PREDICTIONS ON CRICKET

A PROJECT REPORT

Submitted by

KEERTHNA A J 2018115048

PRIYA J 2018115078

submitted to the Faculty of

INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY

CHENNAI 600 025

DECEMBER 2021

ANNA UNIVERSITY

CHENNAI - 600 025

BONA FIDE CERTIFICATE

Certified that this project report titled PREDICTIONS ON CRICKET is the bona fide work of KEERTHNA A J and PRIYA J who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE:

DATE:

DR.D.NARASHIMAN
TEACHING FELLOW
PROJECT GUIDE
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

Dr.S.SRIDHAR
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

Cricket is the most enjoyed game in Asian Subcontinent. Test match, One Day International and T20 are three internationally recognized formats of cricket matches. Cricket is a bat-and-ball game played on a cricket field between two teams of eleven players each. Before a match begins, the team captains toss a coin to decide which team will bat first and so take the first innings. Innings is the term used for each phase of play in the match. In each innings, one team bats, attempting to score runs, while the other team bowls and fields the ball, attempting to restrict the scoring and dismiss the batters. When the first innings ends, the teams change roles; there can be two to four innings depending upon the type of match. Cricket is a sport that contains lots of statistical data like batting and bowling record of the team, an individual player's record, a scoreboard of different matches played, fall of wickets, run rate during the match and many others. Before starting a cricket match, opposite team does some analysis based on this statistical data, to find the weakness of the opposite team. Aim of this project is to predict the total team score of the cricket match in which the average runs that the team can score in a particular match is predicted and to predict the performance of each player. Player performance is shown by displaying a graph of runs that the player can score against each team, runs the player can score against each venue and also the average performance of the particular player when he batted first and when he batted second is calculated. These predictions are done using as random forest algorithm which is a machine learning algorithm. Random Forest algorithm is used because it produces good predictions that can be understood easily. It can handle large datasets efficiently. This algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome in random forest algorithm.

ACKNOWLEDGEMENT

We express our profound sense of gratitude to our guide Dr.D. Narashiman, Teaching Fellow, Department of Information Science and Technology, College of Engineering Guindy, Anna University, for his invaluable support,guidance and encouragement for the successful completion of this project. Our heartfelt thanks to Dr.S.Sridhar,Professor and Head, Department of Information Science and Technology, College of Engineering Guindy, Anna University, for the prompt and limitless help in providing the excellent computing facilities to do the project and to prepare the thesis.

We express our sincere gratitude to our review committee members, Dr.S.Swamynathan, Professor, Dr.J.Indumathi, Professor, Dr.P.Geetha, Assistant Professor, Dr. Selvi Ravindran, Assistant Professor, Ms.S.Kanimozhi, Teaching fellow, Ms.G.Mahalakshmi, Teaching fellow, Ms.Siva shankari, Teaching fellow Department of Information Science and Technology, College of Engineering Guindy, Anna University, for their encouraging guidance and kind supervision given to us throughout the completion of the project. We express our gratitude to the teachers who paved way for doing our project effectively.

KEERTHNA A J

PRIYA J

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	Application Programming Interface
CSS	Cascading Style Sheet
HTML	Hypertext Markup Language
IDE	Integrated Development Environment
IPL	Indian Premier League
KNN	K-Nearest Neighbour
ML	Machine Learning
ODI	One Day International
RMSE	Root Mean Square Error
SVM	Support Vector Machine
URL	Uniform Resource Locator
YAML	Yet Another Markup Language

CHAPTER 1

INTRODUCTION

Cricket is referred as Game of Uncertainty and there is no any precise forecast that a specific team would win in any given conditions.Cricket is the second most popular sports in the world with billions of fans across India, UK, Pakistan, Africa, Australia, etc. It is an outdoor game played on a cricket field at 22-yard rectangular long pitch, between two teams consisting each of 11 players. It is played in three formats namely Test, One Day International and Twenty Over International.

Unlike other sports, cricket stadium's size and shape is not fixed except the dimensions of the pitch and inner circle which are 22 yards and 30 yards respectively. The cricket rules do not mention the size and the shape of the field of the stadium. Pitch and outfield variations can have a substantiate effect on batting and bowling. The bounce, seam movement and spin of the ball depends on the nature of the pitch. The game is also affected by the atmospheric conditions such as altitude and weather. A unique set of playing conditions are created due to these physical differences at each venue. Depending on these set of variations a particular venue may be a batsman friendly or a bowler friendly.

In this project, a method has been proposed in which the final score can be predicted and the performance of each player can be estimated. Random Forest tree algorithm has been used for the prediction

The domain of this project is Machine Learning using some of the Machine Learning algorithms prediction is done. Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn

and improve from experience without being explicitly programmed.It focuses on the development of computer programs that can access data and use it to learn for themselves.The primary aim of machine learning is to allow the computers learn automatically without human assistance and adjust actions accordingly.Tom Mitchell,an American computer scientist who is working on machine learning extensively,states that "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Machine learning algorithms are often categorized as supervised, unsupervised, semi supervised and reinforcement. Supervised machine learning is defined by its use of labeled dataset to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process.This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.Supervised learning problems are categorized into regression and classification problems.Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, Support Vector Machines , decision trees, k-nearest neighbor, and random forest while Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Random Forest: Random Forest is a machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of

ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. The ”forest” it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. **Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.**

1.1 PROBLEM STATEMENT

In the cricket match, all teams want to win, but for winning a match, there are several issues needed to be considered such as selecting some good players in the specific opposition. One player may be very good in one team, but his/her performance against specific country or player may not be good. Again, a specific opposition may have weakness with playing in specific venues. In this project,various statistical data has been analyzed and tried to find out most influencing or favorable conditions and features to be considered for winning. And also cricket matches, like any other game, face challenges like weather, injuries and ground conditions which can affect the quality of a match in terms of performance, So also predicting the performance of each player is done in this project which will help in better team selection.

1.2 OBJECTIVE

Objective of this project is to predict the winning team when two teams are played,by analysing various data, and also to predict each and every player performance which helps in better team selection.Objective also includes using the machine learning algorithms such as Random Forest, Support Vector Machine, Decision Tree, Naive bayes, comparing all these algorithms and finding the efficient algorithm with the highest accuracy also done in this project.

