A REPORT

ON

World Cup's Best Player Prediction

By

Vivek Nimmagadda **15B00288**

Prepared in the partial fulfillment of the
Practice School III Course

AT

DIVYATEJA IT CONSULTANCY SERVICES PVT LTD. I DITCOS I
9th Floor I Vaishnavi Cynosure I Gachibowli I Hyderabad I 500032 I

A Practice School III Station of

BML MUNJAL UNIVERSITY

(May, 2019)

Certificate of Authenticity

CERTIFICATE

This is to certify that Practice School Project of **Vivek Nimmagadda** titled **World Cup's Best Player Prediction** is an original work and that this work has not been submitted anywhere in any form. Indebtedness to other works/publications has been duly acknowledged at relevant places. The project work was carried during **January 7, 2019** to **May 10, 2019** at **Divyateja IT Consultancy Services PVT LTD**.

Signature of PS-III faculty	**Signature of industry mentor/Supervisor**
Name: Dr. Brij Bihari Dubey	**Name:** B.V.S. Sastry
Designation: Assistant Professor	**Designation:** Vice President Operations
(Seal of the organization with Date)	***(Seal of the organization with Date)***

BML MUNJAL UNIVERSITY

PRACTICE SCHOOL – III

JOINING REPORT

Date: January 7, 2019

Name of the Student	Vivek Nimmagadda
Name and Address of the Practice School – III Station	**DIVYATEJA IT CONSULTANCY SERVICES PVT LTD. I DITCOS I 9th Floor I Vaishnavi Cynosure I Gachibowli I Hyderabad I 500032 I**
Location of the Project	Hyderabad, Telangana
Name and Designation of the Industry Guide/ Industry Mentor for the Project	B.V.S. Sastry (Vice President - Operations)
Organization Contact No.	9581576105
Organization E-mail Address	**sastry.bvs@ditcos.com**

ACKNOWLEDGEMENT

I am extremely indebted to **BML Munjal University,** and the **Vice Chancellor and Dean (SoET)** for providing me with this golden opportunity to implement and learn things in a practical approach.

I would also like to thank **Mr. B.V.S. Sastry (Vice President - Operations, DITCOS)** for readily accepting me to join and work in their esteemed organization and also for guiding me throughout my project work in spite of his busy schedule.

I am also very thankful to my faculty mentor **Mr. Brij Bihari Dubey** for his constant supervision and guidance and also for his timely feedback regarding my progress which helped me to traverse this far in completing the project.

I would also like to thank **Mrs. Seema Yadav** for providing us with all the necessary information and important deadlines regarding the project on a timely basis.

Finally, I would also like to express my gratitude to the teaching staff and all my friends and colleagues who have helped me in transitioning an idea into a project.

ABSTRACT

Many statistical techniques which predict the outcome of a cricket match have traditionally used only the total number of runs scored by each team as a base parameter for evaluating a team's performance and predicting future results. However, the total number of runs scored possesses a key random element which leads to large inconsistencies in many games between a team's performance and the number of runs scored or conceded.

The fundamental purpose of this project was to develop an application which can predict the best player and the best team who has the potential to outperform others in the upcoming Cricket World Cup 2019 based on their current form. The project can broadly be classified into three phases where the first phase deals with scraping the individual cricket player's data from a website, second phase is to perform some machine learning algorithms on it and predict the best player and the best team who has the highest probability to win the World Cup and the third phase is to visualize these results. The player data is scraped using UiPath tool, the visualizing is performed on the results achieved from the machine learning models and the GUI is implemented with the help of the Tkinter module of python. The scraped data is meaningfully visualized in order to make the data better understandable to the viewers. This project provides a glimpse of what is to happen in the upcoming tournament based on the outcomes of the previous tournaments.

TABLE OF CONTENTS

1. A Brief Introduction of the Organization's Business Sector

Divyateja IT Consultancy Services Private Limited comes under the category of IT Service Companies. The companies under this sector are mainly devoted to the production and development of products and services which will be used in the manufacturing processes of other companies.

The service sector is not only one of the dominant sectors in India's GDP but has also attracted substantial foreign investments, contributed considerably to exports as well as provided large-scale employment opportunities. India's service sector deals with a wide variety of activities such as transportation, trade, financing, hotel and restaurants, storage and communication, etc. where Information Technology is also one amongst these.

The Indian IT sector has a huge global presence, spreading over 200 cities and 86 countries across the world which is expected to grow even more fiercely in the coming years. With around 5 million people employed, the sector is one of the major employment generation contributors and is also the 4th largest urban employer of women, decreasing gender discrimination and leading to social progress.

Tremendous digital transformation across various industries, growing internet availability, affordable new technologies such as artificial intelligence, machine learning, robotics, and virtual reality and rising demand for cost-effective IT services and are some of the major demand drivers for this sector.

As a result of some of the Government initiatives in recent times such as Digital India, the domestic IT market in India has achieved high growth. There are also significant opportunities for the sector to exploit in terms of regional global expansion because of the increased investments from Japan and China in the Asia Pacific region.

