National College of Ireland

BSc in Computing – Data Analytics

2017/2018

Kuldeep Rawat

X14348026

x14348026@student.ncirl.ie

Indian T-20 Cricket Game Modelling using Data Mining Techniques

Technical Report

Declaration Cover Sheet for Project Submission

Name: Kuldeep Rawat
Student ID: x14348026
Supervisor: Catherine Mulwa

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: Kuldeep Rawat

Date: 13/05/2018

Table of Contents

Executive Summary

Cricket is gaining popularity day by day, normally it is played in Asian countries more than European, even though it was invented by England. The project is based on T20 cricket format only. There are other formats such as one day, test and IPL (Indian Premier League).

This report provides information how the project is developed and what problems can it solve in the world of cricket game. The aim is to develop models using data mining techniques which can help the user to view the analysis that is developed by the cricket analyst. The first stage of this project is to gather dataset from websites such as espncricinfo.com. Then the analysis on matches, player and other aspects of the cricket is completed, for cricket fans discussion about the outcome of cricket matches is very popular and can go on for several minutes. There are many aspects that must be considered about a team such as player health, performance in the previous two matches, playing the position, age, history, and opponent.

This document intends to help the user understand the problems that are identified and the steps are taken to improve the problems related to cricket. The main objective of this project is to change the way people think about a game without knowing the truth or facts about a team, player, and manager. This project not only going to build models but also compare the models and show the facts that can help us predict the outcome of the match. With the help of this project, it can be a benefit for many people such as team manager, people who like betting and fans who want to gain insight into their favorite team or player.

The go for doing this is to extract, clean and perform analysis on data that is gathered and use visualizing tools such Tableau to view analysis for the end users. The users can then run analysis as she/he wishes. In this project, the data mining techniques such as Linear, Multilinear, Polynomial Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Decision Tree is developed. The KNN and SVM produced the best results, KNN (100%) and SVM (85%). The other models are not good as SVM and KNN, still, lots of knowledge can gain from them

1 Introduction

Cricket is a game that is extremely prominent in India and Australia. It initially started in England in the sixteenth century and later spread to different nations, the first international game, however, isn't played in England, it was played between the United States and Canada in 1844. The cricket is played on an oval-shaped playing field, there is no settled measurement for the field yet but the main action takes place on 15-yard territory which is known as the pitch the center of the huge playing field.

There are many different formats of the cricket game such as Test, One Day International (ODI) and T20. The most popular one is T20 as it consists of 20 overs and usually played for 3hours hours whereas test match goes for 5 days and ODI 7-8 hours. There are many different versions of T20 one of the popular ones is Indian Premier League(IPL) which is normally played in India and teams can is formed using the mix of players from other countries so basically an Australian, South African and Indian team player can be in the same team. This project mainly focusses on the T20 international format of the Indian cricket team.

1.1 Project Motivation and Background

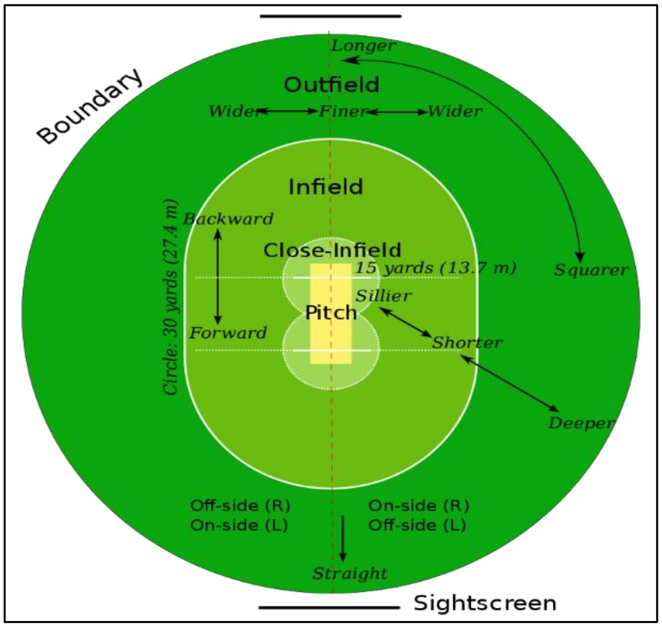


Figure 1: Cricket Pitch

Figure 1 represents the Cricket pitch, which is a game played between two groups of 11 players, where the two groups substitute bowling). A player (bowler) from the bowling group delivers scoring(batting) and defending (a ball to a player (batsman) from the batting group, who should hit it with a bat to score while the rest of the field team (fielders) defend the scoring. Moreover, it is a group activity, the bowler, and batsman, and defenders to some degree, follow up on their own, each completing certain lone activities autonomously, similar to sports baseball.

The idea of doing analysis on cricket generated from personal experience while having conversation talking to a friend about specialization for the 4th year. A question, what do data analyst do? was asked during the talk, the answer was that they perform analysis on certain information. It wasn't quite clear to the friend, therefore, an example of cricket was given, suppose if someone want to predict an outcome of a match, how would they predict it, they would need to know both team history such as player, team performance and then decide which team has higher chance of winning the match. After providing the example it was clear that data analyzing require collecting, extracting and transforming data. Also, show the analysis of charts or graphs to help understand what the result of the analyzing mean.

As a cricket fan decision of prediction model on Indian cricket team is decided, developing this idea will allow many experts to see their player or team weakness and could help them to improve their game and help them achieve what they want. The data from well know cricket website is gathered which provides proper data for cricket matches and a team player. This has the potential to help in the real-world situation if there is a match that is being played and results of the can be determined before the match begins.

A question such as who will win the match? which team has potential to win and why? what can the team manager to improve the game or change the game? There are many benefits for fans and people who like betting on players and teams. The goal of this project will be to address all the question and provide a solution to the problem and the questions. By doing this analysis it can help to better understand cricket and new way of providing better outcome using analysis and comparing them to social media.

The reason for choosing this project is not just because of liking the cricket but to gain more knowledge of cricket while learning how to extract, clean and perform the analysis. As a fan of cricket, the need for exploring the world of cricket is necessary

1.2 Project Scope

The scope of the project is to develop an Indian T-20 Cricket Game Model using data mining technique such as Linear, Multilinear, and Polynomial. Also, SVM, KNN, and Decision Tree. The results will be visualized in Tableau, the dashboard will present the results and provide all the relevant information regarding the matches such as the number of matches played, a probability of a team winning a match, top teams of the T-20 game and Top partnership.

Schedule: The dashboard which contains the information retrieved using data mining technique will be broken down into many small tasks and it will be developed over a period, which will be completed within nine months this includes tasks such as gathering requirements, data extracting and cleaning. Developing models. The tasks such as implementation, evaluation, debugging and finally creating a dashboard to show the result need to be achieved.

1.3 Problem Statement

There are many issues within the cricket field that I would like to address through this project, the most common problem such as inconsistent team selection, constant change in the batting order, lack of match practice for reserve fast bowlers and inexperience spinners. I also want to address problems that are currently faced by Indian cricket team such as poor fielding, different bowling style, conceding too many extra and bowler orders. The analysis will help to overcome these problems by knowing exactly where things are going wrong.

1.4 Aims

1. The first objective of the project is to find a dataset, which includes information about the Indian cricket team and the matches. The data must be real and should contain detailed information based on the number of matches that team India has played against other teams over the past years. The dataset is from ESPNCricInfo.com. website
2. The second objective will be focused on extracting and cleaning the dataset that is relevant to my projects such as matches, players, opponent teams and batsman. This objective is very important for the project as this requires cleaning the data, correct and accurate information must be gathered to proceed. To distinguish the work literature review must be concluded, see whether this project has been completed before or not, if yes what is there that can be improved.
3. The third objective will be to compare the matches, player and team performance in terms of score and result. The outcome this objective will play an important role when doing the analysis.
4. The fifth objective is to transfer the clean data and generate a better dataset to perform the analysis. It will also help me to identify a pattern based on the results and see how the team has performed over the years and what could help them to win the match.
5. The sixth objective is to perform analysis and create prediction and classification model.
6. The final objective will be to complete the documentation with a conclusion and use a visualization tool visualize the result. This will help the users such as cricket player, manager, bettors and cricket fans to predict and gain the better understanding of the cricket matches.

1.5 Project Contributions

1. The results of the reviewed literature are presented.
2. A fully developed and evaluated Linear regression model.
3. A fully developed and evaluated Multilinear regression model.
4. A fully developed and evaluated Polynomial regression model.
5. A fully developed and evaluated Decision tree model
6. A fully developed and evaluated K-Nearest Neighbor model
7. A fully developed and evaluated Support Vector Machine model
8. Visualised results of all the developed models (2-8) above.

1.6 Technologies Used

R studio: R studio used to develop in the implementation of the models, the reason for choosing this is because it widely used by the data analyst, gained knowledge of R language achieve goals of the project as the project require processing of the data. The R studio is open source and can be used for graphics, statistic and loading data. The data that is downloaded from websites in the form of CSV file in Excel can also be used in R studio. R provides various libraries that are inbuilt and can make it easy for a user to perform analysis on certain data.

Python: The Python can be used to write the scripts the which plays an important role in cleaning and merging the data. It is one of the most used software for data analysis. In this project, it is being used to clean, merge and develop prediction models.

Excel: is a spreadsheet tool this allows us to build graphs can help us to see all the data very easily, therefore, it can be used to change data and store the cricket information in a structured way such as CSV.

Tableau: This application will be used for creating, visualizing and integrating the dataset to the dashboard. This is beneficial for the project as it can display all the work very easily, also Tableau will be accessible for users who want to view the analysis that is performed.

GitHub: The GitHub is to allow code to be pushed into the cloud, therefore, all the project can be accessed from anywhere, also keep track of when the code was uploaded and prevents from losing code.

PyCharm: This application is used to run python scripts.

Mendeley: This software is used to do a literature review.

1.7 Definition, Acronyms, and Abbreviation

The outlined Definitions, Acronyms and Abbreviations that are related to the project and the aim of this is to help the user gain a better understanding of the content:

KDD: Knowledge discovery in databases, this is a well-known process of discovering and gaining knowledge from data. It is very useful in data mining which is a technique to prepare data, select data, clean and interpret.

Visualization application: This allows the user to view and display the results. E.g. Tableau.

GitHub: It is web-based Git repository hosting service, which allows the user to save code and share easily with a link.

ESPNcricinfo.com: This is a website used to find datasets for cricket.

Microsoft Excel: This is a spreadsheet program it consists of table, rows, columns and can use for the variety of tasks such as performing analysis on data, solving complex arithmetic and functions.

R Studio: This is a programming application that is used for computing statistics using language R, it free and open source.

MySQL: This allows the user to access and store data. This open-source anyone can download and use it for the database purpose.

Tableau: This is a software which allows user visualize data and solve problems related to the data. Tableau makes it easy for the user to analyze data, easy and convenient way without wasting much time.

Algorithm: This is a step-by-step operation that is used in math and computer to perform a task such as calculating, data processing, also help the user to re-use the same algorithm to do other similar tasks again and again.

Machine Learning: This is the future of computer science, as this is a method of analyzing, it allows machines such as computers to find patterns that can't be easily processed by human minds. It helps to find hidden data and learn from data.

SVM: Support Vector Machine.

KNN: K-Nearest Neighbor.

IPL: Indian Premier League.

1.8 The Structure of This Document

The rest of the technical is structure as follows**:** Chapter 2 presents the literature review of ten past papers, Chapter 3 introduces the scientific methodology approach used and design diagram. Chapter 4 presents requirement specifications, data preparation, and feature selection. Chapter 5, presents implementation, evaluation, and results of Cricket game models which are developed using a multiple data mining techniques.

2 Literature Review on T-20 Cricket Game

2.1 Introduction

The research of related work is required to see whether similar work has been performed on T20 cricket analysis or not. The use of google scholar and IEEE library provided similar work that has been completed by other people. The finding includes the vast amount of work that has been conducted by people from all over the world.

2.2 Review of T-20 Cricket Game

The previous work on T20 cricket game has been completed by four authors, although it is not exactly same as this project because the models are created on the different version of the T20 cricket such as IPL. There are various techniques that are used, the most common one is Naïve Bayes and Decision tree. The following literature is closely related to this project. However, it is not exactly same therefore this project work is extended.

Pranavan Somaskandhan, Gihan Wijesingle and Sampath Degalla (2017), address problems for winning and losing the match by collecting data from all the IPL (Indian Premier League) matches and storing it in a relational database. Naïve Bayes, Support Vector machine algorithm are used to classify the problems. Additionally, Linear, Polynomial, RBF, and Sigmoid are used. Important attributes are identified, then the decision boundary which is generated using Naïve Bayes algorithm is visualized in the chart. This paperwork is well implemented which I believe can be conducted in my project, one of the technique which compares accuracy to a different subset of SVM provide good results. The author additionally mentions work that can be completed in future such as using clustering to divide player into different groups and analyzing player performance.

Gamage Harsha Perera (2015), in his paper ‘Cricket analytics' the data is collected from espncrickinfo.com, the data of test cricket matches is taken, this data set is unique because none of the other authors have performed analysis on three types of cricket format. Compare to others they have used ODI and IPL cricket data whereas this author is using T20, ODI and test cricket format. Clustering is used to group batting average and strike rate also two-player performances are analyzed in detail. This paper consists of the vast amount of information about cricket and many problems are addressed which position should a player play and performance of the player has measured three types of cricket format which provides acceptable results.

Prince Kansal, Pamkaj Kuma and Himanshu Arya (2014), in their paper, Naïve Bayes, multiple regression, and the decision tree are used to predict player price in IPL auction. The best accuracy is achieved by J48, whereas Naïve Bayes did not perform well. The author further adds this analysis can help IPL franchise in an auction to classify a player based on the player past performance.

Ananda B.W. Manage, Stephen M. Scariano and Ceil R. Hallum (2013) in their paper, first principle component is used to rank the top batsman and bowlers of the T20 cricket matches. Similarly, the data is gathered from espncricinfo.com for the year 2013, In this paper, both bowler and batsman performance is calculated and as result top 10 batsman of the year 2013 are revealed. The data is limited to the only year and use of only one technique does not determine the performance of a player, at least 10 matches should be considered.

2.3 Review of Existing Models Developed Using Data Mining Techniques

There are many different types of models developed by authors, the six-literature review below includes usage of Association rule, Naïve Bayes, Support Vector machine, Regression, Decision tree, Linear, Polynomial, and Sigmoid. Naïve Bayes and the Decision tree is considered as one of the best techniques to be used Naïve Bayes produced accuracy of 91%.

Ustsav Jadishbhai Solanki and Prof.Jay Vala (2017), in their paper algorithm called Association rule, is used to identifying meaningful patterns, the data is gathered from three websites espninfo, sports. ndtv and cricbuzz. Like the project conducted by Ujwal UJ, Dr. Antony PJ and Sachin DN (2018), this paper also contains limited data, only 10 matches are taken into the account. Although there is not enough data, the output result of hash apriori algorithm shows the accurate performance of the team players. This research help selecting a balance team member who has the potential to win the matches and make it hard for the opponent team to compete. This paper only focused on ICC (International Cricket Council) championship and the batsman performance, which it is not good enough as more aspects should be covered such as bowler and fielder performance because winning a match is heavily dependent on these aspects.

Ujwal UJ, Dr. Antony PJ and Sachin DN (2018), in their paper the data is gathered from two websites cricsheet and espncricinfo using Python. Google prediction API played an imperative role in choosing the suitable classifier, all the data is saved in Google cloud storage and data training is completed for predictions. The authors further point out, how their project can help the cricket player to change their strategy during the game to make it hard for the opponent to win the match. The project did not consist of enough data, only the 10 matches are observed, the result of the prediction shows 9 out of 10 matches are predicted correctly. The authors additionally added for future scope sentimental analysis can be implemented to gain the better understanding of players mood.

