



Machine Learning In Cricket

By: Moiz Buch

Deliverable 2 Final Year Dissertation

First Reader: Adrian Turcano

Second Reader: Lynne Baille

Declaration

I, Moiz Buch confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signature: Moiz Reyaz Buch

Date: 23/04/2020

Abstract

Sports Analysis is one of the most upcoming fields in job industry. This is because of the availability of the amount of data for the sports industry. Nowadays sports enthusiasts tune in an hour earlier before the game where they watch pundits analyze the game and give their opinion on the team strength and try to give their own prediction about the game. The game of cricket is such a sport where one gets to receive innumerable amount of data because it has a lot of mathematics and statistics in it. This possibility of analyzing a lot of data in sports and especially cricket has happened due to the invention and now rigorous use of machine learning algorithms. However, there have not been many active uses of machine learning algorithms but there are possibilities of even predicting the match outcome using machine learning.

This dissertation will carry out research which machine learning algorithms can be used to find the outcome of a cricket match and to what accuracy can the machine learning predict that outcome. I am interested about learning the match outcome as in cricket the end result matters the most unlike in football where you have a lot of draws in cricket draws are not common and end outcomes matter a lot.

Table of Contents

Contents

Declaration	2
Abstract	3
1. Introduction-	5
1.1 Cricket and its history	5
1.2 The Problem	6
1.3 Personal Motivation Behind the Dissertation	7
1.4 Aim and Objective	8
1.5 Summary	9
2. Literature Review	9
2.1 Objective	9
2.2 Introducing the Literature Review	9
2.3 Machine Learning	10
2.4 Background Cricket	16
2.5 Previous Work on Cricket analysis	21
2.6 Machine Learning with cricket	23
2.7 Summary	27
3. Research Methodology and Requirements Analysis	27
3.0 Objective	27
3.1 Quantitative Data	27
3.2 Research Questions	28
3.3 Non-Functional requirements	30
4. Data Set	31
4.0 Objective	31
The Objective is to discuss the dataset that is being chosen and what is the origin of the dataset and why that particular dataset is being chosen	31
4.1 Data Origin	31
4.2 Data Pre Processing	32
5.0 Design	33
5.0 Objective	34
5.1 Model Components	34
5.2 Program design for implementation	34
6. Implementation	35
6.0 Deciding the Program and Software to use	35
6.1 Testing	36
6.1.1 Cross Validation	36
6.2 Model Implemented	37
7. Conclusion	43
7.1 Survey Results	45
8.0 Evaluation	45
8.1 Main Achievements	45
8.2 Limitations	46
8.3 Future Work	46
9.0 References	48
8. Appendices	50
8.1 Consent Form	50

1.Introduction-

1.1 Cricket and its history

Cricket is known as a Racquet sport that dates back up to the 15th century. It was mainly played in the southern part of England under “club games”, other sports included under the same club games category were Hockey and Golf as all these games had a ball and stick in common. However, cricket wasn’t always the same when it was played earlier. Before there used to be one batsman holding a bat that used to look like a hockey stick and there was one bowler who used to bowl the ball underarm and the batsman defend the ball from a low two stump wicket and the “runs” scored were called notches because the scorers recorded them by nothing tally sticks. It was then in 1760 that the bowling of cricket underwent a revolutionary when bowlers started to rotate their arm and began to pitch the ball on to the wicket. A few years later the shape of the bat had also undergone a revolutionary where before it looked like a hockey stick and then started to look like a modern-day bat where it has a huge cleft but less shoulder edges on the bat. In that period, the bat handle was made of wood but in the 20th century bat handles were made of Sarawak canes imported from Malaysia as the impact of hitting the ball on the bat was withheld strongly from a cane handle rather than a wood handle, so wood handles would break easily and cane handles were strong. It was then in 19th century that cricket started to be played internationally and grew its playing population accordingly mainly in commonwealth countries as it was a sport only played in British controlled nations as the army used to play it as past-time.

Cricket is mainly a team sport played between 2 teams and each team consists of 11 players that play in one match. Additionally, since 2019 many teams have the option of including a substitute player known as the 12th man to play the match in case any injury is occurred for a player. Unlike other team sports such as Hockey, Football and Basketball Cricket does not have many substitutions. The international body recognized in cricket is known as ICC(International Cricket Council) whose

headquarters are based in Dubai, UAE. Additionally there are also sub regional organizations according to each country such as ACC(Asian Cricket Council) that comprises of only Asian cricket playing nations. Furthermore, each cricket playing nation has its own body like any other sport such as Bcci(India), English Cricket Board(England), and ACB(Australia) etc.

Cricket is the second most popular sport in the world as around 800 million people watch the sport all around the world especially tournaments like the ICC Cricket World Cup, Indian Premier league and ICC Twenty20 World cup. Around 700 million viewers are coming from India itself where cricket is played like a religion, below them comes England where viewers watch the most domestic matches and the second most viewers of international matches edging out Australia in 3rd place. [6]. Cricket as a sport provides the viewers a sense of letting of frustration also making of them forget daily life pressures of professional and personal lives. In a country like India, if India plays a match people on the streets, home, cafes or anywhere are like the twelfth man involved in the match for example if India loses one wicket the whole country becomes silent or stand still, whereas on the other hand if the batsman scores a six people start dancing with joy. Each cricket match in the subcontinent countries which consist of India, Bangladesh, Pakistan and Sri Lanka is like a huge religious event that everyone is celebrating. In the west, a country or islands that share the same ideology is West Indies where for them playing a match is like a party night and they have fun.

1.2 The Problem

The Cricket prediction game has its audience gaining interest each year. Through emotional predictions many viewers place predictions on their own team without using any reasoning.

Emotional prediction is when a viewer has a personal connect with the team for example if the person is from New Zealand he would bet for the New Zealand Cricket team. Cricket prediction has become a past time when watching matches just like doing votes on social media. However, many consumers confuse prediction with betting which could also be true. In a country like India where cricket betting is banned it is been recorded that the betting industry is worth \$150 billion that is of illegal money as the law for betting is under review. In other countries such as UAE, England and Australia betting is

legal and it is reported that each One Day international game that has 50 overs each side a total bet of \$200,000 starting up to \$1 million if the game is played between 2 old rivals such as England vs Australia or India vs Pakistan. This year a tournament called Big Bash league which a twenty20 match was sponsored by a company named Bet365.net in which users can predict match outcome and gain points instead of money so that is how it was legal in India as there was no money involved. However, there is no such predictor that uses logic to predict matches and that is the problem of this study.

1.3 Personal Motivation Behind the Dissertation

Since I was 7 years old I have started playing cricket in academies, clubs and at school level as well. Cricket was the only thing that I would think about day and night. Till now along with my studies I try to keep myself fit by being involved in corporate cricket matches based in the UAE. As I grew older and moved to high school I started to look at the game not only from a player's point of view but also as a mathematician/statistician's point of view as I was learning a lot of numbers at that point of time along with physics and so I used to apply everything that I learnt in school scientifically to the sport and from 2011 ICC world cup I got the passion of listening to sport analysis and trying to predict matches on my own. I still remember there was an India vs Pakistan match and one hour before the match started I was avidly listening to the sports analysis where the analyst showed data such as if India bat first they score these many runs and so at the end of the segment just before the match starts the analysts without using any data are asked to predict which team might win. This decision is purely based on the analysts gut feeling without any use of data, this factor used to annoy me a lot as to how and why do they predict like that. Perhaps technology and data sets were not available at the time to try and predict as accurately as possible. Currently I myself is an avid predictor of cricket without any money though just for the fun of it.

Currently studying in the 4th year for my honours degree in Computer Systems at Heriot Watt

University, I am required to do a dissertation research based project that should be done by myself. So, the topic that I have decided to do my research on is mixture of two things that I am interested in, Machine Learning and Cricket. The research topic that I have chosen is how can Machine Learning be used to predict the match outcome in an International Cricket Match.

1.4 Aim and Objective

The aim of the project is to conduct a research based on which machine learning technique is suitable to predict the outcome of an International cricket match that can be further used by sports analyst and sports enthusiast to predict the outcome. The cricket matches that would be chosen to belong to a tournament that has happened a few months back called the ICC Cricket World Cup. I would be comparing the machine learning models and their outcomes to the actual outcomes.

While developing any project one must have objectives to achieve targets when doing the project. The aim of the project is the result of where I want the product to get and the objective is that how would I get there to achieve my aim. Hence, I plan to have SMART objectives for the project.

The definition of smart objectives is the following:

S: stands for specific goals that have a high chance of being achieved.

M: stands for measurable goals, any goal that has a result of success or failure.

A: stand for achievable goals, can the goal be attained and through what process.

R: stands for realistic goals that can be achieved with the available resources.

T: a goal that is time bound.

The following are my goals:

- Evaluate which machine learning techniques can be the most accurate to predict the outcome of the cricket match.
- To decrease the gap between a cricket analyst and cricket information.
- Predicting a tournament correctly using the machine learning technique.

1.5 Summary

This chapter provides an overview of the project by introducing the sport and its history, provides the motivation for doing the project and lastly what are the aims and objectives for the project.

2. Literature Review

2.1 Objective

The main objective of the literature review is to study about the scope of the project. The literature review will give you more information on things related to the topic and some existing projects that have been done related to the topic.

2.2 Introducing the Literature Review

A literature review is the searches and evaluation of the information be it journal, article or a website that one has researched about related to the topic of interest. When searching for how

machine learning can be used to predict matches, a few research papers and articles were found that would be explained in the following section.