1.3 CHALLENGES

The challenges for the proposed work are listed here. Most of the data available are raw data which are needed to be processed before using for training and testing. Datasets had many missing values which required handling of the records.There are many approaches which use one or more machine learning techniques to predict the team score and player performance.Dataset generation is challenging in some cases.Therefore, the generated training data might be less representative or unbalanced.The Duckworth–Lewis technique is a formula for determining the target score for the team batting second in a limited overs cricket match interrupted by weather or other circumstances.As a result, matches based on this method must be filtered because they can't be utilised to predict anything.In addition, numerous important criteria must be considered when predicting team score and player performance.

1.4 APPLICATIONS OF MACHINE LEARNING

Machine learning has introduced new ways to look at technologies. Artificial Intelligence is one of the well-known applications of Machine learning, in which software, computers, and devices performs.Machine Learning are being used in daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Some most trending real-world applications of Machine Learning:

Image Recognition: Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion

Speech Recognition: Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer

speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

Traffic prediction: It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways which is Real Time location of the vehicle form Google Map app and sensors, Average time has taken on past days at the same time.

Product recommendations: Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Self-driving cars: Self-driving cars are one of the most fascinating uses of machine learning. In self-driving automobiles, machine learning is a critical component. Tesla, the world's most well-known automaker, is developing a self-driving vehicle. The car models are taught to recognise people and objects while driving using an unsupervised learning method.

Machine Learning has brought about a positive change in most of the fields. It can also be applied in sports like cricket.Machine Learning can improve the performance and accuracy of players and develop better strategies for the upcoming games. This can be done by predicting the runs scored by a player or the team, the wickets that can be taken and finally predicting the final result of the match. It is always important to select the correct variables so that the prediction is accurate.

1.5 MOTIVATION

The primary motivation behind this project is increasing popularity of cricket matches. This project will be interesting for biggest cricket fans and also for those who always like to guess the results of matches and to create a model that will give a near to accurate projected score emerged. This will help the audience to know what to expect from current match. Wrong projected scores may heighten their expectations which if not fulfilled may lead to disappointment in players.

1.6 ORGANIZATION OF THE REPORT

The rest of the report is organized as follows:

Chapter 2 discusses about the Literature survey which summarizes the related works that had been published related to the current system. It also describes how this system is enhanced to overcome the limitations of the existing system.

Chapter 3 discusses about the System architecture which describes the overall work flow, with detailed explanation of the modules in the architecture diagram.

Chapter 4 deals with the implementation of the work which explains about the details of how this system has been implemented with the algorithms which discusses the procedure and workflow of implementation.

Chapter 5 emphasizes the results of this system with screenshots of the output.

The reference papers and the websites which are referred for this project is given at the end of the report

CHAPTER 2

LITERATURE SURVEY

For our cricket data analysis,a quite number of research papers related to our task had been studied, which is shortly discussed here.

Using Decision Tree algorithm: Md. Muhaimenur Rahman et al used the latest version of the decision tree algorithm that is C5.0 on their own collected data set and successfully got the accuracy of 63.63 percent.In this paper, they analyze various statistical data and tried to find out most influencing or favorable conditions and features to be considered for winning.In this paper, they showed the importance of some features and have found the predicted outcome by using those features on One Day International cricket.Data of ODI matches are collected during the time 2005-2017.Dataset included seventeen features such as “Bangladesh”, “Opposition”, “Venue”, “Toss”, “Day/Day-night”, “Batting”, “Year”, “Bangladesh runs”, “Opposition runs”, “Bangladesh wickets”, “Opposition wickets”, “Bangladesh over”, “Opposition over”, “Bangladesh fall of wicket (overs)”, “Opposition fall of wicket (overs)”, “Bangladesh fall of wicket (runs)”, “Opposition fall of wicket (runs)”. They collected all the records of the matches that were played without any rain interruption. When they collected data, the found many matches that are abandoned, some are half played and some are using (D/L) Duckworth-Lewis method. In addition, they showed the influencing or favorable conditions for the Bangladesh team when the match is in progress. Finally, they tried their best to find the best result using decision tree.The key advantages of this project is it used decision tree algorithm to predict the winning team in One Day International cricket match.Also it used most of the other algorithms to test the accuracy.
Limitations of this project includes

- Only predicting team of winning of One Day International only for Bangladesh team.
- It did not work on other matches like T20 and test matches.
- It did not take account of matches that are played on the basis of Duck-worth Lewis(D/L) method.

[1]

Using Support Vector, CTree and Na¨ive Bays Algorithm: Shilpi Agrawal et al proposed a method of predicting results of Indian Premier League T-20 Matches using Machine Learning. They have studied the problem of predicting the uncertainty of who will win the upcoming IPL match based on the individual competency of each player, coordination and team work of whole team evolving and technique followed by each team in each match. In this paper they proposed a model using machine learning algorithms that can predict winning team based on past data available. They applied three machine learning algorithms namely Support Vector Machine, CTree and Na¨ıve Bayes.In this work,In order to predict IPL T-20 match result, dataset was collected from techgig.com .The dataset comprises details of past 500 IPL matches. Each row in the dataset represent ball by ball details played in both the innings, Collected dataset has raw data table at its initial stage which needs to be pre-processed for removing irrelevant details. Pre-processing stage cleans the dataset by removing those data that are not useful to get results. Data where results have not been declared or marked are removed during this stage.Data provided in dataset is categorical in nature due to which classification is quiet complex. It may also affect classification process resulting in wrong prediction. So all categorical data in dataset except the target attribute (Winner) has been converted into numeric format and normalized on scale basis.Further, suitable data is converted to a numeric form and scale it on three parameters win, loss, and tie. This data

is trained and classified with three classifier SVM, CTree and Na¨ıve Bayes using R Tool.Advantage of this project is it used easily understandable machine learning algorithms like Support Vector Machine, Na¨ıve Bayes and C-Tree for predicting the winning team.. Some of the limitations of this project are

- Very less number of dataset has been used for training and testing.
- Only few machine learning algorithms are used.
- Only raw data has been available which had to be processed.