2. Overview of the Organization

Divyateja IT Consultancy Services Pvt. Ltd. (DITCOS) was incorporated under the constitution in the year 2015 under the leadership of Mr. Pasupuleti Ramana, who has a work experience of more than 2 decades in the IT domain. Other branches of the organization include Singapore and Dubai.

The company has expertise in exploiting the latest trending technologies like Artificial Intelligence, Machine Learning, and providing solutions related to software products, platform using innovative technologies. The services of the company range from Application services to a full ITES (Information Technology Enabled Services), NextGen technologies like IoT, Mobility, Cloud & Big Data. One can take individual services separately on a standalone basis, or as a combination of a full-service oriented solution. They provide practically feasible and cost-efficient solutions to their clients with the required tools and technologies along with the resources to meet their growth objectives. Apart from these technologies, few of the employees of the organization have also expertise in RPA (Robotic Process Automation) which is a very rapidly growing domain in the IT sector currently.

DITCOS is also a strategic partner of a company named Arago for the ASIA PACIFIC region, a leading artificial intelligence company. They assisted Arago in building an AI platform focused on the B2B sector using machine reasoning and analytical components to improve efficiency and promote innovation. Some of the employees of DITCOS from the RPA department had recently worked with a pharmaceutical company which is located in Bangalore named Strides Pharma Science Limited (Strides). They worked on invoice automation where the main aim of the department is to check whether all valid invoices from vendors are approved, processed, and paid properly. Recently, they also signed an agreement with PepsiCo, India where they will be working on another similar invoice automation project.

There are many competitors for DITCOS who provide the same or even more diverse set of solutions to their clients. Some of these include Infosys, Cognizant, Deloitte, Wipro, Tech Mahindra, etc. The total number of employees in the company fluctuates in between 1 - 30.

3. Plan of the Internship Program

I started working with the RPA team as an Automation Intern from February 2019. Duration of my internship is from 7th January 2019 to 10th May 2019. Team deals with the automation of various business tasks and processes at client locations across several different companies located in Bangalore. Some of the companies which they are currently working with are PepsiCo and Strides. They are working with a project related to invoice automation. Previously, I was assigned to the web-development department until February 2019. RPA ensures less manual involvement with high reliability for doing monotonous tasks with very few errors on a regular basis which in turn decreases the total incurred costs. It also helps in procuring new licenses when needed, managing and optimizing the software needed to run a business. Presently, RPA stands out to be a major part of annual revenue for the company, followed by IoT and web development.

I was asked to pursue some certifications related to RPA after I was shifted to this department in order to get some understanding about the topic and after finishing up with those certifications, I was assigned with an internal project where my task was to automate the payroll system of DITCOS. I put in my best efforts into the project and finally automated the entire process by the end of March. I am very thankful to a few of the employees and some of the UiPath forum users for helping me in solving some errors which I encountered during this project. Due to the confidentiality of this project which even consisted of passwords of some of the employees, I couldn't disclose the entire project and as a result with the consent of my respective industry mentor and my faculty in charge, I started working on another project named "World Cup's Best Player Prediction" which utilizes some of the market's leading IT technologies such as RPA, Machine Learning, and Python.

4. Introduction

Cricket has always been one of the most renowned sport not only in India but also throughout many countries in the world and is followed very keenly by a large number of people. From the time of its inception, the fan base for cricket never dropped, particularly in Asian countries ultimately leading to the start of newer cricketing leagues such as the IPL and teams started investing huge sums of money on the players as a result of this popularity. Nowadays each and every team is having their own data analyst who collects all the data related to a particular player and visualizes it in order to assess the player's recent form and fitness which not only helps to understand players better but is also making the task of selectors easy during the screening process. BCCI had also recently announced that they considered the performances of players since the completion of Champions Trophy 2017 for selecting the Indian World Cup 2019 Squad. This collected data is playing a very vital role in driving the teams towards success. Some of the key uses of this data are:

- Formulating match strategies, tactics and player analysis.
- Identifying players individual batting and bowling styles.
- Player acquisition and valuation at the start and end of the tournament, team spending.
- Helps to focus on weak areas of the player.
- Predicting and preventing injury concerns by varying workloads.
- Performance, Match, and League outcome predictions.
- Designing and scheduling the tournaments.
- Odds calculation for professional betting.

5. Motivation

The most important function of a data analyst is to evaluate a team's performance in games and use this data to predict the result of future games. But these outcomes can be very difficult and complicated to predict, with surprises popping-up every now and then.

The possible results for any team taking part in a cricket match are a win, loss, or draw. As a result, it may, therefore, seem quite straightforward to predict the result of a game. Traditional methods have simply used match outcomes to evaluate team performance and build predictive models to predict the outcomes of future games.

However, due to the varying nature of the pitches and the climatic conditions across various countries, there is a random element linked to the total number of runs scored and the total wickets taken during a cricket match. For instance, a team with a good offensive batting record could succumb to a bowling unit where the climate aids more swing bowling, whereas a team with a poor offensive batting strategy could win the match because of the aid their bowlers got from the weather. This makes match outcomes an immature measure of a team's achievements and therefore an incompetent metric on which to predict future outcomes.

