***A PROJECT ON***

# “Cricket Score Prediction”

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF

DIPLOMA IN BIG DATA ANALYSIS



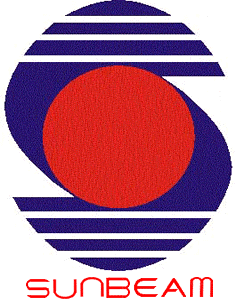
**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

Submitted By:

Prajwal Gursale (80246)

Uday Narkhede (80244)

**Mr.Nitin Kudale Mrs.Manisha Hingne** Centre Coordinator Course Coordinator



**CERTIFICATE**

This is to certify that the project work under the title ‘Cricket Score Prediction’ is done by Prajwal Gursale & Uday Narkhede in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P Mrs. Manisha Hingne

**Project Guide** **Course Coordinator**

Date:

# ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Prajwal Gursale

DBDA Sep 2023 Batch, SIIT Pune

Uday Narkhede

DBDA Sep 2023 Batch, SIIT Pune

.

**TABLE OF CONTENTS**

**Chapter 1**

**PREAMBLE 1**

1.1 Introduction

1.2 Plan of Implementation

1.3 Problem Statement

1.4 Objective of the Program

**Chapter 2**

APPROACH AND DESIGN

2.1 Data Collection

2.2 Data Processing

2.3 Data Visualization

2.4 Model Development and Evaluation

**Chapter 3**

SYSTEM DESIGN

3.1 System Development Methodology

**Chapter 4**

IMPLEMENTATION

**Chapter 5**

RESULTS

**Chapter 6**

GUI

GitHub link

**Chapter 7**

FUTURE SCOPE AND CONCLUSION

**REFERENCES**

**Chapter 1**

**PREAMBLE**

**1.1 Introduction**

Machine Learning is a branch of Artificial Intelligence that aims at solving real-life engineering problems. This technique requires no programming, whereas it depends on only data learning where the machine learns from pre-existing data and predicts the result accordingly. Machine Learning methods have benefit of using decision trees, heuristic learning, knowledge acquisition, and mathematical models. It thus provides controllability, observability, stability and effectiveness.

Cricket is being played in many countries around the world. There are a lot of domestic and international cricket tournaments being held in many countries. The cricket game has various forms such as Test Matches, Twenty20 Internationals, Internationals one day, etc. IPL is also one of them, and has great popularity among them. It's a twenty-20 cricket game league played to inspire young and talented players in India. The league was conducted annually in March, April or May and has a huge fan base among India. There are eight teams which represent eight cities which are chosen from an auction. These teams compete against each other for the trophy. The whole match depends on the luck for the team, player’s performance and lot more parameters that will be taken in to the consideration. The match that is played before the day is also will make a change in the prediction. The stakeholders are much more benefited due to the huge popularity and the huge presence of people at the venue. The accuracy of a data depends on the size of the data we take for analysing and the records that are taken for predicting the outcome.

Cricket is a game played between two teams comprising of 11 players in each team. The result is either a win, loss or a tie. However, sometimes due to bad weather conditions the game is also washed out as Cricket is a game which cannot be played in rain. Moreover, this game is also extremely unpredictable because at every stage of the game the momentum shifts to one of the teams between the two. A lot of times the result gets decided on the last ball of the match where the game gets really close. Considering all these unpredictable scenarios of this unpredictable game, there is a huge interest among the spectators to do some prediction either at the start of the game or during the game. Many spectators also play betting games to win money.

**1.2 Plan of Implementation**

The project can be broken down into 7 main steps which are as follows:

1. Understand the dataset.

2. Clean the data.

3. Analyse the candidate columns to be Features.

4. Process the features as required by the model/algorithm.

5. Train the model/algorithm on training data.

6. Test the model/algorithm on testing data.

7. Tune the model/algorithm for higher accuracy.

**1.3 Problem Statement**

To predict the results of an IPL match using machine learning techniques or algorithms such as Logistic Regression, Gaussian Naive Bayes, K Nearest Neighbours, SVM, Gradient boost algorithm, Decision tree and Random forest.