Hasseb Ahmad, Ali Daud, Yixian Yang and Haibo Hong (2017), in their paper ‘Prediction of rising stars in the game of cricket' focus on using player performance to predict rising star of the cricket. The algorithm Naïve Bayes, Support Vector machine and regression tree used. Compare to other research papers this uses feature evaluation metrics which includes information gain, gain ratio and chi-squared statistic for raking prediction. The dataset for ODI (One Day International) cricket matches is taken, ranging from 2006 to 2013. The paper consists of performing analysis on most of the aspects that reveal lots of information, as talked in previous work by Ustav Jadishbhai Solanki and Prof.Jay Vala (2017), how their work was lacking the use of other aspects such as batsman, bowler, and fielder performance. The author additional state that the rising star can also be predicted in a different format of cricket game such as T20, which the one that my project based on. There many techniques that can be learned and improved and can be applied to any file of the sport.

Tejinder Singh, Vishal Singla and Parteek Bhatia (2016), in their paper 'Score and winning prediction in cricket through data mining' they use of Liner regression on training data set, Naïve Bayes algorithms and Weka for exploring the data. This research has very limited data modeling techniques compare to others, the authors further add that Naïve Bayes achieved an accuracy of 91% whereas Linear regression is compared with an attribute called current run rate which shows that error linear regression is less than the error in current run rate. This research is again quite like the research conducted by Hasseb Ahmad, Ali Daud, Yixian Yang and Haibo Hong (2017), both research paper uses data from ODI cricket matches.

Sannoy Bhattacherjees, Jayakrushna Sahoo and Adrijit Goswami (2015), in their paper, Association rule data mining technique is used to see the performance of cricket player, again the data is gathered from similar website to other which is espncricinfo.com. In this research the R language plays an important part in performing the analysis on each player, the result of this research shows the player who scores low have lower strike rate and the player does not perform well during the away matches. The author further adds, the benefits of this analysis are that team can see which player performs well in certain position and location. Additionally, authors point out this technique is not 100% reliable as it does not provide a clear view, therefore, more analysis needs to be performed on other factors that are related to player performance such as pitch condition, player mood, team performance and opposition team.

P. UmaMaheswari and Dr.M. Rajram (2009) in their paper, principal component analysis on cricket match is performed like research conducted by Ananda B.W. Manage, Stephen M. Scariano and Ceil R. Hallum (2013), but the uniqueness of this research is that, it uses data which is gathered from observing cricket match videos. Individual player performance is analyzed one the example given in the paper is a cricket player Sachin Tendulkar which show there are 80.82% chance of Sachin striking the bouncing ball to the third man.

2.4 Challenges, Issues, and Problems

The challenges that most of the authors have encountered are not being able to accurately predict the result of the matches due to limited usage of the data mining technique, most of the time it is hard to determine which technique should be used and which one has the potential to produce an accurate result and that has the capability to satisfy the user. This is the same case here, where authors did not use a technique that can help them determine the best model for their data. The issues in most of the literature was that the data was very limited, in two of the papers, one by Ustsav Jadishbhai Solanki and Prof.Jay Vala (2017) and other by Ujwal UJ, Dr. Antony PJ and Sachin DN (2018), had issues with predicting accurate and reliable result due to lack of data. The other factors such as using fewer attributes also seem to be one of the major issues thus problems occur due to this reason.

3 Scientific Methodology Approach Used and Design

3.1 Scientific Methodology Used

The scientific methodology used is based on the Knowledge Discovery in Databases (KDD) because it is a process for discovering useful information from a collection of data. To achieve the goals this process is necessary for the success of this project. KDD includes the following approaches (Figure 2):

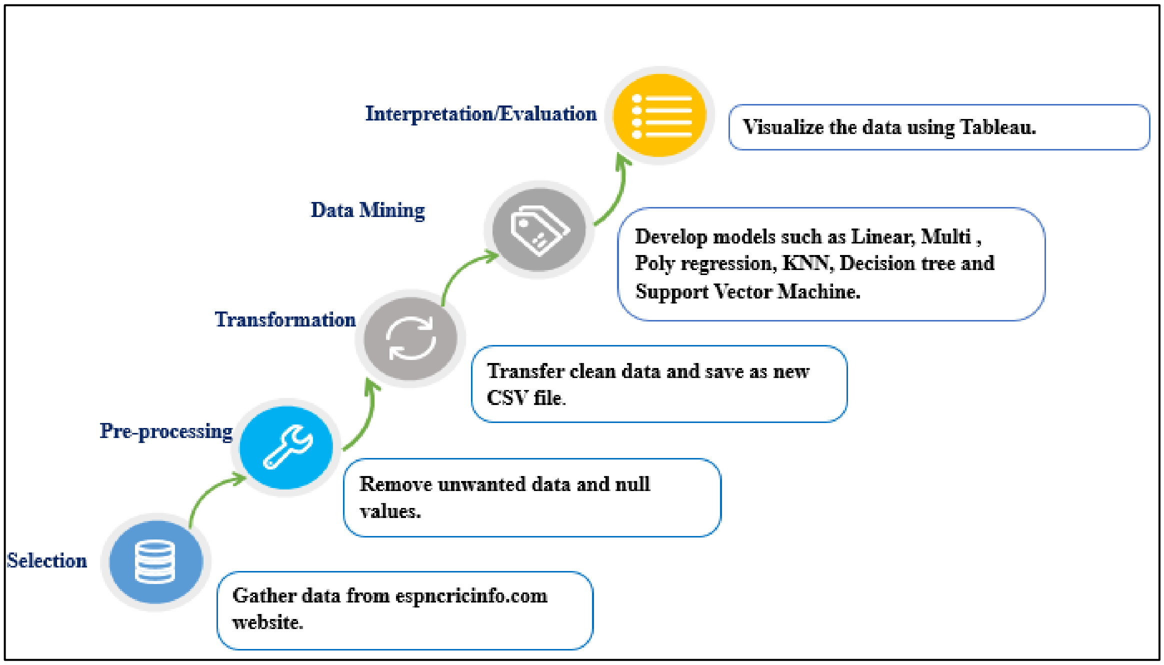

Interpretation/Evaluation
Visualize the data using Tableau.
Data Mining
Develop models such as Linear, Multi , Poly regression, KNN, Decision tree and Support Vector Machine.
Transformation
Transfer clean data and save as new CSV file.
Pre-processing
Remove unwanted data and null values.
Selection
Gather data from espncricinfo.com website.

Figure 2: Scientific Methodology Approach Used

Phase 1-Selection: This phase involves obtaining the T-20 cricket game dataset from espncricinfo.com website. This is very first and important phase of the KDD approach, here it involves finding and selecting the cricket data that is suitable for the project.

Phase 2-Processing: This is the second phase of the KDD approach, where the cricket data that has been collected is then cleaned which means the unwanted data is removed only the data that is required for the project is kept.

Phase 3-Transformation: This is the third step of the KDD approach, as it requires to take the clean set of the data which is produced by the second phase and convert it in the form of better data which can include reducing the dimension and then save it a CSV file.

Phase 4-Data Mining: this very important phase of the data, here the discovering of the important information from large sets of data is derived such as matches won, played and lost by a country. Techniques such as Linear Regression, Multi Regression and Polynomial Regression, Decision tree, KNN and Support Vector Machine is used.

Phase 5-Interpretation: The results are visualized in Tableau.

3.2 Project Design Diagram

The design diagram below (figure 3) describes how the data analyst is going to interact with the dashboard, how the query is sent and returned. It helps us to better understand how the whole process is working and how each component is interacting together.

The benefit of having this diagram is that it will help to stay focus on the behavior of the project.

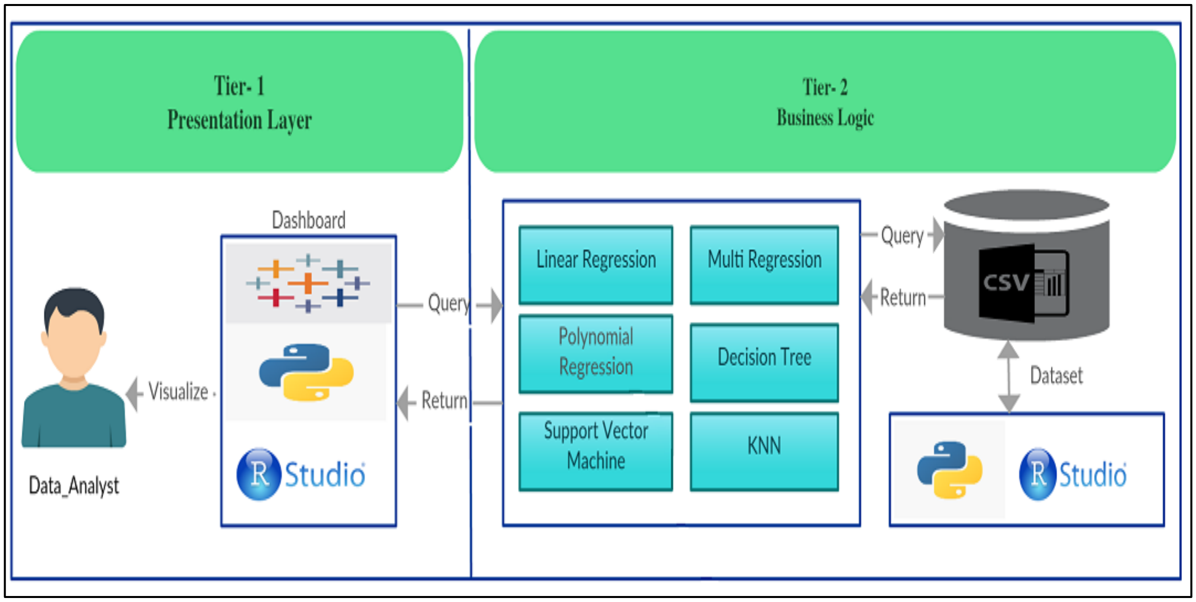


Figure 3: Design Diagram

4 Requirement Specifications, Feature Selection and Data Preparation

4.1 Requirement Specification

This section of the project includes all the requirement that is necessary for the progression, as the project continues, some requirement may be affected due to dropping off a functionality or adding more. The explanation for the changes will be provided.

4.2 User requirements definition

A requirement describes how software/product works and what qualities it must have to attract the users. For this project, the analysis must be performed. The result must be shown in a visualization application such as Tableau in a way so the user can easily interact, this will be achieved by creating a dashboard on Tableau. After completion of the project, the user must be able to access the dashboard, perform analysis on a player, check the best player of the year, predict match result, predict future of a player based on its performance.

To achieve these requirements, analyses on Indian T-20 Cricket match should be performed so the data can be used to determine the performance of a team and probability of winning the match. Also, this will be able to show if there is any relationship between player's performance in matches and result.

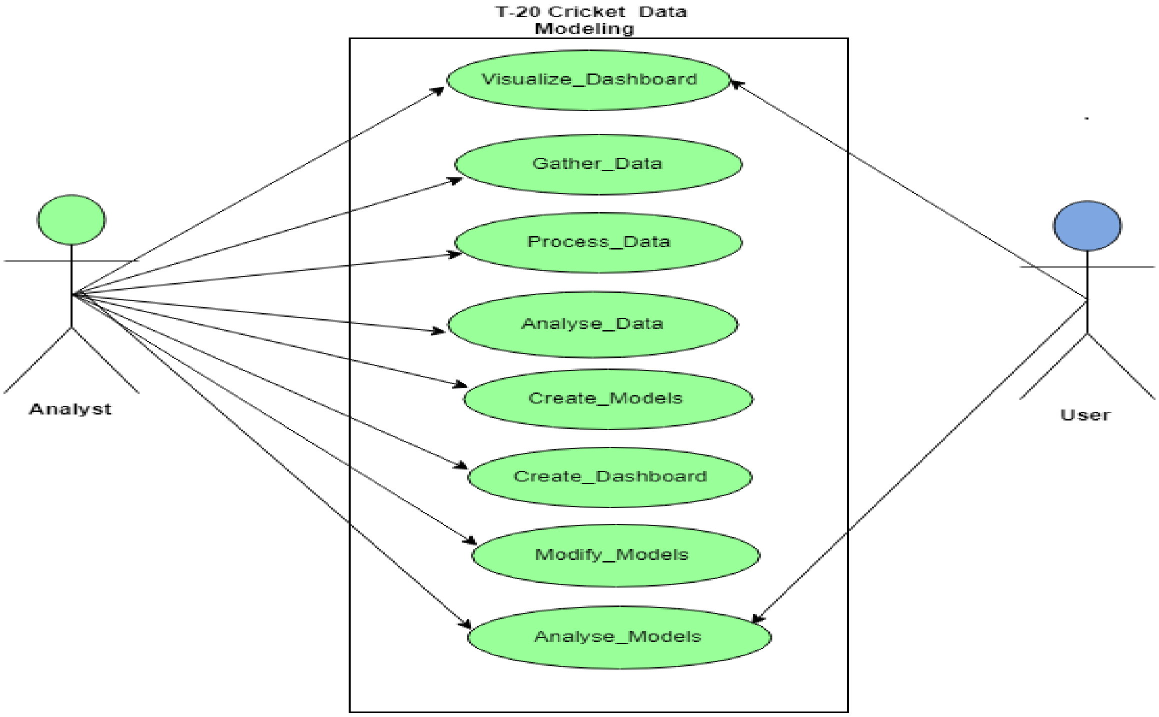
The objective will be to gather datasets on players and the matches, cleaning the datasets and extracting information such as player score, name, a number of matches won/loss and team information. The websites such as espncricinfo.com.

4.3 Functional requirements

The functional requirement for the project is required to complete are the following:

1. The analyst gathers the data from the website.
2. The analyst cleans the data.
3. The analyst analyses the data.
4. The analyst creates prediction data model.
5. The analyst creates a dashboard for the end user with functionalities such data analyzing.

Use case diagram



Requirement 1: Gather Data

Description & Priority

This is the first and the most important requirement of the project because, without data, analysis cannot be performed, therefore gathering the data plays an important role in this project. The data can be fetched using R or Python language.

Use Case

The analyst accesses the websites and develops a script in RStudio to fetch and export the data in form of CSV files.

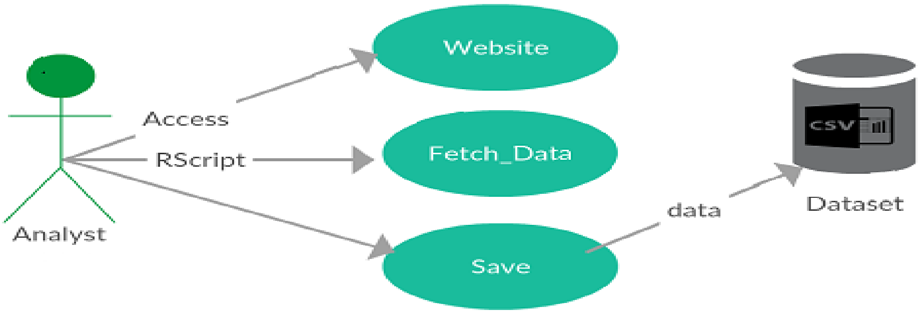
Scope

The scope of this use case is to gather data using website espincricinfo.com.

Description

This use case describes the process of analyst gathering data that is essential for the project via a website using R.

Use Case Diagram



Flow Description

Precondition

The analyst must access the website and R studio to fetch the data using RScripts.

Activation

This use case starts when the analyst accesses the data with the help of internet and programming application such as R/Python.

Main flow

1. The analyst identifies data on the website.
2. The analyst scraps the data using R or Python.
3. The analyst saves the data to a database e.g. Excel.

Exceptional flow

E1: Errors occurs
1. The website gives an error that dataset is corrupted.
2. The analyst identifies the problem.
3. The use case continues at position 1 of the main flow.

Termination

The data from the website is gathered and saved into a database. After completion of the task use case is terminated.

Postcondition

The process goes into a wait state as the datasets are imported and wait for its use.

Requirement 2: Processing Data

Description & Priority

This is the next requirement of the project after the first one, I would consider this as the 2nd important requirement as it involves cleaning the data by eliminating unwanted data such as outliers, null and keeping data that is relevant for performing the analysis.

Use Case

The Analyst retrieves the data that is gathered and eliminate unwanted data.