One of the potential solutions to this problem can be to use in-game statistics to dig deeper than the simple match outcomes. In the past few years, some websites have made the in-depth match statistics available, which opens up the opportunity to consider other parameters as well. This has enhanced the chances of more accurately predicting the total runs scored by a team, removing the random element of scoring runs.

With the recent discoveries in more sophisticated machine learning techniques, better predictive performances are achieved in a wide range of classification and regression problems. The exploration of these algorithms is also leading to the development of even more powerful models which not only helps in predicting the result of a match but also helps in analyzing the performance of players when they are subjected to pressure.

6. Objectives

This project aims to explore more possibilities in getting a deeper understanding of the collected data and also helps assess a team's performance particularly in their home and away games in the World Cups. This has been possible because of the large amount of data which is being recorded in various websites.

Various Machine Learning Classification techniques will be tested on the previously concluded world cup matches and several different hypotheses will be explored in order to maximize the predictive performance of the models and find a model which best suits the data.

In order to visualize these predictions, there are some objectives which need to be fulfilled: Firstly, I need to find some quality data and cleanse it to be used as an input for my models. Suitable data sources must be explored to achieve this task which will, in turn, allow me to have access to a large number of stats to work with, compared to most of the previous research which has fairly been done on the subject where only the end result of a match is taken into consideration.

The main approach will be to build a model for visualizing the win/loss ratio in order to understand a team's performance better and ultimately leads to better predictions in the future. For predicting the results of a match, we will be testing various Classification techniques to obtain the best possible performance.

Parallelly, I will keep track and scrap individual player scores as many international matches are still being played ahead of the world cup. This will allow me to better assess the player's current form and ultimately generates better predictions for the upcoming games.

Another important aspect of this project is to build a suitable Machine Learning training and testing mechanics to be able to test new models, with new features and functionalities, and compare it with other models, which will, in turn, give an idea of how well the data is fitting to various Classification models.

Finally, the model which best fits our data will be assessed against some benchmark prediction methods. A fruitful outcome for the project would be to create a classification model which could predict a future game's result accurately, and a python expression which fetches the best players based on their statistics and returns their names.

7. Challenges

The following are the challenges which I have faced along my journey towards achieving the objectives of the project:

- **Data availability & quality:** Finding a website which consisted of all the required data with the necessary parameters. The previous match data present in most of the leading cricketing websites are automatically archived after a particular period of time and a lot of searching has to be done to fetch that data. On continued research, I was finally successful in finding a website named “howstat.com” which helped the cause.

- **Scraping the data:** Once the required data is found on the website, the next big challenge is to scrap the data using a tool. **Note:** Scrapping can only be done using a tool because manual scraping requires a lot of patience, time and effort. Also, these kinds of repetitive tasks are very monotonous and boring to implement.

- **Understanding of prediction boundaries:** In order to design a model which best suits the dataset and test different hypotheses, we will need to undertake thorough background research of various prediction methods and develop a statistical understanding of different Machine Learning algorithms that can be used for our predictions.

- **Visualizing the plots:** One of the key challenges was to visualize these plots using some kind of a GUI. Finally, I have successfully implemented this with the help of Tkinter (Python).

8. Literature Survey

Ever since the Machine Learning techniques were introduced to predict the outcome of a sports match, a considerable amount of research has been done in this area and this section highlights various citations that have already been done and which were useful to my project.

8.1 Football Result Prediction Using Machine Learning

Previously, only specific parameters were used to be taken into consideration for predicting the outcome of football matches. But there will be many parameters which will determine the total number of goals scored by a team. One such important parameter is the random element which can lead to huge inconsistencies in the predictions made. The meaning and importance of random element have been clearly explained in the project implemented by Corentin Herbinet named 'Predicting Football Results Using Machine Learning Techniques[1]'.

8.2 Sport Result Prediction Using Artificial Neural Networks

Various Machine Learning algorithms are being used for predicting the outcomes of a match. But one of the most powerful and widely used algorithms which have the highest amount of accuracy is artificial neural networks. This technique is mainly gaining popularity as the demand for accurate match predictions are significantly increasing due to the large monetary investments involved in professional betting. The following paper written by Rory P. Bunker and Fadi Thabtah named 'A machine learning framework for sports result prediction[2]' provides a critical analysis of the literature, mainly focusing on the application of Artificial Neural Network (ANN) to predict the result of a match.

8.3 English County Match Result Prediction

With the inauguration of new cricketing leagues across the world, the use of various Machine Learning techniques for predicting the outcome of matches has always been on the rise. The outcome is being particularly predicted for analyzing a player's performance and also for professional betting purposes. The following article named 'Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches[3]' written by Stylianos Kampakis is an optical model which is a combination of simple prediction method with complex hierarchical features.

8.4 UiPath Web Automation

Fetching the player data is only possible with the help of some scraping technique as the data required is relatively large and the fast and easy way is to use the UiPath tool which is free and is also user-friendly. The following blog on web scraping named 'UiPath Web Automation – One Stop Solution To Web Extraction[4]' written by Sahiti Kappagantula gives an overview of the techniques required for web scraping.