2.3 Machine Learning

2.3.1 Introduction to Machine learning

With the innovation of new computer technologies, machine learning of nowadays isn't like the machine learning of the previous era of 20th century. Initially machine learning was created to learn from pattern recognition, hence the theory that computers could learn without the need of being programmed to do the specified tasks. In the late 20th century and early 21st century computer scientists conducting research about computers wished to visualize if the computers will learn from the data given. [17]. The reiteration side of machine learning is the most important aspect because the models are updated with new information and the machine learning is ready to adapt with the new data or information given to its model. [19]. Machine learning depends a lot on previous computations to give reliable predictability as a result. Many people have a misconception of it being a science being newly introduced however it is not true it is a science that is being evolved yearly. [20]

2.3.2 Background of Machine Learning

One can say that Machine learning was invented recently due to its advancements found recently in every aspect of life, however it is not true, machine learning was originally invented in 1959 by a computer scientist who was a master in artificial intelligence and gaming, Arthur Samuel. [18] He was the one who came up with the word "Machine Learning" back then and so is known as one of the pioneers of Machine Learning. One of the best factors in Machine learning is that there is no need for a person specifically program the machine it learns itself based on experiences. [19]. The purpose of machine learning to analyse past data and learn from and solve a given problem using the data. Like most

computer programs machine learning also uses algorithms that will train itself using the data and then make a prediction or decision depending on the amount of data given to learn or train from. [7][9]

2.3.3 Machine Learning Algorithms

The Machine Learning Algorithms are usually divided into 4 different categories:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Semi Supervised Learning
- Reinforcement Learning

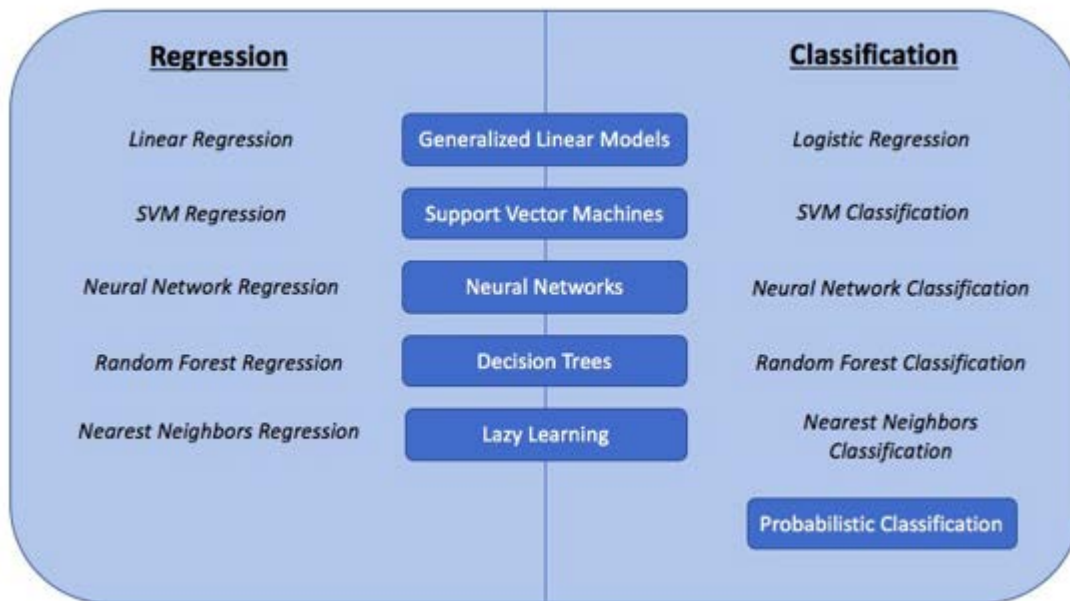
2.3.4 Supervised Learning

In this technique, the algorithm is trained using tagged attributes, “such as an input where the desired output is known.” For example, a cricket match can have data points labelled won or lost. The learning algorithm will get a set of inputs with corresponding correct output, and the

algorithm will eventually learn by comparing the output produced by the algorithm against the actual existing output to find the error or accuracy.[15] The algorithm will then modify itself accordingly. So basically, it will have two sets of data where one data will be the training data that will be used by the algorithm to train or learn from and the second set of data will be the output data produced by the algorithm. This type of algorithms is generally used to predict outcomes.[9] Additionally there is also Classification and regression. Classification occurs when output is a group, while output is a constant number when regression occurs In our case we want to know the match outcome whether the team is win or lose category.

Example of Supervised Learning

- Neural Networks
- Logistic Regression
- Linear Regression
- Nearest Neighbour



Modelling Linear Regression

Modeling Linear Regression and formula have a number of market applications. They are used, for example, to determine industry patterns, and to make predictions and estimates. They can also be used to evaluate the Consumer behavior outcome of price increases. To put it another way, it's a mathematical simulation that helps you to predict and forecast the value of Y based on the various values of X. In fact, you can only have one independent X variable which affects the dependent Y variable. Alternatively, you might have instances where a number of independent variables influence Y.

Simple linear regression means, as you would assume, that there is only one independent X variable which results in changes to different Y values.

Its model / formula is:

$$Y = B_0 + B_1X$$

Where: X – Independent Variable Value, Y – Dependent Variable Value.

B_0 – is a constant (shows the value of Y if the value of $X=0$)

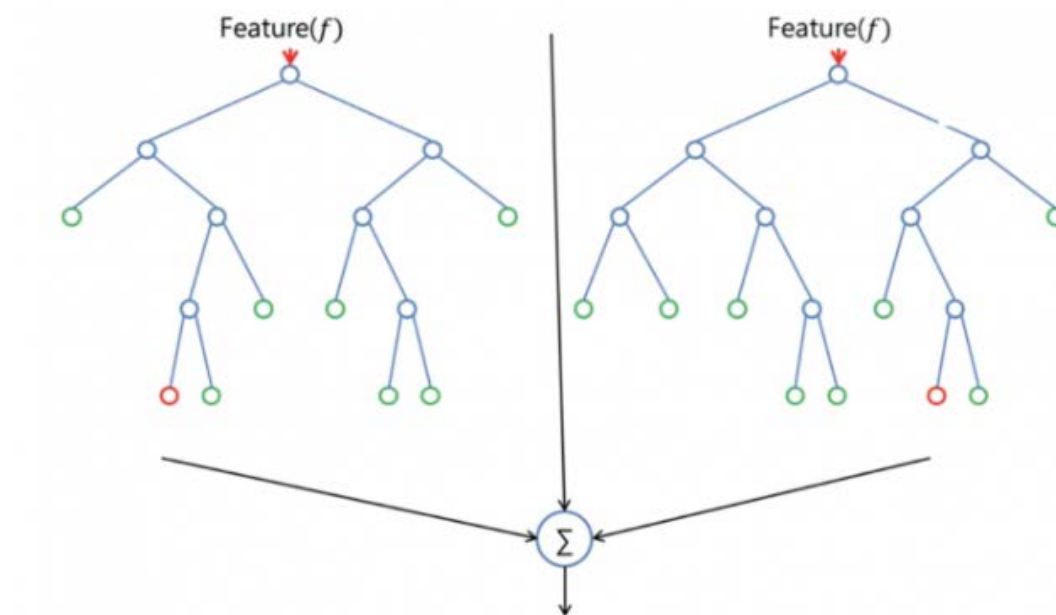
B_1 – the coefficient of regression (Y changes with each unit in X)

Random Forest

Random forest is a scalable, easy-to-use machine learning algorithm that produces great results

most of the time, even without the hyper-parameter tuning. Owing to its simplicity and versatility it is also one of the most used algorithms (it can be used for classification and regression tasks).

Random Forest is an algorithm for supervised learning. The "land" it constructs is an ensemble of decision trees, typically trained with the method of "bagging." The basic concept for the bagging approach is that the cumulative outcome is improved by a mixture of learning models. One big benefit of random forest is that it can be used for problems of classification and regression, which make up the majority of the existing machine learning systems. Let's look at the classification of random forests, as classification is often considered the building block of machine learning. Below is an example shown how 2 trees would appear.



The random forest has about the same hyperparameters as a decision-making tree or classification. Fortunately, a decision tree with a bagging classifier is not required because you can easily use the random forest classifier. You can also manage regression tasks with random forests using the regressor algorithm.

2.3.5 Unsupervised Learning

In this technique, the algorithm uses data that has no labelled attributes. The interesting fact

about this technique is that it is not told the correct outcome. The data only presents scenarios and is supposed to figure out itself what the outcome should be. The main purpose just figure out the data itself and connect the structure. [16] The use of unsupervised learning is various for example to classify consumers of shop that have similar likes or dislikes and can be categorized when presenting the data. [9]

There are several algorithms that use unsupervised learning some of them are:

- Nearest neighbour mapping
- K-means clustering
- Single value decomposition

Generally unsupervised learning can be used to recommend things and are commonly used in organising retailers' data. [8][9]

2.3.6 Semi-Supervised Learning

One of the most unique technique is the semi-supervised learning technique. Overall semi-supervised learning is used for the same applications as supervised learning technique. One of the biggest advantages of using semi supervised technique is that it can analyse data that is both labelled and unlabelled.[17]. Generally, the preference is more towards small data labelled and large data unlabelled, this is because unlabelled data is readily available to get and mostly economically feasible. Semi supervised learning can be used with classification of data, prediction of data but also can be used with regression. Semi-supervised learning can be applied to do many things such as detecting a human's face using a webcam or a front-facing camera on the phone. [8][9]

2.3.7 Reinforcement Learning

Reinforcement learning is generally helpful for robotics, creating games and map navigation applications. The algorithms are dependent risk, sometimes known as trial and error for what action will give the biggest reward. [9] So basically, it is dependent on the action that gives the biggest reward with least amount of risk involved. Reinforcement learning is made up of three main components:

- The agent- the one who will decide what decision to make and will learn with each decision taken
- The environment- anything that the agent is using or interacting with
- The actions- this is what the agent will do

Hence the whole purpose is to take the best action over a given time. For example, we daily use reinforcement learning in our lives through google maps. Google maps can recommend what route is the fastest to take to reach the destination. [7][9]

2.4 Background Cricket

As referred in my introduction section, Cricket is a bat and ball sport that is generally played amongst two teams on a circular ground with a 22-yard pitch right at the centre of the circle. The pitch has two ends and in those two ends there are a set of wickets on both the ends. A set of wickets include 3 wooden sticks and a bail that is placed on top of the wicket. Unlike baseball there is a toss that happens by flipping the coin and whichever team wins the toss has a choice of either batting or bowling first. The main objective of the batting side is to score as many runs as possible by hitting the ball with a cricket bat without letting a wicket go away, meanwhile the main objective of the bowling side is to prevent the batting team from getting the runs also stopping the ball from going to the boundary i.e. 4 runs or 6 runs and getting the batsmen dismissed in cricket terms it is known as “out”. There are multiple

ways of getting dismissed in cricket such as “bowled” where the bowler bowls to the batsman and the batsman misses the ball and the ball hits the stump. The second technique is “caught” where the batsman hits the ball in the air and a person from the fielding side successfully catches it. Another technique is “LBW” this stands for leg before wicket, in this if the batsman is covering the stumps and the ball hits the leg first then the person is given LBW. Another way of getting out is known as “run out” where the fielder throws the ball at either end and if the batsman is short of his crease then he is given run out. Apart from LBW the rest of the modes of dismissals are very like baseball but what isn’t similar is the format.

