Cricket Match Winning Prediction

BY

Maruf Hosen
ID: 171-15-9383
AND

Abdullah Al-Mamun
ID: 171-15-9403

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Dewan Mamun Raza
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

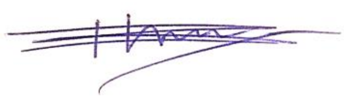
JUNE 3, 2021

APPROVAL

This research based project titled "**Cricket Match Winning Prediction**", submitted by Maruf Hosen, ID No: 171-15-9383 and Abdullah Al-Mamun, ID No: 171-15-9403 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on June 03, 2021.

<u>BOARD OF EXAMINERS</u>

Chairman



____

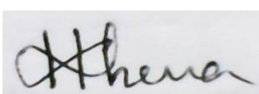
Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



____

Internal Examiner

Most. Hasna Hena

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Sumit Kumar Banshal

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Shah Md. Imran

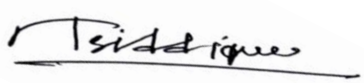
Industry Promotion Expert

LICT Project, ICT Division, Bangladesh

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Shah Md. Tanvir Siddiquee, Assistant Professor, and Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Mr. Shah Md. Tanvir Siddiquee

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Co-Supervised by:

Mr. Dewan Mamun Raza

Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Submitted by:

Maruf Hosen

ID: 171-15-9383

Department of Computer Science and Engineering

Daffodil International University

Abdullah Al-Mamun

ID: 171-15-9403

Department of Computer Science and Engineering

Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to Mr. Shah Md. Tanvir Siddiquee, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" and "*Data Mining*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan**,** Professor and Head**,** Department of Computer Science and Engineering, for his kind help to finish our project and also to other faculty member and the staff of Computer Science and Engineering department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The multi-billion dollar industry is cricket betting. There is also a great incentive for models which can forecast the results of games and overcome bookers' odds. The objective of this thesis was to explore the extent to which the results of cricket matches can be predicted. The English twenty over the county Cricket Cup was the aim competition. About 500 teams and player numbers emerged from the initial features alongside the engineered features. First, the versions with only team features, then all team and player features were optimized. In individual seasons, the result has been tested on the basis of each training during the past season results. The optimum model was a straightforward method of estimation paired with dynamic hierarchical characteristics and a benchmark for the gaming industry was considerably higher. It seems magic to predict the future if a prospective buyer wants to buy the goods in advance or figures out where asset prices are concerned. If we can forecast something's future accurately, we have a huge advantage. This magic and mystery have been only amplified by machine learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

The International Cricket Council (ICC) plays in 106 Member States with about 1 500 million fans worldwide as a sporting cricket (ICC, 2012-2013). But the ten fully-fledged ICC countries and, more particularly, 'the Big Three' of Britain, Australia and India, are a major part of the global finance and concern. Per year the annual demand in cricket gambling, dominated by Asian economies, is estimated to be $10 billion legally and around $40 to $50 billion illicit. Western gaming markets are tightly regulated and controlled, so this is not clear. The Asian betting market could restrict the chance. This is possibly the reason why there have been little work in this field to date. With the size of the betting industry worldwide, there are clear monetary benefits, either from partnering with betting firms, offering forecasts to business gamblers or personal betting, for someone who has access to superior forecasting technologies. In this article, the results of the 20 English Cricket Cup over the years have been predicted by machine learning. In order to interpret data created with various functions, we used a multistage approach. We just dated the first team and then matched the team with the athlete. Sports analytics Machine Learning is a brand new area in computing technology with several challenges. The aim of this research is to design a cricket prediction system for a match, particularly for an IPL match during the match. Various machine learning and mathematical method is adopted to find the best possible result. To allow a comparison of the findings observed, a very common statistical method is used called multiple linear regression.

1.2 Motivation

Cricket is one of the world's most famous sports, seen by most people around the world. Two teams of eleven players each played this game. The prediction of the outcome of a game was identified as a fundamental problem with the development of sport statistical modelling. Cricket is one of the world's best-known team sports. Testing matches, ODIs and T20s are three different formats for the cricket. The T20s, the most common format in the game.

There will be at least 8 to 10 teams during the cricket season and each team participates for a minimum of two occasions for all the other teams. At different venues, matches are held. Toss at the beginning is a key factor in choosing the match winner. You will wish for a field or bat from Toss winning team. The batting team would first attempt to run as many runs as possible to set a goal. In order to win the game in hand, the team that fights second must push the mark. For years, we have seen predicted scores on our television screens at various intervals when playing small overlapping cricket. We are trying to predict the outcome of a cricket match based on different machine learning approaches. Machine learning applications augmented by data mining techniques have been a hot subject worldwide for study, however. With a net fan base of approximately 2.5 billion, cricket is one of the most common sports in Australia, Caribbean, UK and South Asia. The game has enormous support from spectators in over 100 countries, and the people are very keen to predict the game results. The result of a cricket match is determined by several pre-game and in-game attributes. The results of the match are dominated by pre-games attributes such as venue, the track record, innings (first/second), the force of the team etc., the different attributes such as the toss, running pace, wickets, etc.

1.3 Rational of the study

Basically, Cricket is a bat and ball game between 2 teams each with 11 participants. And team has a single inning in which they try to score the most number of runs possible when the other teams are in control. The input stops when the complete limit, which is dependent on the type of the game, is up or the ten batsmen have been rejected. The main aim is to achieve additional runs and hence the deciding element. Cricket game in nature is very volatile. It is hard to accurately forecast the game before the very last minute. Different natural variables that impact play quality, an immense betting market and a massive media presence have offered powerful motivations to model this game from the viewpoint of machine learning. Cricket Regulations are determined by the International Cricket Council (ICC). On the international level, there are 3 universally accepted cricket formats – T20, One Day, and Test Game. The programmed length of the game is the main distinction between all three formats. This changes the number of deliveries that each side plays in their respective innings directly. The cricket test style is the largest and the highest level of the game. The match takes five days for each side to play two innings. A regular trial day includes 3 sessions each lasting 2 hours. ODI is limited in one day international, with 300 deliveries for each team (50 overs). ODI matches generally fall into one of two categories: the Day match or the Day-Night match. T20 is the shortest globally accepted format for this game with 20 team inputs. It's more explosive than the other two formats and more athletic.

1.4 Research Questions

- Why we need this research work?
- What are the outcomes of this research work?
- What are the challenges for this research?
- What are the future scope and plan for this research work?
- Why do we need this research work?

1.5 Expected Output:

The research focuses on Cricket, One Day International or ODIs in the most common format. A varied number of features influences the results of One Day Internationals and it can be forecast like other sports. It is important to find the right qualities or variables that affect the match result. The parameters used and examined for this study, which have been shown to significantly affect the match result. The analytical variables included:

Past performance teams: This element records the historical results of all team matches.

Ground: This is a crucial task when teams have great history of tactical dominance over each other on specific grounds.

Innings: This factor decides which team was first batted and which was second batted. Home Game Advantage: The location function determines whether a certain field is neutral or home for one of the players. Home Game Advantage is obtained. Based on these variables, both classifiers are qualified..

1.6 Report Layout:

In this paper, we addressed a general view of the whole research work in the first chapter. We addressed the history studies relating to these works in the second Chapter. We addressed the methods of analysis in the next chapter. We mentioned which algorithms we were using in the fourth chapter. The experimental findings and conclusions were discussed in the last chapter.

CHAPTER 2

Background Studies

2.1 Introduction:

Machine learning techniques augmented by data mining techniques have been a hot subject worldwide for study, however. Cricket is a common sport with a net core audience in Australia, the South of France, the UK and South Asia. The game is supported by enormous viewers in nations and people are very keen to forecast the game results. The result of a cricket match is determined by several pre-game and in-game attributes. Pre-game characteristics such as location, previous track records, entrances, team strength etc., and different in-game characteristics such as toss, run rate, wickets, batting avg etc. influence in a prevailing way the outcome of a play. Computer intelligence is a promising research field to study and model the Cricket game. Cricket has become an important sport for mathematical research and machine learning because of increased exposure and financial benefits. Cricket's diverse nature and its complicated regulations make the challenge difficult. Due to the gaps in methodology, the different approaches and what was shown in available work is neither quite simple nor recorded. In future studies it would assist if the positive and inconvenience of the present work are well assessed and recorded. The results of a cricket match are affected by several different causes. This document introduces the possibility of using algorithms to forecast the results of an international cricket match. This research has two objectives - to classify influential elements that influence the outcomes of an algorithm and to forecast the outcome of an algorithm based on a cricket match.