8.5 Scraping and Saving the Data Using UiPath

Data scraping is the process of fetching data from a website and storing it in our local system. It's one of the most effective ways to collect data from the internet, and in some cases can also be used to channel that data to other websites. In the following article named 'Data Scraping From A Website And Saving To SQL Server Using UiPath (RPA)[5]' written by Sarathlal Saseendran, the author explains how to fetch the data from a website and store it in a variable to upload the same on to an SQL server database.

8.6 Fitting Logistic Regression to the Dataset

I have used many classification models for predicting the outcome of a match in the upcoming tournament. One among those models is the Logistic Regression which classifies the points into any of the two classes available based on the incomes. An article named 'Evaluating a Classification Model[6]' written by Ritchie Ng better explains the use of Logistic Regression and the confusion matrix.

8.7 Evaluating the Best Classification Model

As I have used many classification models in my project, one of the main objectives is to observe which model best fits the data and which makes better predictions than the other. This article named 'Machine Learning Classifiers[7]' written by Sidath Asiri explains the difference between all these models and the metrics to evaluate the model which fits out data better out of all of them.

8.8 Visualizing the Data

Visualizing the data is always one of the most important phases of predicting the performance of players or even a team. It is a pictorial representation of the scraped data and as we are all well familiar with the fact that visuals can be more appealing compared to the data itself and it's always better to plot and visualize the data. The person who is working with this data must have a clear understanding of it in order to visualize it. The following blog named 'A Statistical Analysis of the two biggest stars in IPL[8]' written by Raj Siddarth describes the importance of visualizing the data and provides us with an idea of different ways in which the data can be visualized.

9. Methodology

As discussed earlier in the abstract, the project can be divided into 3 phases where the last two phases are same for all the three predictions which are predicting the best batsmen, predicting the best bowler and the most probabilistic team to lift the world cup. The methodology followed for scraping the data is very similar for the first two predictions which will be properly explained with the help of flowcharts. Whereas, the data scraping part is a bit different for the team prediction. For simplicity, this section is further divided into subsections each explaining the different phases present in the project.

9.1 Phase 1

Phase 1 of the project deals with scrapping the required player data from the "howstat.com" website. Every player is assigned a particular unique ID in that website and for every row present inside the excel file, the website navigates to a particular player's profile based on that ID and fetches the data and stores it inside a different page of the same excel file for further processing. The process is repeated until it loops through and fetches all the players data. After this first step of scrapping, I found some errors in the data which were then removed by converting the entire data into numbers and the cleaned data is stored inside a CSV file in order to implement the Machine Learning models on it. The same process is repeated for fetching Batsmen as well as the Bowlers data. Whereas individual Teams data is fetched a bit differently. The flowcharts for scraping and eliminating errors are shown below.