[2]

Using linear regression and na¨ıve bays: Tejinder Singh et al developed a model that has two methods, first predicts the score of first innings not only on the basis of current run rate but also considers number of wickets fallen, venue of the match and batting team. The second method predicts the outcome of the match in the second innings considering the same attributes as of the former method along with the target given to the batting team. These two methods have been implemented using Linear Regression Classifier and Naïve Bayes Classifier for first innings and second innings respectively. In both methods, 5 over intervals have been made from 50 overs of the match and at each interval above mentioned attributes have been recorded of all non-curtailed matches played between 2002 and 2014 of every team independently. It has been found in the results that error in Linear Regression classifier is less than Current Run Rate method in estimating the final score and also accuracy of Naïve Bayes in predicting match outcome has been 68 percent initially from 0-5 overs to 91 percent till the end of 45th over. The data has been collected from http://www.espncricinfo. com, where over-by-over data of all the matches are available publicly. The dataset consists of complete matches excluding all the rain-interrupted and rain-abandoned games, played between 2002 and 2014

among the 8 teams namely Australia, India, New Zealand, South Africa, England, Sri Lanka, Pakistan and West Indies. Also it contains the dataset of the matches played on the venues from each country mentioned above. The main purpose of this paper is to make a model for predicting the final score of the first innings and estimating the outcome of the match in the second innings for the ODIs.Key advantage of this project is that they used large number of dataset which is very useful for training part.Another advantage is that they used two different methods like Linear regression for predicting first innings and Naive Bayes for predicting second innings.The project's limitations include the following:

- only linear regression and Naive Bayes machine learning algorithm is used.
- But many more algorithms are available for predicting the outcome.
- The dataset had match details of only specific countries.

[3]

Using linear regression, logistic regression, support vector regression, singular value decomposition: Shivam Tyagi et al used machine learning algorithms to predict the duration of a match in terms of the number of balls expected to be delivered in the match.The prediction of duration of a game will be beneficial for both sport and advertisement industry. The work introduces four different approaches, using historical data, to predict the number of balls in a match. They trained their model on player combinations, player profiles and team strength to achieve high precision. Predictions from this model can help increase the cricket craze in India by making the teams more competitive and thus increasing the overall quality of the tournament. This work will help investors and advertising firms to analyze the upcoming match better and plan their advertisement slots accordingly, creating a win-win environment

for all, players, viewers and advertisers. They used different ML models in on historical data in four different ways and came up with the results to determine the best approach. In their findings Innings, bowler and batsman based approach gave the best RMSE values. so they further realised that using results from all the four approaches can help in making a better prediction. So they developed a tool that gives results for all the four methods, including an option for making a dynamic estimation. They collected the datasets of IPL matches played over the years, data preprocessing has been done in which data was converted in the YAML files into dictionary format by using the python script and the unnecessary columns are removed. After that they did data exploration using matplotlib and seaborn which are the python libraries.The goal of their research was to predict how many balls in a match will be bowled. A closer prediction will enable better planning of advertisement slots and getting maximum viewership.They trained machine learning models, using historical data, and then make predictions for the present -day data. They used different statistics for this type of analysis. These statistics include wins and defeats, player profiles, demographics like location and weather. All these are significant factors in deciding the duration of a match. The decisions to use which machine learning model was based on the characteristics of data and analysis approach. They used a variety of techniques to estimate duration like linear regression, logistic regression, support vector regression, and singular value decomposition, which is a significant benefit of this study.Another advantage includes usage of YAML files as they are easy to work and portable between programming languages. This project has few limitations including

- It did not handle the missing values of dataset.
- Around 400+ values from dataset are removed as they could not handle it.
- If these datasets are handled properly,prediction would have been

more accurate.

[4]

2.1 EXISTING WORK

Md. Muhaimenur Rahman et al done an analysis of Bangladesh One Day International Cricket Data using decision tree algorithm to find out influencing conditions and features to be considered for winning by dividing the analysis into three sections. Shilpi Agrawal et al proposed a model using machine learning algorithms that can predict winning team based on past data available. They used three machine learning algorithms namely Support Vector Machine, CTree and Na¨ıve Bayes and predicted the result. Tejinder Singh et al developed a model to predict the score of the first innings and the score of the second innings separately using linear regression and naive bays algorithm. Shivam Tyagi et al used machine learning algorithms to predict the duration of a match in terms of the number of balls expected to be delivered in the match using linear regression, logistic regression, support vector regression and singular value decomposition.

Limitations of existing works includes less prediction accuracy rate like using SVM gave an accuracy of only 70.91 percent,KNN gave accuracy of 70.3 percent ,Naive Bayes gave accuracy of 64.18 percent and Decision Tree gave accuracy of 73.63 percent.Also the number of datasets used for training and prediction is comparatively less that caused difficulty in training the model.Rather than dealing with the dataset's missing values, they were eliminated, resulting in a smaller number of values in the dataset..Matches that are played on the basis of D/L method are not taken into account for predicting the result.

2.2 PROPOSED WORK

Statistical records about cricket are collected from Cricbuzz,kaggle and ESPNcricinfo comprising the details of past 500 to 600 matches.It consists of details like venue,date of the match,mid,batman,striker and non-striker, batting and bowling team name,runs and wickets for every over, runs scored in last 5 overs and wickets in last 5 overs,total score.The pre-processing stage cleans the dataset by deleting any data that isn't necessary for the results to be obtained. Feature selection is used to reduce the amount of redundant data from the data set.The dataset's categorical data was then transformed to numeric representation using One-hot encoding.Later,using the random forest algorithm training and testing are carried out.Using FLASK framework,a web application has been developed to provide detailed output of the team score prediction.
For player prediction,specific player performance against each team is plotted as bar graph.Player's performance at each venue and players' performance on batting first and batting second using matplotlib are represented using bar chart. Additionally,a line graph is plotted to represent the difference between actual runs scored and predicted runs.This line graph shows the accuracy of the model.

2.3 METHOD OF CITATION

Following are a few methods of citation.

These are some cited journals and can be seen in References [2] , [1], [3], [4]

These websites are cited for better understanding. [5], [6], [7], [8], [9], [10], [11], [12], [13]

In this chapter, Literature survey is discussed in detail. And also discussed about the existing work of the project and the limitations of the existing

work. In the next chapter system architecture and the function modules will be elucidated in detail

CHAPTER 3

SYSTEM ARCHITECTURE

This section deals with the general architecture of proposed model.