2.2 Related Work:

. In this analysis, the effects of a cricket match result due to these various causes, were analyzed across 2 separate ML approaches namely Trees and the Multilayer Perceptron Network. CricAI: Cricket Match Outcome Prediction System was built on the basis of these data. This modeling tool takes into account the proclamation attributes such as land, location (home, out, neutral) and inputs (first/second). [1] The model generated with the most precise tweets from the last 10 overviews. Per model measured the efficiency of the classifiers, and logistic regression and the vector support system showed higher precise results than others. The final research was conducted to find the most famous player in any game and to verify the relationship between the popular player and the man of the game.[2] Applied to align predictions are machine learning algorithms like SVM, Random Tree, Random Forest, and Naive Bayes. Finally, the conclusions were drawn up on the basis of which the algorithm provides the highest precision. Algorithms for decision tree and logistic regression have achieved an accuracy of 87% and 95% respectively.[3] The paper proposes a pattern with two methods: first, the score of the first inputs is predicted not only by the current run rate but also the number of wickets dropped, the location of the matches and the batting squad. The second method forecasts the result of the second input match, taking into account the same characteristics as the previous method and the target provided to the batting team. On both first and second inputs, these two methods were applied with the Naïve Bayes and Linear Regression Classifier. In both methods 5 cycles of 50 over the game were made and all uncurtail matches, played separately between 2002 and 2014, and were registered for each of the above listed attributes. The findings showed that in Linear Run Rate classification, the error in calculating the final score is less than the current rating system, while Naïve Bayes' accuracy in predicting match outcomes was initially 68 per cent from 0-5% to 91 percent to 45th place. [4] This paper analyzes the work already undertaken on the match results forecast in the Cricket domain. This paper follows on from current research and we aim to resolve the missed connections and disadvantages discussed in this paper by the conclusion of our research. [5] We saw scores predicted at varying times on our TV for years watching small cricket overviews. The scores projected are entirely dependent on scored runs and the combined values of various run rates at the end of the inputs. If a team score is 100 at the end of 10 overviews, for instance. Just running speed cannot provide reasonable results, since the inputs may be affected by different variables. By taking existing game data from the cricinfo website, we create a model for T20-format play. [6]

2.3 Research Summary:

The results of a cricket match can be predicted using Machine Learning (ML). Just as with ML other challenges, the toughest difficulty is to find the correct features/attributes. Such characteristics can be used. Any ball, batsmen, bowlers and fields' historical results. The venue of the match, temperature, average team age, the number of players, the results of the previous team and the teams and much more. The results improvement for a given team over a timeframe can also be seen as a positive model. A comprehensive list of discriminating characteristics is very difficult to draw up and we don't know whether many are still significant or not. However, it is because of the complexity of forms and historical facts that a model with the best functionality will not be constructed. We know also the underdogs can and have repeatedly defeated the giants of the game. We will use and refine our predictive model to tap into certain scenarios or anomalies. The difficult issue is to determine what a good model is and how good it is. Currently, cricket matches in One Day International (ODI) are based on the current run rate and can be estimated to represent the amount of runs scored according to the number of over bound matches. Factors such as the number and location of the match are not used. The final result of a match with fair exactness using machine learning would be difficult to predict. You may also estimate a team's final score or the chasing team's chance of winning according to current results.

2.4 Challenges

In this research work we have to face some complexity. We have to find a suitable dataset and find the algorithm to find out which one gives the best accuracy. In this research work we have used Decision Tree, Logistic Regression, KNN, Random Forest, and SVM, Naïve Bayes these six algorithms to compare and find out the best accuracy giving algorithm. We have to preprocess the dataset and make the dataset usable for our research work. We have tried to solve all the challenges and we are still trying to improve our research work and have some better plans according to our work in future.

CHAPTER 3

Research Methodology

3.1 Introduction

Instead of only forecasting player results, the proposed scheme, unlike the existing implementations, concentrates on predicting the best playing 11. This method uses these 6 algorithms, the highest and the most adaptable for this model, which were determined by the Decision Tree, Logistic Regression, KNN, Random Forest, SVM and Naïve Bayes. The player's success must not only be expected, but a decision must be made about whether the player is the best player to be included in the squad, based on recent statistics and other criterion. The model eliminates partial choices entirely and gives you the best option. The method analyzes the available positions within a squad until a player is chosen, i.e. if a batsman and a bowler in the system are selected first the number of places available for batsmen and bowler and for the all-rounders is determined. The judgment process for all plays in the dataset starts from these numbers. The decision trees are built to have a reasonable number of batsmen, bowlers and horsemen. The decision trees are intended. More particulars like the bowler-seam/fast-spin form, hand-haired/left-held batsmen, etc. should also be taken into consideration. The squad can also be equipped with a wicket keeper, who has the highest statistics as a wicket keeper. The model may also be learned to decide on the success of the expected squad, according to the cumulative analyzes of the performance of individual players who can almost posturize the circumstances on the ground. This will encourage not only players but also their team and coaches to see when and how their team needs to concentrate on achieving positive match results. Various additional characteristics such as site analyzes and paying/team results may be introduced in the future. And to see how long and how many matches a single player reliably plays without any injuries, etc, a summary analysis of player health cold is added. For domestic tourneys where few players are kept or on contracts for a certain period of years, the suggested framework may be very helpful. In such situations, the analyzer has difficulty studying the records of each player manually and finding out whose players and retained players will make the right squad.

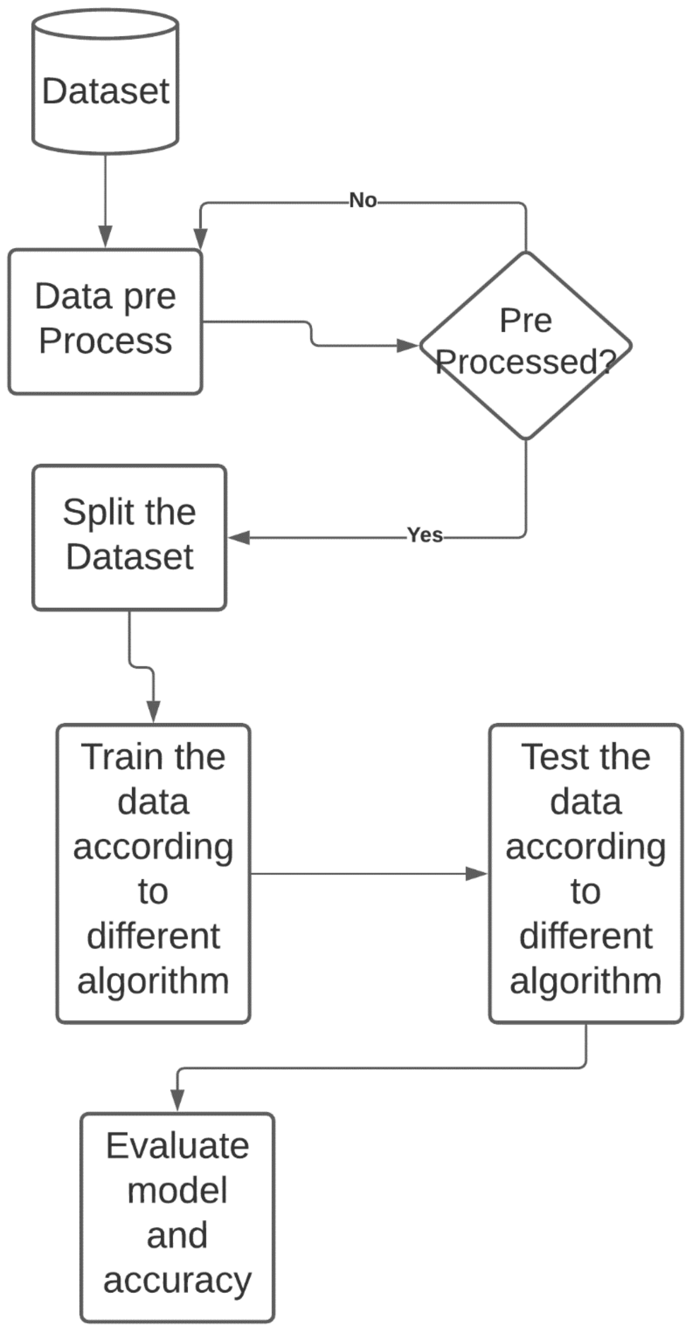


Figure 3.1.1 Workflow

3.2 Research Subject and Instrumentation

The method of detecting cricket match results with the logical and analytical process was initially addressed. The model of machine learning needs a large GPU pc and other instrument setup. A list of the appropriate tools for this model is now provided below.

Hardware and Software:

- Intel Core i7 9^{th} generation
- 2 TB HDD
- 12GB RAM
- 240GB SSD

Development Tools:

- Windows 10
- Python 3.9
- Pandas
- Seaborn
- Numpy
- Scikit-learn
- Matplotlib

3.3 Data collection and Data preprocessing

We obtained data from any of the data are from cricinfo, cricbuzz, and cricsheet. Our data are therefore almost reliable. Then we vacuum and rearrange them. We updated the data set a bit according to our requirement. We would subsequently use the dataset for each algorithm. As we say already 6 algorithms, but we have and process the data properly to adapt the data in our model for each algorithm to function differently.

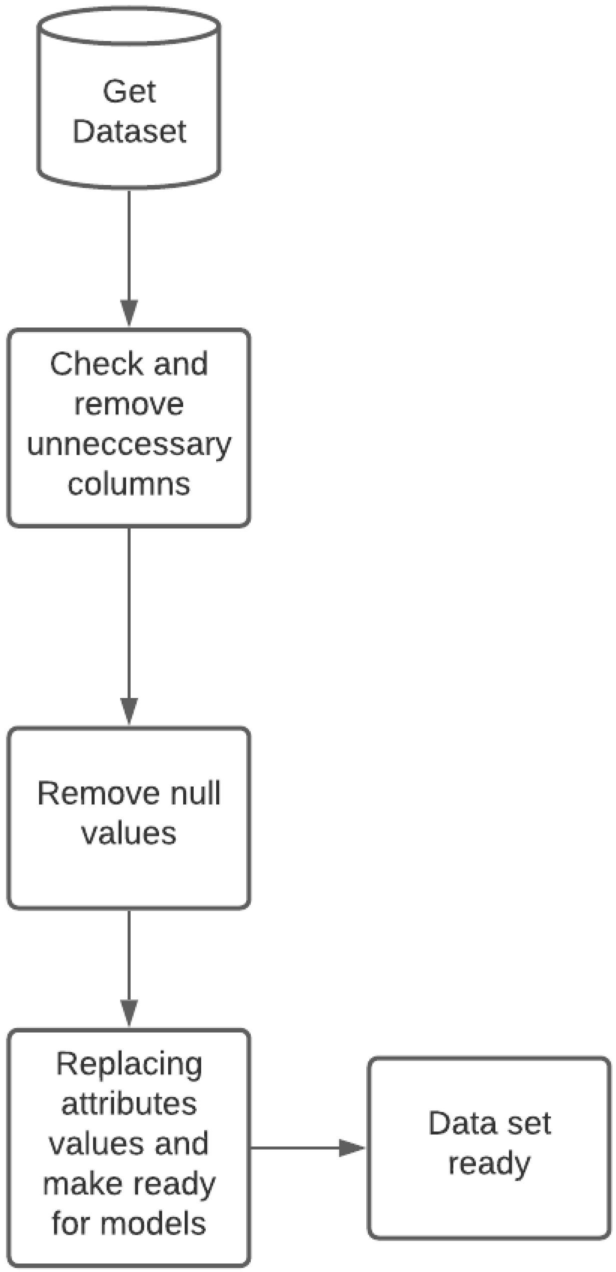


Figure 3.3.1: Dataset pre-process steps

3.3. a Get Dataset:

As we said before, we gather data from various sources, but researchers use different kinds of data sets. Then we have to carefully choose the data set because we want to make a cricket match with a data set with many algorithms. That is why we need to carefully select the dataset and correctly process the data collection.