We have used 17 features which are as follows: season, city, date, team1, team2, toss\_winner, toss\_decision, result, dl\_applied, winner, win\_by\_runs, win\_by\_wickets, player\_of\_match, venue, umpire1, umpire2 and umpire3

**1.4 Objective of the Project**

The main objective of this project is to give the team players information about how each venue makes a difference to the game. And give feedback of how the players can improve their own performance in each game. And also give have a better planning of how the match should be played overall by the whole team regardless of the toss decision.

**Chapter 3**

**APPROACH AND DESIGN**

**3.1 Data Collection**

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data, we collect to develop practical machine learning solutions.

Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. From those patterns, you build predictive models using machine learning algorithms that look for trends and predict future changes.

The Indian Premier League's official website is the principal basis of data for this project. The data was web scrapped from the website and kept in the appropriate format using a python library called beautiful soup. The dataset has the columns regarding match-number, IPL season year, the place where match has been held and the stadium name, the match winner details, participating teams, the margin of winning and the umpire details, player of the match etc. Indian Premier League was only 11 years old, which is why, after the pre-processing, only 577 matches were available. Here, some of the columns may contain null values and some of the attributes may not be required for match winner prediction which is discussed in data preprocessing.

**3.2 Data Preprocessing**

**3.2.1 Data cleaning**

There are some null values in the dataset in the columns such as winner, city, venue etc. Due to the presence of these null values, the classification cannot be done accurately. So, we tried to replace the null values in different columns with dummy values.

**3.2.2 Choosing Required Attributes**

This step is the main part where we can eliminate some columns of the dataset that are not useful for the estimation of match winning team. This is estimated using feature importance. The considered attributes have the following feature importance.

**3.3 Data Visualization **

The data which has been collected is used for visualizing for the better understanding of the information. Matplotlib Library is used here for visualizing the graphs  The data visualization is necessary to understand the solution in a better way. The below graphs were drawn based up on the previous seasons of the IPL matches.

**3.4 Model Development and Evaluation**

Here, we have developed a generic model and applied all classification methods. The data is split into training data and test data, we train the model using certain features and use it to predict the testing data, then we calculate the performance of the system.

The various classification models used are:

Logistic Regression, Gaussian Naïve Bayes Classifier, KNN (K Nearest Neighbor) algorithm, Support Vector Machines, Gradient Boost Algorithm, Decision Trees and Random Forest Classifier. Among these methods the Random Forest and Decision tree has given good results

**Chapter 5**

**SYSTEM DESIGN**

Design is a meaningful engineering representation of something that is to be built. It is the most crucial phase in the developments of a system. Software design is a process through which the requirements are translated into a representation of software. Design is a place where design is fostered in software Engineering. Based on the user requirements and the detailed analysis of the existing system, the new system must be designed. This is the phase of system designing. Design is the perfect way to accurately translate a customer’s requirement in the finished software product. Design creates a representation or model, provides details about software data structure, architecture, interfaces and components that are necessary to implement a system. The logical system design arrived at as a result of systems analysis is converted into physical system design.

**5.1 System development methodology**

System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps which is used for product progress. The development method followed in this project is waterfall model.

**5.1.1 Model phases**

The waterfall model is a successive programming improvement process, in which advance is seen as streaming relentlessly downwards (like a waterfall) through the periods of Requirement start, Analysis, Design, Implementation, Testing and upkeep.

**Prerequisite Analysis**: This stage is worried about gathering of necessity of the framework. This procedure includes producing record and necessity survey.

**Framework Design**: Keeping the prerequisites at the top of the priority list the framework details are made an interpretation of into a product representation. In this stage the fashioner underlines on calculation, information structure, programming design and so on.

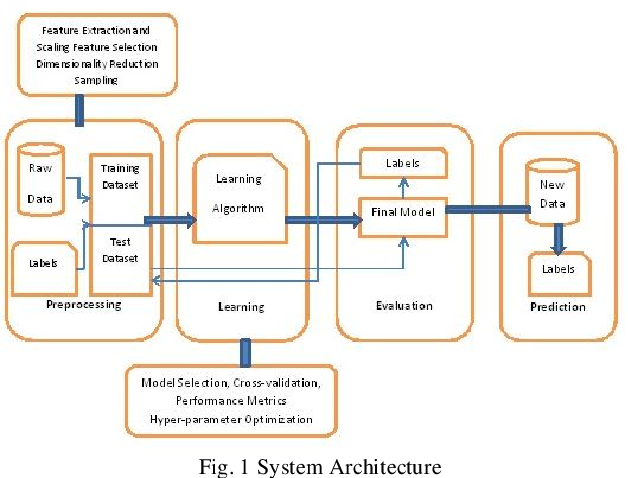
**Coding**: In this stage developer begins his coding with a specific end goal to give a full portray of item. At the end of the day framework particulars are just changed over into machine coherent register code.