2.4.1 Formats of Cricket

There are 4 main formats of cricket. The first format introduced in cricket was a test format where the two teams must bat twice and bowl twice in 5 days. However, it is not necessary for the team to bat and bowl twice it is only if the team that has already batted once and lost all 10 wickets then bowl and get the other team “all out” then only they can bat the second time. However, there is no time bound in Test cricket. A team can bat for 5 days continuously if they want to draw the match. To win the test match each team needs to do the batting and bowling twice. The team that is bowling second and the team that is batting second have two objectives. The bowling team should get them all out and the batting must achieve the target. The second format of the game is called one day international. The One day international has gone through an evolution of its own. Initially started on in the early 1970s ODIs (One Day internationals) were introduced to shorten the length of a cricket match. Initially ODIs were to be played in 60 overs. However, in 1992 ICC Cricket World Cup in Australia it was then decided that the maximum number of overs for each side will be 50 overs as many ICC officials thought that 60 overs was way too long for the viewers to watch in the stadium and at home. Since then One Day Internationals have been played in 50 overs. Each side is given 50 overs to bat and bowl. This format is time bound unlike test cricket. So even if there were

no dismissals in the 50 overs the batting side must stop batting and bowling side must stop bowling and they switch roles. The rest of rules are the same such as achieving targets, restricting runs and the modes of dismissals.

The Third Format introduced in cricket is called the Twenty20 format. This format was initially introduced by the ICC to lure other countries such as America, China, Russia, Japan and other non-cricket playing nations to start playing cricket. This was the main reason in 2007 ICC decided to start playing Twenty20 matches and host different Twenty20 world cups. However, after a few years only a couple of countries supported the Twenty20 initiative for the rest of countries like China, Russia and other European Nations Cricket was still slow and boring. So Twenty20 format is just like 50 over ODI format except for the number of overs being played are 20 per each side.

The Fourth and the most recent format being introduced in cricket is T10. This format is currently under testing stages where only a few domestic tournaments and international leagues are implementing this format to see if it is going to work for the viewers. If the

format is a success ICC is trying for Cricket to be an Olympic sport in the next Olympics and this is the format that is going to be implemented. So just Twenty20, T10 is just 10 overs each side with rest of the rules same is Twenty20.

2.4.2 Viewership in Cricket

As already mentioned Cricket is the second most popular sport in the world with over 1 billion followers of the sport. However, each format has a huge gap in viewership. Test Cricket is the least watched format in the sport of cricket with only 800,000 Test cricket followers those also comprise of 60% to be from England, 20% India and 20% rest of the countries. However, in a country like Australia every the 26th of December there is a Test cricket match held at the Melbourne Cricket Ground known as Boxing Day Test, this is more of a cultural thing for Australia going on for over more than a century. Overall the viewership in Test Cricket is less. The factor why it is less is also to do with bilateral series. Not every fan watches all the matches of all the countries and Test Cricket is only played in bilateral series. However sometimes there are series like India vs Australia that have a huge viewership all over the world, in 2017 the gross viewership of the first 3 test matches was about 1.1 billion viewers. [12]

Coming to One Day Internationals, in 2019 ICC cricket world cup Over 1 billion people tuned in to watch the India vs Pakistan of this year's world cup with 90% of the population from India and the rest of them from all over the world. However, there were also other matches that reached the 600 million viewership mark such as the World Cup Final between England vs New Zealand but the record till date was the 1 billion viewership match of India and Pakistan due to their bilateral rivalry as well. Furthermore, One Day Internationals do well only when the tournament is sponsored by ICC (the main cricket council) such as the Champions Trophy and the ICC Cricket World cup. The world cup takes place after every 4 years just like Football. [14]

However consistently the most watched format has been the t20 format. This format has changed the game for ICC and cricket in terms of revenue and viewership. The average viewership of t20 game reaches up 2.5 million viewers if both the teams are from top 10 rankings. Additionally, t20 format has brought many different leagues all over the world that gives international players the opportunity of earning more money and because the viewership of such tournaments is high all stakeholders generally end up getting a lot of money. [13]

Furthermore, t20 is generally watched the most, on average population makes the most prediction in this format rather than an ODI match this is because a shorter format with a lot of fun and excitement and as the viewership is also a lot results in more prediction polls taking place especially in different t20 leagues taking place all over the world.

In an article written on Digital Sports by Dr Masoumeh Mercer, he has mentioned that Cricket has brought on many new developments in terms of technologies Hawkeye review which shows and animation of the ball where it has pitched and if it going to hit the stump or not and brought on Umpire Decision Review system known as “DRS”. The implementation of such technologies has given opportunities for newer technologies to take shape that would result in being beneficial for the viewers for i.e. instant reviews of run out from different angles, novel statistics and very recently introduced match predictions. The match predictions are using historical data and different analytics that would show the viewers who in the end has a higher likelihood of winning matches. Cricket was such a sport that always used analysis since the start of live broadcasting came in however with artificial intelligence and especially machine learning growing year by year it has been possible to make live predictions. [10]

On Twitter fans avidly follow match predictions and has become a trending topic and even now when cricket fans meet the most common question asked, “who do you think is going to win today’s match?” In this year’s Cricket World cup 2019 there was a Cricket expert for Sky Sports Broadcasting who was also previously a popular cricket named Brendon McCullum and he had predicted that Bangladesh would be last in the points table, however Bangladesh ended up beating South Africa and coming 6th in the points table, additionally many other experts predicted South Africa coming in to the semi-final but they were one of the first few teams to be eliminated from the tournament and so this made people question the expert panel and their prediction and the lack of evidence for it. Generally, fans who watch sports analysis etc. want to know accurate prediction just for fun. [10]

2.5 Previous Work on Cricket analysis

When searching on google there were plenty of research papers based on the topic of Predicting a cricket match using Machine Learning technique and each paper used a different technique than the other and this is what gave me a better confidence on the implementation of my project. However, there were only a few organizations that was indirectly predicting cricket data using historical data sets.

CricAlgorithmics

CricAlgorithmics is data analysis tool specifically made for cricket (as the name suggests) that can help and predict what the outcome of the next ball is. Additionally, it gives the overall graphic data of where the wickets fell and which over the runs came the most. Below is the graphic prediction of win percentage of Pakistan over England in the ICC cricket world cup.

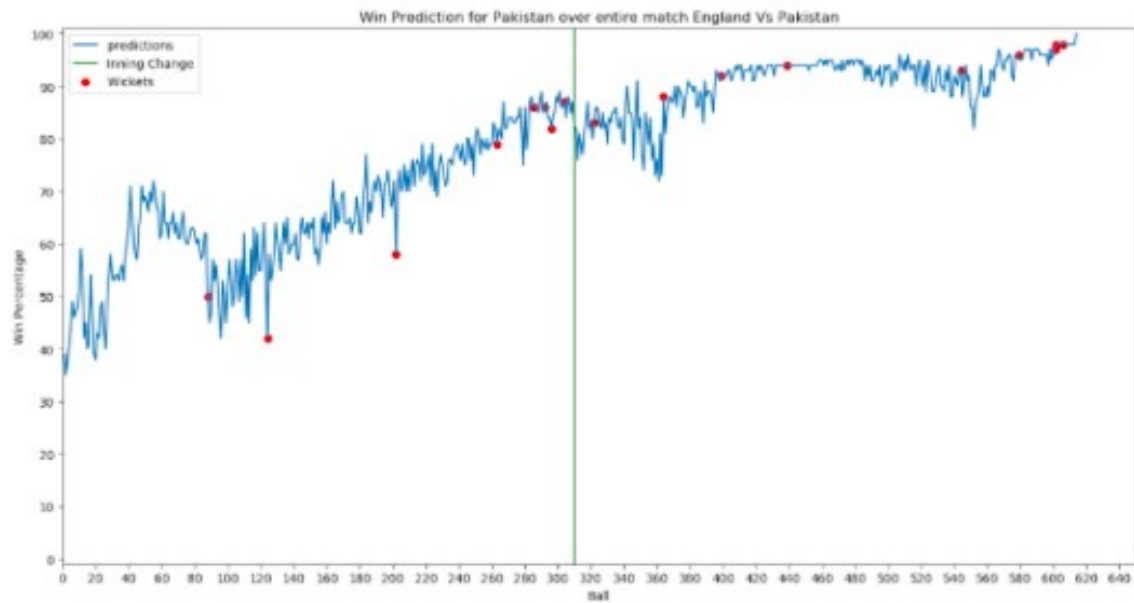


Figure1 image of Pakistan vs England Win predictor- [10]

Similarly, below is the graph prediction for India vs New Zealand semi-final that clearly shows that it is predicted, India will lose the match.[10]