3.3.b Check and drop unnecessary columns:

For data preprocessing it is really critical that redundant columns are removed. We must verify whether or not there is a null value. The null value simply lowers the accuracy ratio, so that we exclude the null value, and if there is no null value it will allow us adequately to train the data set and give us a more reliable outcome. And we must delete redundant columns so we don't have to incorporate all the columns into our model.

3.3.c Removing null values:

As we already said, the deletion of the null value from a data set is necessary because it won't be properly conditioned by this null value until we delete the null values while using various algorithms during training. And often the algorithm model produces less reliable results for these null values. Often data may be incomplete, or the data may not be entered correctly to delete those items. If we delete all null values correctly, we will see the higher precision when we evaluate the algorithm for this dataset.

3.3. d Purified data:

Now with all the data processing steps such as checking and deleting unnecessary columns, removing null value, the data we get is eventually pea processed or cleaned data. This results are cleaner than the initial data. The original information could have several types of items, but the uniform look is more easily trained for machines following preprocessing. In fact, the pre-process data is the training we provide to the system.

Here is the dataset before processing, there are: ground stats, ODI match total, match result from 2010 to 2020 and ranking based on 2019 world cup.

	A	B	C	D	E	F	G	H	I
1	Ground	Mat	Won	Wkts	Balls	Ave	RPO		
2	Eden Gardens, Kolkata - India	4	4	72	2297	30.01	5.64		
3	Feroz Shah Kotla, Delhi - India	4	4	75	2331	23.85	4.6		
4	Melbourne Cricket Ground - Australia	15	15	217	8482	35.28	5.41		
5	Saurashtra Cricket Association Stadium, Rajkot - India	2	2	26	1200	44.73	5.81		
6	Adelaide Oval - Australia	10	10	157	5645	30.97	5.16		
7	Nehru Stadium, Kochi - India	3	3	46	1567	30.08	5.29		
8	Brisbane Cricket Ground, Woolloongabba, Brisbane - Austra	8	8	123	4189	29.84	5.25		
9	JSCA International Stadium Complex, Ranchi - India	5	4	68	2522	33.86	5.47		
10	Boland Park, Paarl - South Africa	3	3	51	1711	28.98	5.18		
11	Sydney Cricket Ground - Australia	16	14	209	7876	36.05	5.74		
12	Diamond Oval, Kimberley - South Africa	3	3	40	1514	33.17	5.25		
13	Bellerive Oval, Hobart - Australia	6	6	94	3192	33.69	5.95		
14	Punjab Cricket Association IS Bindra Stadium, Mohali, Chan	5	5	67	2957	45.61	6.2		
15	Senwes Park, Potchefstroom - South Africa	2	2	30	1164	31.16	4.81		
16	Himachal Pradesh Cricket Association Stadium, Dharamsala	4	4	56	1987	29.71	5.02		
17	W.A.C.A. Ground, Perth - Australia	12	12	177	5918	29.36	5.26		
18	Manuka Oval, Canberra - Australia	7	7	110	4026	38.41	6.29		
19	Seddon Park, Hamilton - New Zealand	16	15	215	8165	36.03	5.69		
20	McLean Park, Napier - New Zealand	8	8	110	4299	34.27	5.26		
21	National Cricket Stadium, St George's, Grenada - West Indie	7	6	87	3346	35.06	5.47		
22	Eden Park, Auckland - New Zealand	12	10	187	5603	27.88	5.58		
23	Sharjah Cricket Stadium - U.A.E.	31	31	470	15659	27.45	4.94		
24	Mangaung Oval, Bloemfontein - South Africa	3	3	48	1448	29.79	5.92		
25	ICC Academy, Dubai - U.A.E.	23	22	326	12214	32.28	5.16		
26	SuperSport Park, Centurion - South Africa	12	11	159	5826	35.65	5.83		
27	The Wanderers Stadium, Johannesburg - South Africa	9	9	122	4663	38.38	6.02		
28	Kingsmead, Durban - South Africa	9	9	132	4820	33.93	5.57		
29	Mahinda Rajapaksa International Cricket Stadium, Sooriyav	12	10	138	5340	39.56	6.13		
30	Willowmoore Park, Benoni - South Africa	3	3	40	1488	32.8	5.29		

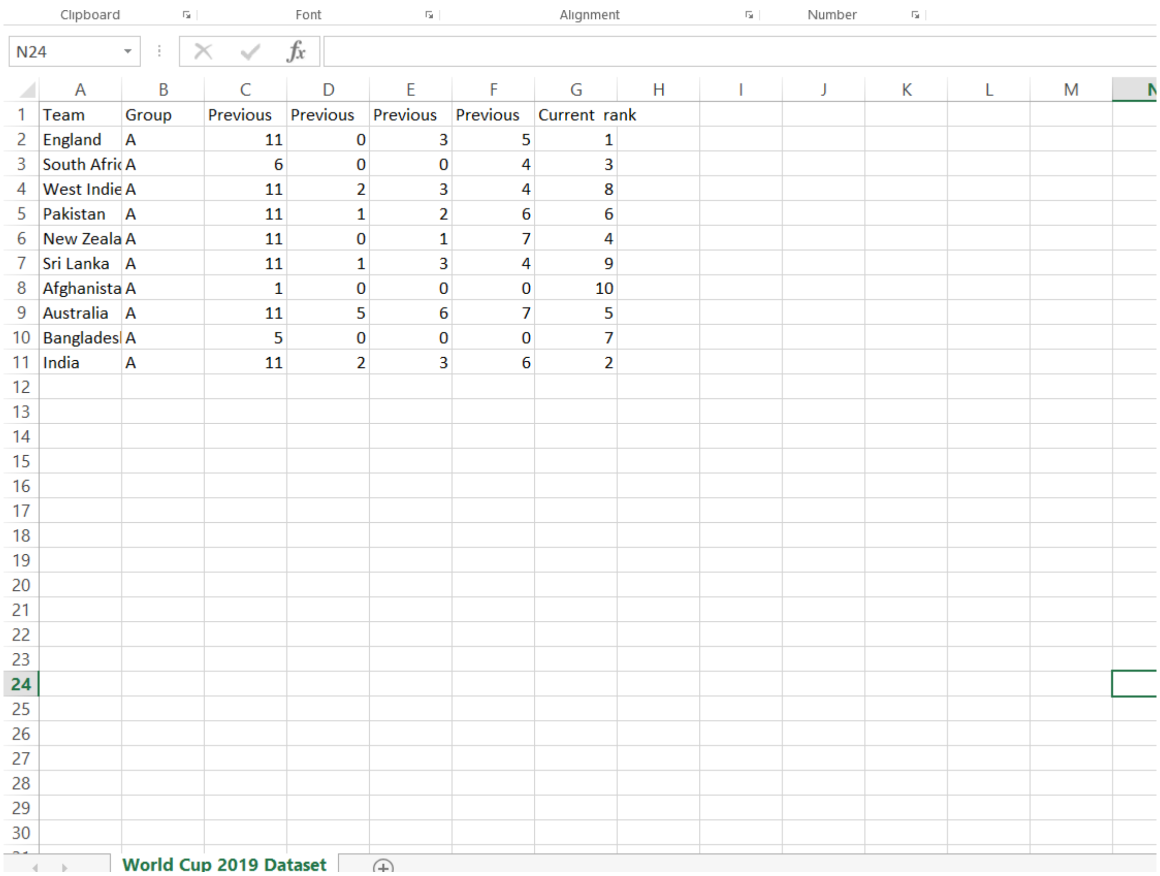
date	Team_1	Team_2	Margin
4-Jan-10	Bangladesh	Sri Lanka	7 wickets
5-Jan-10	India	Sri Lanka	5 wickets
7-Jan-10	Bangladesh	India	6 wickets
8-Jan-10	Bangladesh	Sri Lanka	9 wickets
10-Jan-10	India	Sri Lanka	8 wickets
11-Jan-10	Bangladesh	India	6 wickets
13-Jan-10	India	Sri Lanka	4 wickets
22-Jan-10	Australia	Pakistan	5 wickets
24-Jan-10	Australia	Pakistan	140 runs
26-Jan-10	Australia	Pakistan	40 runs
29-Jan-10	Australia	Pakistan	135 runs
31-Jan-10	Australia	Pakistan	2 wickets
5-Feb-10	New Zealand	Bangladesh	146 runs
7-Feb-10	Australia	West Indies	113 runs
8-Feb-10	New Zealand	Bangladesh	5 wickets
9-Feb-10	Australia	West Indies	8 wickets
11-Feb-10	New Zealand	Bangladesh	3 wickets
12-Feb-10	Australia	West Indies	
14-Feb-10	Australia	West Indies	50 runs
16-Feb-10	Kenya	Netherlands	6 wickets
16-Feb-10	Afghanistan	Canada	1 run
18-Feb-10	Kenya	Netherlands	80 runs
18-Feb-10	Afghanistan	Canada	4 wickets
19-Feb-10	Australia	West Indies	125 runs
21-Feb-10	India	South Africa	1 run
24-Feb-10	India	South Africa	153 runs
27-Feb-10	India	South Africa	90 runs
28-Feb-10	Bangladesh	England	6 wickets
2-Mar-10	Bangladesh	England	2 wickets