**Usage:** The execution stage includes the genuine coding or programming of the product. The yield of this stage is regularly the library, executables, client manuals and extra programming documentation.

**Testing**: In this stage all projects (models) are coordinated and tried to guarantee that the complete framework meets the product prerequisites. The testing is worried with check and approval.

**Support**: The upkeep stage is the longest stage in which the product is upgraded to satisfy the changing client need, adjust to suit change in the outside environment, right mistakes and oversights beforehand undetected in the testing stage, improve the proficiency of the product.

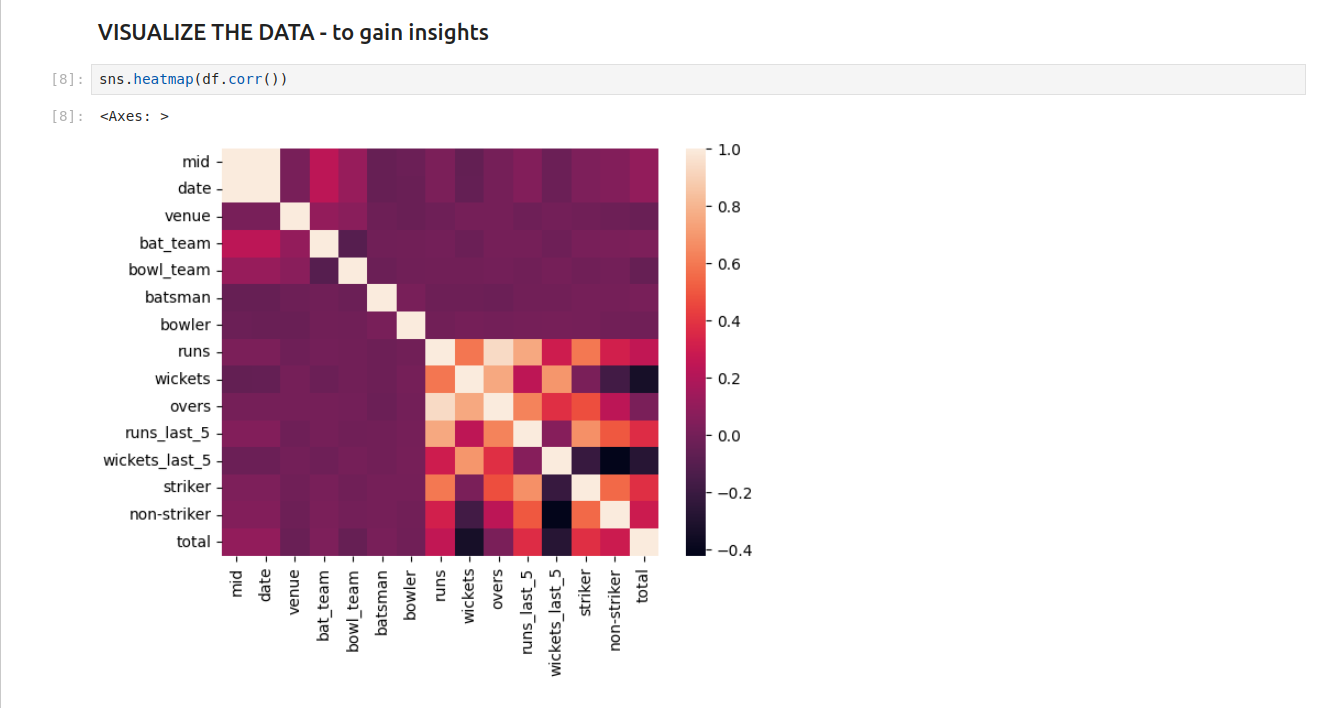
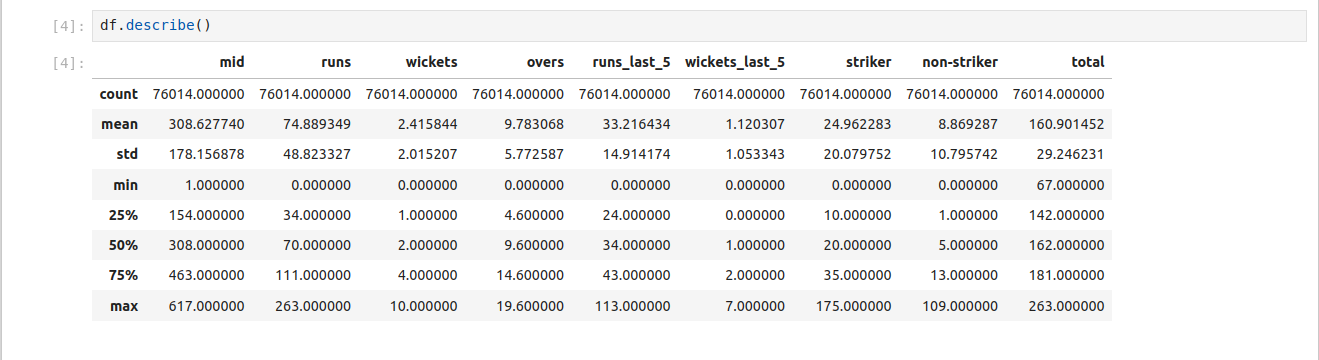
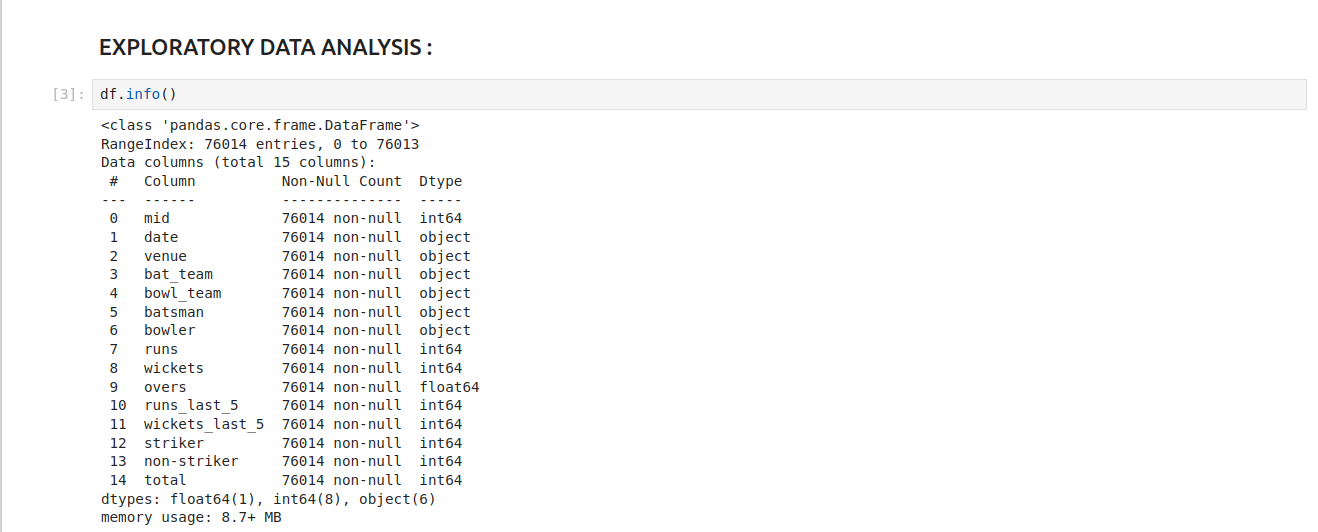
**5.1.2 System Architecture**