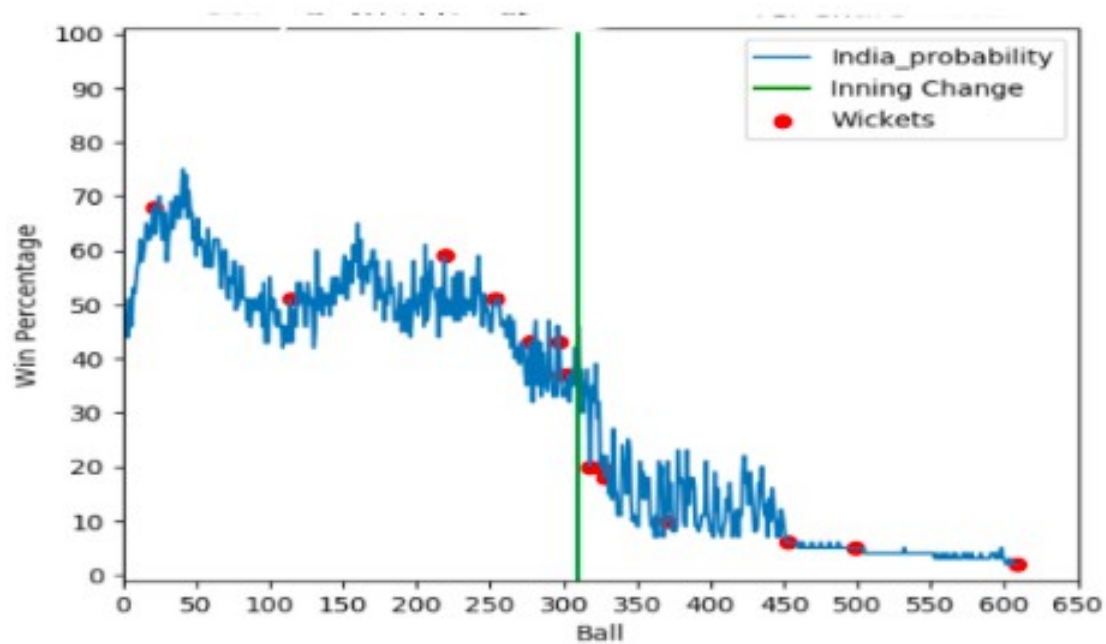


Figure 2 image of India vs NewZealand Win predictor- DigitalSport[10]

TvConal

Another organization that does something similar but only works for the viewers of Singapore and its name is TvConal. TvConals generally specialise over sports analytics, their portfolio includes Football, Boxing and Cricket. They are still expanding their work to achieve the most accuracy in terms of predictions as prediction is all about outcome accuracy. For the Singapore viewers TvConal displayed their heart beat and the predicted win percentage for every match additionally TvConal also analyses the prediction of each player's performance based on the situation. [10]

Dream11

A company that was based out of TvConals idea of predicting player performance is Dream11. This organization is based out in India and it is like a Fantasy cricket team where a player is trying to predict and create the best 11 cricketers of each match and if the chosen cricketers perform well in real life then the players will get points. The fantasy team strategy relies entirely on current data sets but not historical records. It doesn't remember past performances. The way Dream 11 recommends predicting the player is through the player that has gained the most points in the database of Dream 11 it has nothing to do with past year figures. [10]

2.6 Machine Learning with cricket

One of the research papers that was thoroughly interesting and research upon was written by Kalanka Jayanath [1]. The purpose of this paper is just to analyse the ODI cricket predictors

such as coin toss, venue Day/ night game. This research paper discusses the use of 'CART' classification and regression trees through which it can predict the outcome of the ODI game and the factors that matter the most when it comes to prediction. Initially he started off by comparing the CART model to other machine learning techniques such as Random Forest Vector, Naïve bayes, and Logistic Regression approach. Throughout the study there were a lot of interesting analysis such as how Home Field advantage does exist statistically, in that it shows that the team that plays a home venue match is likely to perform better than away. Additionally, in the study he also proves that if one has won the toss is more likely to win the match as well. These two predictors go for most of the teams except for England and West Indies. For England, the study found out that even if England loses the toss it doesn't make a difference for them they will likely win the match if played at home. However, for West Indies this was something surprising it is said that even if west indies play at home and win the toss it is not a huge factor for them. So, the study concluded that if west indies play at home then it's not a huge win factor. Then the study went on to suggest that every One Day International World Cup should be played in the West Indies as over there all the countries are going to be neutral no home advantage.

Another interesting study that came to my attention and was meticulously research upon is written in 2016 and was submitted in a European Conference on Machine Learning and Practice of Knowledge in Discovery Databases written by Madhan Jhanwar.[2]. The idea is very close to what is being implemented now on many sports analysis channels. The study is about how to predict a cricket match using the Team composition approach. The idea of the study is to make the best possible 11 of each and then compare the strength of Team A vs Team B whichever team has a greater Team Strength will win the match. When looking for the composition of 11 players in a team all the career statistics and recent score of each player is being downloaded as data set. If the player is a batsman, average runs in ODI, number of centuries, number of half centuries, strike rate, recent scores, no. of innings played, no. of wickets taken and economy. To train the data, set a machine learning technique by the name

of K-nearest neighbour(K-NN) was used. The accuracy of K-NN was recorded to be 71%.

[2]. Then for each country the accuracy was calculated and the lowest percentage of accuracy was to be for south Africa which was close to 50% this is also since South Africa is the second most team that has changed a team. It has changed 3 players every match on average. This why an inconsistent accuracy might be getting obtained. Overall this study was also a good way of trying to predict a One Day International match however I thought that it was letting out many factors such as venue and coin toss from its data set to improve its training. Further, there was a Research Paper that was intriguing and a little different from the rest was about Predicting the outcome of a Twenty20 match. This paper uses a new library and algorithm to predict the outcome of the cricket match it is known as Adaboost. In this paper, the researchers first classify all the players and statistics and then also have the whole team data set but do not take in to account the venue, Day/night match or toss. Again, this is more of player based strategy where players matter more than external factors. In the end the accuracy of Adaboost turned out to be 62%. [3]. Additionally, the study also concentrates on other statistics such as it demonstrates an ROC curve graph which shows the threshold of the adaboost system. Furthermore, the study also explores on the batsman and bowling ranking system. The Research paper implemented it's on ranking system and evaluated it against the ICC player ranking system. At the end when the bowlers and batter's ratings were generated, they're being used to calculate the match outcome. The only disadvantage found in this study was the limited data set used to train data from 2010-2014 was used in a research that started in 2016. If the data was up to date maybe a better accuracy rate would've been met.

Another interesting study was of predicting a players' performance in one day international cricket using machine learning. This was a conference paper published by Niravkumar Pandey and Kalpdram Passi.[4] The study goes in to detail about how 3 different machine learning algorithms can be used to predict any player's performance be it a bowler, batsman

or a wicketkeeper. The machine learning algorithms used are Naïve Bayes, Decision Trees, Support Vector Machine and Random Forest. Additionally, there were different attributes given to different bowlers, batsmen and keepers. For batting it was based on No.Of Innings played, No. of Centuries, Batting Average, Strike Rate, Zeros, and No. Of Half Centuries. For Bowling the attributes were based on No. Of Innings, Overs bowled, Bowling Average, Bowling Strike Rate and Four/Five Wicket Haul. If the player was a keeper it would be the same attributes as a batsman but additionally adding No. Of catches and No. Of stampings. Furthermore, all the players be it a bowler or a batsman were also evaluated in terms of their previous performances in the venue and against the opposition. Additionally, then each player was then given a rating according to his attribute the higher the number the higher the rating. In the end after conducting all the analysis of all the 4 machine learning algorithms the paper concluded that Random Forest was the best machine learning algorithm to predict the player's performance. The advantage of conducting such a research is that it can also help the players to train differently against other oppositions and if the player is the right choice to pick in the playing 11. If the player has performed poorly against a team in the same venue then statistically he should not be picked as the team might have a higher chance of losing the match. So overall a thought-provoking study. [4]

An interesting PHD research that I came across is written by Ananda Bandulasiri. It is a research paper that talks about predicting the outcome of an ODI(one day international) match using machine learning. An interesting factor about the paper is that it goes in detail about justifying the fact that why Duckworth Louis, that is the current system being used by the ICC when the match is stopped during play, is a faulty system and should be changed. Duckworth louis scores are calculated with the use of run rate per over and the wickets of each team remaining. In this paper, Ananda uses Logistic regression to predict the outcome of the match by looking at the historical data for each team. For e.g. if Australia is playing Sri

Lanka in Sri Lanka's home ground in a day and night match and Sri Lanka has won the toss there is more statistical evidence that when there is a day and night match taking place in Sri Lanka and Sri Lanka has won the toss that Sri Lanka would win the match. Furthermore, there is greater evaluation of this data with the use ROC curve that shows that if a Logistic Regression is used as a predictor it has a higher accuracy than an DuckWorth Louis and Run Rate system being currently used in cricket.

2.7 Summary

Overall when conducting the research on my topic I have found many resources that could help me out in doing my dissertation and some professionals have already done their research in to this and there is also a huge possibility that they might already be developing this product so that it could be used for better sport analysis or any other activities.

3. Research Methodology and Requirements Analysis

3.0 Objective

The objective of this chapter is to cover the requirement specifications of the software system that will be created. Additionally it will also cover the topic of quantitative and qualitative research for this project.

3.1 Quantitative Data

Quantitative data is a method data collection to collect only numerical data. Throughout this project I carried out a lot of Qualitative data collection through statistics modelling and analysis. When analysing the data, the data helped in

figuring out patterns and eventually help in coming to conclusions. Additionally, Quantitative data used in helping find future work if needed.

Qualitative Data

Qualitative data is a type of method collection that does not involve any numbers when collecting data. It is more of collection information through interviews, observations and documents. In this project, I plan to interview a couple of coders who have done a similar software in football so that I can ask them the algorithms that they have used to create their software and interview betting enthusiasts who regularly place betting in matches.

Additionally, I would be conducting a deeper research on the different types of machine learning algorithms that can be useful for prediction. This is because I have never studied machine learning before and a deeper research work would build a good foundation for me and I can better understand what each algorithm does and how.

3.2 Research Questions

- Can machine learning be used to predict an outcome of the whole match?
- Which machine learning technique is the most accurate to predict the cricket match?
- What attributes can increase the accuracy of the prediction?
- What are the datasets that are needed to predict a cricket match?