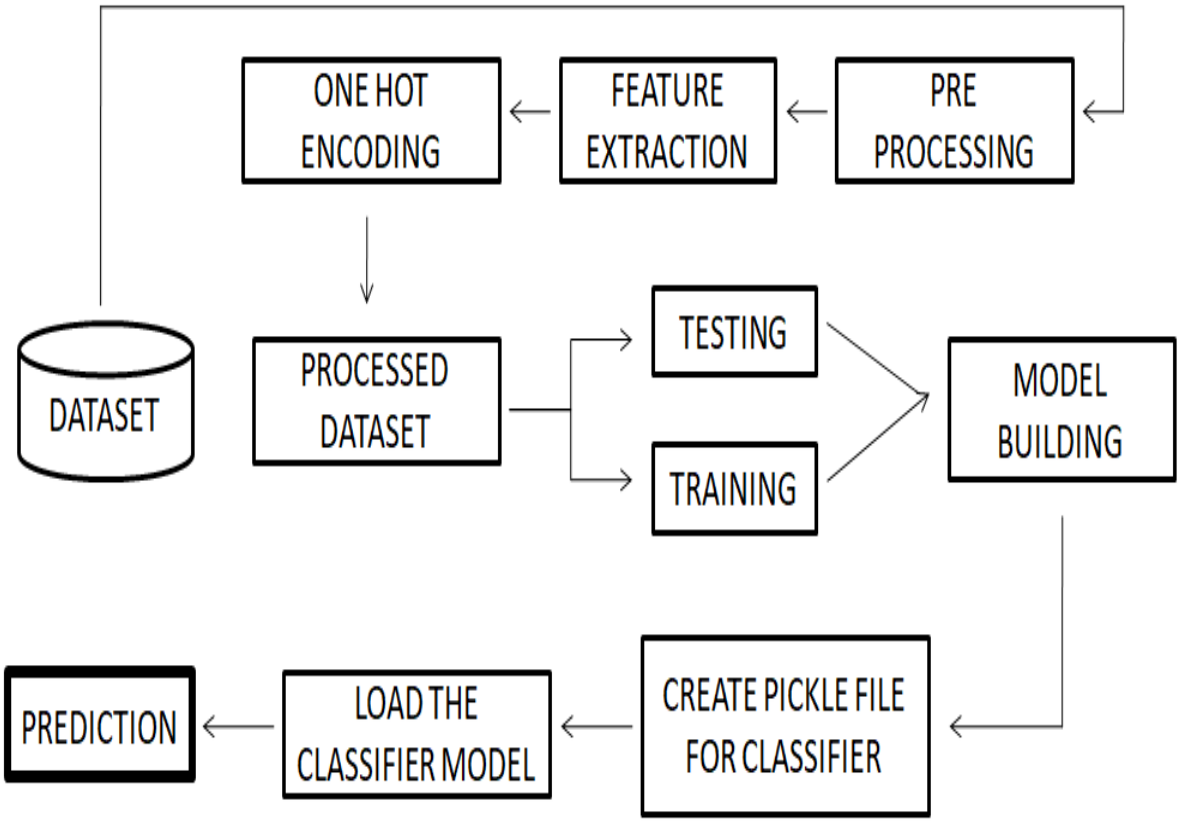

Fig 3.1 Architecture of proposed model

3.1 FUNCTION MODULES

3.1.1 Dataset and Preprocessing

Dataset that is required for our project is collected from Cricbuzz, ESPNcricinfo and kaggle. The dataset comprises details of past 500 to 600 matches. It consists of details like venue,date of the match,mid,batman,striker and non-striker, batting and bowling team name,runs and wickets for every over, runs scored in last 5 overs and wickets in last 5 overs,total score. Each row in the dataset represent details for every over. By using these features,the most influencing or favorable conditions and features for winning a match are being predicted. Data preprocessing is a process of preparing the raw data and make it suitable for a building and training Machine Learning models. In this project,Pre-processing stage cleans the dataset by removing the data that are not useful to get results. This stage cleans the dataset by removing those data that are not useful to get results.Attributes like venue,mid,batman,striker and non-striker are dropped and also first 5 overs data are removed for a better prediction.

3.1.2 Feature Selection

Feature selection is primarily focused on removing non-informative or redundant predictors from the model. It is done for keeping only the most relevant variables from the original dataset. Feature selection helps to reduce the amount of redundant data from the data set. In the end, the reduction of the data helps to build the model with less machine's efforts and also increase the speed of learning and generalization steps in the machine learning process. This stage filters the consistent teams. The teams which are standard(currently playing) are filtered by removing the defunct teams. This makes the model to predict even more accurately. This helps in improving the accuracy. Feature selection also reduces the overfitting problem and reduces training time.

3.1.3 One Hot Encoding

Conversion of data is done because Machine learning models require all input and output variables to be numeric Data of batting team and bowling team in dataset are categorical in nature, using categorical features in machine learning are difficult and also training categorical features are quite complex. It may also affect classification process resulting in wrong prediction. In this step categorical data in dataset has been converted into numeric format using One-hot encoding, One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. One-hot encoding changes every single name of the batting and bowling team into a single column, 0 indicates non existent while 1 indicates existent.

3.1.4 Training and Classification

For training and classification purpose, Random Forest Algorithm have been used with the proposed model which is a supervised learning algorithm and works with the concept of ensemble learning and also this runs efficiently on a large dataset and gives better accuracy than other algorithms.

3.2 WORKFLOW OF THE MODEL

The Flask Framework, a Python intrinsic framework, is used to develop a web application. The command ”pip install flask” in the terminal is used to install Flask into our IDE (Pycharm). After that, importing a flask module is required in a new PyCharm project, and the code for our prediction is written in the run() method of the Flask class, which executes the application on the local development server. After running the script, the output will include a URL, which will direct you to our webpage. The dataset has been loaded into the IDE (pycharm), and some changes need to be made to it, for which the pandas module from Python has been utilised. Columns that aren’t important

to our forecast are eliminated after importing pandas. The useless columns are saved in a variable, and the irrelevant columns are eliminated using the pandas drop() function. Giving the main elements is required in this forecast, thus only the teams that are actively playing are filtered, and all other defunct teams are deleted. After that, the first 5 overs from the dataset were eliminated for a better prediction, as our dataset contains details for each over such as 0.1 over, 0.2 over, 0.3 over, and so on. As a result, the data from the first five overs has been erased. The data is then converted from category to numeric using One-hot encoding, which is done using the pandas function get-dummies(). The names of the batting and bowling teams are categorical data in our dataset, so they are converted to numeric representation using One-hot encoding. The columns are rearranged to place the 'total score' column last, allowing the dataset to be separated for testing and training. Matches played after 2017 are subjected to testing, whereas matches performed prior to 2017 are subjected to training. The random forest regression model from scikit-learn was imported, the model was instantiated, and the training data was used to fit the model. The next stage is to determine how good the model is, which is done by making predictions on the test features and then comparing them to the known answers.