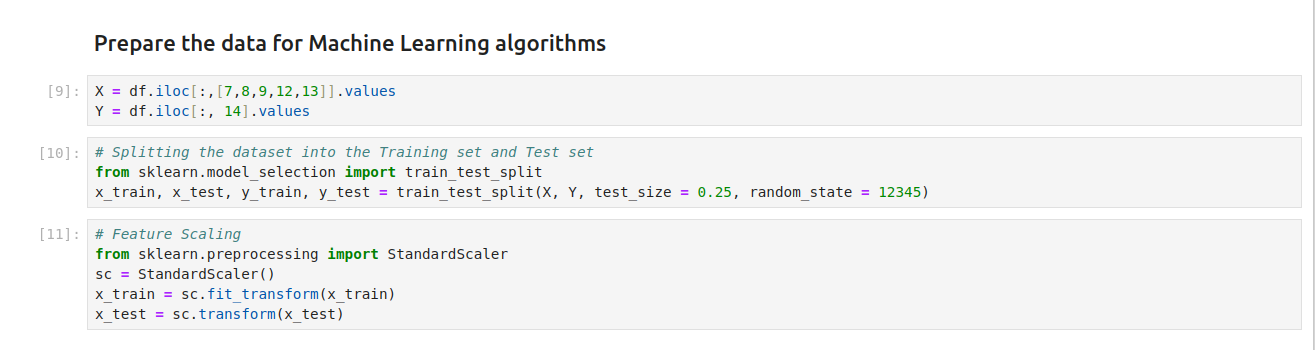


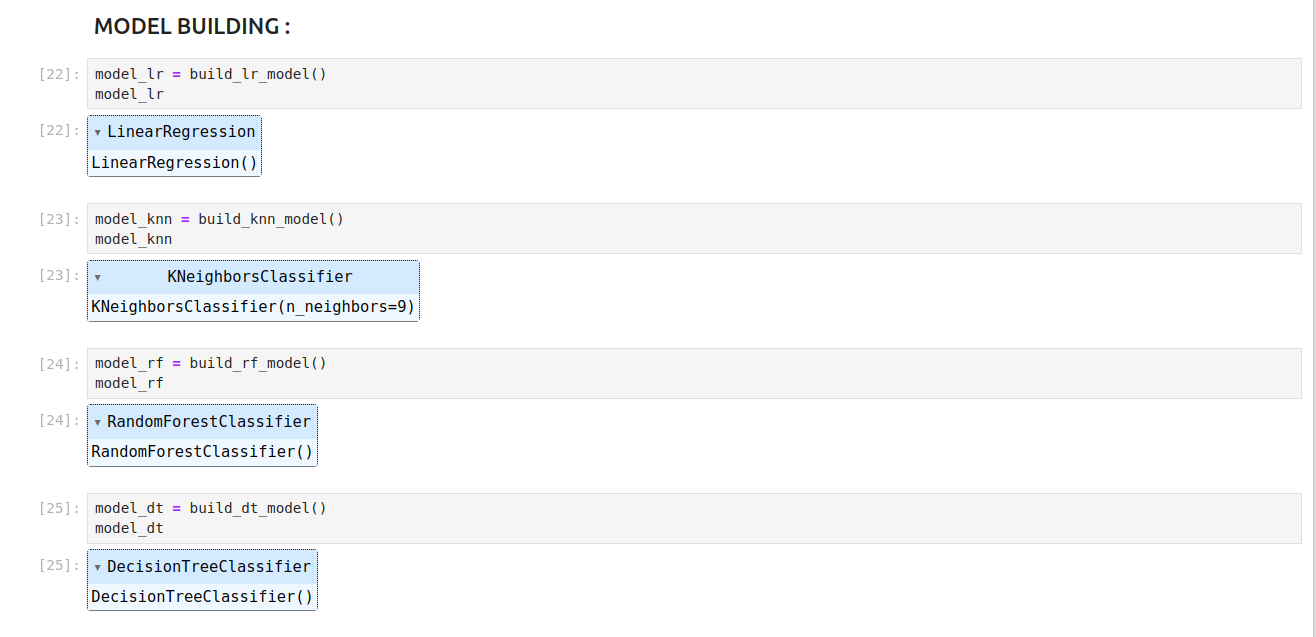
## **Chapter 6**

## IMPLEMENTATION

## 





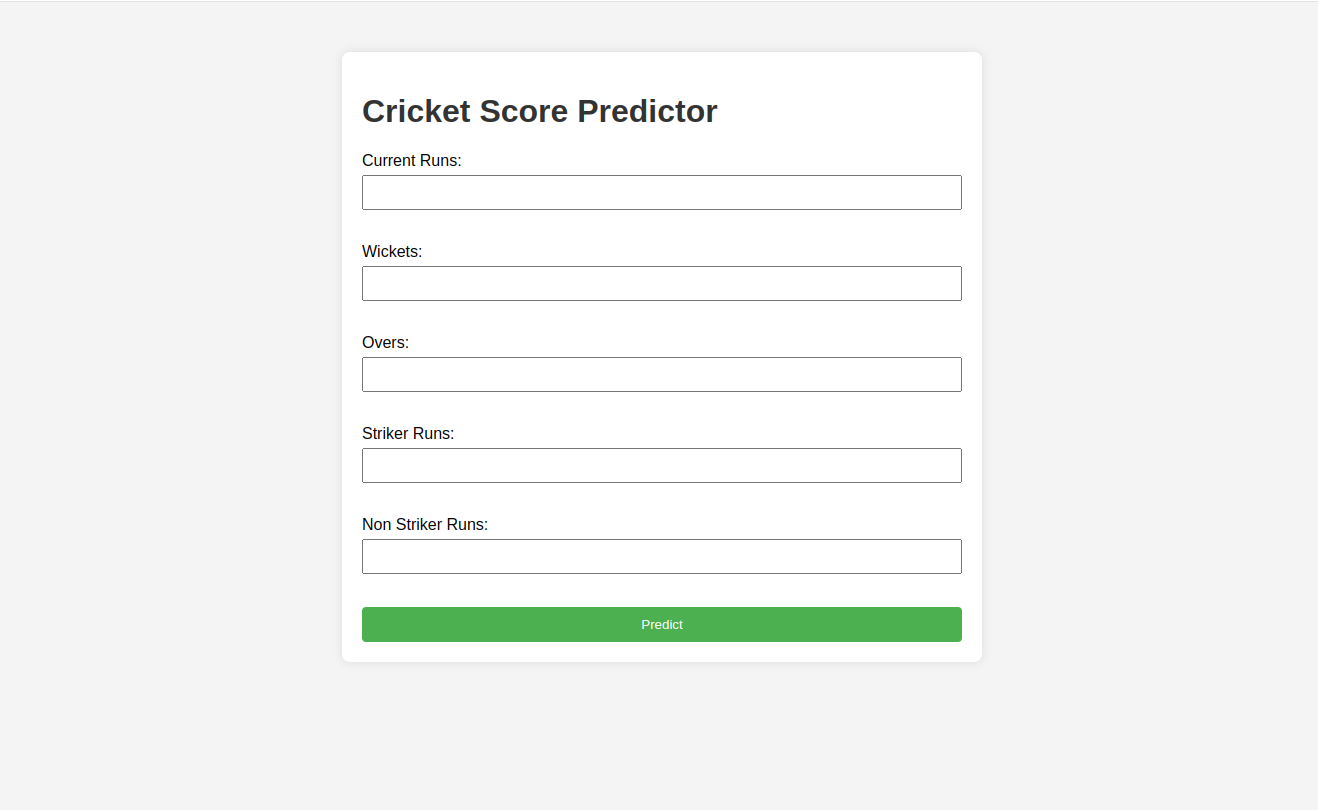


## 

## **Chapter 6**

## GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools



**GitHubLink:**

## **Chapter 7**

**FUTURE SCOPE AND CONCLUSION**

Selection of the best team for a cricket match plays a significant role for the team’s victory. The main goal of this paper is to analyse the IPL cricket data and predict the players’ performance. Here, three classification algorithms are used and compared to find the best accurate algorithm. The implementation tools used are Anaconda navigator and Jupyter. Random Forest is observed to be the best accurate classifier with 89.15% to predict the best player performance.

This knowledge will be used in future to predict the winning teams for the next series IPL matches. Hence using this prediction, the best team can be formed. This project opens scope for future work in the field of cricket and predicting other important things like best team of players, best venue, best city, best fielding decision to win a match.

**REFERENCES**

* 1. Veppur Sankaranarayanan, Vignesh and Sattar, Junaed an Lakshmanan,”Auto-play: A Data Mining Approach to ODI Cricket
* 2. Simulation and Prediction”,SIAM Conference on Data Mining, 2014 K. A. A. D. Raj and P. Padma, ”Application of Association Rule
* 3. Mining: A case study on team India”, 2013 International Conference on Computer Communication and Informatics, 2013
* 4. Tim B. SWARTZ, Paramjit S Gill and S. Muthukumarana,”Modelling and simulation for one-day cricket”, Canadian Journal of Statistics, 2009, Vol 37, No 2, pp-143-160