Functionality	Priority	Achieved Yes/No
The system should be able to read the dataset	Must have	Yes with the use of Pandas I was able to achieve this functionality to read the dataset
The system should be able to predict which team will win the match.	Must have	Yes after using the random forest classifier and logistic regression I was able to achieve this requirement.
The system should be able to show the accuracy of each algorithm	Must Have	Yes after implementing the algorithm I had trained and tested the algorithm using the data . After training and testing I tested the accuracy and it presented it.
The system should be able to edit the dataset i.e. deleting the dataset if not needed.	Could have	Yes, I was able to replace all the null data that was there to NA and in some casing removing it from out of the picture.
The system should be able to work on multiple devices.	Should have	No, unfortunately Jupyter Notebook only works with Macbook OS and Windows OS
The system should manually select the type of algorithm needed to perform the prediction.	Must have	Partially achieved as at the end only 1 algorithm can only be used in my project.
The system should be able to present various types of graphs for quantitative data collection.	Must Have	Yes I was able to present bar graphs and other tables to present my data work and additionally use numpy to calculate, mean and Standard Deviation.
The system should be able to save the data predicted	Could Have	Yes, the advantage of working with Jupyter Notebook was that it was automatically saving data on the cloud.

3.3 Non-Functional requirements

NFR	Priority	Achieved Yes/No
The system should be able to perform at least 2 predictions concurrently	Could have	Yes, it depends on the line of code that you implement if you chose to predict two matches at the same time it would work.
Should be able to present any error with the data.	Must have	Yes. If there are any errors with the data the algorithm would not work and may gave inaccurate calculations
The machine should have responsive result in prediction.	Must have	Yes, it predicts matches very swiftly.
The software should not lose any data during the usage	Must Have	Yes, as already mentioned when working with Jupyter Notebook no data was lost due to being on cloud.
The capacity of the software should not exceed 250 mb	Must have	Yes, the capacity of the software is around 60mb.

4.Data Set

4.0 Objective

The Objective is to discuss the dataset that is being chosen and what is the origin of the dataset and why that particular dataset is being chosen.

4.1 Data Origin

We received a data collection of the 'Kaggle Indian Premier League Server' database from Kaggle Data Science. This database has been made available to the public and combine data in a functional database from three separate sources:

- Match scores, lineups, events: <http://cricinfo.com/>, <http://cricbuzz.com/>

It includes the following data

1. Id
2. Season
3. City
4. Date
5. Team 1
6. Team 2
7. Toss_winner
8. Toss_decision
9. Result
10. DL_applied
11. Win_by_runs
12. Win_by_wickets
13. Player of Match
14. Venue

15. Umpire 1

16. Umpire 2

17. Umpire 3

The dataset that is being used consists of 636 IPL matches that were played from the year 2008 to 2017. It has about data of 14 teams in the IPL. Just to make it clear the IPL is a Indian Premier League that is played in India in t-20 format.

4.2 Data Pre Processing

An significant step in our model is to analyze and pre-process the data in a format that we can use to test and train different models.

To do this, three pre-processing steps were taken: Part of the information we needed was initially CSV in the database and included all corresponding events including city, venue, DI score etc. We developed a python script for extracting and reading this data in new tables connected to the table 'Matches' using a foreign key mapping to a match identifier. Below is an excerpt from the Python for the dataset:


```

In [2]: import numpy as np
import pandas as pd

In [3]: matches = pd.read_csv(r'C:\Users\dladclnt\Downloads\Cricket-Prediction-master\matches.csv')

In [4]: matches.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 636 entries, 0 to 635
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   id                     636 non-null   int64  
1   season                 636 non-null   int64  
2   city                   629 non-null   object  
3   date                   636 non-null   object  
4   team1                   636 non-null   object  
5   team2                   636 non-null   object  
6   toss_winner            636 non-null   object  
7   toss_decision          636 non-null   object  
8   result                 636 non-null   object  
9   dl_applied             636 non-null   int64  
10  winner                 633 non-null   object  
11  win_by_runs            636 non-null   int64  
12  win_by_wickets         636 non-null   int64  
13  player_of_match        633 non-null   object  
14  venue                   636 non-null   object  
15  umpire1                 635 non-null   object  
16  umpire2                 635 non-null   object  
17  umpire3                 0 non-null     float64
dtypes: float64(1), int64(5), object(12)

```

Some data elements have been set to NULL, which means that certain unusable rows have been removed and we have imputed the full amount of data in other instances. For eg, in some games, the value of the possession was missing, so in this case we entered a shorter dataset. There were many instances where the city that was mentioned in the dataset was missing. Hence we first read the dataset and then filled NA that means not available to it.

```

In [5]: matches[pd.isnull(matches['winner'])]
matches['winner'].fillna('Draw', inplace=True)

```

Above is the excerpt from the python to find out the data that was null.

5.0 Design

5.0 Objective

Throughout this section, we present the general design and choices of my model.

My model consists of a mixture of several regression algorithms and classifications, which produce various metrics that are ultimately used in our classification model (for matching results) and regression model.

5.1 Model Components

Toss Winner – In my model I have calculated the amount times a team has won the Toss and then tried to see that if the team that has won the toss has it won the match as well.

Toss Decision – If the team has won the toss what have they decided to do. There can be only 2 decisions, they can either bat first or ball first.

Classification Model- This model will use the data of Toss Winner, Toss Decision, Venue and city and the two teams taking part in the cricket match. With the use of all these values we shall be able to predict the winner of a cricket match.

Regression Model – This model will use the data of Win by Runs, Win by Wickets and Man of the match. The results will then be used to generate the match outcome.

5.2 Program design for implementation

There are certain programming languages that can be used for implementing machine learning and data analysis of this research project and they 'R' and Python. In Python within itself there are 2 options using an IDE. A software framework that offers extensive facilities for computer programmers for software creation is an integrated creation environment (IDE). An IDE typically includes a source code editor, create automation software and debugger at least. Many IDEs, like NetBeans and Eclipse, do not provide the compiler, interpreter or both necessary;

others, like Lazarus and SharpDevelop, don't. Additionally Pycharm is an IDE that is mainly used for developing Python projects.

Jupyter Notebook is a computer collaborative web-based platform that generates Jupyter notebook documents (formerly IPython Notebooks). The word "notebook" can convey a lot of different entities together, particularly Jupyter's web application, Jupyter's Python web server, and Jupyter's context-dependent document format. A Jupyter Notebook document is a JSON document that uses a versioned scheme and contains a list of input / output cells, that can contain code, text, mathematics, plots and rich media, which typically end with ".ipynb" extension. This document is an input / output file.

6. Implementation

6.0 Deciding the Program and Software to use

When the work on this project had started, in the initial I had a personal inclination of making the project use 'R' the language due to its data analysis abilities and strong mathematical use, additionally I had already implemented a coursework that was done using 'R' in the university for Data Modelling and Analysis Coursework. So I was already aware of the regression model implementation in 'R' however I wasn't too keen on implementing the machine learning part in 'R' due to limited resources availability for both primary and secondary resources.

This is the reason why I went ahead and chose Python as program to implement my research work and another benefit with Python is its efficient coding requirements and also data visualization factor which was available in R would available in Python as well.

Even in Python I went ahead and chose the Jupyter Notebook as the final piece of code that could help me in my research work. This is because that Jupyter Notebook has an online Python Web server capability and can be accessed anywhere with the same account so there is less loss in data and it also functions just like another regular Python IDE i.e. Pycharm.

Additionally there were a couple of Libraries used to do this project

Pandas – This library was used for data analysis in order to read data and create a few data visualized graphs.

Numpy – Numpy helped in calculating the average, mean and standard deviation of the data for my research.

Scikit Learn- This was one of the most important tools that can be used for machine learning as this library has most of the machine learning algorithms such as logistic regression, Random forest and K- nearest number.

6.1 Testing

We will discuss various test methods and methods in this section, which we have used as a basis for optimizing our model.

6.1.1 Cross Validation

Cross validation is a way to prevent overflowing where a model suits very well but can not be used as a data that has not been used before cross validation, so that a distinction takes place between the training data set and a test data set. Cross-validation is a way to avoid overfitting.

The training set is used for training the model, while the test set measures the predictive efficiency of the model. Cross validation runs the training and test process several times with a different part of the data set used as test data for each iteration to ensure a large number of samples for which to train the model.

The downside of using cross-validation to train and evaluate a new model is that the number of cross-validation iterations increases the length of the testing. It should, however, be sacrificed in general to speed model training to achieve a more powerful, less overfitting final model.

The Testing Metric that was used for this project was accuracy.

$$\frac{n_{true}}{n_{total}} Acc =$$

where:

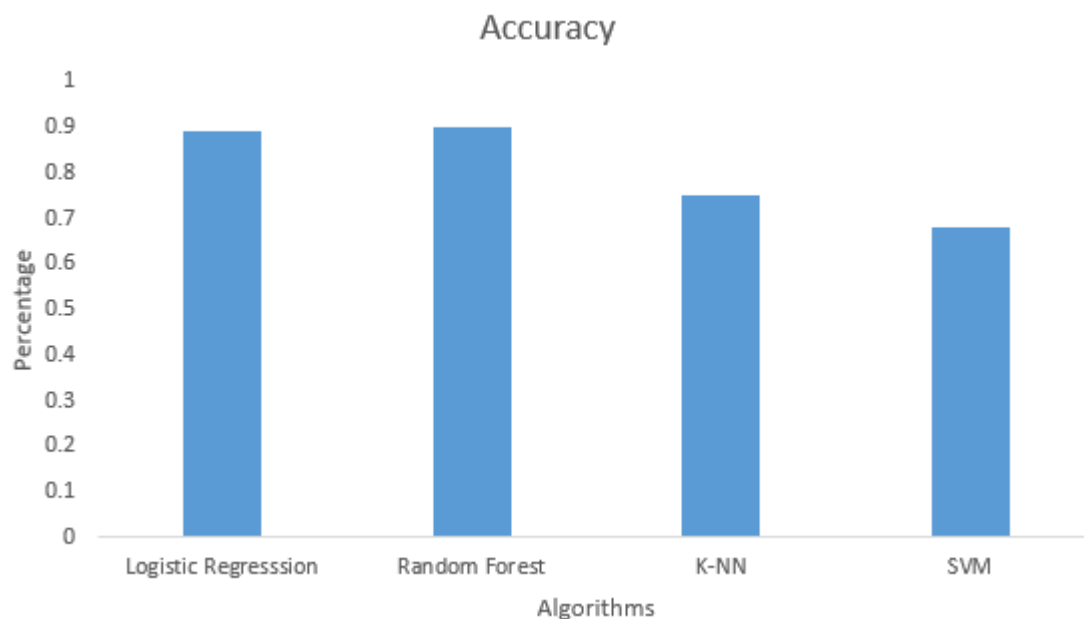
n_{true} is the number of examples that the classifier has correctly predicted

n_{total} is the total number of examples

Accuracy is a simple measure that helps us to understand what outcome our classification model has correctly predicted by looking at what proportion of examples. The accuracy varies

from 0 to 1, and increases efficiency for greater accuracy. But the accuracy metric lacks some critical data to measure the efficiency of our classifier.