The dataset for player prediction was taken from Kaggle newline (deliveries.csv and matches.csv). After loading the dataset into the IDE, the user must enter the player's name and team name as user input. The next step is to retrieve the batsman data for the player that the user has entered from the deliveries.csv file. Unique team names are filtered from the matches.csv dataset, and the team name entered by the user must be removed from the filtered teams. The data is iterated with the corresponding opponent teams and venue once the player team has batted first. The venue and team data are checked against the dataset, and if they match, their match id is added to the matches dataframe. The batsman data contains information on the player who was entered. Batsman runs are calculated when this data id matches the match id found in the matches

dataframe. A list is made with the team name, venue, batted first, batted second, balls, and runs. Second batting follows the same procedure. The batting first and second lists are combined and utilised to make predictions. The 'runs scored against each team','runs scored against each venue','batting first versus batting second', and lastly the 'actual runs and forecasted runs' graphs are plotted as output. The players projected runs when they played with other teams, as shown in the runs scored against each team graph. The runs scored against each venue graph shows how the player performs when playing on specific venues. Compare the runs scored when they bat first vs. when they bat second in the batting first vs. batting second graph. The accuracy of the constructed model may be seen in the last graph, which shows actual runs vs expected runs.

3.3 USER INTERFACE FLOW OF THE SYSTEM

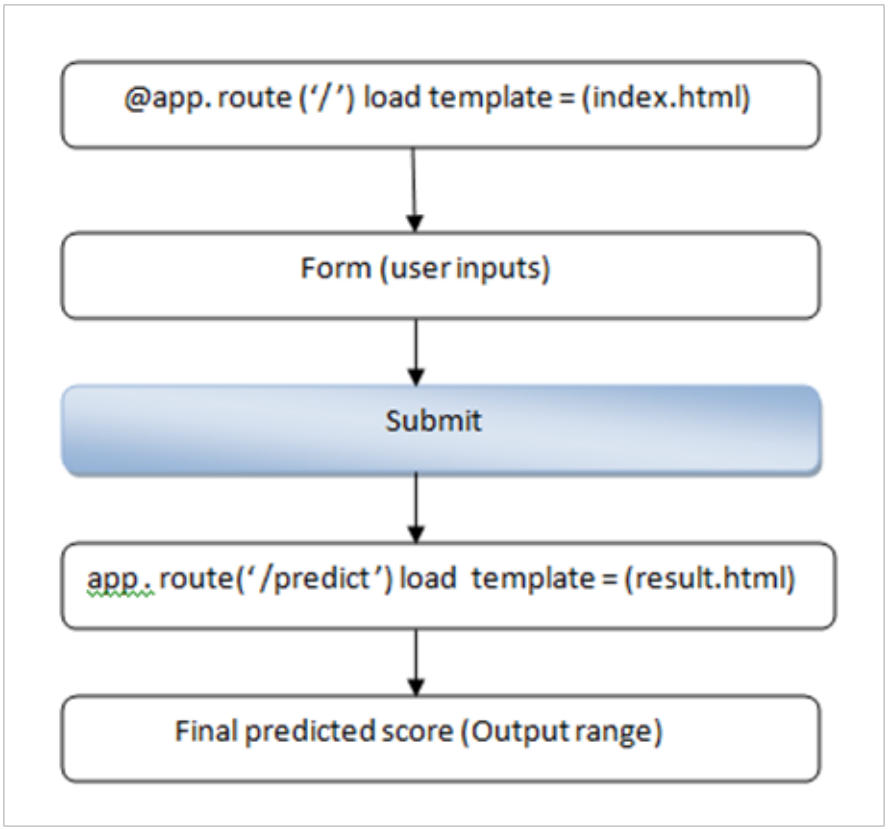

Fig 3.2 User Interface flow of the system

The Flask framework is used for the development of this project. In the Flask framework, the routing of the pages is done based on the URLs. If the browser finds the '/' in the URL then routing the user to the home page will be done. On the homepage, main form will be available. In that form, user inputs are taken. The user inputs include, Name of the batting team, name of the bowling team, number of runs scored, number of overs bowled, number of wickets taken, number of runs scored in the previous 5 overs, and number of wickets taken in the previous 5 overs. Once the user hits predict score button then the form is submitted. After submission of the form, the model comes into the picture in the backend. The inputs are compared with the historical data and then a score is predicted. After the submission of the form, the user is redirected to the '/result' URL i.e. the user is redirected to the result page where the user can see the actual predicted score. On the prediction page, the user will be getting the output in the form of a range i.e. from the lower bound to the upper bound. The below given picture shows the output for the cricket score prediction.

Random Forest Tree:

Random forest tree is an ensemble and supervised learning algorithm.It can be used for both Classification and Regression problems in Machine Learning. This

creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions,and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. It is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

3.4 RANDOM FOREST ALGORITHM

Algorithm 1: Pseudo code for the random forest algorithm

To generate c classifiers:
for i = 1 to c **do**
  Randomly sample the training data D with replacement to produce D_i
  Create a root node, N_i containing D_i
  Call BuildTree(N_i)
end for

BuildTree(N):
if N contains instances of only one class **then**
  return
else
  Randomly select x% of the possible splitting features in N
  Select the feature F with the highest information gain to split on
  Create f child nodes of N , N_1 ,..., N_f , where F has f possible values (F_1 , ... , F_f)

System architecture diagram is illustrated in this chapter and all the components were explained explicitly. In the next chapter, Implementation of the project will be talked through and also the outputs of the project will be illustrated.

CHAPTER 4

IMPLEMENTATION

Proposed methodology includes Dataset Collection, Pre-processing of collected dataset, Feature extraction from raw data, conversion of categorical data into numerical data using One-hot encoding method, partitioning of samples into training and test samples, training and classification.

4.1 WORKING ENVIRONMENT

The Programming language used in this project is PYTHON 3. Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. The project was carried out in Pycharm IDE. PyCharm is a dedicated Python Integrated Development Environment providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive Python, web, and data science development. Scikit-learn is used in this project which is a machine learning framework. Scikit-learn is a Python library used for machine learning.The framework is built on top of several popular Python packages, namely NumPy, SciPy, and matplotlib.

4.2 LIBRARIES AND ALGORITHM

Kaggle Dataset

Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment. Also this platform is trusted

by some of the largest data science companies of the world such as Walmart, Facebook and Winton Capital. On Kaggle, data scientists get exposure and a chance to work on problems faced by big companies in real-time. For the player predicition both the dataset has been downloaded from the kaggle which nearly had 15,000 data.

Python Flask

Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. It is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It gives developers flexibility and is a more accessible framework to build a web application quickly using a Python file.

Random Forest algorithm

Random Forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it can be utilised for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complex problem and increase the model's performance. "Random Forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset," according to the name. Instead than relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions.

NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Using NumPy in Python gives functionality comparable to MATLAB since they are both interpreted,[20] and they both allow the user to write fast programs as long as most operations work on arrays or matrices instead of scalars.

