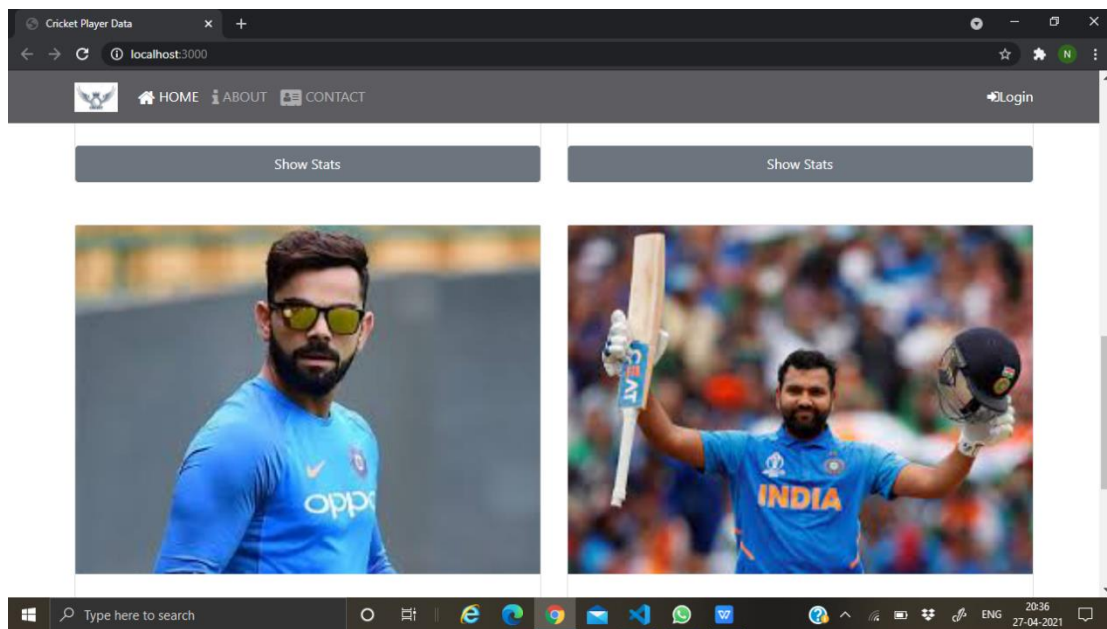


Fig 6.3 Home Page-iii

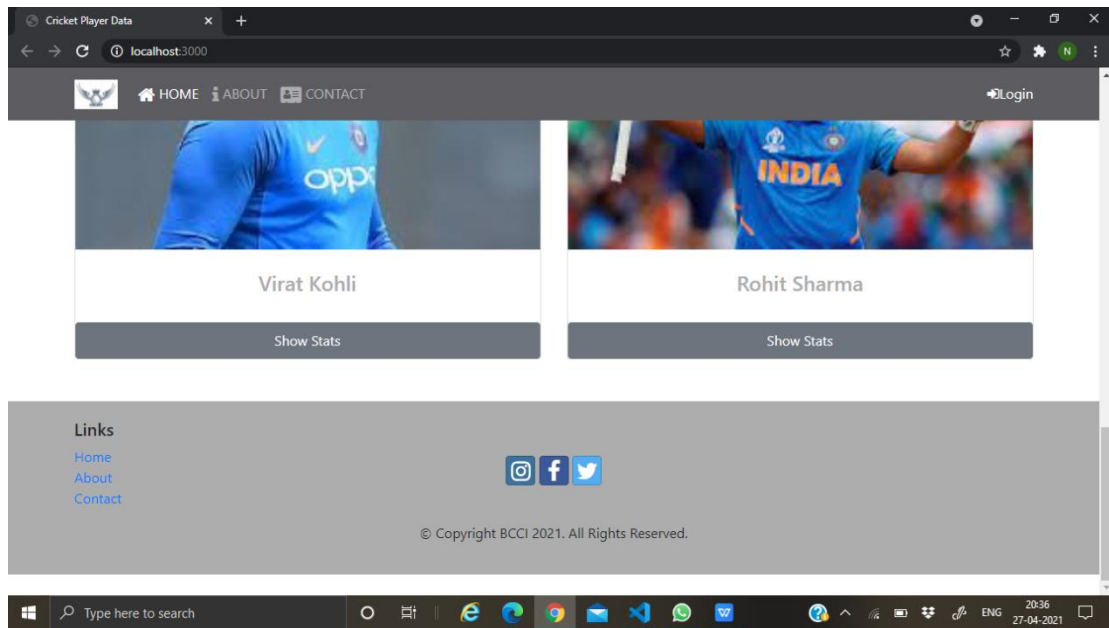


Fig 6.4 Home Page-iv

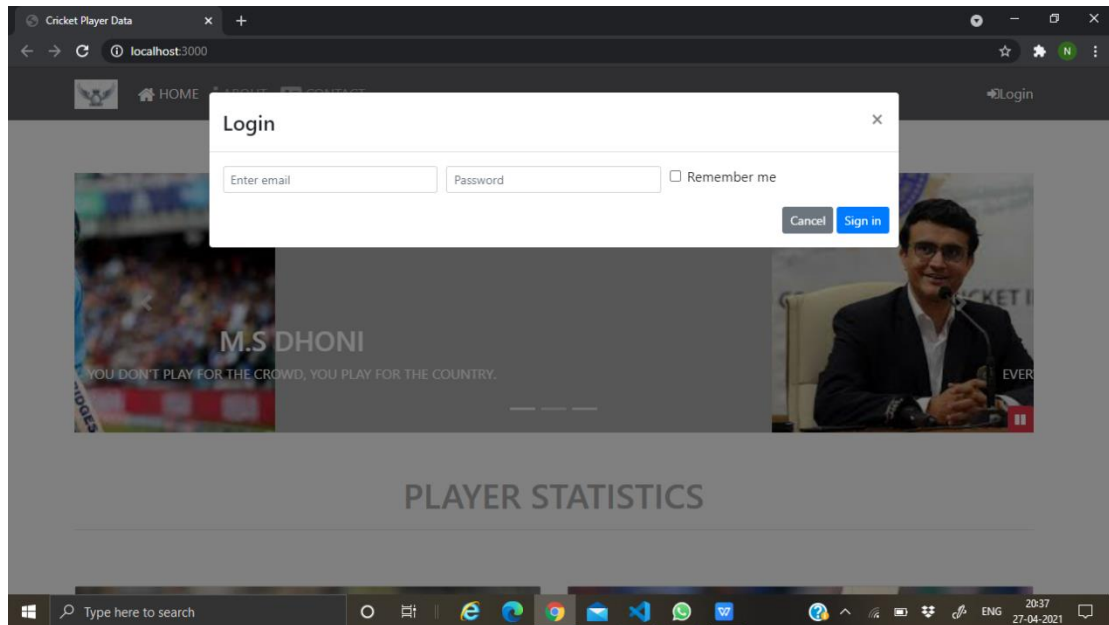


Fig 6.5 Login Page

Cricket Player Data

localhost:3000

HOME

Login

Players's Stats

Batting

MatchType	50	100	St	Ct	6s	4s	SR	BF	Ave	HS	Runs	NO
listA	114	60	0	175	NA	NA	NA	NA	45.54	200*	21999	55
firstClass	116	81	0	186	NA	NA	NA	NA	57.84	248*	25396	51
T20Is	0	0	0	1	0	2	83.33	12	10.00	10	10	0
ODIs	96	49	0	140	195	2016	86.23	21367	44.83	200*	18426	41
tests	68	51	0	115	69	NA	NA	NA	53.78	248*	15921	33

Bowling

MatchType	10	5w	4w	SR	Econ	Ave	BBM	BBI	Wkts	Runs	Balls	Ini
-----------	----	----	----	----	------	-----	-----	-----	------	------	-------	-----

Fig 6.6 Batting Stats

Cricket Player Data

localhost:3000

HOME

Login

Players's Stats

Batting

T20Is	0	0	0	1	0	2	83.33	12	10.00	10	10	0
ODIs	96	49	0	140	195	2016	86.23	21367	44.83	200*	18426	41
tests	68	51	0	115	69	NA	NA	NA	53.78	248*	15921	33

Bowling

MatchType	10	5w	4w	SR	Econ	Ave	BBM	BBI	Wkts	Runs	Balls	Ini
listA	0	2	4	50.8	4.97	42.17	5/32	5/32	201	8478	10230	N/A
firstClass	0	0	NA	107.1	3.45	61.74	NA	3/10	71	4384	7605	N/A
T20Is	0	0	0	15.0	4.80	12.00	1/12	1/12	1	12	15	1
ODIs	0	2	4	52.2	5.10	44.48	5/32	5/32	154	6850	8054	27
tests	0	0	0	92.1	3.52	54.17	3/14	3/10	46	2492	4240	14

Fig 6.7 Bowling Stats

Chapter 7

Results and Discussions

Currently the GUI is implemented where the data is fetched and displayed. This whole data will be used in the analysis and the training of the predictor models and for the prediction system altogether which will be further implemented. The model will be based on our own formula to consider different tournaments and different parameters altogether. The Rising Stars of the Indian Cricket will be displayed in all the domains i.e., batting, balling and all-rounders.