6.2 Model Implemented



In the machine learning research I tried to implement 4 different machine learning algorithms separately at first . When training the algorithms and testing it with the data I had chosen, I found out particularly that Random Forest and Logistic Regression were the highest accuracy rate, which is quite obvious in the above diagram. In the end I went on to implement one algorithm , one Random Forest as a classifier. This would further produce a much accurate prediction for the cricket matches.

In [98]: matches.describe()

Out[98]:

	id	season	team1	team2	toss_winner	dl_applied	winner	win_by_runs	win_by_wickets	umpire3
count	636.000000	636.000000	636.000000	636.000000	636.000000	636.000000	636.000000	636.000000	636.000000	0.0
mean	318.500000	2012.490566	5.540881	5.511006	5.371069	0.025157	5.309748	13.682390	3.372642	NaN
std	183.741666	2.773026	3.329169	3.341677	3.293140	0.156726	3.288726	23.908877	3.420338	NaN
min	1.000000	2008.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	NaN
25%	159.750000	2010.000000	3.000000	3.000000	2.000000	0.000000	2.000000	0.000000	0.000000	NaN
50%	318.500000	2012.000000	5.000000	5.000000	5.000000	0.000000	5.000000	0.000000	4.000000	NaN
75%	477.250000	2015.000000	9.000000	8.000000	7.250000	0.000000	8.000000	20.000000	7.000000	NaN
max	636.000000	2017.000000	13.000000	13.000000	13.000000	1.000000	14.000000	146.000000	10.000000	NaN

Above is a table I have posted that describe details of my data. This is done using the numpy library such as counting the number of data I have for each column, additionally for each

column I have also shown the mean, standard deviation, quartile ranges, minimum value of the column, and also the maximum value of the column. This was done in order to visualize the data and start training and testing the model.

```
encode = {'team1': {'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13},
          'team2': {'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13},
          'toss_winner': {'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13},
          'winner': {'MI':1,'KKR':2,'RCB':3,'DC':4,'CSK':5,'RR':6,'DD':7,'GL':8,'KXIP':9,'SRH':10,'RPS':11,'KTK':12,'PW':13,'Draw':14}}
```

```
dictationValue = encode['winner']
```

```
dictationValue
```

```
{'MI': 1,
 'KKR': 2,
 'RCB': 3,
 'DC': 4,
 'CSK': 5,
 'RR': 6,
 'DD': 7,
 'GL': 8,
 'KXIP': 9,
 'SRH': 10,
 'RPS': 11,
 'KTK': 12,
 'PW': 13,
 'Draw': 14}
```

In the above screenshots I have done two tasks first in encode I have given each team a number that I would then refer to instead of the team name. In the second part of the code I have declared the dictationValue will be encode that uses the winner line in the above encode part. Hence the result that I got was the above dictationValue.

```
print(list(dictationValue.keys())[list(dictationValue.values()).index(1)])
```

```
MI
```

Then I make an amendment in the dataset on python where I write the command to print the

dictationValue where the index is 1 which means the first team and then it prints out MI.

```
matches = matches[['team1','team2','city','toss_decision','toss_winner','venue','winner']]
```

```
df = pd.DataFrame(matches)
df.describe()
```

	team1	team2	toss_winner	winner
count	636.000000	636.000000	636.000000	636.000000
mean	5.540881	5.511008	5.371069	5.308748
std	3.329169	3.341677	3.293140	3.288726
min	1.000000	1.000000	1.000000	1.000000
25%	3.000000	3.000000	2.000000	2.000000
50%	5.000000	5.000000	5.000000	5.000000
75%	9.000000	8.000000	7.250000	8.000000
max	13.000000	13.000000	13.000000	14.000000

df

	team1	team2	city	toss_decision	toss_winner	venue	winner
0	10	3	Hyderabad	field	3	Rajiv Gandhi International Stadium, Uppal	10
1	1	11	Pune	field	11	Maharashtra Cricket Association Stadium	11
2	8	2	Rajkot	field	2	Saurashtra Cricket Association Stadium	2
3	11	9	Indore	field	9	Holkar Cricket Stadium	9
4	3	7	Bangalore	bat	3	M Chinnaswamy Stadium	3
...
631	7	3	Raipur	field	3	Shaheed Veer Narayan Singh International Stadium	3
632	8	3	Bangalore	field	3	M Chinnaswamy Stadium	3
633	10	2	Delhi	field	2	Feroz Shah Kotla	10
634	8	10	Delhi	field	10	Feroz Shah Kotla	10
635	10	3	Bangalore	bat	10	M Chinnaswamy Stadium	10

636 rows × 7 columns

Then I further make amendment to the dataset where my first command is to only use the 7 columns of team 1, team 2 , city, toss_decision, toss_winner, venue and winner. Hence instead of 17 columns we use only 7 related columns. So this procedure is called data mining.

```
perm1 = df['toss_winner'].value_counts(sort=True)
```

```
perm1
```

```
1      85
2      78
7      72
3      70
9      68
5      66
6      63
4      43
10     35
13     20
8      15
11     13
12      8
Name: toss_winner, dtype: int64
```

Then in order to further train my model I then try to create an array where I can only see the list of toss winners. On the left hand side the numbers that you see are the numbers I used to define in encoding and dictationValue and on the right hand side the number that you see are the amount of tosses that have been won by the team.

In order to further simplify this I have also made it easier in my model so that you may be able to understand as well. As can be seen belowi simplified the dictationValues to be indexed in to the team names.

```
#No of toss winners by each team
for idx, val in perm1.iteritems():
    print('{} -> {}'.format(list(dictationValue.keys())[list(dictationValue.values()).index(idx)],val))
```

```
MI -> 85
KKR -> 78
DD -> 72
RCB -> 70
KXIP -> 68
CSK -> 66
RR -> 63
DC -> 43
SRH -> 35
PW -> 20
GL -> 15
RPS -> 13
KTK -> 8
```

Now below I will be posting the second correlation to my machine learning and predicting

model that is the match winners.

```
perm2=df['winner'].value_counts(sort=True)
```

```
#No of match winners by each team
for idx, val in perm2.iteritems():
    print('{} -> {}'.format(list(dictationValue.keys())[list(dictationValue.values()).index(idx)],val))
```

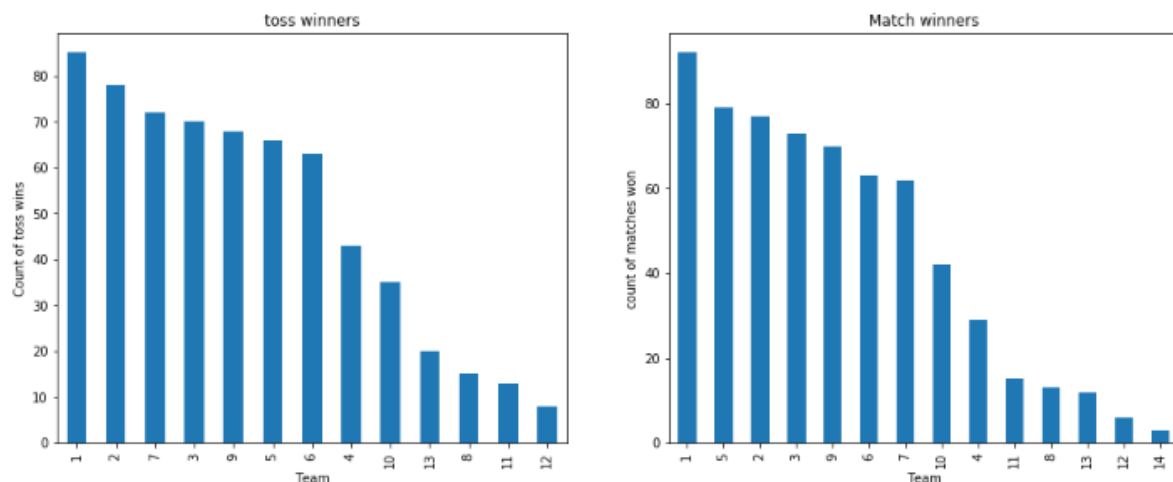
```
MI -> 92
CSK -> 79
KKR -> 77
RCB -> 73
KXIP -> 70
RR -> 63
DD -> 62
SRH -> 42
DC -> 29
RPS -> 15
GL -> 13
PW -> 12
KTK -> 6
Draw -> 3
```

Further I then evaluate using the same formula I used for perm1 and get how many matches each team has one and displayed it according to descending order.

```
import matplotlib.pyplot as plt
fig = plt.figure(figsize=(16,6))
ax1 = fig.add_subplot(121)
ax1.set_xlabel('Team')
ax1.set_ylabel('Count of toss wins')
ax1.set_title("toss winners")
perm1.plot(kind='bar')

ax2 = fig.add_subplot(122)
perm2.plot(kind = 'bar')
ax2.set_xlabel('Team')
ax2.set_ylabel('count of matches won')
ax2.set_title("Match winners")
```

Text(0.5, 1.0, 'Match winners')



Then as you can see the outcome when comparing both the bar graphs there is a clear correlation between toss winners and match winners, hence we can come to the conclusion that if a team has won the toss they are more likely to win the match and so this set of data can

further help us in improving our predictor.

df							
	team1	team2	city	toss_decision	toss_winner	venue	winner
0	10	3	14	1	3	23	10
1	1	11	25	1	11	16	11
2	8	2	27	1	2	25	2
3	11	9	15	1	9	11	9
4	3	7	2	0	3	14	3
...
631	7	3	26	1	3	27	3
632	8	3	2	1	3	14	3
633	10	2	9	1	2	8	10
634	8	10	9	1	10	8	10
635	10	3	2	0	10	14	10

636 rows x 8 columns

In order to train and test the machine learning algorithm I need to change the other data to values as well so because the scikit only recognizes values and we need scikit to implement the Random forest algorithm.