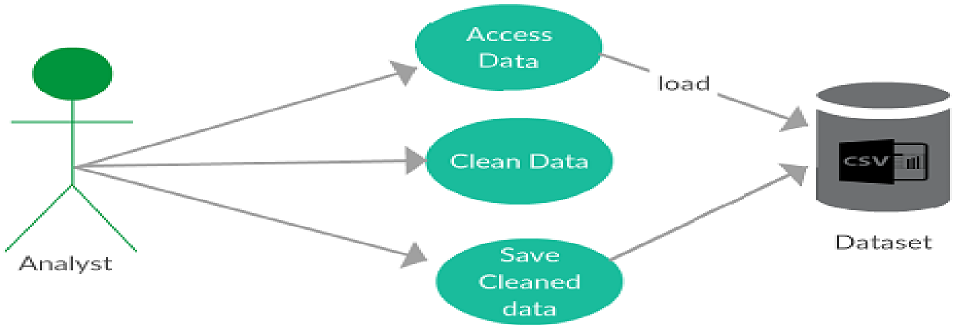
Scope

The scope of this use case is to clean and remove the data that is not required and the data that has potential to prevent us from achieving an accurate result.

Description

This use case describes how the data is cleaned for the project to be successful.

Use Case Diagram



Flow Description

Precondition

The dataset that is gathered must be in a wait state that is to be loaded using Rstudio or Python.

Activation

This use case starts when an Analyst access the programming application such as R or Python.

Main flow

1. The Analyst access the programming application.
2. The Analyst loads the dataset from database e.g. Excel.
3. The Analyst starts cleaning the data with the help of programming application and creates new data that is needed.
4. The Analyst saves the cleaned data to the database.
5. The Analyst exits the programming application.

Exceptional flow

E1: Unable to retrieve a dataset

1. The system is not able to load dataset.
2. The Analyst tries to find the cause of the problem and correct it

3. The use case continues at position 2 of the main flow.

Termination

After the cleaning is performed the system terminates.

Postcondition

The process goes into a wait state as the new dataset wait for its retrieval.

Requirement 3: Analyse data

Description & Priority

This requirement plays an important role when it comes to analyzing data, I would consider this requirement to be level 3 priority as it comes, after all, another step such as extracting, cleaning and mining. The purpose of this requirement is to analyze and understand the data.

Use Case

The analyst accesses the database and starts to analyze the data and record the data.

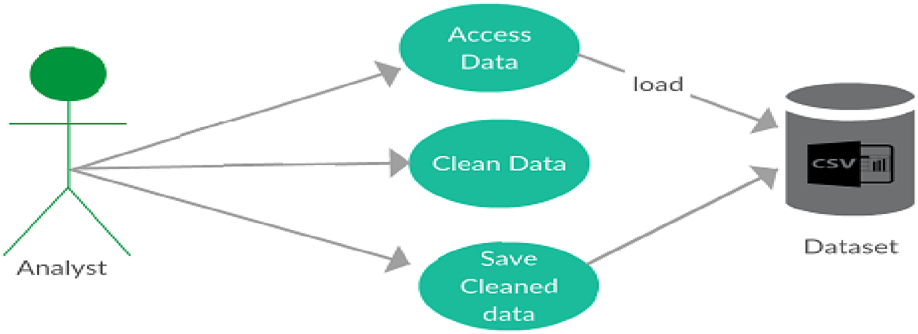
Scope

The scope of this use case is to perform analysis and record the results.

Description

This use case describes how the analyst performs the analysis on the data using a programming language.

Use Case Diagram



Flow Description

Precondition

The process is in initialization mode waiting for the analyst to load data from the database.

Activation

This use case starts when an Analyst access the programming application, loads the data from the database.

Main flow

1. The analyst loads the data.
2. The analyst runs analysis on the dataset using programming application.
3. The analyst records the result of the analysis.
4. The analyst saves and exits the programming application.

Exceptional flow

E1: Unable to examine data
1. The application is unable to perform analysis due to unknown reasons.
2. The analyst checks the database, script and correct errors.
3. The use case continues at position 2 of the main flow.

Termination

The analysis is performed successfully and recorded. The use case terminates.

Postcondition

The process goes into a wait state for next step to be performed.

Requirement 4: Create Models

Description & Priority

This is the next requirement of the project after the analyzing the data, I would consider this as the 4th important requirement as it involves developing a model using machine learning algorithms.

Use Case

The analyst accesses the data that and start performing machine learning algorithms.

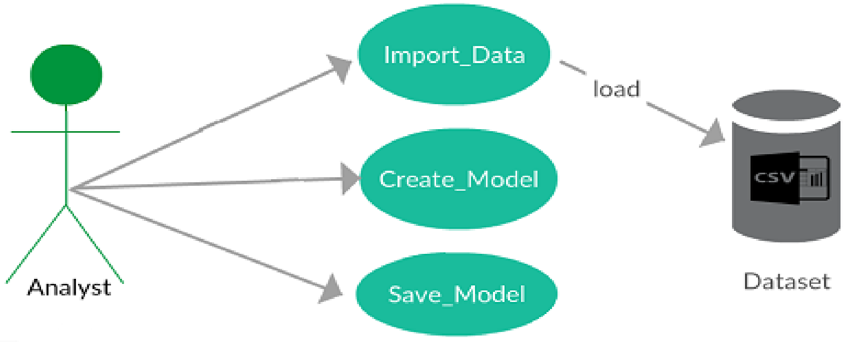
Scope

The main goal of this use case is to create prediction and classification model for Indian cricket team and the opponents.

Description

This use case describes the how analyst performs a different machine learning algorithm to create models.

Use Case Diagram



Flow Description

Precondition

The dataset that is cleaned and merged must be in a wait state that is to be loaded.

Activation

This use case starts when an analyst accesses the programming application such as R and Python.

Main flow

1. The analyst accesses the programming application.
2. The analyst loads the dataset from database e.g. Excel.
3. The analyst starts applying algorithms with the help of programming application and creates models.
4. The analyst saves the scripts for future use.
5. The Analyst exits the programming application.

Exceptional flow

E1: Incorrect algorithm used

1. The programming application cannot output results due to the incorrect algorithm used.
2. The analyst checks the fault and fixes it.
3. The use case continues at position 3 of the main flow

Termination

After the analyst exits the programming application.

Requirement 5: Visualize

Description & Priority

This requirement is important for both user and analyst point of view as it can be used to visualize and make a prediction of data that is analyzed. I would consider this requirement to be level 5 priority because it is the final step and before concluding this it requires all other four requirements to be completed.

Use Case

The analyst accesses the clean and merged data from the database and import the data into Tableau.

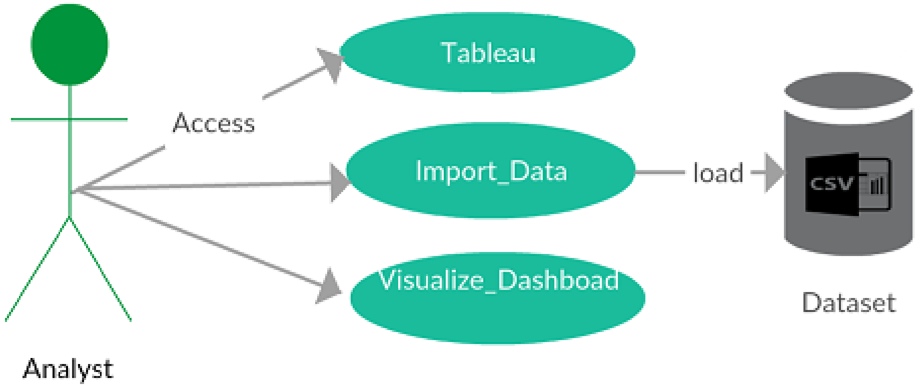
Scope

The scope of this use case is to predict the performance of a team, player, and others that are related to this data via visualization.

Description

This use case describes how analyst uses its knowledge to predict and visualize the information related to team performance.

Use Case Diagram



Flow Description

Precondition

The analyzed dataset is in waiting for state and ready to be used for the last step of the project.

Activation

This use case starts when analyst access the application and import data from the database.

Main flow

1. The analyst import data from the database.
2. The analyst uses Tableau for visualization.
3. The analyst documents the results of the visualization.
4. The analyst sends the data to dashboard for the user to view.
5. The analyst saves the data and exits the application.
6. Analyst exits the dashboard.

Termination

The analyst terminates the application when the task of visualization is completed and documented.

4.4 Non-Functional requirements

In this section, the non-functional requirement is discussed for the project.

4.4.1 Performance/Response time requirement

This requirement is very highly prioritizing by many businesses as they want to keep their customer happy and informed. The dashboard that is created for the user should run smoothly and should not take more than 1-2 seconds to respond to the user's demands e.g. if the user wants to perform analysis for listing the top player, it should be quick and easy.

4.4.2 Availability requirement

As dashboard, will be connected to the internet, therefore, there should be any downtime, which it should always be available to access from anywhere and anytime. E.g. the when the user search for viewing the dashboard it should show up and doesn't give a message that it is not available.

4.4.3 Recover requirement

This requirement is the high priority for the project, make sure the data is saved in two places one on hardware and other in a cloud, in the event of hardware failure or server errors. The second storage can be accessed and data can be retrieved. The backup of all the data is necessary.

4.4.4 Robustness requirement

This requirement is good to have as it makes sure everything is working properly. E.g. if one part of the script gives an error, it doesn't depend on just one script, therefore, it switches to another script that has similar capability to perform a certain task.

4.4.5 Security requirement

The project should be secure, to provide security for the project, I will be using filter feature which is available on Tableau, it allows the analyst to select what information can the user view and from which region e.g. analyst can select only region India can view detail report of Indian cricket player.

4.4.6 Reliability requirement

The data set is gathered from secured website espncricinfo.com, as we know they are updated from time to time and maintained. When the new data is updated about a recent cricket match or a player, the aim of this project should be to provide reliable and accurate figures that can be trusted by the users.

4.4.7 Maintainability requirement

The project requires being maintained as the datasets can cause some small error which can be easily corrected. The dashboard should be simple and easy to maintain.

4.4.8 Portability requirement

The project should be portable, the dashboard should be available for different devices such as mobile phone and computers. By doing this user can access it anywhere and any device.

4.4.9 Extendibility requirement

This project needs to expandable, as the data changes over a period and the new way of performing the task is developed therefore algorithm that can be extended is very useful. Overall the project should be able to cater for changes.

4.4.10 Reusability requirement

This requirement is very important, a different type of analysis will be performed on different data set but the concept and the algorithms don't change much, therefore, a python script and R script can be used again and again.

4.4.11 Resource utilization requirement

Hardware such as a computer, with the internet access, backup storage such as cloud-based GitHub, Google drive will be used and Tableau for visualizing.

4.4.12 Environmental requirement

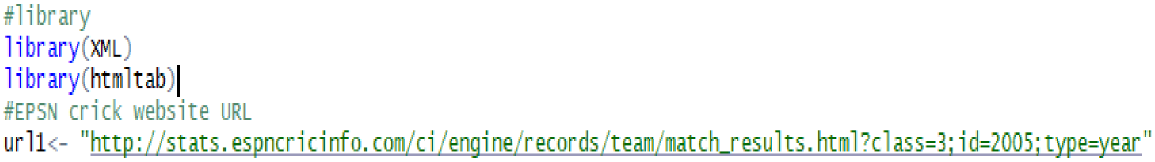
The project needs to be developed in the way so it does not have the effect on the environment e.g. physical, social and organizational. The dashboard should be accessible for everyone and privacy of the client is necessary. To help the user better understand the analysis a support must be provided by having help and support option.

4.4.13 Usability requirement

The project aim should be clearly focused on the usability, as this is where the user will be interacting with data. The simple option for selecting data and performing analysis should be provided to the user so it doesn't confuse or prevent them from using the dashboard.

4.5 Data Preparation

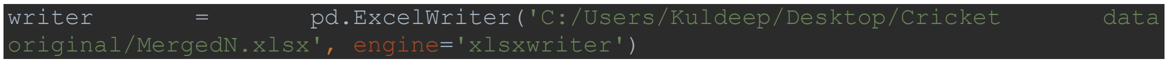
The data is downloaded from espncricinfo.com website using RScript and it is exported in the form of CSV files.



```
#library
library(XML)
library(htmltab)
#EPSN crick website URL
url1<- "http://stats.espncricinfo.com/ci/engine/records/team/match_results.html?class=3;id=2005;type=year"
```

Figure 4: Fetching data from the website

The PyCharm is then used to discover and combine the data. Exporting the merged files using PyCharm.



```
writer = pd.ExcelWriter('C:/Users/Kuldeep/Desktop/Cricket data
original/MergedN.xlsx', engine='xlsxwriter')
```

To make sure data is cleaned, the cleaning is performed in again RStudio to further investigate and remove unwanted columns. After reading the data in RStudio it contains 41 columns including country, team matches played, points ratings etc. next unwanted columns, an outlier, and null values are removed. Also, the summary of the data is checked. Then the file is again saved as a clean file. To understand more about the data, the density plot code is run to see the flow of the data, also Cullen and Frey's graph is used to see if there is any imbalance class.

4.6 Data Analysis and Feature Selections

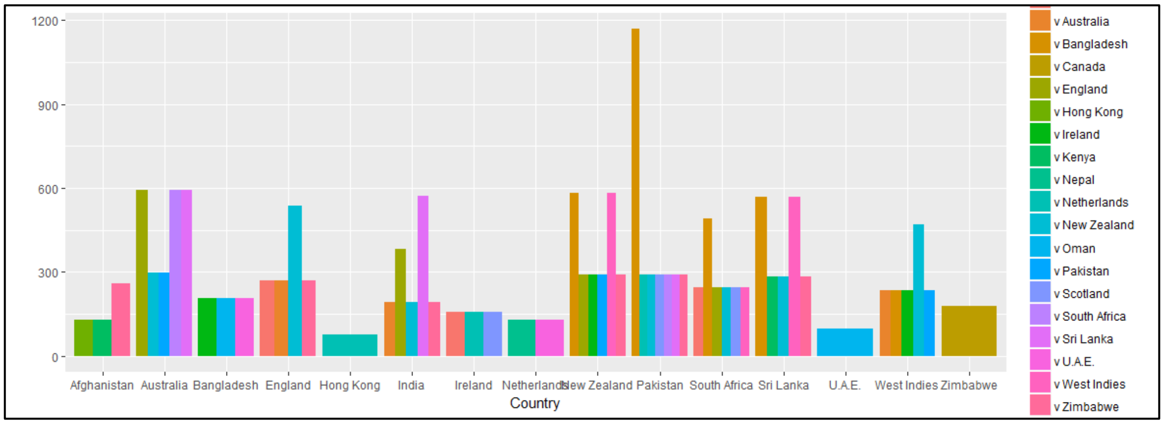


Figure 5: Country and its Oppositions

The chart in (figure 5) shows countries and opposition they have played with. The country who has played many matches with one another can be seen clearly which is Pakistan opposition Australia. The second highest is other such as New Zealand, Sri Lanka and lowest is the Hong Kong, UAE who has not played many matches with other countries.