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL database tables or queries, and Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Pyplot is a Matplotlib module which provides a MATLAB-like interface. Matplotlib is designed to be as usable as MATLAB, with the ability to use Python, and the advantage of being free and open-source.

4.3 IMPLEMENTATION

For the Cricket score prediction, A web application is created using Flask Framework which is a inbuilt framework in python.There are many modules or frameworks which allow building webpage using python like a bottle, Django, flask, etc. But the real popular ones are Flask and Django, In these two framworks flask is easier to learn and understand than django, also flask is considered more “Pythonic” than Django. Flask is an API of Python that allows us to build up web-applications,flask provides us with tools, libraries and technologies that allow us to build a web application. Flask is installed into our IDE (Pycharm) using a command ”pip install flask” in the terminal. After that in a new pycharm project importing a flask module is mandatory and the code is written for our prediction, run() method of Flask class runs the application on the local development server. After running the script, the output will contain an URL,that is where our webpage will be visible.

Dataset is loaded into the IDE(pycharm), Some modification should be done in dataset for which pandas module from python has been used. Pandas is a Python library used for working with data sets. Pandas can clean messy data sets, and make them readable and relevant. After importing pandas,columns are removed which are irrelevant for our prediction. The columns which are irrelevant are stored in a variable and using drop() function in pandas the irrelevant columns are removed. In this prediction , giving the important features are mandatory so only the teams which are currently playing are filtered and all other defunct teams are removed (teams like ’Pune Warriors’, ’Deccan Chargers’). After this,the first 5 over data from the dataset has been removed, this is done to get a good prediction since our dataset has details for each over like 0.1 over, 0.2 over , 0.3 over and so on. So first 5 overs data has been removed. After this conversion of data from categorical to numeric is done using One-hot encoding, this is done using a function get-dummies() from pandas. Categorical data in our

dataset are batting and bowling teams name, so that it is changed into numeric format using One-hot encoding. Rearranging the columns are done to put the 'total score' column at last so that dataset can be split for testing and training. Matches played after 2017 are taken for testing and matches that are played before 2017 are taken for training.

For training and testing the data, splitting data into training and testing sets. During training, the model had been let to 'see' the answers, in this case the **total score**, so it can learn how to predict the total score from the features, the model's job is to learn the relationship during training. Then, when it comes to evaluating the model,it had been asked to make predictions on a testing set where it only has access to the features and not the answers, Because we do have the actual answers for the test set,these predictions can be compared to the true value to judge how accurate the model is. Generally, when training a model,data is randomly split into training and testing sets to get a representation of all data points. Training the model is simple using Scikit-learn. The random forest regression model from scikit-learn has been imported,the model has been instantiated , and fitted the model on the training data.The next step is figuring out how good the model is and so predictions on the test features are made and then they are compared to the known answers.

For the User Interface, code is written in HTML. Web applications mainly use HTML to display information for the visitor, so HTML files had been incorporated in our app, which can be displayed on the web browser. Flask provides a render-template() helper function that allows use of the Jinja template engine.render-template() lets us render HTML template files that exist in the templates folder, This will make managing HTML much easier by writing HTML code in .html files as well as using logic in the HTML code.These HTML files, (templates)can be used to build the application pages, such as the main page and the pages on prefernces. In addition to the templates folder, Flask

web applications also typically have a static folder for hosting static files, such as CSS files, JavaScript files, and images the application uses. Using that,a style.css style sheet file had been created to add CSS to our application and images are also added for the background of our web application.

For player prediction, dataset has been downloaded from Kaggle (deliveries.csv and matches.csv).First step is to load the dataset into the IDE and then user must give the player name and player team name as user input.Next step is to extract the batsman data from deliveries.csv dataset for the player the user has entered.From the matches.csv dataset,unique team names are filtered and then the team name which user has entered must be dropped from the filtered teams.Taking data where player team has batted first and it is iterated with respective opponent teams and venue.The venue and team data are compared with the dataset and if they are found same ,their match id will be added to matches dataframe.The batsman data has the details about the entered player's data .When this data id matches with the match id that is in the matches dataframe, then batsman runs are calculated.A list containing team name,venue,batted first batted second,balls and runs are created.The same process is done for second batting.Batting first and batting second lists are summed up and used for prediction.

Four graphs are plotted as output with 'runs scored against each team' Fig 4.1,'runs scored against each venue' Fig 4.2,'batting first vs batting second' Fig 4.3 and finally the 'actual runs and predicted runs'.Runs scored against each team graph shows the players predicted runs when they played with other teams(Fig(1)).Runs scored against each venue graph depicts the runs when the player plays on specific venues(Fig(2)).In batting first vs batting second graph,comparison between the runs scored when they batted first and batted second(Fig(3)).Last graph,actual runs vs predicted runs shows the accuracy of the developed model.

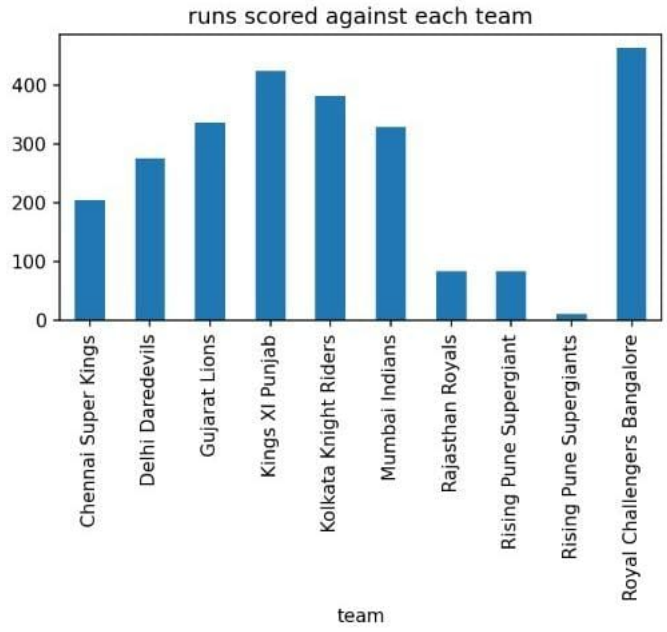

Fig 4.1 Runs scored against each team

Fig 4.1 shows the graph of runs that is scored against each team by a player. This output is for the player whose name is entered in the user input (example : DA Warner). This shows DA Warner's runs when he plays against the teams which are mentioned in the graph. These team names are the names which are available in the dataset. In this graph, it is evident that DA Warner scores high runs when he plays against 'Royal Challengers Bangalore' and scores low runs when he plays against 'Rising Pune Supergiants'. Apart from 'Royal Challengers Bangalore' he also scores good runs when he plays against 'Kings XI Punjab' and 'Kolkata Knight Riders'. These details can be inferred from the Fig 4.1