results

U22

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Score	Overs	RPO	Target	Inns	Result	Oppositior	Ground	Start Date	Match_ID	Country	Country_ID
2	412	250	48.3	5.15		1	won	v India	Kolkata	3-Jan-13	ODI # 331!	Pakistan	7
3	680	165	48	3.43	251	2	lost	v Pakistan	Kolkata	3-Jan-13	ODI # 331!	India	6
4	413	157	48.5	3.21	168	2	lost	v India	Delhi	6-Jan-13	ODI # 331(	Pakistan	7
5	681	167	43.4	3.82		1	won	v Pakistan	Delhi	6-Jan-13	ODI # 331(	India	6
6	117	198	40	4.95	306	2	lost	v Australia	Melbourne	11-Jan-13	ODI # 331;	SriLanka	8
7	1076	305/5	50	6.1		1	won	v Sri Lanka	Melbourne	11-Jan-13	ODI # 331;	Australia	2
8	682	316/9	50	6.32	326	2	lost	v England	Rajkot	11-Jan-13	ODI # 331ι	India	6
9	836	325/4	50	6.5		1	won	v India	Rajkot	11-Jan-13	ODI # 331ι	England	1
10	118	172/2	40.1	4.28	171	2	won	v Australia	Adelaide	13-Jan-13	ODI # 331!	SriLanka	8
11	1077	170	46.5	3.62		1	lost	v Sri Lanka	Adelaide	13-Jan-13	ODI # 331!	Australia	2
12	683	285/6	50	5.7		1	won	v England	Kochi	15-Jan-13	ODI # 332(	India	6
13	837	158	36	4.38	286	2	lost	v India	Kochi	15-Jan-13	ODI # 332(	England	1
14	119	75/6	20	3.75	75	2	won	v Australia	Brisbane	18-Jan-13	ODI # 332:	SriLanka	8
15	1078	74	26.4	2.77		1	lost	v Sri Lanka	Brisbane	18-Jan-13	ODI # 332:	Australia	2
16	684	157/3	28.1	5.57	156	2	won	v England	Ranchi	19-Jan-13	ODI # 332;	India	6
17	838	155	42.2	3.66		1	lost	v India	Ranchi	19-Jan-13	ODI # 332;	England	1
18	277	208	46.2	4.48		1	lost	v New Zea	Paarl	19-Jan-13	ODI # 332:	SouthAfric	3
19	551	209/9	45.4	4.57	209	2	won	v South Af	Paarl	19-Jan-13	ODI # 332:	Newzealac	5
20	120	14/0	3.2	4.2	223	2	n/r	v Australia	Sydney	20-Jan-13	ODI # 332‹	SriLanka	8
21	1079	222/9	50	4.44		1	n/r	v Sri Lanka	Sydney	20-Jan-13	ODI # 332‹	Australia	2
22	278	252	49.1	5.12	280	2	lost	v New Zea	Kimberley	22-Jan-13	ODI # 332!	SouthAfric	3
23	552	279/8	50	5.58		1	won	v South Af	Kimberley	22-Jan-13	ODI # 332!	Newzealac	5
24	121	215	48.3	4.43	248	2	lost	v Australia	Hobart	23-Jan-13	ODI # 332(	SriLanka	8
25	1080	247/5	50	4.94		1	won	v Sri Lanka	Hobart	23-Jan-13	ODI # 332(	Australia	2
26	685	258/5	47.3	5.43	258	2	won	v England	Mohali	23-Jan-13	ODI # 332;	India	6
27	839	257/7	50	5.14		1	lost	v India	Mohali	23-Jan-13	ODI # 332;	England	1
28	279	264/9	50	5.28	261	2	won	v New Zea	Potchefstr	25-Jan-13	ODI # 332ι	SouthAfric	3
29	553	260/9	50	5.2		1	lost	v South Af	Potchefstr	25-Jan-13	ODI # 332ι	Newzealac	5
30	686	226	49.4	4.55		1	lost	v England	Dharamsal	27-Jan-13	ODI # 332!	India	6

ODI_Match_Totals



Team	Group	Previous	Previous	Previous	Previous	Current rank
England	A	11	0	3	5	1
South Afric	A	6	0	0	4	3
West Indie	A	11	2	3	4	8
Pakistan	A	11	1	2	6	6
New Zeala	A	11	0	1	7	4
Sri Lanka	A	11	1	3	4	9
Afghanista	A	1	0	0	0	10
Australia	A	11	5	6	7	5
Banglades	A	5	0	0	0	7
India	A	11	2	3	6	2

Fig 3.3.2: Before Pre-processing data

Here is the dataset after processing, there are: 2019 world cup squads for all the teams, ground stats, ODI match total, match result from 2010 to 2020 and ranking based on 2019 world cup.

	Player	ID	Country
33	Mashrafe Mortaza (c)	56007	Bangladesh
34	Shakib Al Hasan (vc)	56143	Bangladesh
35	Tamim Iqbal	56194	Bangladesh
36	Liton Das	536936	Bangladesh
37	Mushfiqur Rahim (wk)	56029	Bangladesh
38	Mahmudullah	56025	Bangladesh
39	Mohammad Mithun	269237	Bangladesh
40	Sabbir Rahman	373538	Bangladesh
41	Mehedi Hasan Miraz	989761	Bangladesh
42	Soumya Sarkar	436677	Bangladesh
43	Rubel Hossain	300619	Bangladesh
44	Mohammad Saifuddin	629070	Bangladesh
45	Mosaddek Hossain	550133	Bangladesh
46	Mustafizur Rahman	330902	Bangladesh
47	Abu Jayed	410763	Bangladesh
124	Dimuth Karunaratne (c)	227772	SriLanka
125	Dhananjaya de Silva (vc)	465793	SriLanka
126	Angelo Mathews	49764	SriLanka
127	Avishka Fernando	784369	SriLanka
128	Lahiru Thirimanne	301236	SriLanka
129	Kusal Mendis	629074	SriLanka
130	Kusal Perera (wk)	300631	SriLanka
131	Thisara Perera	233514	SriLanka
132	Isuru Udana	328026	SriLanka
133	Jeffrey Vandersay	370040	SriLanka
134	Jeevan Mendis	49700	SriLanka
135	Milinda Siriwardana	222354	SriLanka
136	Lasith Malinga	49758	SriLanka
137	Suranga Lakmal	46919	SriLanka

WC_players

	Ground	Span	Mat	Won	Tied	NR	Runs	Wkts	Balls	Ave	RPO
2	Eden Gardens, Kolkata - India	2013-2017	4	4	0	0	2161	72	2297	30.01	5.64
3	Feroz Shah Kotla, Delhi - India	2013-2019	4	4	0	0	1789	75	2331	23.85	4.6
4	Melbourne Cricket Ground - Australia	2013-2019	15	15	0	0	7656	217	8482	35.28	5.41
5	Saurashtra Cricket Association Stadium, Rajkot - India	2013-2015	2	2	0	0	1163	26	1200	44.73	5.81
6	Adelaide Oval - Australia	2013-2019	10	10	0	0	4863	157	5645	30.97	5.16
7	Nehru Stadium, Kochi - India	2013-2014	3	3	0	0	1384	46	1567	30.08	5.29
8	Brisbane Cricket Ground, Woolloongabba, Brisbane - Australia	2013-2018	8	8	0	0	3671	123	4189	29.84	5.25
9	JSCA International Stadium Complex, Ranchi - India	2013-2019	5	4	0	1	2303	68	2522	33.86	5.47
10	Boland Park, Paarl - South Africa	2013-2018	3	3	0	0	1478	51	1711	28.98	5.18
11	Sydney Cricket Ground - Australia	2013-2019	16	14	0	2	7535	209	7876	36.05	5.74
12	Diamond Oval, Kimberley - South Africa	2013-2018	3	3	0	0	1327	40	1514	33.17	5.25
13	Bellerive Oval, Hobart - Australia	2013-2018	6	6	0	0	3167	94	3192	33.69	5.95
14	Punjab Cricket Association IS Bindra Stadium, Mohali, Chandigarh - India	2013-2019	5	5	0	0	3056	67	2957	45.61	6.2
15	Senwes Park, Potchefstroom - South Africa	2013-2015	2	2	0	0	935	30	1164	31.16	4.81
16	Himachal Pradesh Cricket Association Stadium, Dharamsala - India	2013-2017	4	4	0	0	1664	56	1987	29.71	5.02
17	W.A.C.A. Ground, Perth - Australia	2013-2017	12	12	0	0	5197	177	5918	29.36	5.26
18	Manuka Oval, Canberra - Australia	2013-2016	7	7	0	0	4226	110	4026	38.41	6.29
19	Seddon Park, Hamilton - New Zealand	2013-2019	16	15	0	1	7748	215	8165	36.03	5.69
20	McLean Park, Napier - New Zealand	2013-2019	8	8	0	0	3770	110	4299	34.27	5.26
21	National Cricket Stadium, St George's, Grenada - West Indies	2013-2019	7	6	0	1	3051	87	3346	35.06	5.47
22	Eden Park, Auckland - New Zealand	2013-2017	12	10	1	1	5214	187	5603	27.88	5.58
23	Sharjah Cricket Stadium - U.A.E.	2013-2019	31	31	0	0	12905	470	15659	27.45	4.94
24	Mangaung Oval, Bloemfontein - South Africa	2013-2018	3	3	0	0	1430	48	1448	29.79	5.92
25	ICC Academy, Dubai - U.A.E.	2013-2019	23	22	0	1	10524	326	12214	32.28	5.16
26	SuperSport Park, Centurion - South Africa	2013-2019	12	11	0	1	5669	159	5826	35.65	5.83
27	The Wanderers Stadium, Johannesburg - South Africa	2013-2019	9	9	0	0	4683	122	4663	38.38	6.02
28	Kingsmead, Durban - South Africa	2013-2019	9	9	0	0	4480	132	4820	33.93	5.57
29	Mahinda Rajapaksa International Cricket Stadium, Sooriyawewa, Hambantot	2013-2017	12	10	0	2	5460	138	5340	39.56	6.13