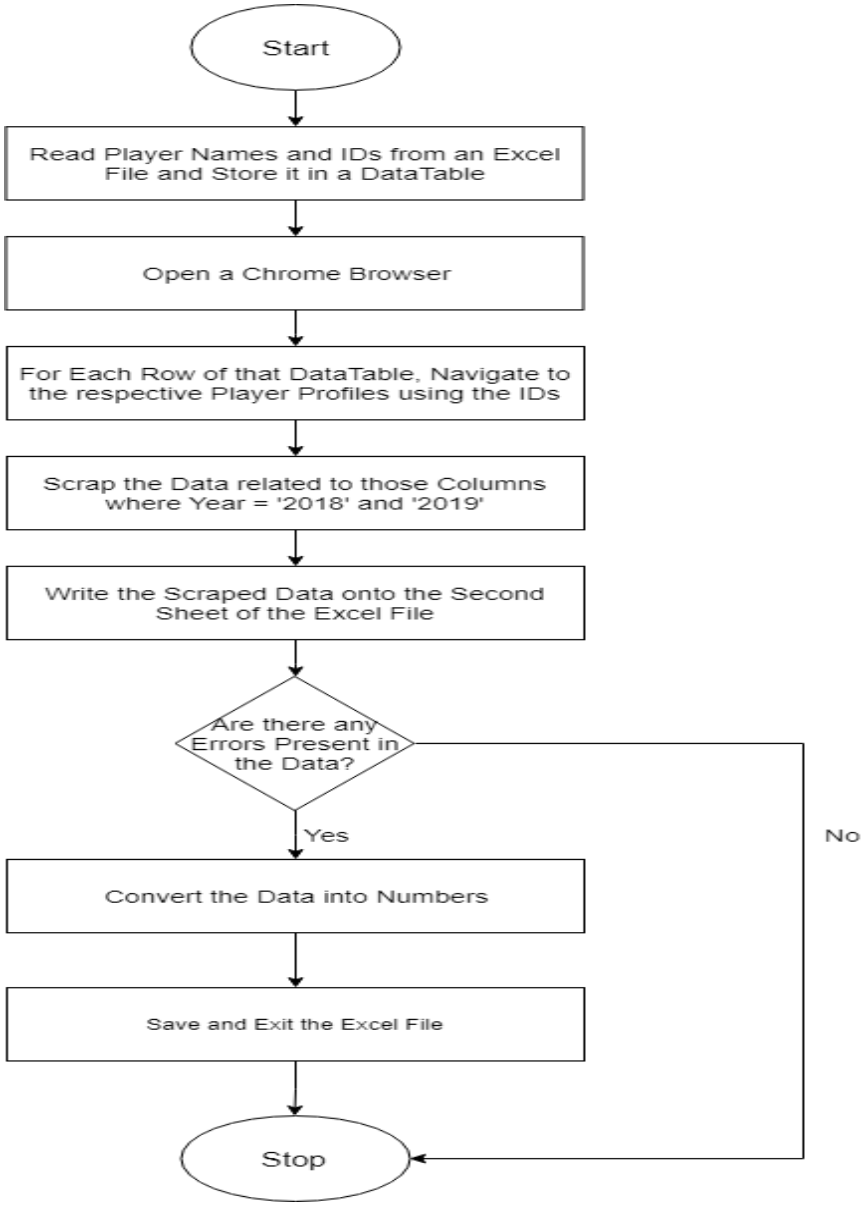

Figure 9.1

Scraping and Eliminating Errors in the Data

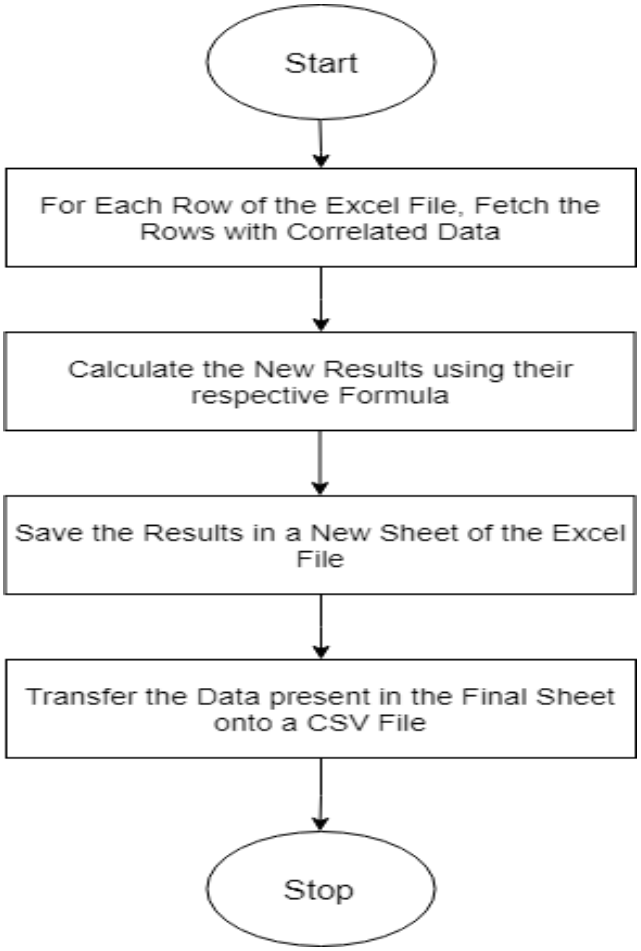

Figure 9.2
Combining Correlated Data and Storing in a CSV

9.2 Phase 2

Phase 2 of the project deals with retrieving the data and storing it in a data frame for performing various algorithms on it. Firstly, I wanted to visualize the batsmen based on their batting averages and their total runs scored in order for the viewers to get a better understanding of which batsmen are performing better. I have used K-Means clustering for this plotting and also used the elbow method to find the optimum number of clusters. Next, I visualized the batsmen who scored 4 or more centuries throughout the span of this one and a half years. Then I plotted a pie graph which shows the total percentage of runs a batsman scored along with their names. Finally, I used a python command which loops through the data frame and fetches a player who scored most runs with the highest average who will ultimately be the probable player for the best performer award. Similar measures were followed even for predicting the best bowlers. But, for predicting the best team I have used various Machine Learning classification techniques such as Logistic regression, SVM, K-Nearest Neighbors, Naive Bayes, Decision Trees and Random Forests on the previously concluded world cup matches in order to predict the results of the upcoming tournament.

9.3 Phase 3

Phase 3 deals with the visualization part which is done with the help of the Tkinter library of python. All the Machine Learning algorithms were separately written inside their respective functions. The Tkinter root window consists of a canvas where the previously created plots will be displayed. Separate buttons are present at the bottom which calls their respective functions. There's also a button which opens a new window and displays the image of the best player along with their names whenever it's clicked.