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import metrics
from sklearn.model_selection import train_test_split
def classification_model(model, data, predictors, outcome):
    model.fit(data[predictors], data[outcome])
    predictions = model.predict(data[predictors])
    accuracy = metrics.accuracy_score(predictions, data[outcome])
    print("Training Model Accuracy", accuracy*100)
    training_x, testing_x, training_y, testing_y = train_test_split(data[predictors], data[outcome], test_size = 0.25, random_state=0)
    '''print(training_x)
    print(testing_x)
    print(training_y)
    print(testing_y)'''
    model.fit(testing_x, testing_y)
    predictions_ = model.predict(testing_x)
    accuracy_ = metrics.accuracy_score(predictions_, testing_y)
    print("Testing Model Accuracy", accuracy_*100)
```

In the above code I have implemented the Random Forest Classifier model and trained to the prediction outcome, the data being used in the model is the modified data with all the values.

```

model = RandomForestClassifier(n_estimators=100)
outcome_var = ['winner']
predictor_var = ['team1', 'team2', 'venue', 'toss_winner', 'city', 'toss_decision']
classification_model(model, df, predictor_var, outcome_var)
df.head(7)

```

C:\Users\dladclnt\anaconda3\lib\site-packages\ipykernel_launcher.py:8: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

Training Model Accuracy 89.15094339622641

C:\Users\dladclnt\anaconda3\lib\site-packages\ipykernel_launcher.py:19: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

Testing Model Accuracy 96.22641509433963

Then after implementing the RandomForest Classifier model we came with a very high

Training Model Accuracy of 89.2 and Testing Model Accuracy of 96.2

After getting my accuracy done now it is the time of the predictor to take place and how is that going to work? Please see the below screenshot