The screenshots of the data being fetched and displayed are attached in the working. This whole players data will be then processed and then sent in for the further analysis. Model will be designed in such a way that data of different formats will be considered altogether, and accurate predictions will be made.

Chapter 8

Conclusions and Future Scope

The analysis and the prediction part will be done in further development. Accuracies will be tested, and predictions will be made.

Similar systems can be implemented by different countries for their upcoming stars using data from IPL, Big Bash League etc.

Further modules can be added in the system where an optimum team could be formed with the data from the RSs predictions. This can be possible with hybrid models and further work on the selection process.

Appendix

- Espncricinfo dataset: The dataset contains more than thousand professional cricket players all over the world through all timeline scraped from the ESPN official website. Contains 181 columns with unique ID, Description for some and all batting bowling scores from Tests, ODI, List A, First_Class, T20I, T20. There are over 180 useful feature points of every player however there are a lot of missing data as well in some of the players.

This dataset will be used for the older data.

- CricAPI: This api is used for fetching the live data of the players. All the stats of the players and their pictures as displayed on the homepage are being fetched from this api. All the data from cric api is the live and most recent data, with all the matches in consideration.

Literature Cited

- [1] H. Ahmad, A. Daud, L. Wang, H. Hong, H. Dawood and Y. Yang, "Prediction of Rising Stars in the Game of Cricket," in *IEEE Access*, vol. 5, pp. 4104-4124, 2017, doi: 10.1109/ACCESS.2017.2682162.
- [2] Daud, A., Ahmad, M., Malik, M.S.I. *et al.* Using machine learning techniques for rising star prediction in co-author network. *Scientometrics* **102**, 1687–1711 (2015). <https://doi.org/10.1007/s11192-014-1455-8>
- [3] Gholam R. Amin, Sujeet Kumar Sharma, Measuring batting parameters in cricket: A two-stage regression-OWA method, *Measurement*, Volume 53, 2014, Pages 56-61, ISSN 0263-2241, <https://doi.org/10.1016/j.measurement.2014.03.029>.
- [4] Satyam Mukherjee, Quantifying individual performance in Cricket — A network analysis of batsmen and bowlers, *Physica A: Statistical Mechanics and its Applications*, Volume 393, 2014, Pages 624-637, ISSN 0378-4371, <https://doi.org/10.1016/j.physa.2013.09.027>
- [5] Khot A., Shinde A., Magdum A. (2021) Rising Star Evaluation Using Statistical Analysis in Cricket. In: Tripathy A., Sarkar M., Sahoo J., Li KC., Chinara S. (eds) *Advances in Distributed Computing and Machine Learning. Lecture Notes in Networks and Systems*, vol 127. Springer, Singapore. https://doi.org/10.1007/978-981-15-4218-3_31
- [6] Daud, A., Song, M., Hayat, M.K. *et al.* Finding rising stars in bibliometric networks. *Scientometrics* **124**, 633–661 (2020). <https://doi.org/10.1007/s11192-020-03466-w>
- [7] Passi, Kalpdrum & Pandey, Niravkumar. (2018). Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning. 111-126. 10.5121/csit.2018.80310.
- [8] Tripathi A, Islam R, Khandor V, Murugan V. (2020). Prediction of IPL matches using Machine Learning while tackling ambiguity in results. *Indian Journal of Science and Technology*. 13(38): 4013-4035. <https://doi.org/10.17485/IJST/v13i38.1649>
- [9] A. Balasundaram, S. Ashokkumar, D. Jayashree and S. Magesh Kumar, "Data mining based Classification of Players in Game of Cricket," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 271-275, doi: 10.1109/ICOSEC49089.2020.9215413.
- [10] Aleemulla Khan P., Thirupathi Rao N., Bhattacharyya D. (2020) Prediction of Cricket Players Performance Using Machine Learning. In: Fiaidhi J., Bhattacharyya D., Rao N. (eds) *Smart Technologies in Data Science and Communication. Lecture Notes in Networks and Systems*, vol 105. Springer, Singapore. https://doi.org/10.1007/978-981-15-2407-3_20

Acknowledgements

We would like to express our special thanks of gratitude to our project guide Asst Prof. Aniket Kore for their able guidance, support and suggestions which helped us in completing this project. We would also like to extend our gratitude to our Principal, Dr. Hari Vasudevan and the Head of the Computer Engineering Department, Dr. Meera Narvekar for providing us with all the facility that were required for completion of this project.