J11

	A	B	C	D	E	F
1	date	Team_1	Team_2	Winner	Margin	Ground
2	4-Jan-10	Bangladesh	Sri Lanka	Sri Lanka	7 wickets	Dhaka
3	5-Jan-10	India	Sri Lanka	Sri Lanka	5 wickets	Dhaka
4	7-Jan-10	Bangladesh	India	India	6 wickets	Dhaka
5	8-Jan-10	Bangladesh	Sri Lanka	Sri Lanka	9 wickets	Dhaka
6	10-Jan-10	India	Sri Lanka	India	8 wickets	Dhaka
7	11-Jan-10	Bangladesh	India	India	6 wickets	Dhaka
8	13-Jan-10	India	Sri Lanka	Sri Lanka	4 wickets	Dhaka
9	22-Jan-10	Australia	Pakistan	Australia	5 wickets	Brisbane
10	24-Jan-10	Australia	Pakistan	Australia	140 runs	Sydney
11	26-Jan-10	Australia	Pakistan	Australia	40 runs	Adelaide
12	29-Jan-10	Australia	Pakistan	Australia	135 runs	Perth
13	31-Jan-10	Australia	Pakistan	Australia	2 wickets	Perth
14	5-Feb-10	New Zealand	Bangladesh	New Zealand	146 runs	Napier
15	7-Feb-10	Australia	West Indies	Australia	113 runs	Melbourne
16	8-Feb-10	New Zealand	Bangladesh	New Zealand	5 wickets	Dunedin
17	9-Feb-10	Australia	West Indies	Australia	8 wickets	Adelaide
18	11-Feb-10	New Zealand	Bangladesh	New Zealand	3 wickets	Christchurch
19	12-Feb-10	Australia	West Indies	no result		Sydney
20	14-Feb-10	Australia	West Indies	Australia	50 runs	Brisbane
21	16-Feb-10	Kenya	Netherlands	Kenya	6 wickets	Nairobi (Gym)
22	16-Feb-10	Afghanistan	Canada	Afghanistan	1 run	Sharjah
23	18-Feb-10	Kenya	Netherlands	Netherlands	80 runs	Nairobi (Gym)
24	18-Feb-10	Afghanistan	Canada	Canada	4 wickets	Sharjah
25	19-Feb-10	Australia	West Indies	Australia	125 runs	Melbourne
26	21-Feb-10	India	South Africa	India	1 run	Jaipur
27	24-Feb-10	India	South Africa	India	153 runs	Gwalior
28	27-Feb-10	India	South Africa	South Africa	90 runs	Ahmedabad
29	28-Feb-10	Bangladesh	England	England	6 wickets	Dhaka
30	2-Mar-10	Bangladesh	England	England	2 wickets	Dhaka

results

M30

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Result	Margin	BR	Toss	Bat	Oppositior	Ground	Start Date	Match_ID	Country	Country_ID			
2	418	won	85 runs		lost	1st	v India	Kolkata	3-Jan-13	ODI # 3315	Pakistan	7			
3	692	lost	85 runs		won	2nd	v Pakistan	Kolkata	3-Jan-13	ODI # 3315	India	6			
4	419	lost	10 runs		lost	2nd	v India	Delhi	6-Jan-13	ODI # 3316	Pakistan	7			
5	693	won	10 runs		won	1st	v Pakistan	Delhi	6-Jan-13	ODI # 3316	India	6			
6	121	lost	107 runs		lost	2nd	v Australia	Melbourne	11-Jan-13	ODI # 3317	SriLanka	8			
7	1096	won	107 runs		won	1st	v Sri Lanka	Melbourne	11-Jan-13	ODI # 3317	Australia	2			
8	694	lost	9 runs		lost	2nd	v England	Rajkot	11-Jan-13	ODI # 3318	India	6			
9	852	won	9 runs		won	1st	v India	Rajkot	11-Jan-13	ODI # 3318	England	1			
10	122	won	8 wickets	59	won	2nd	v Australia	Adelaide	13-Jan-13	ODI # 3319	SriLanka	8			
11	1097	lost	8 wickets	59	lost	1st	v Sri Lanka	Adelaide	13-Jan-13	ODI # 3319	Australia	2			
12	695	won	127 runs		won	1st	v England	Kochi	15-Jan-13	ODI # 3320	India	6			
13	853	lost	127 runs		lost	2nd	v India	Kochi	15-Jan-13	ODI # 3320	England	1			
14	123	won	4 wickets	180	lost	2nd	v Australia	Brisbane	18-Jan-13	ODI # 3321	SriLanka	8			
15	1098	lost	4 wickets	180	won	1st	v Sri Lanka	Brisbane	18-Jan-13	ODI # 3321	Australia	2			
16	696	won	7 wickets	131	won	2nd	v England	Ranchi	19-Jan-13	ODI # 3322	India	6			
17	854	lost	7 wickets	131	lost	1st	v India	Ranchi	19-Jan-13	ODI # 3322	England	1			
18	283	lost	1 wickets	26	lost	1st	v New Zea	Paarl	19-Jan-13	ODI # 3323	SouthAfric	3			
19	560	won	1 wickets	26	won	2nd	v South Af	Paarl	19-Jan-13	ODI # 3323	Newzealac	5			
20	124	n/r	-		lost	2nd	v Australia	Sydney	20-Jan-13	ODI # 3324	SriLanka	8			
21	1099	n/r	-		won	1st	v Sri Lanka	Sydney	20-Jan-13	ODI # 3324	Australia	2			
22	284	lost	27 runs		won	2nd	v New Zea	Kimberley	22-Jan-13	ODI # 3325	SouthAfric	3			
23	561	won	27 runs		lost	1st	v South Af	Kimberley	22-Jan-13	ODI # 3325	Newzealac	5			
24	125	lost	32 runs		won	2nd	v Australia	Hobart	23-Jan-13	ODI # 3326	SriLanka	8			
25	1100	won	32 runs		lost	1st	v Sri Lanka	Hobart	23-Jan-13	ODI # 3326	Australia	2			
26	697	won	5 wickets	15	won	2nd	v England	Mohali	23-Jan-13	ODI # 3327	India	6			
27	855	lost	5 wickets	15	lost	1st	v India	Mohali	23-Jan-13	ODI # 3327	England	1			
28	285	won	1 wickets	0	lost	2nd	v New Zea	Potchefstr	25-Jan-13	ODI # 3328	SouthAfric	3			
29	562	lost	1 wickets	0	won	1st	v South Af	Potchefstr	25-Jan-13	ODI # 3328	Newzealac	5			
30	698	lost	7 wickets	16	lost	1st	v England	Dharamsal	27-Jan-13	ODI # 3329	India	6			

Team	Group	Previous	Previous	Previous	Previous	Current rank
England	A	11	0	3	5	1
South Afric	A	6	0	0	4	3
West Indie	A	11	2	3	4	8
Pakistan	A	11	1	2	6	6
New Zeala	A	11	0	1	7	4
Sri Lanka	A	11	1	3	4	9
Afghanista	A	1	0	0	0	10
Australia	A	11	5	6	7	5
Banglades	A	5	0	0	0	7
India	A	11	2	3	6	2

World Cup 2019 Dataset

Fig 3.3.3: After Pre-processing data

```
In [48]: results.head()
         print(results.to_string())
```

```
                 date      Team_1     Team_2     Winner     Margin                  Grou
nd
0             4-Jan-10  Bangladesh  Sri Lanka  Sri Lanka  7 wickets                  Dha
ka
1             5-Jan-10       India  Sri Lanka  Sri Lanka  5 wickets                  Dha
ka
2             7-Jan-10  Bangladesh      India      India  6 wickets                  Dha
ka
3             8-Jan-10  Bangladesh  Sri Lanka  Sri Lanka  9 wickets                  Dha
ka
4            10-Jan-10       India  Sri Lanka      India  8 wickets                  Dha
ka
5            11-Jan-10  Bangladesh      India      India  6 wickets                  Dha
ka
6            13-Jan-10       India  Sri Lanka  Sri Lanka  4 wickets                  Dha
ka
7            22-Jan-10   Australia   Pakistan  Australia  5 wickets               Brisba
ne
8            24-Jan-10   Australia   Pakistan  Australia   140 runs                 Sydn
```

Figure 3.3.4: Total dataset from 2010 to 2020

CHAPTER 4

Implementation and Algorithms

4.1 SVM (Support vector machine):

A separate hyperplane can demonstrate SVM as a classifier of discrimination. SVM will, in many cases, be a familiar master's rule. It is also used in problems of grouping. In the models of the SVM, each object can be seen as a degree in the n-dimensional field worth each element (where n is many features). The Associate can be well understood as an indicator of the 2 groups. Using this evaluation, we can understand the SVM classification process.

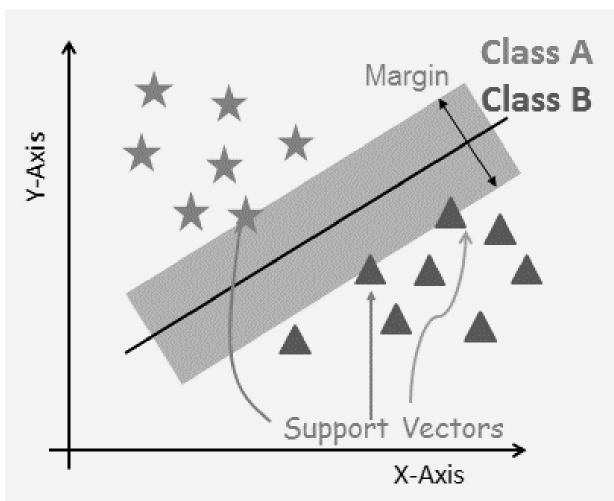


Figure 4.1.1: SVM figure

The main objective is the maximum possible isolation of the collection of data. The feature can be defined as a margin of any of the nearest data points. The goal is to decide the optimal margin between supportive data collection vectors for a hyperplane. In the steps mentioned, SVM searches for the hyperplane maximum:

Construct hyperplanes that distinguish the classes more effectively. Left-hand figure of three hyperplanes, yellow, blue and orange. The higher black and white grade is wrong, but black line divides the two grades correctly.