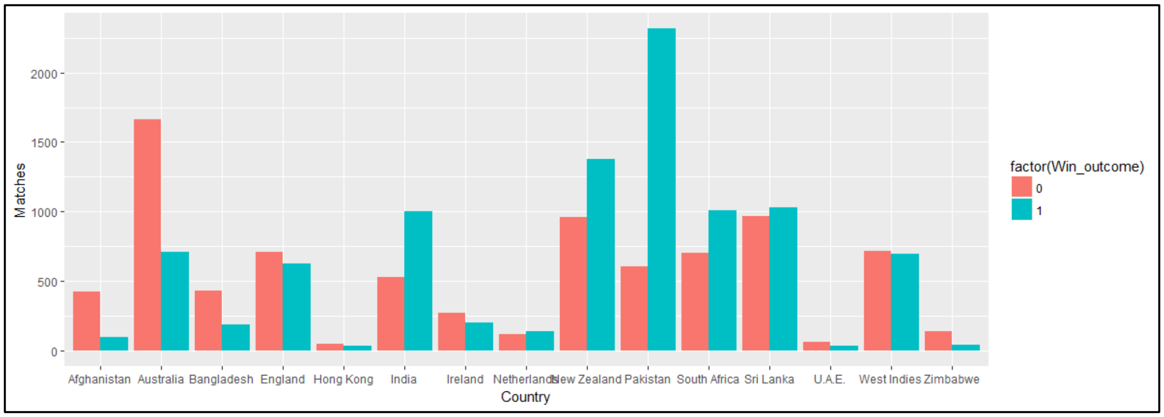
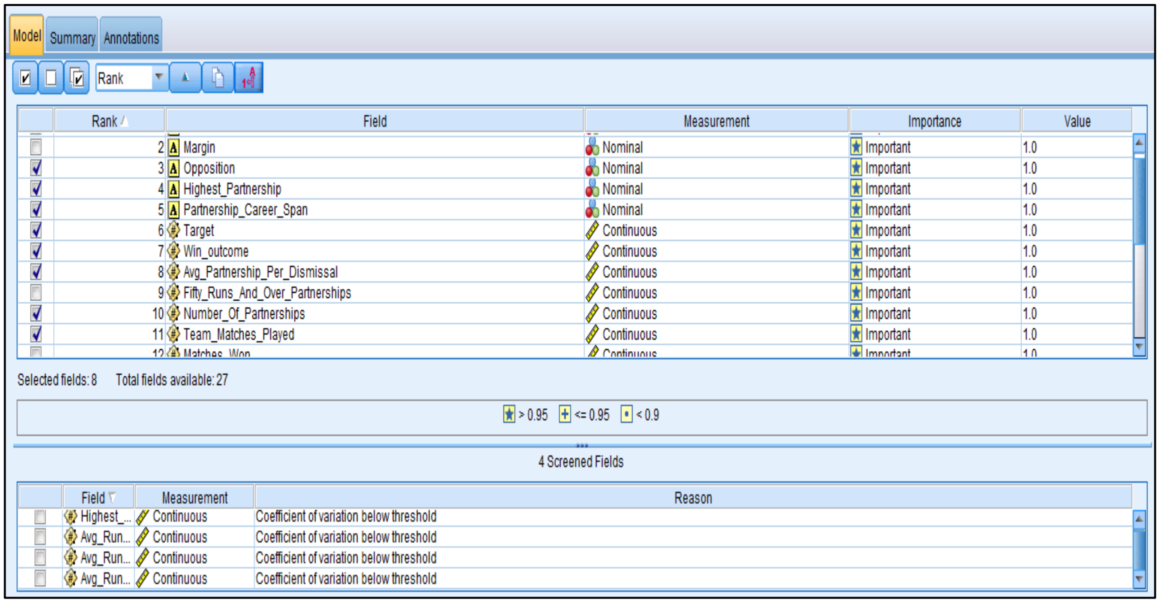


Figure 6: Winner and Team Matches Played

The winning team and matches the team has played are shown in the graph above. Winning is considered as 1 and losing 0. We can see from (figure 6) top team such as Pakistan, Australia, and New Zealand has played and won many matches. The other team who have won the very small amount of match can be seen in countries such as Hong Kong, UAE, and Zimbabwe. After analyzing the data, the features are selected. The SPPS is also used to gain more understanding about the feature selection and to confirm and test the selected features.



	Rank	Field	Measurement	Importance	Value
☐	2	A Margin	Nominal	★ Important	1.0
☑	3	A Opposition	Nominal	★ Important	1.0
☑	4	A Highest_Partnership	Nominal	★ Important	1.0
☑	5	A Partnership_Career_Span	Nominal	★ Important	1.0
☑	6	Target	Continuous	★ Important	1.0
☑	7	Win_outcome	Continuous	★ Important	1.0
☑	8	Avg_Partnership_Per_Dismissal	Continuous	★ Important	1.0
☐	9	Fifty_Runs_And_Over_Partnerships	Continuous	★ Important	1.0
☑	10	Number_Of_Partnerships	Continuous	★ Important	1.0
☑	11	Team_Matches_Played	Continuous	★ Important	1.0
☐	12	Matches_Won	Continuous	★ Important	1.0

Selected fields: 8   Total fields available: 27

★ > 0.95 + <= 0.95 • < 0.9

4 Screened Fields

	Field	Measurement	Reason
☐	Highest_...	Continuous	Coefficient of variation below threshold
☐	Avg_Run...	Continuous	Coefficient of variation below threshold
☐	Avg_Run...	Continuous	Coefficient of variation below threshold
☐	Avg_Run...	Continuous	Coefficient of variation below threshold

Figure 7: Feature Selection in SPSS

Before proceeding to the next step which involves developing models, SPSS is used to check what type of model are suitable for the data.

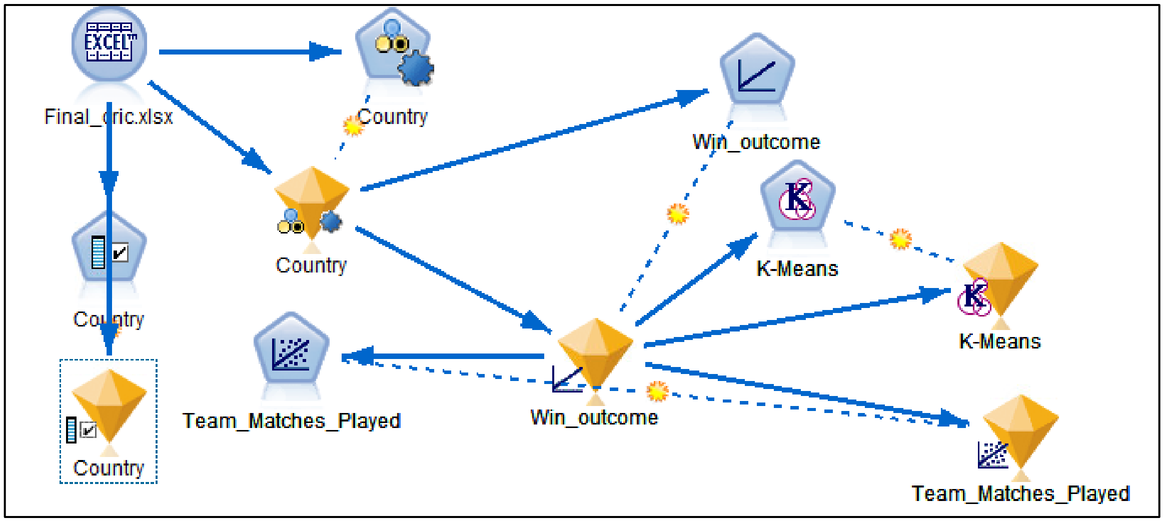


Figure 8: SPSS Testing Suitable Model

The (figure 8) shows how data is loaded in the form of xlsx to determine the features, an auto classifier is used to see the types of model that can be performed on the dataset. Testing the model is checked here to avoid failure of the model in the later stage.

A various analysis is performed using SPSS such as the test for normality, checking which data would be suitable for a type of attributes. The result suggested for developing model were Logistic Regression, K_Means, and CHAD. Although, the further test was performed on Linear Regression, Multilinear and Decision tree. Finally, the decision is made to develop six models.

5 Implementation, Evaluation, and Results of T-20 Cricket Models

5.1 Introduction

In this section of the paper, all the six model that is developed implemented and evaluated using the results.

5.2 Linear Regression Prediction Model

Linear regression is very popular predictive modeling technique that is widely used by the data analyst. There is the much linear regression, this is considered as the best one. Before going into more details about linear regression, regression analysis must be understood. Linear analysis is like modeling technique, which is also used to find a relationship between variables. In this project, for example, it can be performed on matches played and rating to visualize when one or more matches are played the rating goes up or down. The reason for using this is that it provides analysis on the relationship between two attributes. The linear model can be performed on continuous or discrete variables. The line which is seen in the regression model is called best fit straight line. The equation Y = a+b*X+e is used and it only has one independent variable. For obtaining best line the formula below is used, which minimize the sum of squares.

$$\underset{w}{min} \, ||Xw - y||_2{}^2$$

5.2.1 Implementation

This is the very first model which is created by using Python, the first step involves installing the packages such as glob, os, pandas, np and then they are imported within the script to successfully run the code. The cleaned data is loaded into the data frames starting from 0 to all the way to 12. To see the data has been loaded with small code such as "print (df8.head)())" is used. The merged data which is named as "Merge4" in the python script data is split into training (20%) and testing (80%) randomly. The next step involves feature scaling which basically computes mean and standard deviation that can be used to see average matches won by the team and their ratings. Overall the main aim of this is to view the distribution whether it is normal or not normal if it is not normally distributed or skewed further scaling is completed on the train and test dataset. After having that completed, next step is to develop a prediction model which can help to predict future value. The liner regression code is run on the training and test, then standard root mean square is calculated to see how well it has performed. Finally, two attributes are chosen one “Matches played” and other “Rating” to see whether the number of matches has a relationship with a rating. Also, the scatter plot is created to visualize clearly.

5.1.2 Evaluation and Results

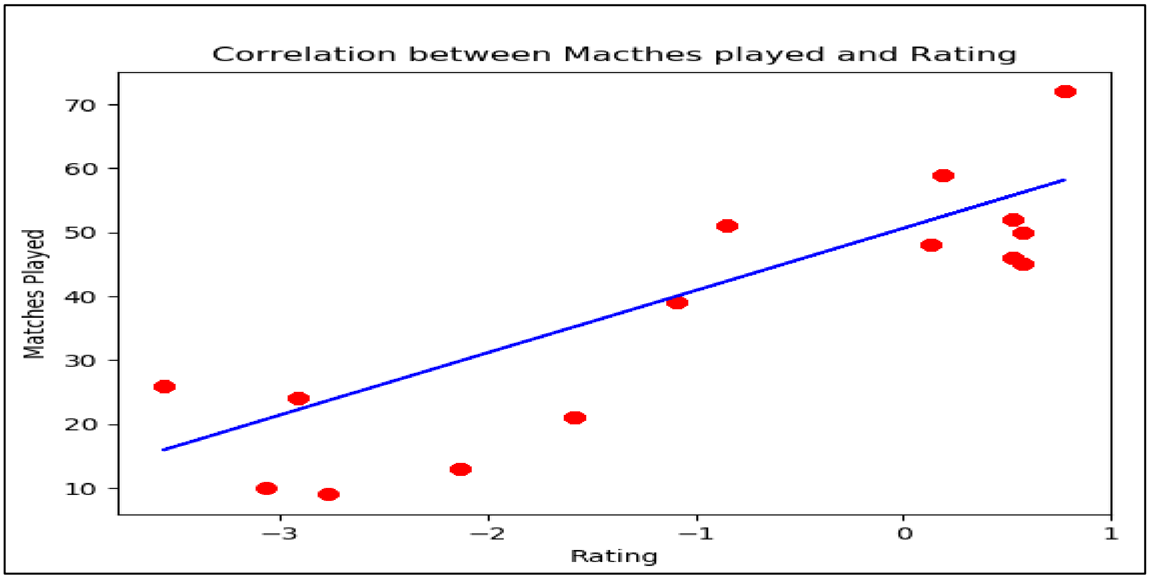


Figure 9: Linear Regression

The figure 9, shows the number of matches played and the relationship between the ratings. We can see the more matches the team plays, the rating increases whereas the less matches the rating decreases. From looking at the diagram we can see there are some outliers. This model can be furthered developed to produce better results. For evaluation the root mean square is calculated which is 49.64, this shows the spread of the data, which ok but not good. More improvement is needed, the prediction needs to train and tested again for producing the better results.

5.3 Polynomial Regression Model

This is another very popular regression used by the data analyst, it is calculated using the equation, y =a+b*x^2, however, the best fit line in this one is curved not a straight like linear. The (figure10) below shows how to read the charts properly, as we can see the first on the left is underfit, in the middle just right and the last one on the right is overfit.

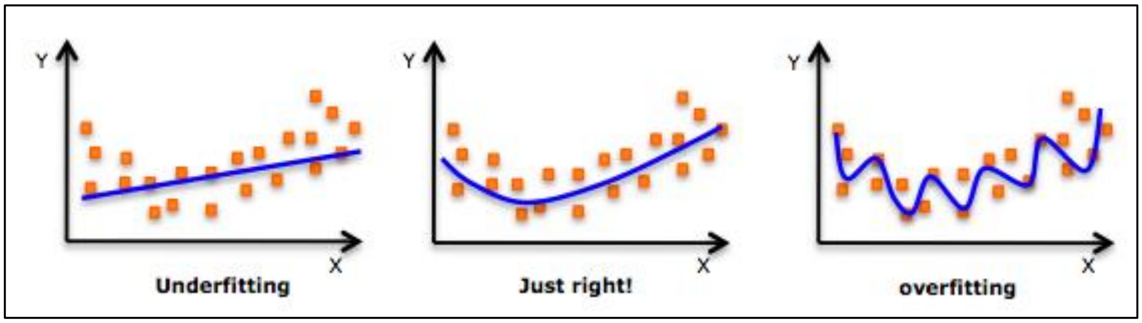


Figure 10: Polynomial charts example

5.3.1 Implementation

This is the second model developed in Python, again like the linear regression model the packages are installed, the working directory is set and data is loaded to perform the polynomial regression. Before adding the code for polynomial regression, little data exploration is completed to understand data and determine which attribute can be used. Again the "Merged4" is used, therefore data is split into test (20%) and train (80%) randomly. The next step like the above one feature scaling is used to compute and see a mean and standard deviation, that tell us the distribution of the data. Finally, the prediction model polynomial is developed by running the code for polynomial regression and then the code for fitting the data is run and finally the visualization using the plot diagram is used.

5.3.2 Evaluation and Results

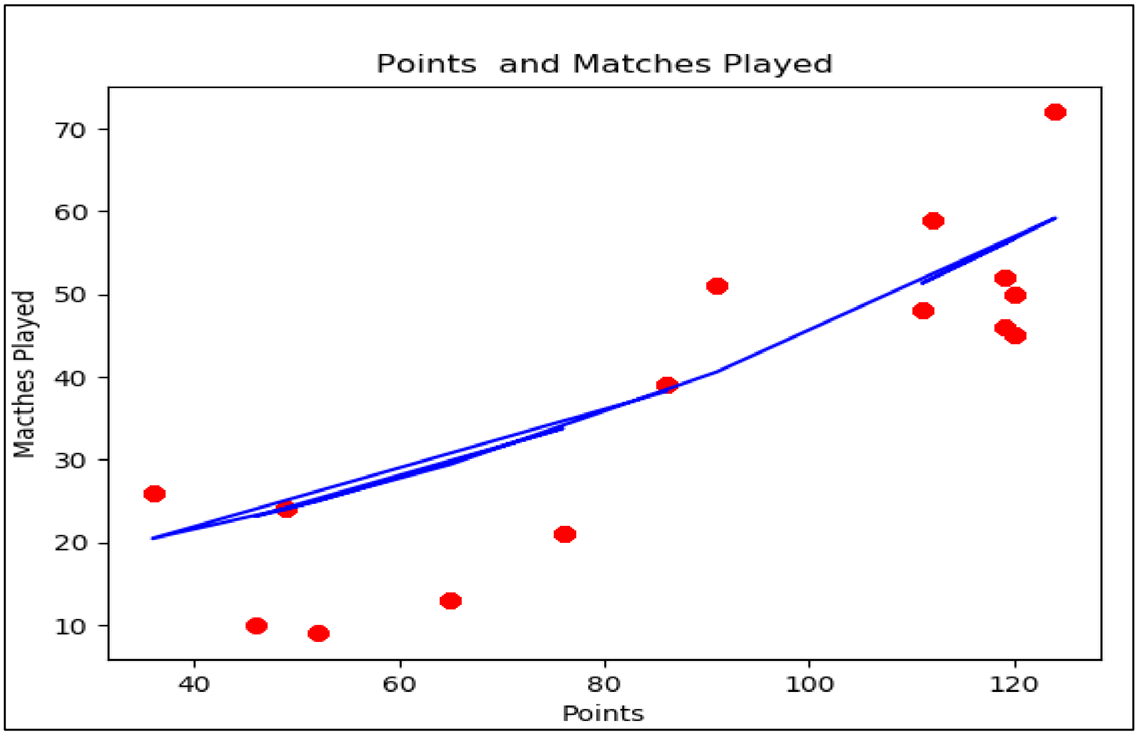
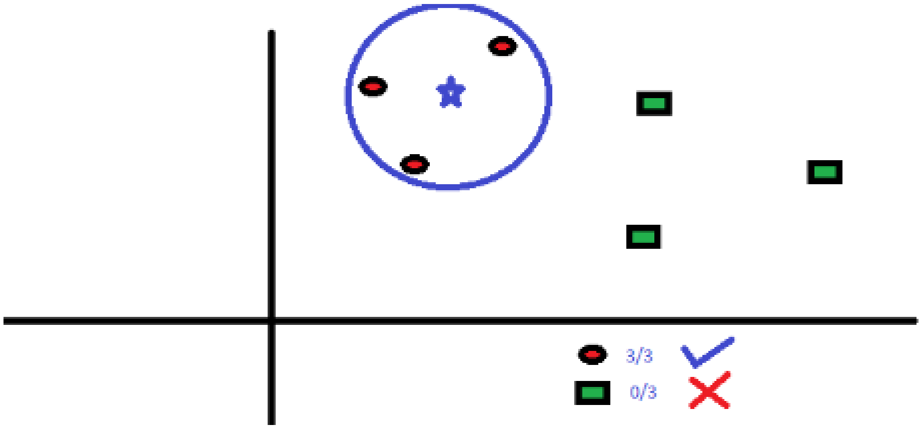


Figure 11: Polynomial Chart

The (figure 11) above shows correlation between matches played and team points, we can clearly see the data is under fitted, not much information can have determined from the chart above, it shows a number of the matches are played by team members the points tend to increase like the ratings which are common in real-world situation because the more match the team plays and wins its points are likely to increase whereas if fewer matches are played the points tend to decrease. There is some correlation between those to attributes due to underfitting of the data it is not clear therefore further work needs to complete.