10. Outcomes

A fully functional python application has successfully been implemented which displays some buttons to be clicked in a Tkinter canvas. Each of these buttons is associated with a function which is called whenever the button is clicked. The application is able to accurately predict and plot various visualizations present in the project. Along with this, the application was able to recommend the best batsmen, best bowler, and most successful team out of all based on their recent performances. Apart from plotting the predictions, images are also displayed in the Tkinter canvas of a new root window and aside from the GUI, various Machine Learning techniques are also used to predict the probable result of a match when any two top eight teams face off against each other in the upcoming tournament. My model can predict the outcome of a match with a probability equal to 63.8% which is not great but at the same time not bad either which can further be improved by taking some useful parameters into consideration. Finally, Mac OS platforms have a tendency to not display the background colors of a Tkinter button which has been thoroughly examined and eliminated.

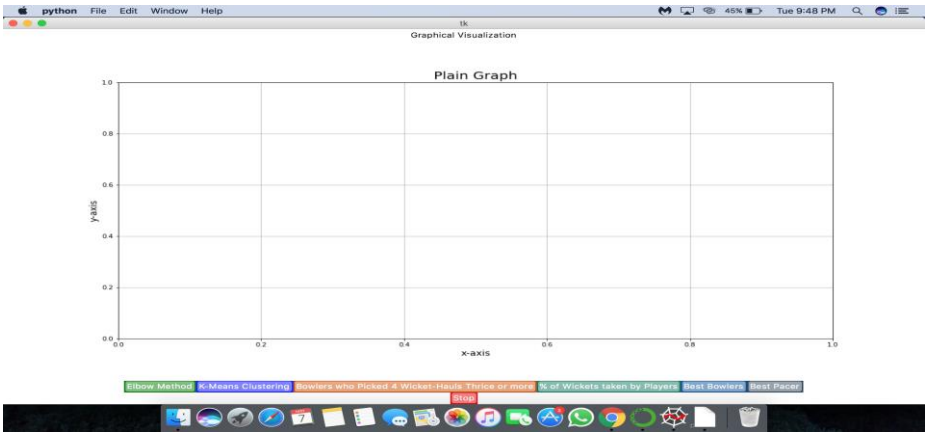

Figure 10.1
Player Prediction GUI

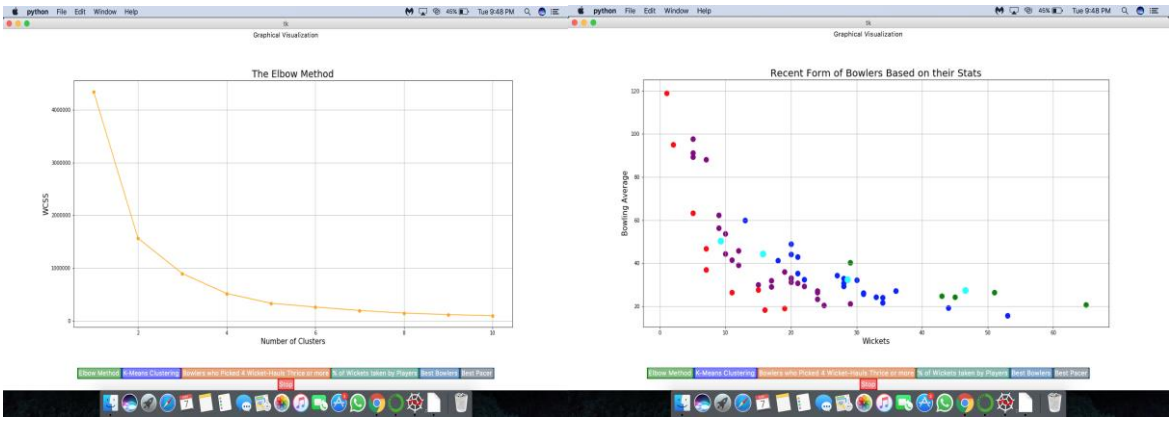

Figure 10.2
Elbow Method(left) and K-Means Clustered Data(Right)