For full separation, choose the right hyperplane from the nearest distance.

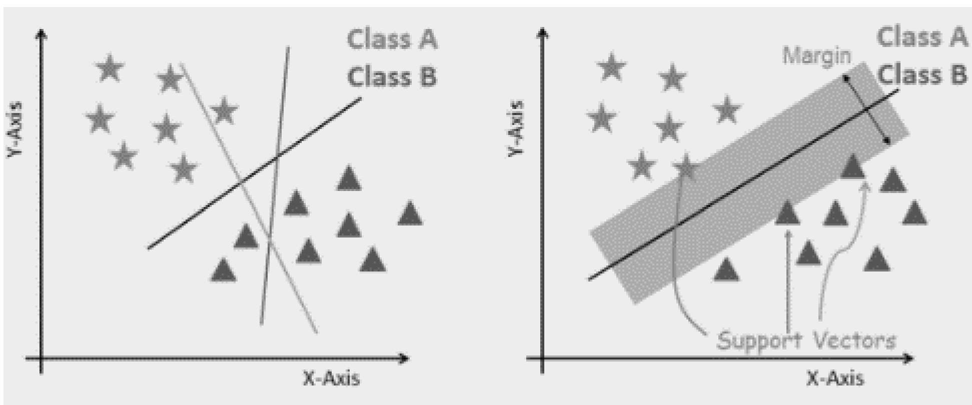


Figure 4.1.2: How svm works

There are several problems concerning linear hyperplane that cannot be solved, as seen below, during the management of nonlinear and indiscriminate planes (left-hand side).

In the example, SVM uses a backpropagation algorithm to move the space into a larger dimensional space, as seen on the right-hand side. The data points are seen in the x-axis and z-axis (Z is squared with x and y: z=x^2=y^2). These points could now be easily distinguished by linear discrimination.

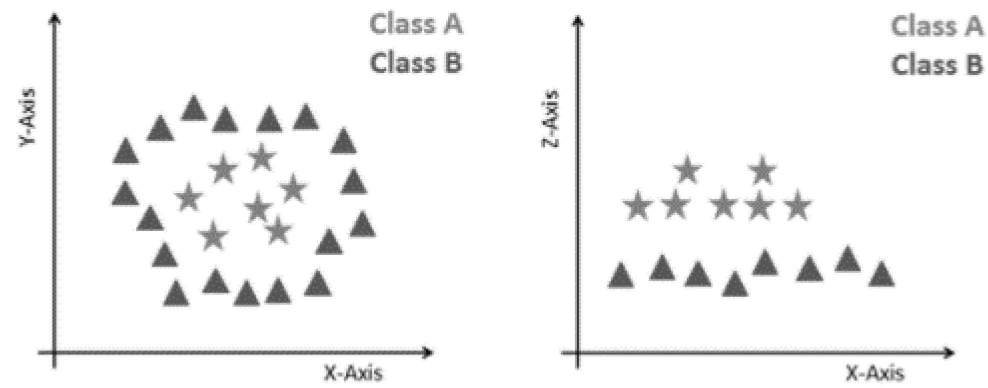


Figure 4.1.3: Managing nonlinear and indistinguishable planes

The Svm classifier is used in action by the kernel. The data input space can be translated into form by a kernel. SVM uses a technique for the kernel trick. The kernel takes a small entrance room and makes it bigger. In other words, it can be supposed that it turns an unrelated problem into separable ones by adding more dimensions. It is most useful for non-linear separation problems. You can make a precise classification by using the kernel trick.

Any two acts taken can be used as a linear kernel as a regular dot product. The product between the two vectors represents the number multiplied by every input pair. It would be shown as mathematical expression:

$$\mathbf{K(x, xi) = sum(x * xi)}$$

A more prominent linear kernel form will be the polynomial kernel. The polynomial kernel distinguishes between both nonlinear and curved input space.

$$\mathbf{K(x,xi) = 1 + sum(x * xi)^\wedge d}$$

Where d is the level of polynomial. d=1 is similar to that a functional realization. The grade must be set manually in the learning algorithm.

The radial kernel is a prevalent kernel function often used for support vector classification. RBF will restore infinite spatial dimensions.

K(x,xi) = exp(-gamma * sum((x – xi^2))

In this case, gamma is a constant of 0 to 1. A greater gamma value equals the training data collection, which results in needless adjustment. A strong predictable value is Gamma=0.1. Gamma=0.1. The gamma value must be specified independently in the learning algorithm. SVM groups have fair accuracy and quicker estimation in comparison with the Navy Bayes algorithm. They use less memory and use a subset of instructional points in decision-making. The SVM suits well for a large and powerful separation margin.

4.2 Random Forest Classifier:

Indeed, random forest algorithm are mostly a monitoring research algorithm. And can be used along with grouping for regression. The most robust and quick to use algorithm might be this one. In a forest there are trees. Plus they have leaves, cooler is a woodland. Random forests produce randomly selected samples for the decision-making trees, accurate per tree and by votes choose the best solution. The value of the function is also very well indicated. Random forests use a wide variety of applications, involving systems for recommendations, image processing and feature selection. It's being used to categorize loyal borrowers, to diagnose and forecast fraud. The Random Forest algorithm is based on the key features selected for a data set.

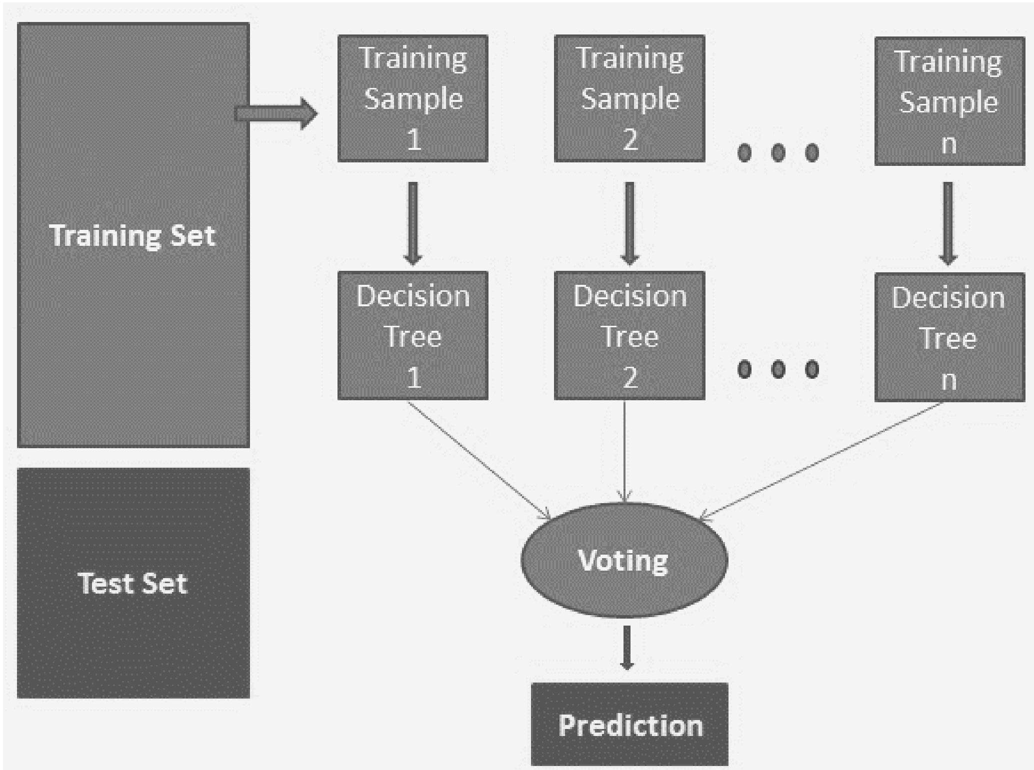


Figure 4.2.1: Steps of how Random forest classifier works

1) Choose a random sample dataset.

2) Create a decision tree and create a projection for each sample from each decision tree.

Check each predicted result to hold a referendum.

3) Choose the result as the general projection of the most votes.

The number of decision-making trees involved in this process makes it an unbelievably detailed and stable method for the random woods. There is no question of over-supply. The main reason is that all forecasts take the average and remove the forecasts. Both algorithms can be used for the classification and regression problems. Random forests will also handle the outliers. Median values are used in two ways to override the continuous variables and to calculate the average weighted approximation of missing values. Will obtain the relative value of the function to assign parameters for the classifier. The procedure suggested for random wood used the general approach of bootstrap for tree topics. In the case of a workout data set X=x1... x n and the answers Y=y1,...,yn boosting selects the random sample and substitutes the workout set by the sample tree. After processing, x' unsightly sample estimates can be done by averaging x' for the whole structure of the tree on x':

$$\hat{f} = \frac{1}{B}\sum_{b-1}^{B} f_b(x')$$

4.3 K nearest Neighbor (KNN):

The KNN is one of the most simple, easily interpreted, flexible and cutting-edge algorithms. The KNN is an easy-to-interpret. KNN is used to recognize and recognize economics, biology, political science, handwriting and visuals.