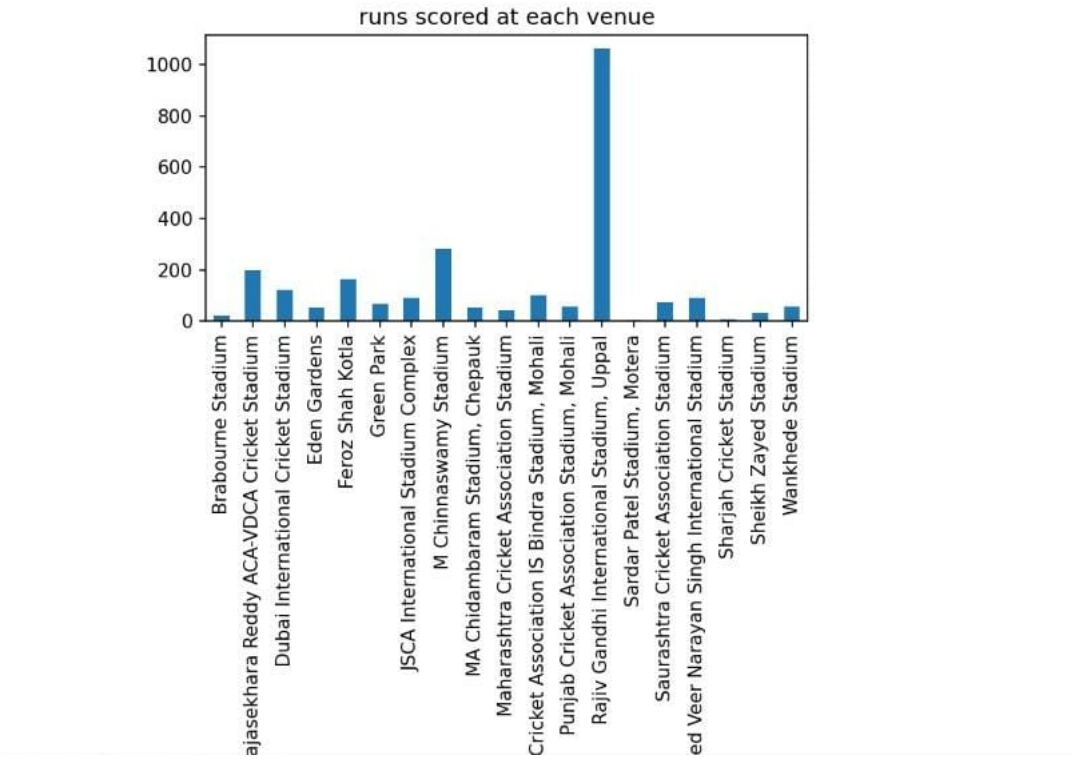

Fig 4.2 Runs scored at each venue

Fig 4.2 shows the graph of runs that is scored at each venue by a player. This output is for the player whose name is entered in the user input (example : DA Warner). This shows DA Warner's runs when he plays at the venues which are mentioned in the graph. These venue names are the names which are available in the dataset. In this graph, it is understood that DA Warner scores high runs when he plays at the venue 'Rajiv Gandhi International Stadium, Uppal' and scores low runs when he plays at the venue 'Sharjah Cricket Stadium'. He gives a moderate performance in the 'M Chinnaswamy Stadium'. It is shown that DA Warner is not good at many venues. These details can be inferred from the Fig 4.2

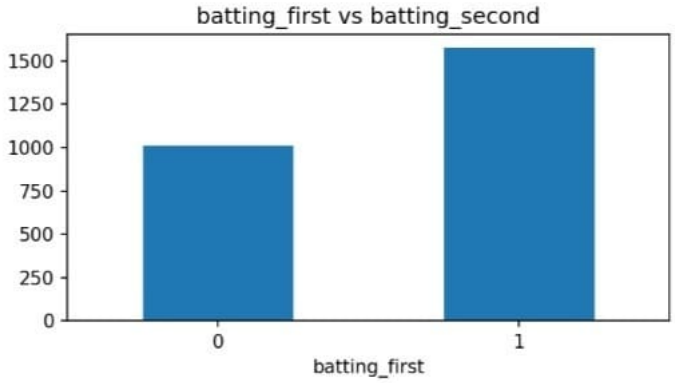

Fig 4.3 Batting first and batting second

Fig 4.3 shows the graph of a player when he is batted first and batted second. In cricket, If the team is uncertain about the nature of the pitch or simply wants to play safe, they often bat first. If the opposition bowling is strong, batting first is often considered a good option. The captain opts to bat second if he is confident that their team can successfully chase any total. Once the target is known, the team does not have to worry about setting a winnable score. Another advantage of batting second is during day-night One Day International games, played under lights. This graph shows when DA Warner's performance is good if he batted second. 0 represents batting first and 1 represents batting second.

In this chapter the implementation is discussed in detail. The next chapter will be dealing with the results of the work and the performance analysis.

CHAPTER 5

RESULTS AND PERFORMANCE ANALYSIS

This section emphasizes on the achieved results and subsequently discusses the findings of proposed model. For cricket score prediction, dataset is collected from cricbuzz and kaggle it has 600 data of cricket matches played over past years. Giving the main features is mandatory in this forecast, thus only the teams that are currently playing are filtered, and any other defunct teams (teams like 'Pune Warriors', 'Deccan Chargers') are deleted.As a result, the data from the first five overs has been erased.Because the dataset has details for each over, such as 0.1 over, we can make a solid prediction. Following that,One-hot encoding is used to convert categorical data to numeric data.This is accomplished with the help of the pandas function get-dummies().The names of the batting and bowling teams are categorical data in our dataset, so they are converted to numeric representation using One-hot encoding. The columns have been rearranged in order to place the 'total score' column at last so that the dataset may be divided into testing and training sections.For the score prediction, in our web app we have to enter the batting team name, bowling team name, overs, runs, wickets, runs scored in previous 5 overs and wickets taken in previous 5 overs. After that the predicted score will be displayed.