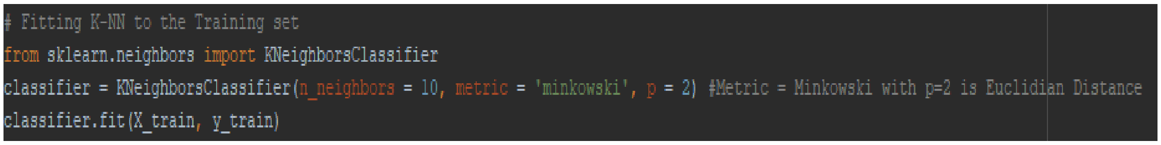
5.4 K- Nearest Neighbors

This is one of the most used algorithms in data analytics field due to its ability to perform classification and regression prediction. It very powerful and yet easy to interpret, basically it works by detecting the nearest attributes and presumes which ever attribute is closest to a group they belong to same class. Setting a k value plays an important role as it only checks the data within that field. Setting k value too low and too high can cause problems. Setting a K-values depends on the dataset but normally it should be between 3 and 10. For example in the diagram below k is set to 3 so it only looks within a circle, where if we were to set it to 6 it would expand the circle.



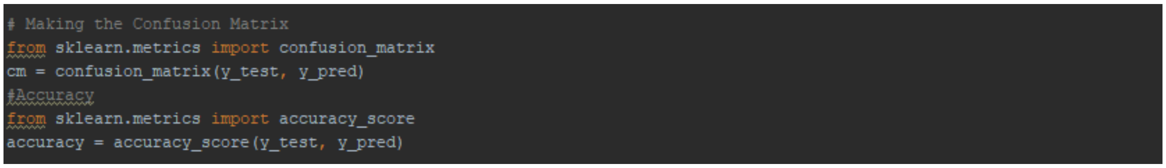
5.4.1 Implementation

This is the third model developed using Python, as a first step the data is loaded, the attributes on which KNN is to be performed are selected, the data is then split into the train (20%) and test (80%). The feature scaling is used again and the Euclidean distance is calculated to see the distance between the test and train data. After this, prediction on the test data is encoded to see the classification of the closest values and then finally, the accuracy an of train and test is completed.



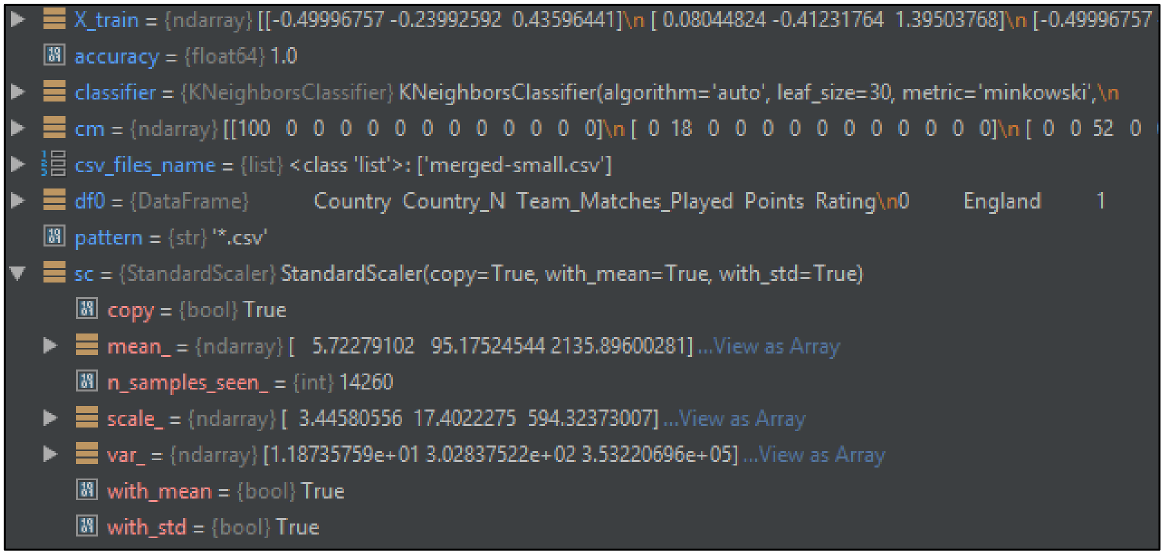
```
# Fitting K-NN to the Training set
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 10, metric = 'minkowski', p = 2) #Metric = Minkowski with p=2 is Euclidian Distance
classifier.fit(X_train, y_train)
```

Code for confusion matrix is completed, no table is generated:



```
# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
#Accuracy
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred)
```

5.4.2 Evaluation and Results



```
▶ X_train = {ndarray} [[-0.49996757 -0.23992592  0.43596441]\n [ 0.08044824 -0.41231764  1.39503768]\n [-0.49996757
  accuracy = {float64} 1.0
▶ classifier = {KNeighborsClassifier} KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',\n
▶ cm = {ndarray} [[100   0   0   0   0   0   0   0   0   0   0   0   0]\n [  0  18   0   0   0   0   0   0   0   0   0   0   0]\n [  0   0  52   0
▶ csv_files_name = {list} <class 'list'>: ['merged-small.csv']
▶ df0 = {DataFrame}        Country  Country_N  Team_Matches_Played  Points  Rating\n0       England        1
  pattern = {str} '*.csv'
▼ sc = {StandardScaler} StandardScaler(copy=True, with_mean=True, with_std=True)
    copy = {bool} True
  ▶ mean_ = {ndarray} [   5.72279102   95.17524544 2135.89600281] ...View as Array
    n_samples_seen_ = {int} 14260
  ▶ scale_ = {ndarray} [  3.44580556  17.4022275  594.32373007] ...View as Array
  ▶ var_ = {ndarray} [1.18735759e+01 3.02837522e+02 3.53220696e+05] ...View as Array
    with_mean = {bool} True
    with_std = {bool} True
```

Figure 12: KNN Result

The code above in figure shows the accuracy is 1.0 which mean it is 100% accurate and mean of the attributes are 5.77222 and 95.17 etc. the sample which KNN is performed is on 14260 of the train data. The reason why it produces 100% accuracy is that not the data was like one another. Also, the k value is set to 10, if we reduce the k value a different result will be shown.

5.5 Support Vector Machine

The Support Vector Machine is gaining popularity after Naïve Bayes, Random Forest, this is a supervised machine learning algorithm. This is one of the best algorithms as it can classify and it has similar functionality as regression which is mentioned above in heading 5.2. The hyperplane is used to separate the classes. Like the graph shown in (figure 13) below, that is how the two classes are differentiated using hyperplane, there are many others, it can be divided into the two classes, which is completed by adding one or more hyperplanes. The advantage of using SVM is that it works well with the clear margin. The disadvantage is that it does not perform well on a large dataset or if the data has too many outliers.

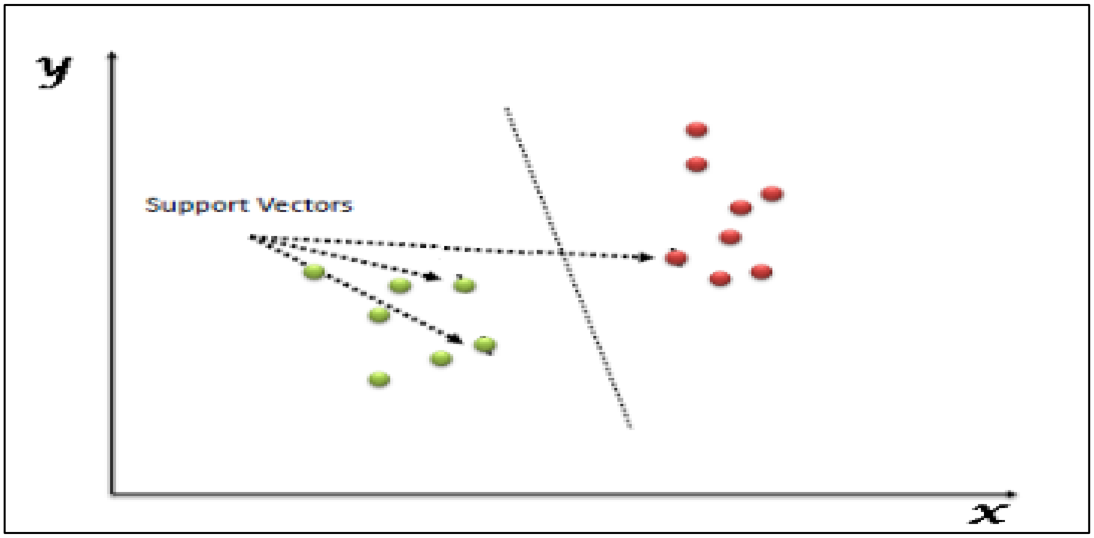
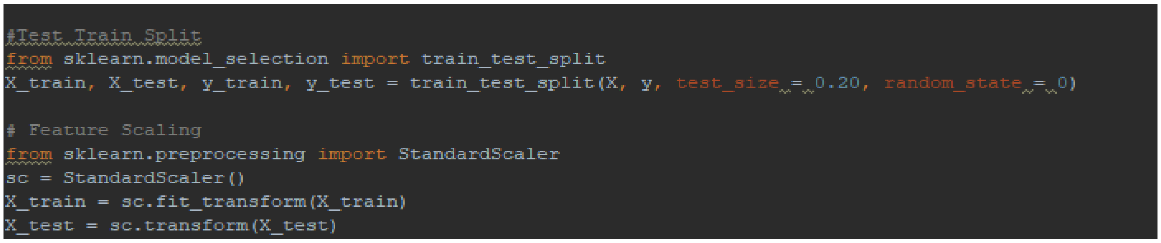


Figure 13: SVM chart

5.5.1 Implementation

This is the fourth model, which is created using Python, where the data is loaded, then it is split into a train (20%) and test (80%), feature scaling is completed then the Linear Support Vector Machine is used for classification to compare more of elements. Predict linear classification code is encoded to see the classification on a sample. Polynomial Support Vector Machine is also used to get the better understanding of the data and how it is classified from the different point of view. Again, prediction code is encoded to see the classification, then finally, to output the result confusion matrix is developed and accuracy is shown.

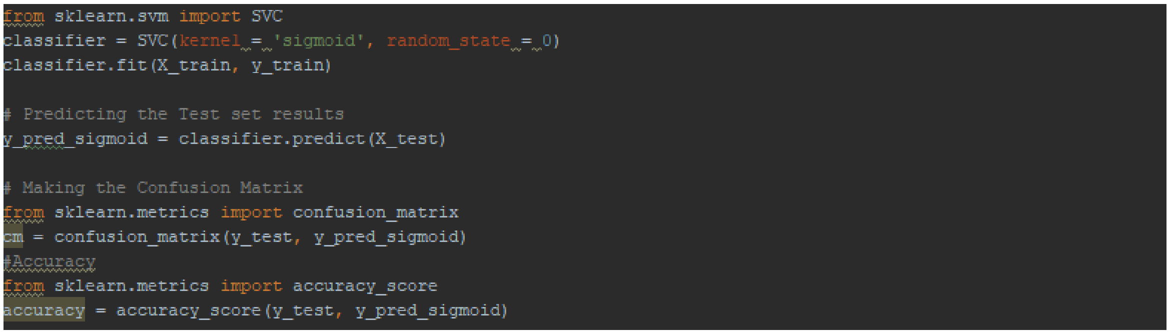
Code for feature scaling and training:



```
#Test Train Split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 0)

# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Similar problem in this the confusion matrix table is not shown:

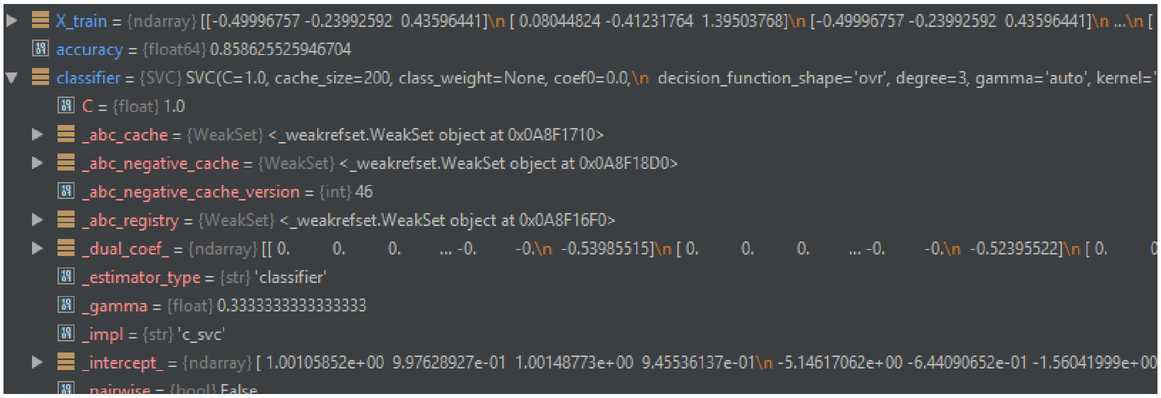


```
from sklearn.svm import SVC
classifier = SVC(kernel = 'sigmoid', random_state = 0)
classifier.fit(X_train, y_train)

# Predicting the Test set results
y_pred_sigmoid = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred_sigmoid)
#Accuracy
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred_sigmoid)
```

5.5.2 Evaluation and Results