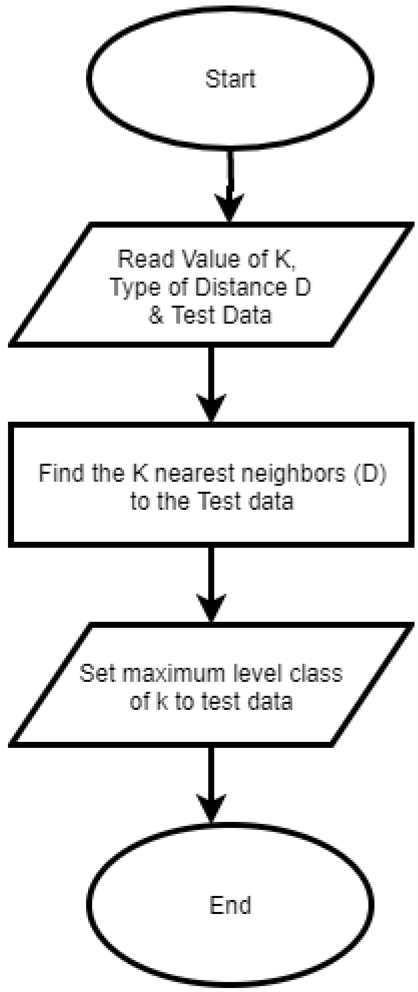


Figure 4.3.1: KNN flowchart

Euclidean metric equation:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \cdots + (x_n - x'_n)^2}$$

Probability:

$$P(Y = j | X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j)$$

In credit ratings, financial companies can predict consumer credit. Banking institutions anticipate whether credit disbursements are safe or risky. In political science, potential nominees in two groups can vote or not. In regression and classification problems, the KNN algorithm is used. KNN based on a feature similarity approach. The study algorithm is non-parametric and uncomfortable. Non-parametric means are not taken into account for the associated data distribution. The data collection represented the model setup. This is also very helpful if the majority of real-time data sets fail to comply with statistical scientific standards. Ignorant algorithm guarantees that the model does not include training data centers. All data used in the preparation testing phase. This reduces teaching and research costs and speed. Time and memory checking means expensive. In the worst case, it would take more time to scan for all data points and to search for information stored in all data points. K in KNN is the number of direct neighbors. The most important element is the number of neighbors. K is usually a weird number if the number of classes is 2. The algorithm is the nearest algorithm for K=1. This is the simplest situation. Assume that P1 is the label to anticipate. You first noticed the closest point to P1 and the label next to P1.

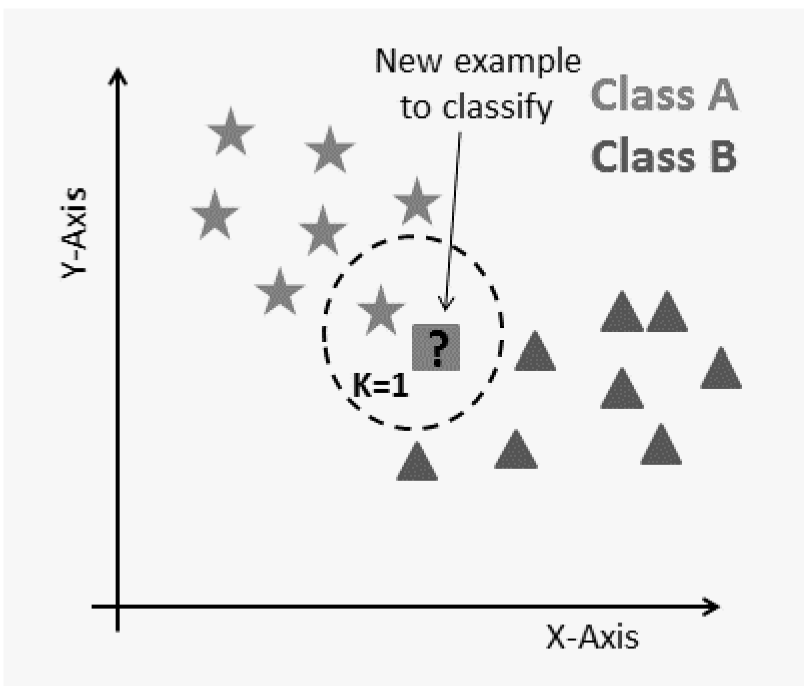


Figure 4.3.2: The process of KNN algorithm

P1 is the label to be planned. Secondly, you select the k closest to P1 and then you classify points by the voting majority of their k neighbors. The provision shall be taken for the most votes for each object for its class and for the class. Distance dimensions like Euclidean, Hamming and Manhattan and Minkowski are used for the distance between points when the closest points are closest to each other. KNN shall take the following basic steps:

a) Estimated range

b) Near neighborhood

c) Labeling vote

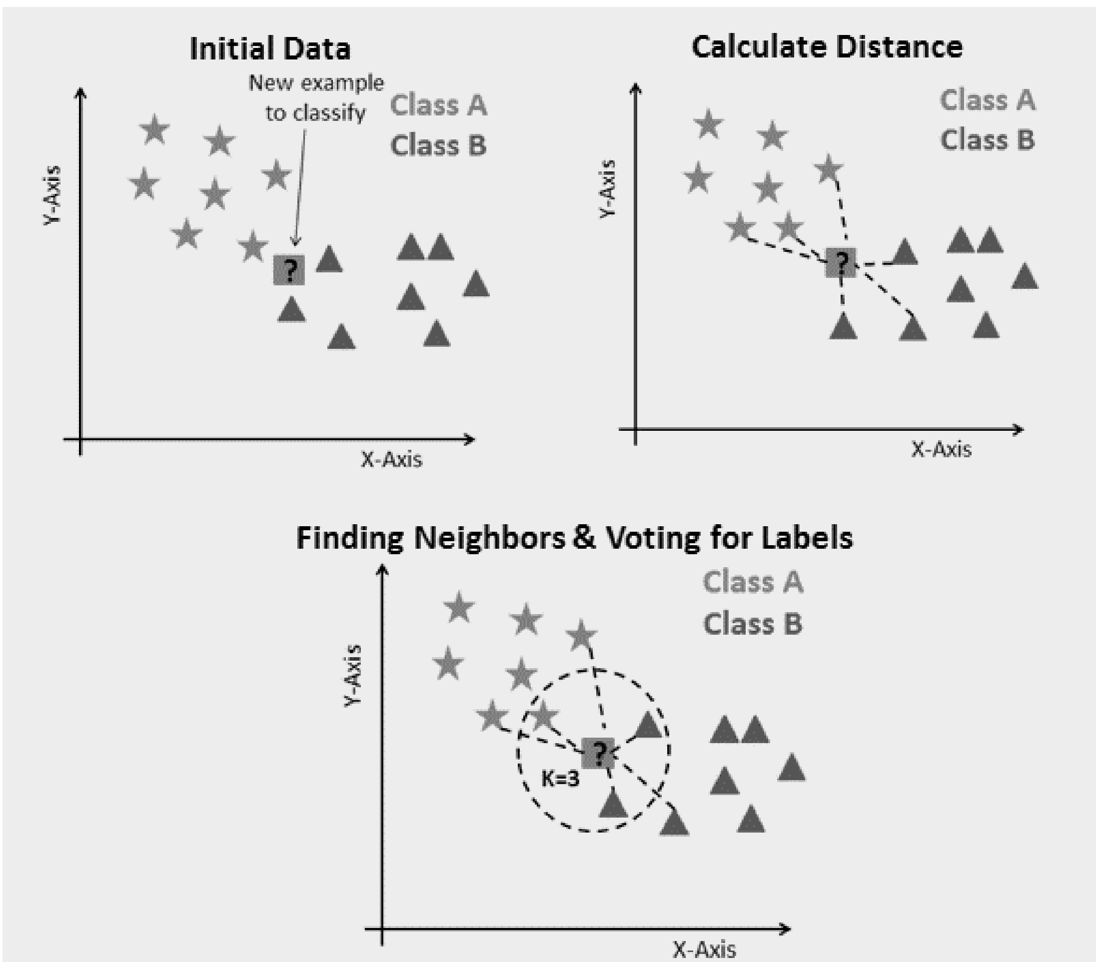


Figure 4.3.3: The process of KNN algorithm

K as a vector governing projection model. Both data sets don't have flawless neighbors. Each dataset has its own criteria. The disturbance would have a greater effect on the outcome with a small number of neighbors, which would make a substantial number of neighbors computationally costly. Research also has shown that the smoother decision cap with a larger number of nodes is the most flexible and has a low bias but a high difference, and a large number of neighbors means a smaller but more biased variation. Generally, data analysts get an unlikely amount only if the number of parties is even.

$$\mathrm{d}(\mathrm{p},\mathrm{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots (q_n - p_n)^2}$$

$$\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

4.4 Naive Bayes:

The Naive Bayes rating algorithm is the most simple and efficient of a large range of knowledge. Naive Bayes classification is used successfully in a variety of applications, including spam handling, document classification, and emotional perception and recommendation systems. The probability theorem for undefined prediction in Bayes is included.

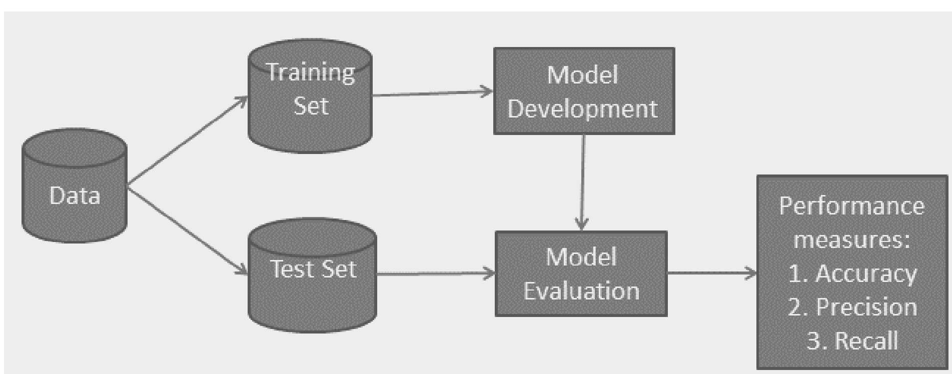


Figure 4.4.1: Naïve Bayes working process

The division comprises a review and assessment process in two phases. During the learning process, the grader trains its model on a specified data set and monitors its performance during the evaluation phase. Effects assessed for various metrics including accuracy, error, accuracy and memory. Naive Bayes is a mathematical classification system based on theorem from Bayes. This is one of the simplest managed algorithms. Naive Bayes is the fast, exact and trustworthy algorithm. On big datasets, Naive Bayes high precision and speed classifiers are feasible.