For the player performance prediction. The model dataset originally is of two parts, one is matches dataset and other one is deliveries dataset. In match dataset, 636 data are available which has 17 columns are available namely, matchid, season(year), city where the match has conducted, date, team1 , team2, toss winner, toss decision, result, dl('Duckworth-lewis System) applied, winner, win by runs, win by wickets, player of match, venue, umpire1 and umpire2. In deliveries dataset, 15,000 data are available which has the details of the

match id, inning, batting team, bowling team, over, ball, batsman, bowler, wide runs, byeruns, legbye runs, no ball runs, penalty runs, batsman runs, extra runs, total runs.The dataset consists of players' details comprising of authentic indian players to most of the foreign cricket players.They are the cricket players who played between the year of 2008 to 2017.These cricketers took part majorly as compared to others. It has minute details like non striker details,penalty runs,super overs and so on.These minute player details has been helpful in predicting the total score and player performance effectively.

"Batsman" column from the deliveries dataset is used in extracting the bastman data for the player whom the prediction has to be done. team1 and team2 column from the matches dataset are the names of the teams between whom the match was conducted. team1 is used to filter the unique team names and the player's own team name is removed for the prediction.Then,batting first and batting second score are calculated to compare the performance of the player when he batted first and second. Runs are calculated by iterating the teams names and the unique venue, when the iterated venue and venue from the dataset are equal and when the iterated team name and team names from the dataset are equal,their match id are saved in an object called "matches".By comparing the match id with the 'Batsman data' ,runs are calculated.

For the player performance, runs scored against each team, runs scored against each venue, batting first and batting second graphs are shown. In runs scored against each team, it shows the player's performance when he played with each team and expected high runs or low runs can be inferred. Similarly ,runs scored against each venue ,shows the player performance on each venue. In batting first and batting second graph,average performance of the particular player when he batted first and when he batted second is shown. For training 75 percentage of dataset are given and 25 percentage of data are tested.

Random forest algorithm is used for the prediction. For training and testing, train-test-split technique is used which is a technique used for evaluating the performance of an ML algorithm. The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train-test-split() function.The function takes a dataset as input and returns the dataset split into two subsets. Ideally,the original dataset can be splitted into input (X) and output (y) columns, In this case X has features like team name, venue, batting-first-second, balls and Y has runs. Then X and Y are splitted for testing and training (X-train, X-test, Y-train, Y-test). After this using random forest classifier the final output is predicted. In this work, historical data has been collected from real cricket matches and useful features have been extracted after pre-processing of data. Further, suitable data is converted to a numeric form and scaled it. This data is trained and classified with Random forest classifier using Scikit-Learn. Comparison between predicted runs and actual runs is shown in Fig 5.1. It shows that using random forest algorithm for this prediction the predicted runs are higher than the actual runs.

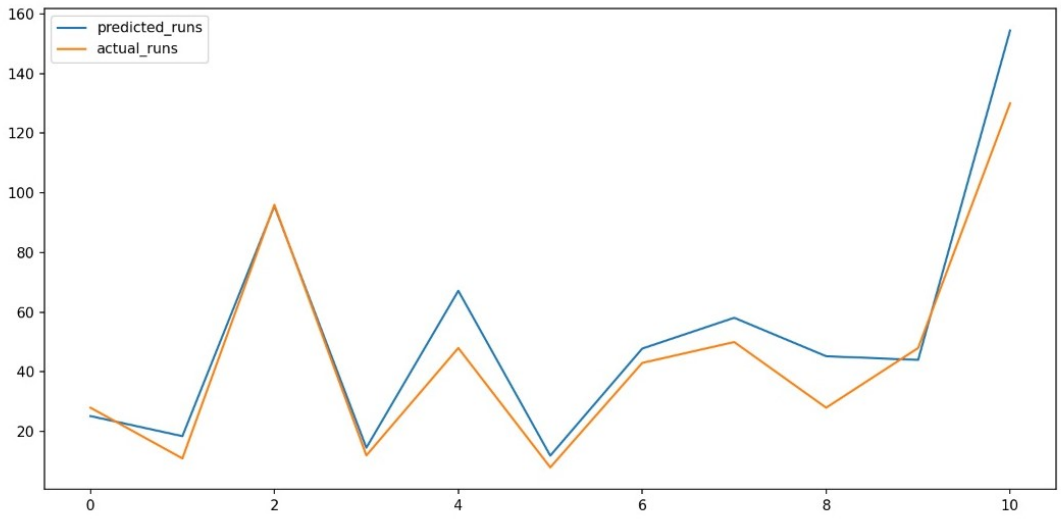

Fig 5.1 Predicted runs and Actual runs

Fig 5.1 shows the graph for the comparison between predicted runs and actual runs. The blue coloured line in the graph shows that is a plotting of predicted runs and the orange coloured line in the graph shows that is for actual runs. It is understood that the random forest model predicts the runs higher than the actual runs. In the next phase, this prediction will be done using few more machine learning algorithms namely support vector machine, naive bays and decision tree and the comparison of the accuracy will be discussed in detail.

REFERENCES

[1] M.O. Faruque Shamim M.M. Rahman and S.Ismail. An analysis of bangladesh one day international cricket data: A machine learning approach. *International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 49:190–194, 2018.

[2] S. P. Singh S. Agrawal and J. K. Sharma. Predicting results of indian premier league t-20 matches using machine learning. *8th International Conference on Communication Systems and Network Technologies (CSNT)*, 58:67–71, 2018.

[3] Parteek Bhatia Tejinder Singh, Vishal Singla. Score and winning prediction in cricket through data mining. *2015 International Conference on Soft Computing Techniques and Implementations-(ICSCTI)*, 78:500–507, 2015.

[4] Sarath Chandra Makkena Swayam Swaroop Mishra Vishnu S. Pendyala Shivam Tyagi, Rashmi Kumari. Enhanced predictive modeling of cricket game duration using multiple machine learning algorithms. *International Conference on Data Science and Engineering (ICDSE)*, 110:10.1109/ICDSE50459.2020.9310081, 2020.

[5] Dataset. https://www.kaggle.com/.

[6] Details about cricket. https://www.espncricinfo.com/.

[7] Learning Flask. https://www.mygreatlearning.com/blog/everything-you-need-to-know-about-flask-for-beginners/.

[8] Machine-Learning. https://www.javatpoint.com/machine-learning.

[9] Model-Building. https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1.

[10] Pre-Processing. https://www.javatpoint.com/data-preprocessing-machine-learning.

[11] Pickling. https://www.analyticsvidhya.com/blog/2021/08/quick-hacks-to-save-machine-learning-model-using-pickle-and-joblib/.

[12] Setting up flask in Pycharm. https://www.jetbrains.com/help/pycharm/creating-flask-project.html.

[13] Sci-kit. https://scikit-learn.org/stable/.