```
▶ X_train = {ndarray} [[-0.49996757 -0.23992592  0.43596441]\n [ 0.08044824 -0.41231764  1.39503768]\n [-0.49996757 -0.23992592  0.43596441]\n ...\n [
  accuracy = {float64} 0.858625525946704
▼ classifier = {SVC} SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,\n  decision_function_shape='ovr', degree=3, gamma='auto', kernel='
    C = {float} 1.0
  ▶ _abc_cache = {WeakSet} <_weakrefset.WeakSet object at 0x0A8F1710>
  ▶ _abc_negative_cache = {WeakSet} <_weakrefset.WeakSet object at 0x0A8F18D0>
    _abc_negative_cache_version = {int} 46
  ▶ _abc_registry = {WeakSet} <_weakrefset.WeakSet object at 0x0A8F16F0>
  ▶ _dual_coef_ = {ndarray} [[ 0.        0.        0.        ... -0.        -0.\n  -0.53985515]\n [ 0.        0.        0.        ... -0.        -0.\n  -0.52395522]\n [ 0.        0
    _estimator_type = {str} 'classifier'
    _gamma = {float} 0.3333333333333333
    _impl = {str} 'c_svc'
  ▶ _intercept_ = {ndarray} [ 1.00105852e+00  9.97628927e-01  1.00148773e+00  9.45536137e-01\n -5.14617062e+00 -6.44090652e-01 -1.56041999e+00
    _pairwise = {bool} False
```

Figure 14 SVM Results

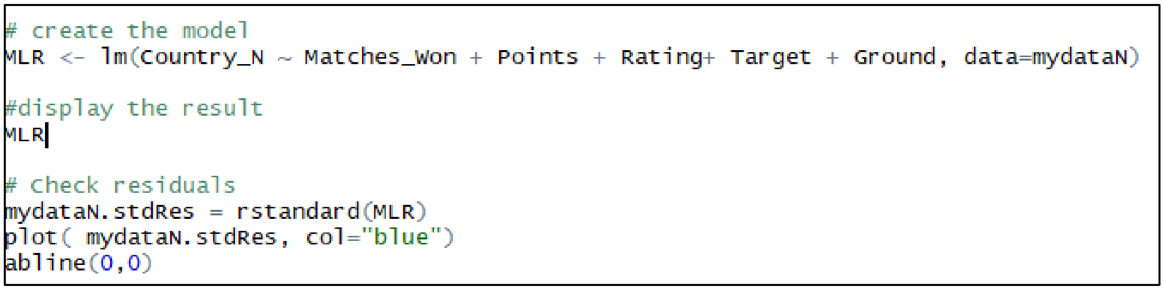
The (figure14) above shows a result of the SVM with the accuracy of 0.86% which is very good, this means we are sure that 86% the prediction would be correct and the two classes which are win_outcome and country_N are classified correctly.

5.6 Multilinear Regression Model

The Multilinear regression is same as linear regression the only difference is that it has more than one independent variable.

5.6.1 Implementation

This is the fourth model that is developed using R, the file this time is read in RStudio, the correlation between the attributes is checked e.g. country, matches won, points, ratings. The next step involves pairing the correlation and create a scatter plot to see the relationship between the listed attributes. The model is then created, after that the residuals are checked and they are plotted to gain the better understanding when all the step of above is completed a quick code for checking the normality of is encoded to see whether the results are normal or not.



```
# create the model
MLR <- lm(Country_N ~ Matches_Won + Points + Rating+ Target + Ground, data=mydataN)

#display the result
MLR

# Check residuals
mydataN.stdRes = rstandard(MLR)
plot( mydataN.stdRes, col="blue")
abline(0,0)
```

Figure 15: MLR model and checking residuals

5.6.2 Evaluation and Results

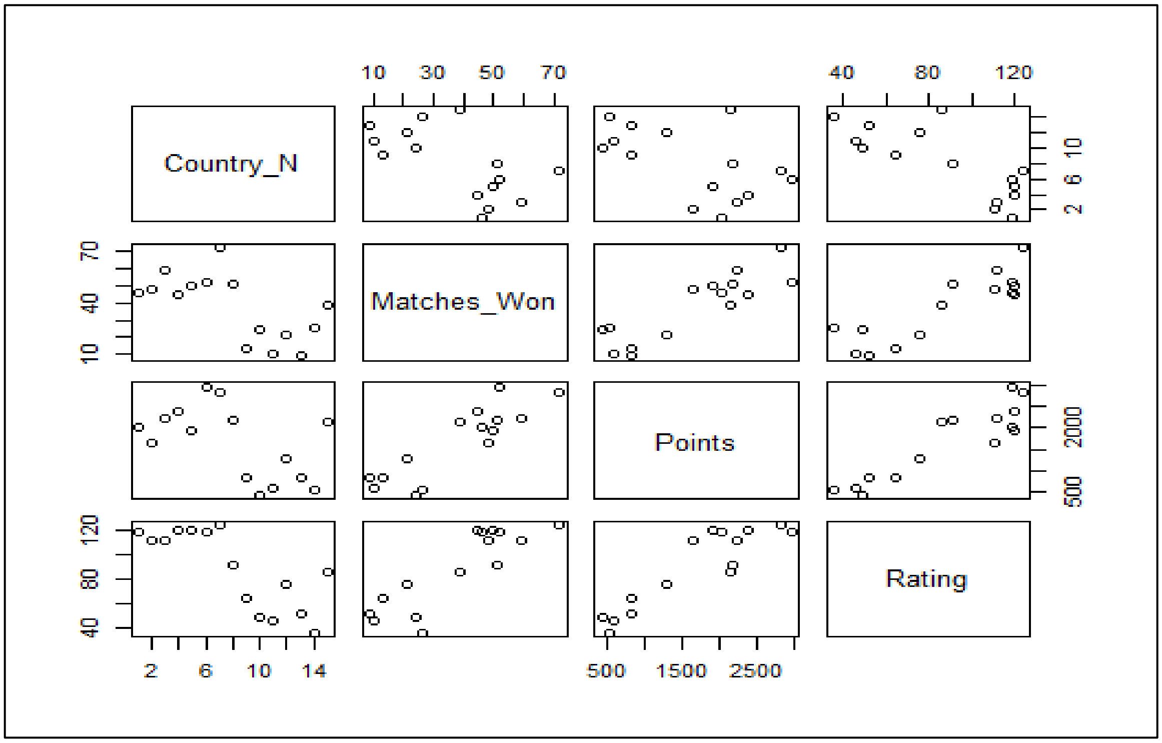
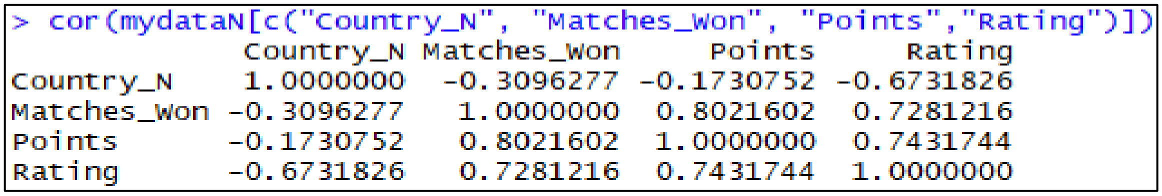
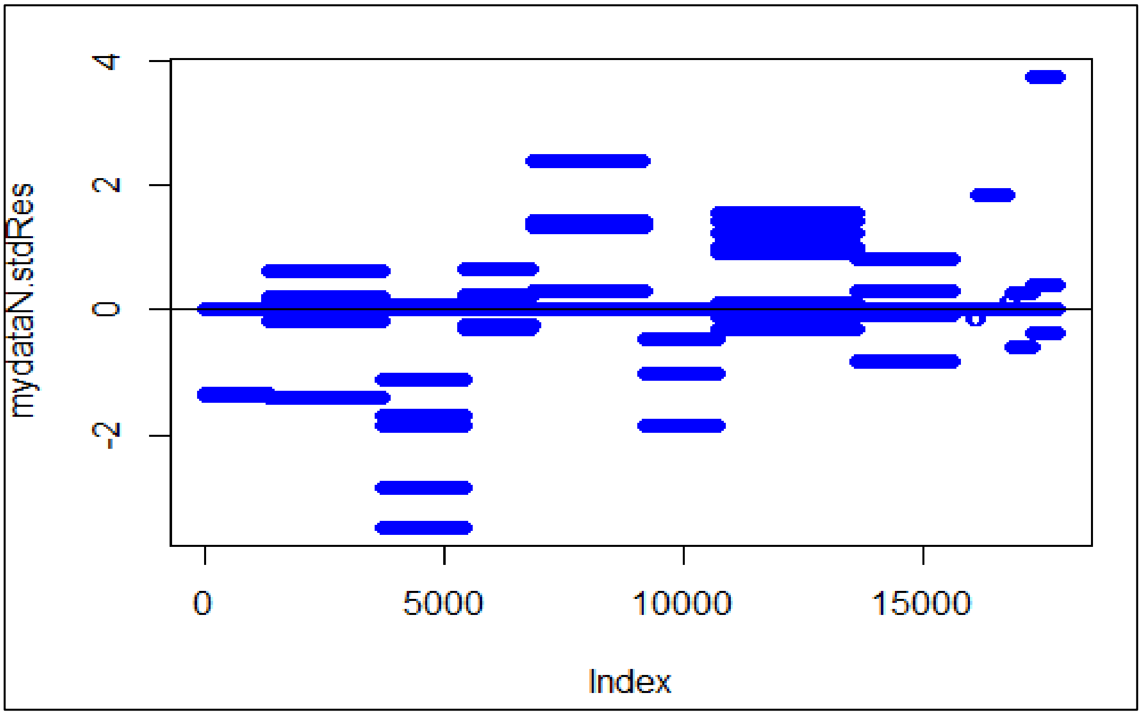


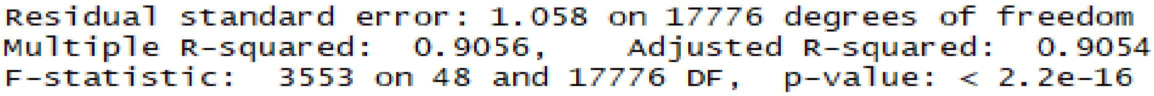
Figure 16: MLR correlation



```
> cor(mydataN[c("Country_N", "Matches_Won", "Points","Rating")])
              Country_N Matches_Won     Points     Rating
Country_N     1.0000000  -0.3096277 -0.1730752 -0.6731826
Matches_Won  -0.3096277   1.0000000  0.8021602  0.7281216
Points       -0.1730752   0.8021602  1.0000000  0.7431744
Rating       -0.6731826   0.7281216  0.7431744  1.0000000
```

In the (figure16) above the high correlation between Matches_won and Points, Rating is found. Three of these are highly related to each other which makes sense if a team wins the matches the points and rating is likely to go up. There is no correlation between Country and other groups, points are again highly correlated with a rating. Poor correlation between country and other groups.





```
Residual standard error: 1.058 on 17776 degrees of freedom
Multiple R-squared:  0.9056,	Adjusted R-squared:  0.9054 
F-statistic:  3553 on 48 and 17776 DF,  p-value: < 2.2e-16
```

The p-value is less than alpha value of 0.05 which means the there is a significant difference between country, matches won, points and rating. Therefore, the null is rejected the null hypothesis state that there is no difference between the groups and are in favor of alternative hypothesis which state that there is a significant difference between the country, rating, points and matches won.

5.7 Decision Tree

The decision tree is normally used to classify problems. Basically, it works by pre-defining what needs to be completed that is why it is called supervised algorithm. It helps us to solve problems related to making decisions by providing multiple solutions to the given problem. It consists of root node which is at the very top, test node and leaf for the result of the test. It is very fast, effective and less time-consuming. This can be used to avoid making the errors when solving the problem. The (figure 17) below shows how it splits the problems into the subproblem for making the decision.

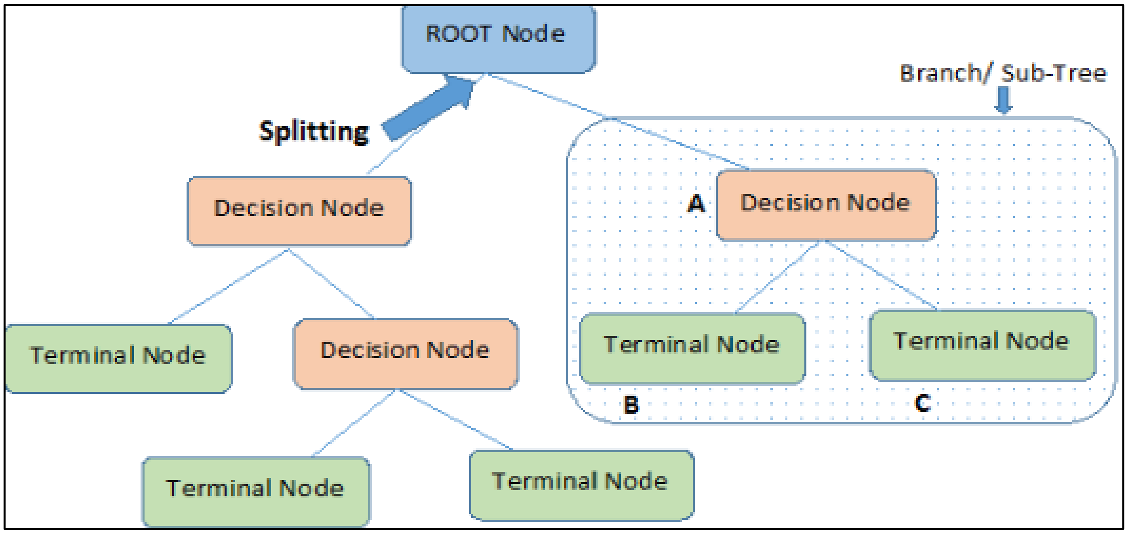


Figure 17: Decision Tree Example

5.7.1 Implementation

This is the sixth and the last model which is also developed using RStudio, the very first step in this includes installing and running the libraries such as C50, which necessary for running and developing the model, then data is normalized, min and mix technique is used for normalizing the data. Then the bar plot is created to see classification distribution of the attribute "Win_outcome". After the normalization process the data is split into a train (20%) and test (80%), model is trained on column "6" of the data and finally the model is developed using the Rpart library and the confusion matrix is also developed. To improve the model, it is trained and tested again.

5.7.2 Evaluation and Results

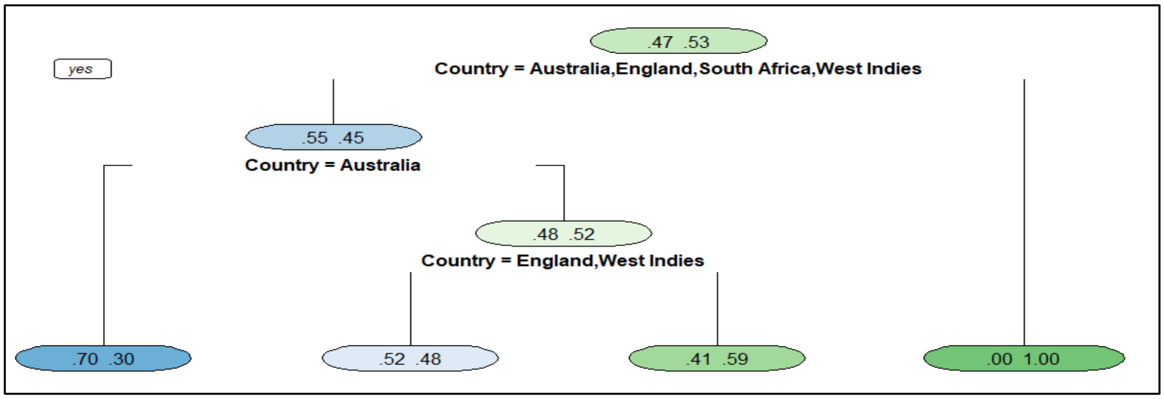
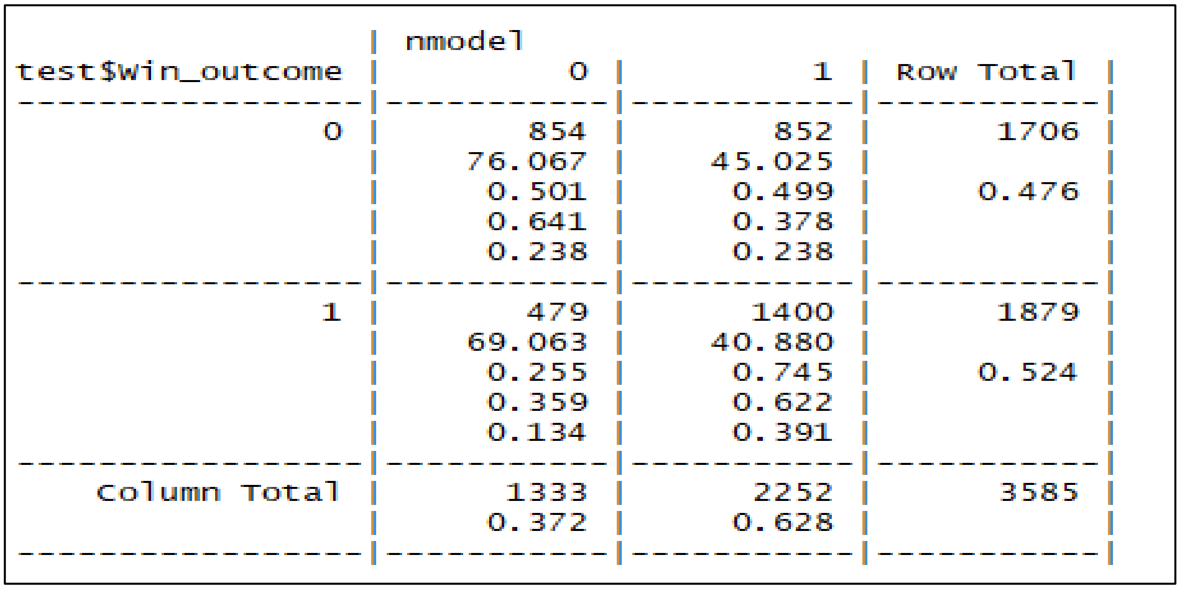

yes
.47 .53
Country = Australia,England,South Africa,West Indies
.55 .45
Country = Australia
.48 .52
Country = England,West Indies
.70 .30
.52 .48
.41 .59
.00 1.00

Figure 18: Decision Tree