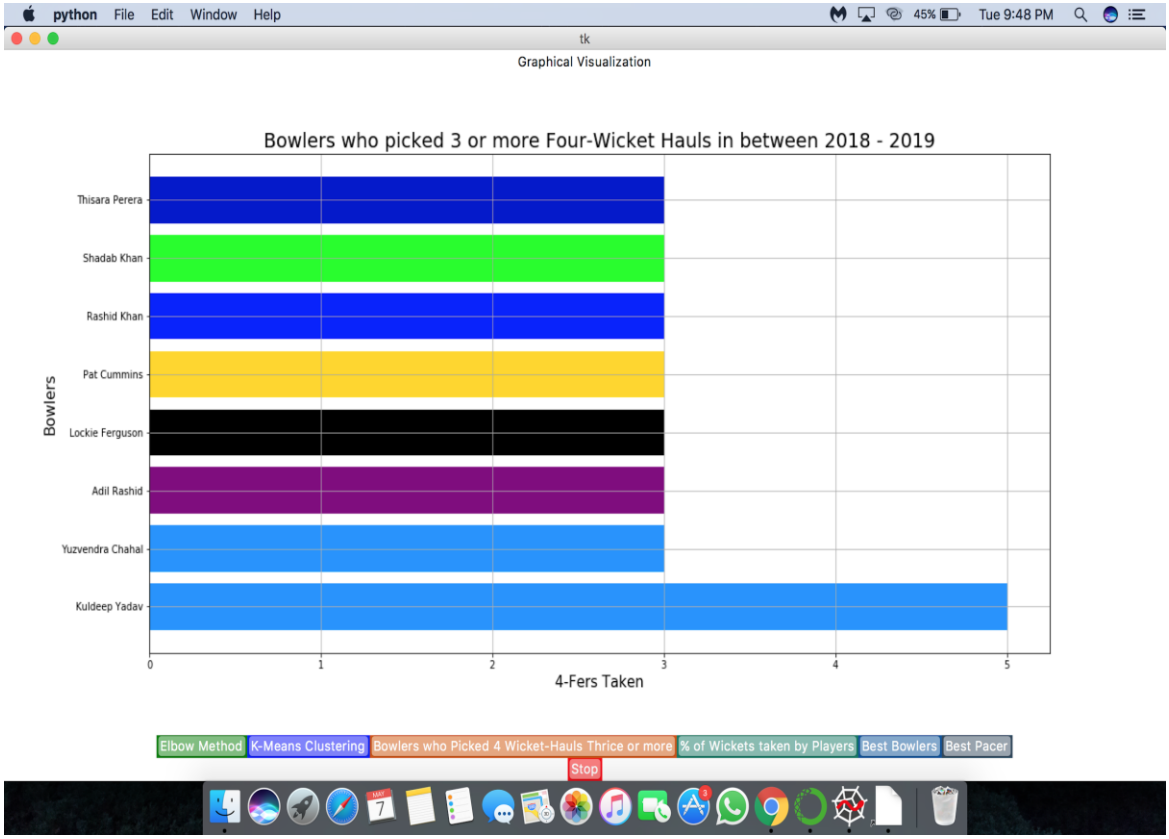

Figure 10.3
4 Wicket Hauls of Bowlers

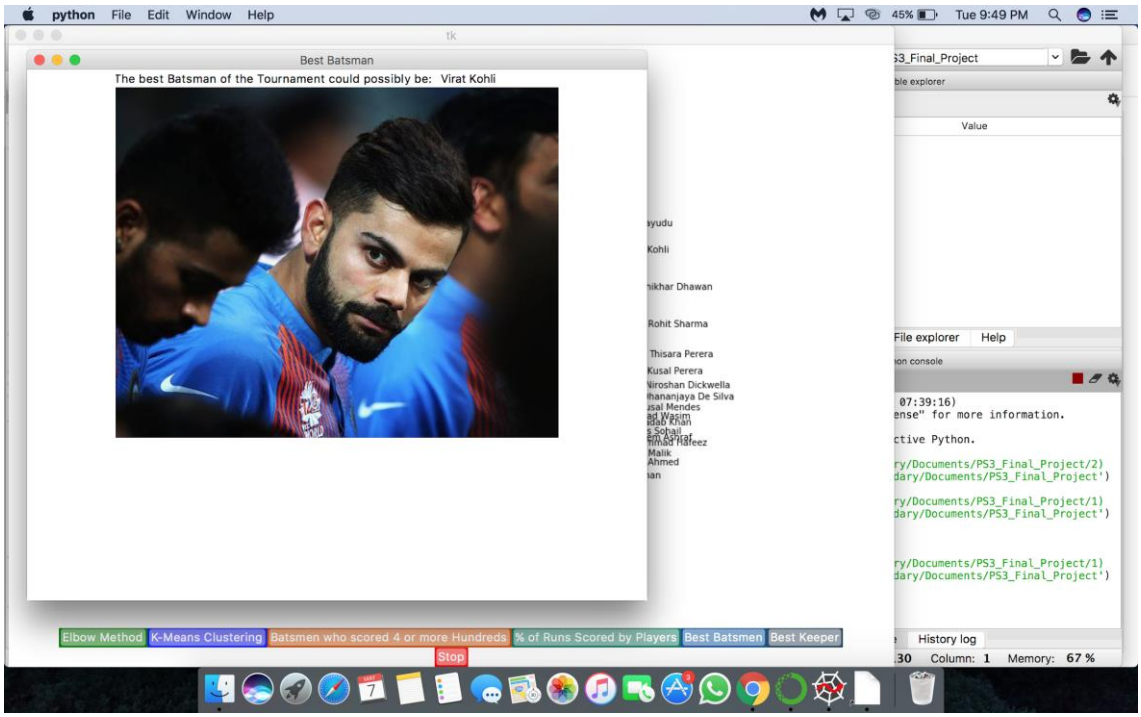

Figure 10.4
Predicting Best Batsmen

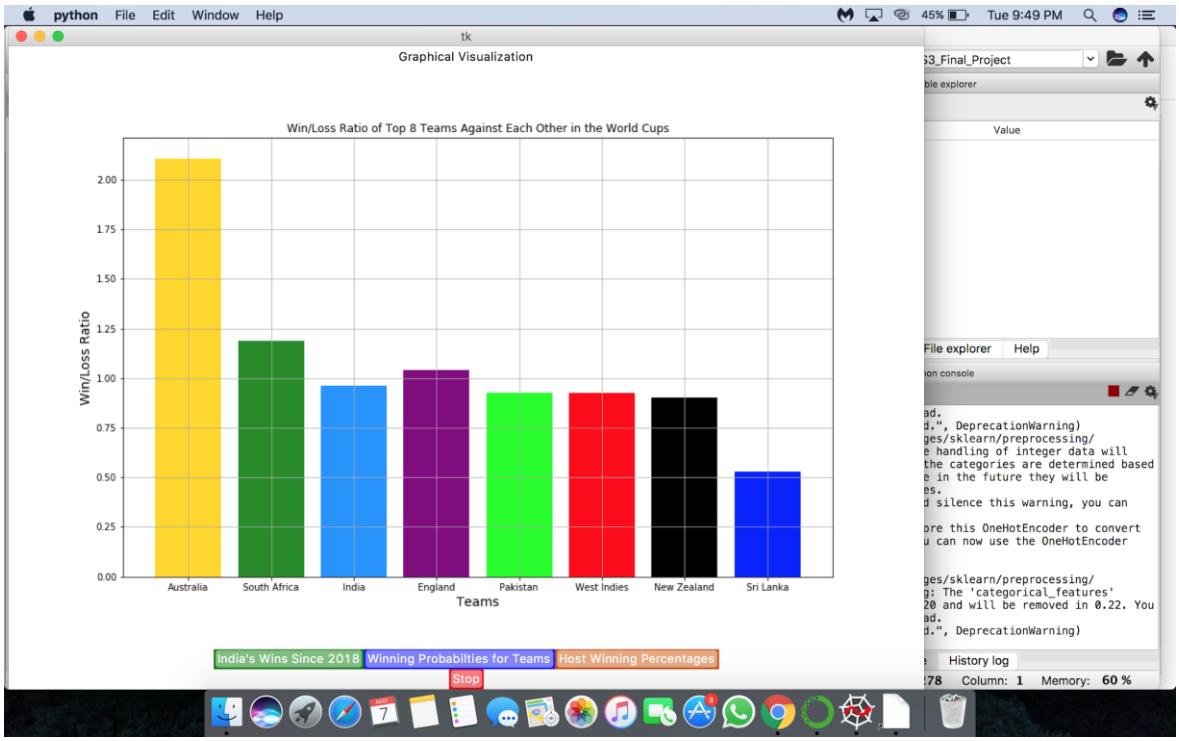

Figure 10.5
Best World Cup Performer

11. Future Scope

The use of various Machine Learning algorithms for predicting the performance of a team has shown great promise and potential for future scope. Although the use of complex algorithms has been fairly limited in my project due to limited time availability, this technology could be exploited even further in the coming future. Some of the improvements which can be made to improve the project are listed below:

11.1 Improved Data

Accurate predictions can only be made if we have large amounts of data to work with. The player data used in the project is only limited to the total matches he played throughout the year and the total runs he scored in those matches. Whereas in order to build a model with more predictive power, a lot more data is required such as the nature of the pitch or whether the team is batting 1st or chasing the target and many more parameters can be taken into consideration.

11.2 Using a Better Classification Model

One way to achieve even higher predictive accuracy is to use a better classification technique such as the Artificial Neural Networks which can take even more parameters as inputs and generate the required outputs.

11.3 GUI Enhancements

The user interface could also be improved by using Django which supports many rich illustrations, unlike Tkinter which is bound to simple interfaces.