```

team1='DD'
team2='RCB'
toss_winner='DD'
input=[dictationValue[team1],dictationValue[team2],'14',dictationValue[toss_winner],'2','1']
input = np.array(input).reshape((1, -1))
output=model.predict(input)
print(output)
print("Predicted winner : "+list(dictationValue.keys())[list(dictationValue.values()).index(output)])

```

[7]
Predicted winner : DD

In the above picture I have declared 2 teams facing each other and also declared that the toss was won by a team in this case DD. Now the input model will the DictationValue but the outcome will be the above model predictor which will use past data and see the teams full history and bring out the outcome with the dictated value and the team name. In this case team number 7 that is DD which can be crosschecked on the above screenshot I posted for dictationValue. Hence I would say that model has succeeded.

7. Conclusion

- Can machine learning be used to predict an outcome of the whole match?

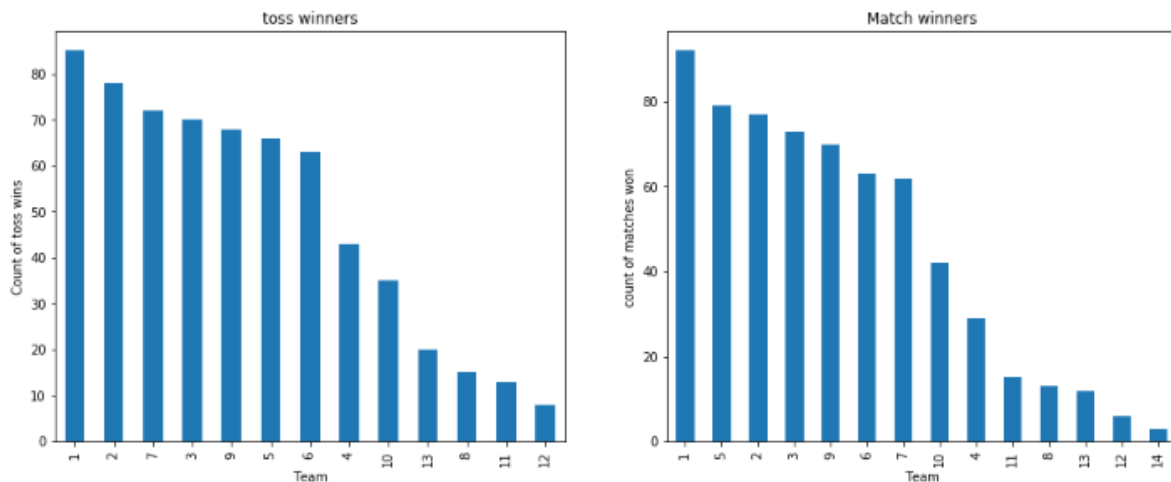
To Answer my first research question after working through most of the machine learning models and predicting the outcome of couple of t20 IPL matches it can be said that yes machine learning can be used to predict the outcome of the whole cricket with accuracy being close to 90% and also depended on the dataset that the machine learning model is provided.

- Which machine learning technique is the most accurate to predict the cricket match?

To answer this question I would say that it may depend from data to data. The data that I used for the machine learning models it provided me with the fact that Random Forest is the best classifier to work as it gives a Trained accuracy of 89.2% and test accuracy of 96%. But then again it would depend on data to data.

- What attributes can increase the accuracy of the prediction?

The attributes that I personally found out that would improve the accuracy of the prediction would be the team that has won the toss, additionally what would also be really important is the amount of time the team has won matches.



As can be seen above there is correlation between the team that has won the toss and the teams that have won matches. The machine learning model should also be smart enough to catch the past history of matches played between the teams.

- What are the datasets that are needed to predict a cricket match?

The datasets that are really necessary to create a perfect prediction is the teams, the city where the match is played, which team won the toss, how many runs were won by, how many wickets was the match won by and who won the match.

7.1 Survey Results

There was a survey that I had done after I finished my product. After conducting the survey I found out that not many users are using poll predictors to predict matches which was definitely not my opinion as I definitely always use poll predictors to predict matches. However the feedback I received was positive because even though many users don't use poll predictors they gave the opinion that they would still use my machine learning model to predict matches and one user also said that it would depend on the accuracy and model I'm using to predict. This opinion is quite motivating for me and I would like to take this project forward with me and work on it more.

8.0 Evaluation

8.1 Main Achievements

Throughout this project there were many achievements that were gained when started to bring the code in to the jupyter notebook. The very first thing that I achieved by doing the dissertation is do python. Before doing this dissertation project I had very limited knowledge about python and how I can implement it in a very large scale. Then I took the initiative of learning python in 2 weeks. There were classes that I would attend every day at night as the classes were based in the United States and they were online. Then second part is learning machine learning alongside python. I never knew that the human mind had such power to learn a lot of things side by side and adapt very quickly when needed to adapt. But I should also thank the internet as without the internet I would be nothing, the internet has enabled me to learn more swiftly and use up resources correctly and effectively. Moreover completing this project is the biggest achievement I ever had as I was working side by side due to financial circumstances I had to start working and do this project simultaneously. The amount of sleepless nights and hardwork I had to put is unimaginable and it has made me achieve time management skills and maturity in an assignment.

8.2 Limitations

The limitations that I was facing while doing this project were many. As I was working in a full time there were many aspects that I could not complete in my project. I wanted to convert my project in to a working python application but due to a lot of errors in the end I had to give up and time was an issue. Additionally there were many errors when I was implementing different machine learning algorithms which was very time consuming. Another huge limitation was the dataset. I was intending on getting the dataset for ICC cricket world cup 2019 but the official dataset was not available and what ever was available could not be relied on. So I had to shift to a dataset that was official and reliable and so I chose the cricket dataset of IPL as it was reliable and found easily as well. Further due to the covid-19 my communications with the professor reduced eventhough we did have an option of using online meetings, I was in a habit of visiting the professor personally during his office hours as he was generally free and talk to him on updates however due to the virus even that reduced.

8.3 Future Work

For the future there is a lot to talk about from my end. Honestly for future work I would like to make gui application that is available to everyone where there is a dataset for 3 different format of the cricket game and the users get to choose which format, which team, and what algorithm do they want to implement in order to predict the match. I want to make it publicly available to the children so that I also broaden their knowledge about sports analytics and something that is very interesting for them to play around with and also I would keep it opensource so that any one can edit and customize it according to their preference.

Overall I would like to conclude to say that working on this project was a dream come true where I was implementing two things I love cricket and computers and I couldn't end my college life in any better fashion than this. A huge thank you to Professor Adrian for helping me

out in all the stages of the project and being so patient with me. Additionally a huge thank you to professor lynne baillee for all her feedback from the first deliverable it helped.

9.0 References

1. Jayalath, Kalanka. (2017). A Machine Learning Approach to Analyze ODI Cricket Predictors. *Journal of Sports Analytics*. 10.3233/JSA-17175.
2. Jhawar, Madan & Pudi, Vikram. (2016). Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach.
3. Singhvi, Arjun, et al. "Predicting an Outcome of a t20 Match." *Cs.wisc.edu*, pages.cs.wisc.edu/~shruthir/Documents/MachineLearning_Final_Report.pdf.
4. Passi, Kalpdram & Pandey, Niravkumar. (2018). Increased Prediction Accuracy in the Game of Cricket Using Machine Learning. *International Journal of Data Mining & Knowledge Management Process*. 8. 19-36. 10.5121/ijdkp.2018.8203.
5. Bandulasiri, Ananda. "Predicting the Winner in One Day International Cricket." (2008).
6. Ians, IANS. "Star India Target 700 Million Fans for IPL 2018." *Ipl – Gulf News*, Gulf News, 7 Nov. 2018, gulfnews.com/sport/cricket/ipl/star-india-target-700-million-fans-for-ipl-2018-1.2158986.
7. Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". *IBM Journal of Research and Development*. 3 (3): 210–229. CiteSeerX 10.1.1.368.2254. doi:10.1147/rd.33.0210
8. Russell, Stuart J.; Norvig, Peter (2010). *Artificial Intelligence: A Modern Approach* (Third ed.). Prentice Hall. ISBN 9780136042594.
9. "Machine Learning: What It Is and Why It Matters." *SAS*, www.sas.com/en_ae/insights/analytics/machine-learning.html.
10. Mercer, Masoumeh Izadi. "Exploring Cricket Predictions and Emerging Trends." *Digital Sport*, <https://digitalsport.co/exploring-cricket-predictions-and-emerging-trends>.
11. Terry, David (2000). "The Seventeenth Century Game of Cricket: A Reconstruction of the Game" (PDF). *The Sports Historian*, No. 20. London: The British Society of Sports History.

- pp. 33–43. Archived from the original (PDF) on 27 November 2009. Retrieved 2 May 2016.
12. Ashley-Cooper, F. S. (1900). "At the Sign of the Wicket: Cricket 1742–1751". *Cricket: A Weekly Record of the Game*. Cardiff: ACS. pp. 4–85. Retrieved 8 September 2017.
 13. (18 April 2018). "IPL 2018: New Rules allow Aditya Tare to keep wickets for injured Ishan Kishan during MI vs RCB match". <http://www.timesnownews.com/sports/cricket/ipl/article/ipl-2018-new-rules-allow-aditya-tare-to-keep-wickets-for-injured-ishan-kishan-during-mi-vs-rcb-match/218530> , *Times Now News*. Retrieved 4 May 2018
 14. Mitchell, T. (1997). *Machine Learning*. McGraw Hill. p. 2. ISBN 978-0-07-042807-2.
 15. Harnad, Stevan (2008), "The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence", in Epstein, Robert; Peters, Grace (eds.), *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Kluwer
 16. Schermer, B. W. (2007). *Software agents, surveillance, and the right to privacy: A legislative framework for agent-enabled surveillance* (paperback). **21**. Leiden University Press. pp. 140, 205–244. hdl:1887/11951. ISBN 978-0-596-00712-6. Retrieved 2012-10-30.
 17. Champandard, A. J.: *AI Game Development: Synthetic Creatures with learning and Reactive Behaviors*. New Riders, USA (2003)
 18. *Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning. Cambridge, Mass.: MIT Press. ISBN 978-0-262-03358-9.*
 19. Marcus Callies; Wolfram R. Keller; Astrid Lohöfer (2011). *Bi-directionality in the Cognitive Sciences: Avenues, Challenges, and Limitations*. John Benjamins Publishing. pp.73–. ISBN 90-272-2384-X.
 20. Jones, Stephen K. (2009). *Brunel in South Wales*. III: Links with Leviathans. Stroud: The History Press. p. 56. ISBN 9780752449128.
 21. "7+ Sample Survey Consent Forms." *Sample Survey Consent Form - 6+ Documents in PDF, Word*, www.sampletemplates.com/sample-forms/survey-consent-form-

template.html.

8. Appendices

8.1 Consent Form

(This is a sample consent form taken from agnes university) [21]

Sample Consent Form for Online Surveys

[Note that this is a *sample* consent form for student researchers and should be altered to accurately reflect the *individual study*. Faculty researchers should make the obvious modifications to remove student references.]

You are invited to participate in a web-based online survey on [Machine Learning in Cricket]. This is a research project being conducted by [Moiz Reyaz], a student at Heriot Watt University.

PARTICIPATION

Your participation in this survey is voluntary. You may refuse to take part in the research or exit the survey at any time without penalty. You are free to decline to answer any particular question you do not wish to answer for any reason.

BENEFITS

You will receive no direct benefits from participating in this research study. However, your responses may help us learn more about the need for machine learning predictors in cricket.

RISKS

There are no foreseeable risks involved in participating in this study other than those encountered in day-to-day life.

OR

There is the risk that you may find some of the questions to be sensitive.

OR

There is the risk that some questions may cause emotional discomfort.

OR

The possible risks or discomforts of the study are minimal. You may feel a little [uncomfortable/embarrassed/sad/tired etc] answering [personal/sensitive/many/...] survey questions.

ELECTRONIC CONSENT: Please select your choice below. You may print a copy of this consent form for your records. Clicking on the “Agree” button indicates that

- You have read the above information
- You voluntarily agree to participate
- You are 18 years of age or older

☐ Agree

☐ Disagree

Q1

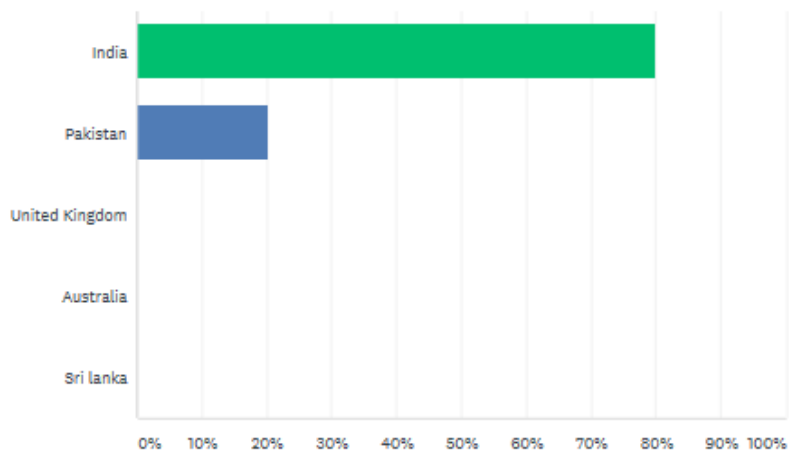


Customize

Save as ▼

Where are you from?

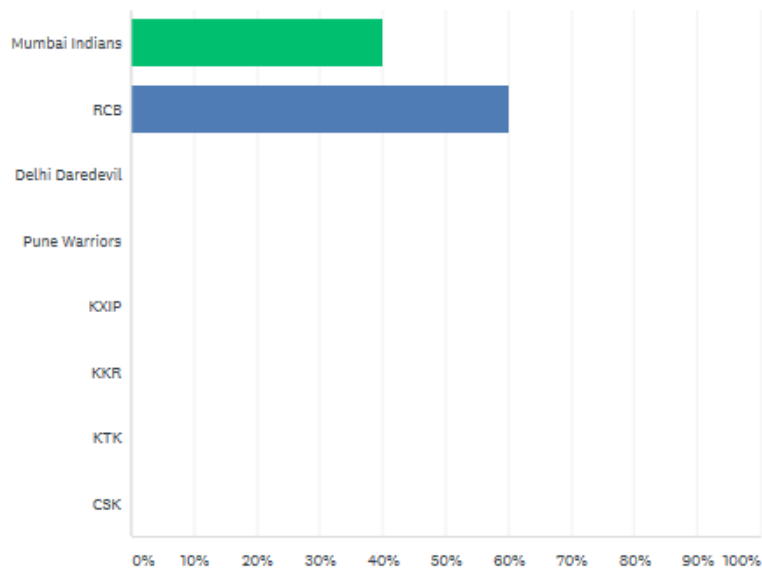
Answered: 5 Skipped: 0



ANSWER CHOICES	RESPONSES	
▼ India	80.00%	4
▼ Pakistan	20.00%	1
▼ United Kingdom	0.00%	0
▼ Australia	0.00%	0
▼ Sri Lanka	0.00%	0
TOTAL		5

What Indian Premier League Cricket team do you support?

Answered: 5 Skipped: 0



ANSWER CHOICES	RESPONSES	
▼ Mumbai Indians	40.00%	2
▼ RCB	60.00%	3
▼ Delhi Daredevil	0.00%	0
▼ Pune Warriors	0.00%	0
▼ KXIP	0.00%	0
▼ KKR	0.00%	0
▼ KTK	0.00%	0
▼ CSK	0.00%	0
TOTAL		5

Why do you support the Cricket Team?

Answered: 5 Skipped: 0

RESPONSES (5) WORD CLOUD TAGS (0)

🔒 Sentiments: OFF ☐



NEW!

Introducing Sentiment Analysis

Detect the feeling and sentiment behind written responses.



[Watch a demo](#)



Try it!

Premier plan only

[UPGRADE](#)



Apply to selected ▾

Filter by tag ▾

Search responses



Showing 5 responses

4/23/2020 9:11 PM

[view respondent's answers](#)

[Add tags ▾](#)



Hometown

4/23/2020 12:03 AM

[View respondent's answers](#)

[Add tags ▾](#)



My favorite players play for the team

4/22/2020 10:39 PM

[View respondent's answers](#)

[Add tags ▾](#)



Because Mumbai has the best team and is closest alternative to gujarat

4/22/2020 8:40 PM

[View respondent's answers](#)

[Add tags ▾](#)




I am from that city

4/22/2020 8:11 PM

[View respondent's answers](#)

[Add tags ▾](#)

Q6

 Save as ▼



Would you accept the decision of the machine learning software and use it to predict the poll?

Answered: 5 Skipped: 0

RESPONSES (5)

WORD CLOUD

TAGS (0)

 Sentiments: OFF 




NEW!

Introducing Sentiment Analysis

Detect the feeling and sentiment behind written responses.

 [Watch a demo](#)

↑ 

Try it!

Premier plan only

[UPGRADE](#)

☐ Apply to selected ▼ Filter by tag ▼

Search responses  

Showing 5 responses

☐ Why not, depends on the methodology and reliability
4/23/2020 9:11 PM

[View respondent's answers](#) [Add tags](#) ▼

☐ Yes
4/23/2020 12:03 AM

[View respondent's answers](#) [Add tags](#) ▼

☐ No, I would rather watch the game and make predictions instinctively
4/22/2020 10:39 PM

[View respondent's answers](#) [Add tags](#) ▼

☐ Yes for sure