```
 1) root 8000 3790 1 (0.4737500 0.5262500)
  2) Country=Australia, England, South Africa, West Indies 6833 3043 0 (0.5546612 0.4
453388)
    4) Country=Australia 2368 706 0 (0.7018581 0.2981419) *
    5) Country=England, South Africa, West Indies 4465 2128 1 (0.4765957 0.5234043)
     10) Country=England, West Indies 2750 1325 0 (0.5181818 0.4818182) *
     11) Country=South Africa 1715 703 1 (0.4099125 0.5900875) *
  3) Country=New Zealand 1167    0 1 (0.0000000 1.0000000) *
```



```
                 | nmodel
test$Win_outcome |         0 |         1 | Row Total |
-----------------|-----------|-----------|-----------|
               0 |       854 |       852 |      1706 |
                 |    76.067 |    45.025 |           |
                 |     0.501 |     0.499 |     0.476 |
                 |     0.641 |     0.378 |           |
                 |     0.238 |     0.238 |           |
-----------------|-----------|-----------|-----------|
               1 |       479 |      1400 |      1879 |
                 |    69.063 |    40.880 |           |
                 |     0.255 |     0.745 |     0.524 |
                 |     0.359 |     0.622 |           |
                 |     0.134 |     0.391 |           |
-----------------|-----------|-----------|-----------|
    Column Total |      1333 |      2252 |      3585 |
                 |     0.372 |     0.628 |           |
-----------------|-----------|-----------|-----------|
```

Figure 19: Decision Tree Confusion Matrix

Looking at the figure above we can see 479 out of 1333 "No/0" was incorrectly classified as "Yes/1" and 854 "Yes" are incorrectly classified as "0/No". this leads to the problem which is called true negative win outcome. The correctly classified are 1400 times winner therefore 479 is the false negative. The result needs to be improved.

6 Conclusion and Recommended Future Work

This report shows how different data mining techniques can be used to predict the winning outcome of the cricket game, this data is fetched from the cricket website called espncricinfo.com, due lack of data size the data is artificially generated to produce a certain size. The challenges were encountered during deciding the model to be used. Both R and Python are used to create the models, also removing the noise from the data was one of the challenging parts of this project. The difficulty in creating model occurred due to lack of correct data once again, the result shown above are not correctly justified due to this problem. The Support Vector Machine (86%) and KNN (100%) produced better results compared to other models such as Decision Tree, Linear, Multilinear and Polynomial Regression. The factors such as team matches played and win outcome does affect rating and points that are gained by the team. The study does not correctly predict win outcome clearly but, we can presume from this study if this is conducted on a cleaned dataset the models do have the ability to produce the better and acceptable results. The finding of this is mainly dependent on the three model which performed good results, the study reveals that cricket win outcome is affected many factors such as team players, number matches it has played and the number matches won by the team.

If the team performance is consistent it is likely that player of the team will be happy therefore they likely to perform better, thus rating and points increases, whereas if the team who does not perform well tend to have a low rating and points which affect the players. Correlation between a team playing matches and rating and points increasing was found in Multi Linear Regression.

For the future work, the first thing that needs be considered is a good selection of the data because in this project the data was not realistic. The second aim should be to have various data that includes attributes such as male and female cricketers, therefore a comparison can be made. If I were to do this project again I would consider performing more model that is suitable for the data to produce a better result. The Twitter analysis can also be performed to see the player or team emotions and the reaction of cricket fans during the matches. Next time I would also keep in mind that evaluation part need be properly evaluated to give the reader better understanding of the problem that is being addressed and solution that is provided.

The future work also needs to be a focus on writing the document properly as well coming up new ideas that can be added to the project such as creating a website that allows the user to access matches and perform analysis based on the data that is provided to him/her. The website can include functionality, that will teach the user how cricket is played before he/she wants to perform the analysis. This means anyone who is not even a cricket fan but they are keen to learn about cricket, they learn with the help of the website which will allow them to gain more understanding of the cricket matches and become interested in it.

Overall the I learned a lot from doing this project and will focus on improving the data analyst skill and produce a good result using the data mining techniques effectively.

References

1. Analyticsvidhyacom. 2015. Analytics Vidhya. [Online]. [3 February 2018]. Available from: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
2. Dezyrecom. 2017. *DeZyre.* [Online]. [28 October 2017]. Available from: https://www.dezyre.com/article/big-data-analytics-the-new-player-in-icc-world-cup-cricket-2015/89
3. Espncricinfocom. 2018. Cricinfo. [Online]. [2 October 2018]. Available from: http://stats.espncricinfo.com/ci/engine/records/index.html?class=3
4. Gamage Harsha Perera (2015). ResearchGate. [Online]. [14 March 2018]. Available from:https://www.researchgate.net/publication/261227104_Application_of_Association_Rule_Mining_A_case_study_on_team_India
5. Hasseb Ahmad, Ali Daud, Yixian Yang and Haibo Hong (2017). [Online]. [14 March 2018]. Available from: https://ieeexplore.ieee.org/document/7878604/
6. Ibmcom. 2017. *Ibmcom.* [Online]. [28 October 2017]. Available from: http://www.ibm.com/analytics/us/en/technology/spss/
7. Ieeeorg. 2018. Ieeeorg. [Online]. [14 March 2018]. Available from: https://ieeexplore.ieee.org/Xplore/home.jsp
8. Machinelearningmasterycom. 2014. Machine Learning Mastery. [Online]. [3 Feb 2018]. Available from: https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/
9. Ncirlie. 2018. Ncirlie. [Online]. [3 Feb 2018]. Available from: https://moodle.ncirl.ie/course/view.php?id=1024
10. Oraclecom. 2018. Oraclecom. [Online]. [25 Feb 2018]. Available from: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm
11. Prince Kansal, Pamkaj Kuma and Himanshu Arya (2014). [Online]. [14 March 2018]. Available from: https://ieeexplore.ieee.org/document/7019707/
12. Pranavan Somaskandhan, Gihan Wijesingle and Sampath Degalla (2017). [Online]. [14 March 2018]. Available from: https://ieeexplore.ieee.org/document/8300399/
13. P. UmaMaheswari and Dr.M. Rajram (2009). [Online]. [25 Feb 2018]. Available from: https://ieeexplore.ieee.org/document/4809163/
14. Rroijcom. 2017. *Rroijcom.* [Online]. [28 October 2017]. Available from: https://www.rroij.com/open-access/an-overview-of-knowledge-discovery-databaseand-data-mining-techniques.php?aid=48833

15. Sannoy Bhattacherjees, Jayakrushna Sahoo and Adrijit Goswami (2015). [Online]. [14 March 2018]. Available from: https://ieeexplore.ieee.org/document/7306757/

16. Scikit-learnorg. 2018. Scikit-learnorg. [Online]. [25 Feb 2018]. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

17. Softwareadvicecom. 2017. *Softwareadvicecom.* [Online]. [28 October 2017]. Available from: https://www.softwareadvice.com/bi/tableau-alternatives/

18. Softwareadvicecom. 2017. *Softwareadvicecom.* [Online]. [28 October 2017]. Available from: https://www.softwareadvice.com/bi/tableau-alternatives/

19. Tejinder Singh, Vishal Singla and Parteek Bhatia (2016). [Online]. [25 Feb May 2018]. Available from: https://ieeexplore.ieee.org/document/7489605/

20. Ujwal UJ, Dr Antony PJ and Sachin DN (2018). Ripublicationcom. [Online]. [14 March 2018]. Available from: https://www.ripublication.com/ijaer18/ijaerv13n5_96.pdf

21. Ustsav Jadishbhai Solanki and Prof.Jay Vala (2017). Statsfuca. [Online]. [14 March 2018]. Available from: https://www.stat.sfu.ca/research/defences/Abstracts/1157-1164/1157-HarshaPerera.html

7 Appendix

7.1 Project Proposal

7.1.1 Objectives

Objective 1: For the success of this project the first objective would be to find dataset of cricket matches that are played by certain teams and player over the years. I will be using the websites that are well known and provide correct information about cricket such as insights which is powered by espncricinfo.com and it provides two decades of historical data.

Objective 2: The second objective will be focused on extracting and cleaning the dataset that is relevant to my project such as matches, players, teams. This objective is very important for my project as this requires cleaning the data, therefore, I would need to make sure, the correct and accurate information is gathered to proceed.

Objective3: My third objective will be to compare the matches, player and team performance in terms of score and result. The outcome this objective will play an important role when doing the analysis.

Objective 4: The fourth objective will to identify a pattern based on the results and see how the team has performed over the years and what could help them to win the match.

Objective 5: My fifth objective will be to create a survey on Twitter and Facebook. As well as gathering tweets from twitter, I will also have results from the survey this will make the analysis unique and help me to predict the result of a match accurately.

Objective 6: The sixth objective will be used to compare the information that I have gathered from social media, own analysis and analysis that is performed by a firm/agency.

Objective 7: The final objective will be to complete the documentation with a conclusion and see how the analysis could help cricket team to improve their performance and change the game.

7.1.2 Background

The idea of doing analysis on cricket generated from my personal experience when I was talking to one of my friend about my specialization for the 4th year. I was asked a question what do data analyst do and my answer was that they perform analysis of certain information. It wasn't quite clear to him, therefore, I gave him an example of cricket, suppose you want to predict an outcome of a match, how would you predict it, you would need to know both team history such as player and team performance and then decide which team has a higher chance of winning the match. After giving him this example it was clear to him that data analyzing require collecting, extracting and transforming. Also, show the analysis of charts or graphs so we can understand what the result of the analyzing mean.

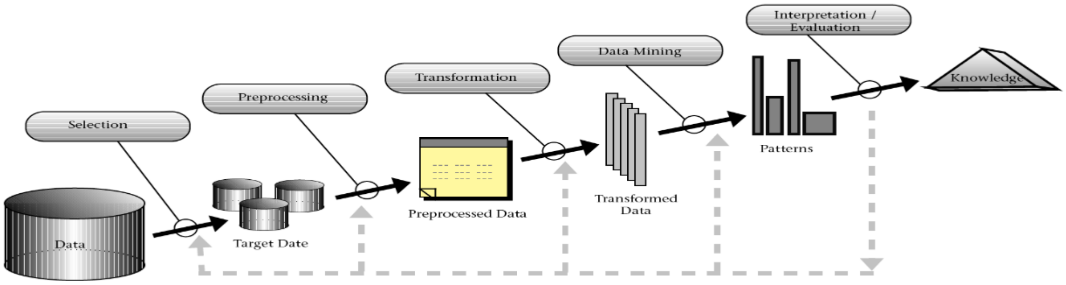
As I am a cricket fan therefore I decided, why not do this for my final year and add something to it to make it bit more interesting.

Developing this idea will allow many experts to see their player or team weakness and could help them to improve their game and help them achieve what they want. I will be gathering data from many wells know cricket website that provides proper data for cricket matches and a team player. I would be able to help in the real-world situation if there is a match that is being played and results of the can be determined before the match begins, as I will be taking factors such as player and team performance, also what people think from all over the. The social media, Facebook, Twitter will be used to compare data analysis against the data that is gathered from cricket websites.

A question such as who will win the match? which team has potential to win and why? what can the team manager to improve the game or change the game? There are many benefits for fans and people who like betting on players and teams. The goal of this project will be to address all the question and provide a solution to the problem and the questions. By doing this analysis it can help to better understand cricket and new way of providing better outcome using analysis and comparing them to social media.

7.1.3 Technical approach

I will be using KDD (Knowledge Discovery in Databases) because it is a process for discovering very useful information from a collection data, therefore it will help me to discover data for the cricket.



KDD includes the following approaches:

1. Selection: This is very first and important phase of the KDD approach, here it involves finding and selecting the data information that is suitable for the project and that will help me to achieve the goal of the project.

2. Processing: This is the second phase of the KDD approach, where the data that has been collected it cleaned which mean the data that is not needed is eliminated and only the data that is required for the project are kept.

3: Transformation: This is the third step of the KDD approach, as it requires to take the clean set of the data which is produced by the second phase and convert it in form of better data which can include reducing the dimension.

4: Data Mining: this very important phase of the data, here the discovering of the important information from large sets of data is derived. There are many techniques to be used because the data mining uses many mathematical analyses to show patterns.

5: Interpretation: this stage requires to interpret data from the records. After doing the analysis I will use this method to read and interpret data.

7.1.4 Special resource required

R Studio: It is required for fetching and exploring the data.
Python: This is required for cleaning and creating the model.
Tableau: This is required to visualize the prediction model.
MySQL: It is required to store the cricket data.

7.1.5 Technical details

R Studio: I will use the R studio to extract the data from the web and to perform some analysis.

Python: The Python will be used to clean, perform cross-validation and feature scaling. It will also be used to create a prediction model for the project.

SPSS: This application is very popular which has the capacity to perform well, using the less complex technique. So, this will help me selecting the right model for specific data.

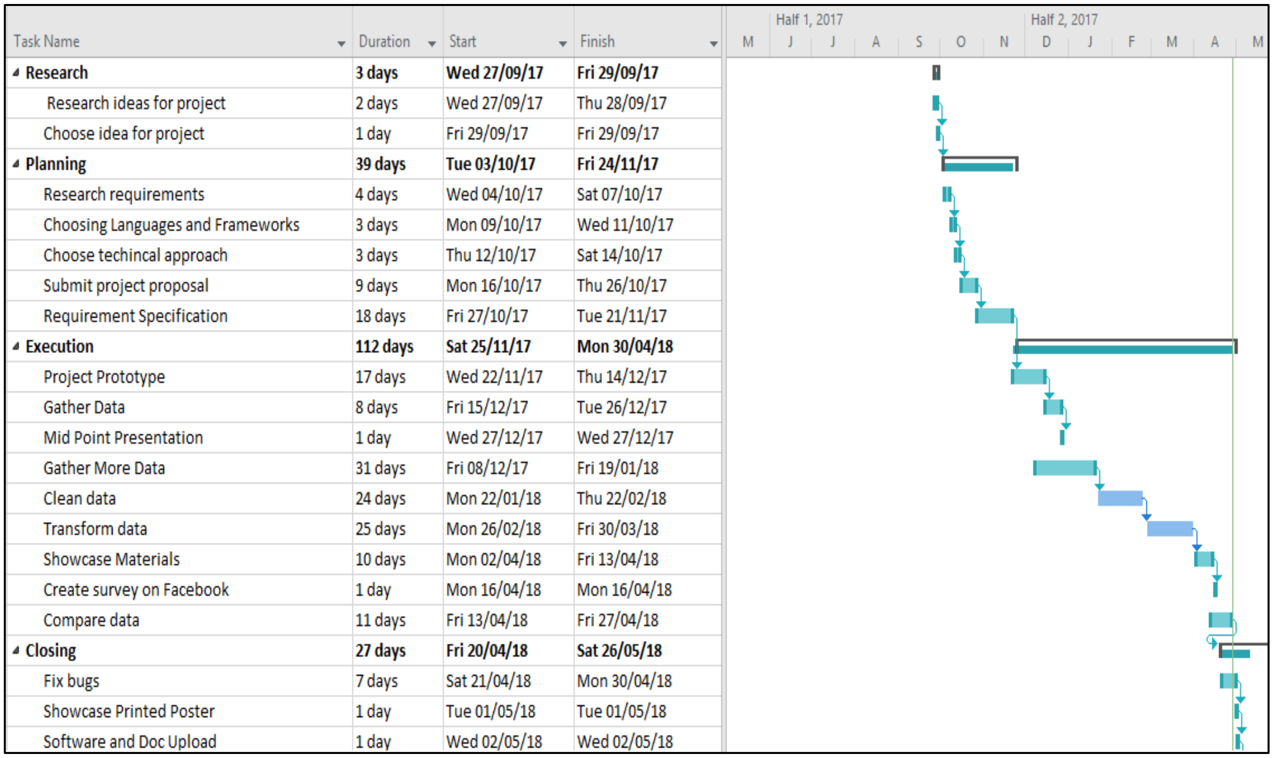
Excel: It is a spreadsheet tool this allows us to build graphs that can help me see all the data very easily, therefore, it can be used to change data and store the cricket information in a structured way.

SQL Server: It is very popular for easy accessing and storing data, I will be using it too I access all the data from where ever I want.

Tableau: I will be using this tool to create and visualize the result of the output, it will be very beneficial to display all the work very easily.

7.1.6 Project Plan

TASK	PRIORITY	DAYS	START DATE	END DATE	DUE DATE
Project proposal	Medium	9	17/10/2017	27/10/2017	27/10/2017
Requirement specification	Medium	18	01/11/2017	20/11/2017	24/11/2017
Gather data	Medium	Throughout project			
Project prototype	High	17	04/11/2017	25/11/2017	02/12/2017
Mid-point presentation	Medium	1	05/12/2017	05/12/2017	05/12/2017
Analyze the data	Low	Throughout project			
Showcase materials	Low	10	23/03/2018	05/04/2018	06/04/2018
Software & document upload	High			12/05/2018	13/05/2018
Presentation	Medium	6	14/05/2018	20/05/2018	22/05/2018
Showcase	Low			30/05/2018	30/05/2018



Task Name	Duration	Start	Finish
Research	**3 days**	**Wed 27/09/17**	**Fri 29/09/17**
Research ideas for project	2 days	Wed 27/09/17	Thu 28/09/17
Choose idea for project	1 day	Fri 29/09/17	Fri 29/09/17
Planning	**39 days**	**Tue 03/10/17**	**Fri 24/11/17**
Research requirements	4 days	Wed 04/10/17	Sat 07/10/17
Choosing Languages and Frameworks	3 days	Mon 09/10/17	Wed 11/10/17