11.4 Additional Functionalities

A new functionality can also be added where the project could automatically collect the data from the respective website and update the excel file if any of the information is modified on the website.

12. References

[1] Corentin Herbinet, "Predicting Football Results Using Machine Learning Techniques". *A project submitted in partial fulfillment of the requirements for the Joint Mathematics and Computing MEng of Imperial College London,* June 20, 2018.

[2] Rory P. Bunker and Fadi Thabtah, "A machine learning framework for sport result prediction*". A Research Article from Applied Computing and Informatics*, September 19, 2017, Volume 15 Issue 1.

[3] Stylianos Kampakis and Bhiksha Raj, "Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches". *A Research Article submitted in partial fulfillment of the requirements for Computer Science of University College London*, 2015, Document 1511.05837.pdf.

[4] Sahiti Kappagantula, "UiPath Web Automation – One Stop Solution to Web Extraction". *A blog listed on Edureka,* March 16, 2019.

[5] Sarathlal Saseendran, "Data Scraping from a Website and Saving to SQL Server Using UiPath (RPA)". *An Article listed on Sharpcorner,* November 12, 2018.

[6] Ritchie Ng, "Evaluating a Classification Model". *A Research Article listed on ritchieng.com,* 2017.

[7] Sidath Asiri, "Machine Learning Classifiers". *A Research Article listed on towardsdatascience.com,* June 11, 2018.

[8] Raj Siddarth, "A Statistical Analysis of the two biggest stars in IPL (Virat Kohli and MS Dhoni)". *A Research Article listed on Kaggle,* 2017.