Choose techincal approach	3 days	Thu 12/10/17	Sat 14/10/17
Submit project proposal	9 days	Mon 16/10/17	Thu 26/10/17
Requirement Specification	18 days	Fri 27/10/17	Tue 21/11/17
Execution	**112 days**	**Sat 25/11/17**	**Mon 30/04/18**
Project Prototype	17 days	Wed 22/11/17	Thu 14/12/17
Gather Data	8 days	Fri 15/12/17	Tue 26/12/17
Mid Point Presentation	1 day	Wed 27/12/17	Wed 27/12/17
Gather More Data	31 days	Fri 08/12/17	Fri 19/01/18
Clean data	24 days	Mon 22/01/18	Thu 22/02/18
Transform data	25 days	Mon 26/02/18	Fri 30/03/18
Showcase Materials	10 days	Mon 02/04/18	Fri 13/04/18
Create survey on Facebook	1 day	Mon 16/04/18	Mon 16/04/18
Compare data	11 days	Fri 13/04/18	Fri 27/04/18
Closing	**27 days**	**Fri 20/04/18**	**Sat 26/05/18**
Fix bugs	7 days	Sat 21/04/18	Mon 30/04/18
Showcase Printed Poster	1 day	Tue 01/05/18	Tue 01/05/18
Software and Doc Upload	1 day	Wed 02/05/18	Wed 02/05/18

Figure 20: Project Plan using Gantt Chart

7.2 Monthly Journals

7.2.1 Introduction

I am Kuldeep Rawat and Today is 29th September, which is the first day of writing this journal. I am in 4th year of my degree and studying data analytics as the specialization for the final year at National College of Ireland.

7.2.2 September 2017

My achievements

This month I looked at many websites, which shows the recent ideas and trending products and service. I am required to come up with an idea based on data analytics. Finding a new idea from data analysis point of view was bit tough because I have never created a project which requires analyzing data, therefore, lots of research was put into it. After doing my research, I finally decide to choose an idea for the final project on Cricket T20 matches. I knew this idea was not innovative at all but I just wanted to choose a topic that I am interested in and whose data is easily available from the internet. I also prepared a little plan for presenting this idea on 2nd October 2017 to the three judges.

My reflection

I felt, there was the benefit of doing lots of research for coming up with an idea to solve problems in the real-world situation. However, I was not confident enough to present my idea to someone else because I did not have a clear picture of how I was going to achieve the goal of the project.

Intended changes
Next month, I will try to do more research in cricket filed and see if anything can be added to the idea to make it more interesting.

Supervisor meeting
No supervisor is assigned yet.

7.2.3 October 2017

My Achievements
This Month on 2nd October, I presented my idea for the final project to the three judges. The idea was based on the analysis of cricket world cup 2019 when I was presenting this idea to the three judges they were not really interested in it as it was not new or innovative. I knew this before going to the three judges, but I had a plan which was to add few things to make it different, as I expected the judges did not agree to accept it, therefore, I also added few things such creating a web-app after doing an analysis of the cricket match. The judges suggested me to do analysis on recent match rather than the world cup and added instead of creating a voting system on the web-app retrieve data from Facebook, Twitter and then compare it to the analysis that you'll be performing.

I agreed with what they were saying, this talk went for approximately 12 minutes. After coming out of the room I was relieved.

5th October 2017
I have completed the bit of research how I am going to proceed with an idea and what tools to use.

10th October 2017
Today I started working on the project proposal I checked the example of the project proposal which was given on the Moodle. After writing my objectives for the project proposal, I decided to meet my supervisor to learn about the structure of the proposal and what else can write to improve it. I emailed my supervisor on the 19th for the meeting.

20th October 2017
Today I had met with my supervisor at 1:00 pm, she described many aspects of cricket that I can work on and helped me understand the project proposal structure. The benefits of meeting her were, that I was able to expand my idea and make it more interesting by adding few more functionalities. Also, got helped with improving the proposal.

22nd October 2017
As discussed with my supervisor, I started to write about the project proposal, in the that I wrote about objectives of the project, background, technical approach, project plan and technical details. All this information is relevant to my project.

25th October 2017

I did not finish writing the project proposal, therefore, I tried to complete it, at this stage, I had everything done except for the project plan and Evaluation.

26th October 2017
Today I created a project plan using Microsoft Project and completed it. I also wrote about evolution.

27th October 2017
I uploaded the project proposal on the Moodle but made a mistake to not press the submit button, therefore it was uploaded but not submitted. On the next day when opened Moodle, I saw the assignment was overdue by 14 hours, after viewing it I pressed the submit button and it was uploaded and submitted. I still feel that It can be still improved and I will start working on the next assignment.

My Reflection
This month I put lots effort in the project, looking back I knew I had to do well to succeed in completing this project. Although there was lots of work pressure from other subjects, which constantly stops me from focusing properly on this project.

Intended Changes
I intend to focus more on this project as it is worth 20 credits. I know there are other subjects too, but I need to create a plan for this project so it can be completed easily and effectively without having too much workload.

Supervisor Meeting
Supervisor Name: Catherine Mulwa
Date of Meeting: 22/10/2017
Item discussed:

1. Project proposal structure.
2. Type of model that I should consider.
3. New functionalities.
4. Type of data and where to get from.

Action Items:

1. Complete project proposal.
2. Do research about types of functionalities.
3. Check out cricket website for data.
4. Learn how to gather data using R or Python.

7.2.4 November 2017

My Achievements

This month, I worked on requirement specification document, which took longer than expected. Finally, I could finish the document and uploaded it on time. In this document, I added more information about what exactly my project is about, also listed all the technologies and described the process that needs to be completed for the project. I would say this month was one of the busiest months as I had to finish requirement specification, technical report and prepare for CA for other subjects.

Although, it was the busy month for I had positive feelings towards the CA and project. The requirement specification was due on the 24th and then the technical report was due on the 30th. The technical report was easy to do as it required us to just take the work from requirement specification and modify it, which I did without any issues. After visiting my supervisor, I was able to improve the document layout and there were few errors such as incorrect use case, heading for some section were not written properly, I also prepared presentation slides and was able to fetch a small data of cricket players and used Rstudio to demo the work to the examiner.

My Reflection

The amount work that I was putting in was not making me tired at all, although, it was increasing my energy to do more and more work. Submitting all the work on time gave me beautiful rewarding feeling.

Intended Changes

Next month I will continue to follow the KDD process which I decided to use in my project, I hope to gather more data from the website and start cleaning code. I also believe technical report needs to be updated so I will make sure change the document also.

Supervisor Meeting

Supervisor Name: Catherine Mulwa
Date of Meeting: 21/11/2017 and 28/11/2017
Item discussed:

1. Requirement specification document structure.
2. Use case.
3. How to draw architecture diagram.
4. What will be required for the presentation?

Action Items:

1. Complete the document.
2. Change the Use case and architecture diagram.
3. Complete prototype.
4. Create error-free presentation slides.

7.2.5 December 2017

My Achievements

During the start of the month, it was a little stressful month for me because I got ill therefore I was not able to attend the presentation which was due on the 6th. However, during the middle of this month, I was recovered fully and was able to complete my assignments and attend Introduction to R, CA. At this stage I did not know the date for my presentation because I was waiting for mail from NCI360 in regard to the extension of the presentation date, few days I received an email on the 19th that my application has been approved and straight away I emailed Eammon and Catherine to arrange a presentation slot for me. The reply received from Catherine was that my presentation was going to be in January and date cannot be confirmed. While all this happening, I was also preparing for my exams in January.

My Reflection
This month was the bit a stressful month for me as I was bit sick and there was a lot going on in college, however, I kept trying hard and did not lose confidence.

Intended Change
I did not work much on the project this month but for next month I would look forward to starting analyzing the data make some small changes that are required.

Supervisor Meeting
There was supervisor meeting this month because I was ill for few days during the start of the month and then I also CA and Christmas holidays.

7.2.6 January 2018

My Achievements
The end of this month was the beginning of the new semester and the start of the month was spent preparing and then doing the exams. After the exam was completed, I was happy and did not focus too much on the project for few days as I wanted to relax and enjoy the feeling of joy and the end of one semester. I met my supervisor and discussed next few tasks that I should focus on, I also spent some time doing the research on an algorithm that can be applied to my data. The outcome of the research, I found there are many techniques such as regression, classification, and clustering that can be applied to the project. I discovered SPPS modeler which can help me generate models for my project. I learned how to use it and applied it to the cricket data. Finally, on 25th I did my presentation which went really and I received lots of good feedback from my supervisor and Lisa Murphy (Second examiner).

My Reflection
This month I got back on track with all the positive energy and was satisfied with my presentation results.

Intended Change
After doing the presentation, there were many feedbacks given by my supervisor and Lisa Murphy which I believe was very helpful for me such as keeping the idea related to the data analyst and using Tableau to visualize graph easily and quickly, also making sure to double check my document before uploading because there were many small errors in the document. I believe these feedbacks were valuable and should be considered.

Supervisor Meeting
Supervisor Name: Catherine Mulwa
Date of Meeting: 10/01/2018 and 30/01/2018
Item discussed:

5. Mid-point presentation feedback.
6. Feature selection.
7. Prediction model and classification model.
8. Problems encounter and solution.

Action Items:

5. Focus on the feedback that is provided.
6. Start developing models.
7. Continue working on the document.

7.2.7 February 2018

My Achievements

This month, I was in the second week of my final semester, over the last few weeks I worked on gathering the data from various websites, I looked at many videos and tutorial on YouTube and data camp. Today I found a helpful website which consists lots of cricket data, the name of the website espncricinfo.com although, I have visited this website before but never really thought it had so much information in it.

February 2nd, 2018

On this day I received an email from FDM about video interview. I researched about the company and clicked on the practice video interview link that was provided within the email. This was completely new for as I have never completed a video interview and have no idea how it works and what kind of questions will be asked or how long will it go for.

February 5th, 2018

I planned to meet Helen Conway to get my CV approved and to ask her about the video interview, she approved my CV and said it was very well done. When I talked to her about the interview she suggested me to look up few presentations that were on Moodle and a link to sparks where I could practice the interview.

February 8th, 2018

I was at work on this day and when I came home 9:00 pm I quickly changed my clothes and put on a shirt to look professional for the interview. The interview was quite hard because I had to very less time to come up with an answer and less time to answer the question. This was my first interview, which I think I did not do well. I prefer the face-to-face as I do well and express myself to the employer wherein the video interview, I did not get much time to explain.

February 10th, 2018

I did an online aptitude test for the company Virgin Media. This test went well.

February 23rd, 2018

Today I had a 3rd meeting of the month with my supervisor at 1:00 PM. She explained what I need to next, she suggested me to work on the prediction model and write a literature review on cricket. Find at least 10 papers between 2002-2018.

February 25 to 28, 2018

I uploaded my assignment for data mining on 25th. Overall this month I gathered all the data that was required, merged and cleaned. I also moved on to the next phase of the project which was to create analyses and create models.

My Reflection

This month was the very easy productive month for me as I had the data cleaned and merged, all my literature review was completed and a linear model was started.

Intended to Change

I need to have clear picture where I am going, therefore, need to update my plan and set few targets to effectively achieve them.

Supervisor Meeting

Catherine has been supportive over the past few months and always guided me in the right direction. This month I had three meeting with her.

Supervisor Meeting

Supervisor Name: Catherine Mulwa
Date of Meeting: 08/02/2018, 15/02/2018 and 27/02/2018
Item discussed:

9. Literature Review
10. More about prediction model and classification model.
11. Solution to the errors.

Action Items:

8. Complete Literature review.
9. Fix errors.
10. Develop more models.

7.2.8 March 2018

My Achievements

This month, I created the showcase poster and developed two models. I also wasted most of my time fetching the twitter data because I was thinking of performing some sentimental analysis on the cricket matches as well as prediction models. I did not focus too much on creating the model because I already had some research completed and wanted to start after finishing some sentimental analysis on the Twitter data. After talking to my supervisor and one of a friend about the project I decided to leave the idea of fetching the data from Twitter as it was going to take too long to download the data. Catherine agreed to this and told me to focus on the supervised data that you have gathered from the cricket website.

My Reflection

This month I was very active but I think I wasted most of time and energy in doing the work that should not have been completed and it was not suitable for my project. I was lucky to discover and stop before I move too close to the project deadline.

Intended change

I did finish two of the regression model and need to make few changes to improve it.

Supervisor Meeting

The supervisor meeting throughout the year has been great and it was very helpful and full support of supervisor was provided with very useful tips which keep us on track. I feel confident enough to complete my project on time.