Phase 1: Pre-value assessments of some class marks

Phase 2: Find the opportunity for each attribute of the class

Phase 3: Place this value and decide the probability for future use of the Bayes Formula.

Phase 4: Determine which class is most likely as input is in the higher probability class.

The Naive Bayes classification scheme assumes that the power of a certain function in a particular class is independent of other characteristics. For example, a loan seeker is appropriate or not depending on employment, previous experiences, credit age and place of business. And while these characteristics are interdependent, they are viewed separately. This assumption makes it possible to program and is thus considered naive. This is called liberation of the class.

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

P (h): hypothesis likelihood (regardless of the data). This is called the previous h choice.

P (D): the opportunity for data (regardless of the hypothesis). This is known as the prior chance.

P (h|D): high-life hypothesis provided the data of D. This is called a subsequent chance.

P (D|h): mere data option d, since the hypothesis h is applicable. There is a subsequent probability.

4.5 Logistic Regression:

LR is a binary-class prediction statistical method. The outcome or variable is in essence dichotomous. That means only two classes can be taught. TRUE = 1 or FALSE = 0. Logistics regression is similar to linear regression. We may assume that as an overall linear construct. Linear Regression Equation.

$$y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_n X_n$$

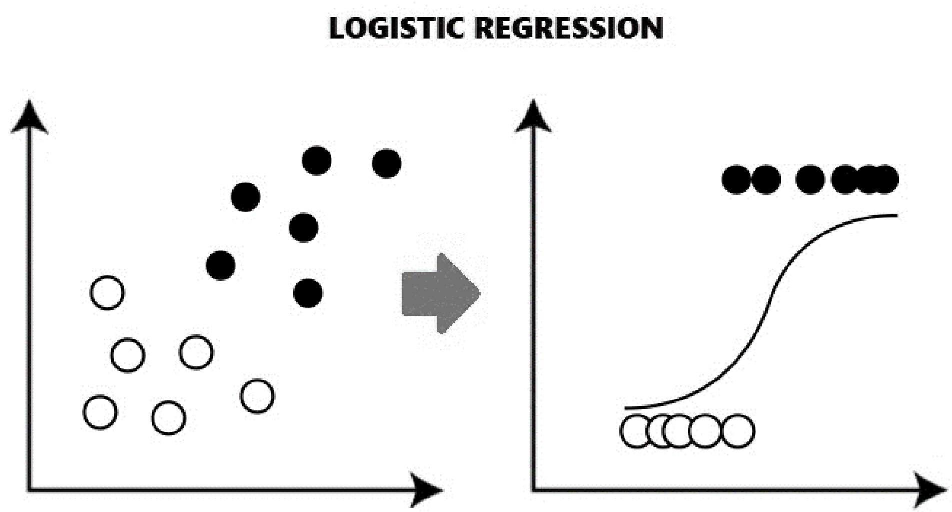


Figure 4.5.1: Logistic Regression

Sigmoid function: $p = \frac{1}{1+e^{-y}}$

Final equation after applying sigmoid function:

$$p = \frac{1}{1 + e^{-(0+\beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_n X_n)}}$$

4.6 Decision Tree:

A decision tree is a flow-chart tree structure in which the internal node represents a process (or attribute), a branch is a decision-making right, and a result each node of a leaf represents. The highest node of a decision tree is called the root node. It is dependent on the attribute's meaning. It divides the tree into regular partitions recurrently. You will make your choices in this fluid diagram shape. It is a show that imitates people's thinking conveniently, like a diagram. This is why policy makers are easy to grasp.

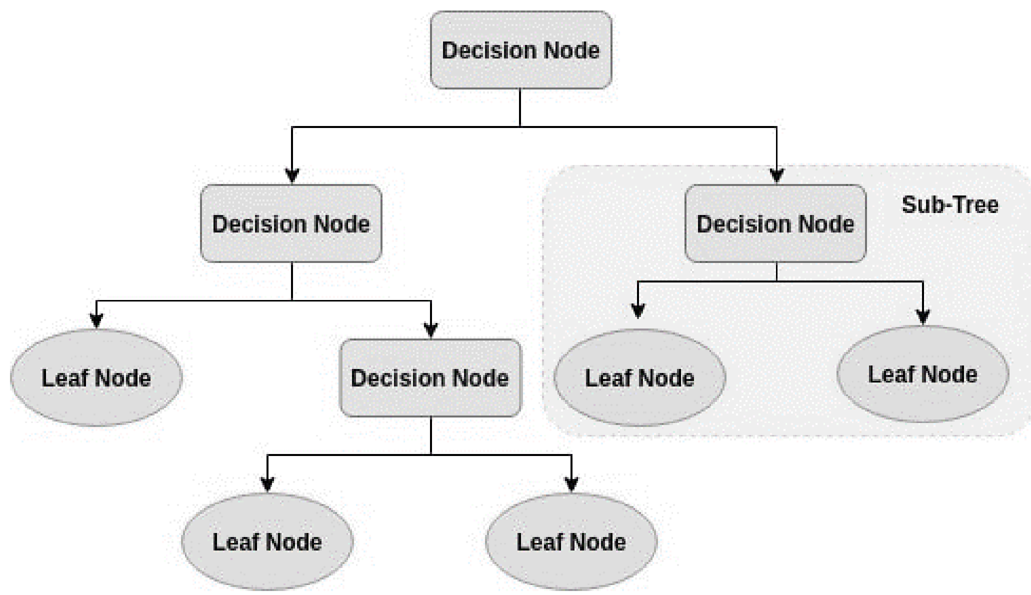


Figure 4.6.1: Decision tree

Tree Decision is a white storage box ML algorithms. The algorithms such as the neural grid have no black box structure and have an inherent decision - making logic. It has a shorter training time than the neural network algorithm. The time complexity of the decision tree has a significant effect on the number of records and the number of features included in the data. The decision tree is a non-parametric process that is not based on probability expectations. Decision trees can accommodate high-dimensional effects with reasonable precision. To separate documents by using an ASM choose the best function. Create an attribute of the decision node and divide it into a thread. Starts building a tree by repeating this process recursively for every child until there is one condition: Both tuples are of the same value. There are no more features left. No more occurrences.

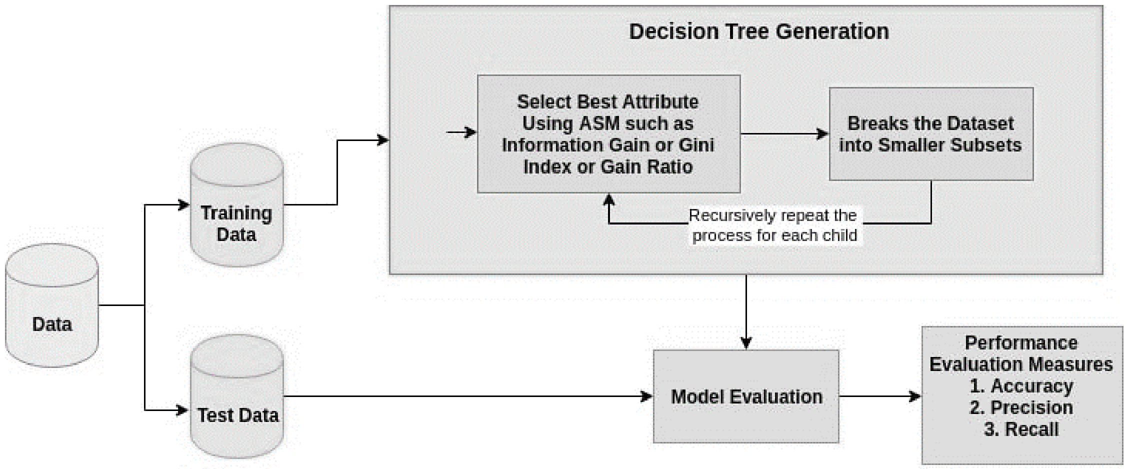


Figure 4.6.2: Working process of Decision Tree

Attribute Selection Measure:

The selection measure of the characteristics is a heuristic approach for choosing the partitioning criteria in the best way possible. The splitting rules are also known as it aids in the detection of splitting points for tuples at a specific node. By describing the given dataset, ASM lists a rank for each feature(s). The best score is chosen as a dividing characteristic. For an underway attribute, split points for branches must also be established. Knowledge benefit, gain ratio and Gini index are the most commonly used performance behavior.

Information Gain:

The entropy term developed which measures the insertion impurity. Entropy in physics and mathematics is considered randomness or impurity of the unit. In a group of instances in information theory, it refers to impurity. The decrease in entropy is the reward. The data gain is dependent on the values of the attribute, which determines the difference between entropy and average entropy after a dataset partition. The ID3 (Iterative Dichotomies) algorithm takes the advantage of the information.

$$Info(D) = \Sigma_{i=1}^{m} P^i \, log_2 \, pi$$

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} X \; Info(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$\Delta \, \mathrm{Gini}(\mathrm{A}) \; = \; \mathrm{Gini}(\mathrm{D}) - Gini_a(D)$$

CHAPTER 5

Experimental Results and Discussion

5.1 Introduction

We will discuss the experimental results in this segment. All six algorithms we are going to use have been discussed. We can now see how well those algorithms were doing with precision. And all six algorithms compare the accuracy.

The steps we take to finish this study are:

Step 1: Dataset Collection

Step 2: Pre-processing data set

Step-3: Import various libraries and other items required

Step 4: Divide our dataset

Step 5: Build all 6 algorithms for models

Step 6: Train the 6 different algorithms for machine learning

Step-7: Find all the algorithms for accuracy

Step-8: Compare the best outcome of which algorithm

This are the steps we take to finish this study.

5.2 Experimental Results

We know that no computer can provide us with 100% output. Likewise, we should train our system and adjust those parameters to make it more precise and properly train. However, the accuracy from various algorithms is very high.

Below are some photographs that reflect our study in a short form.

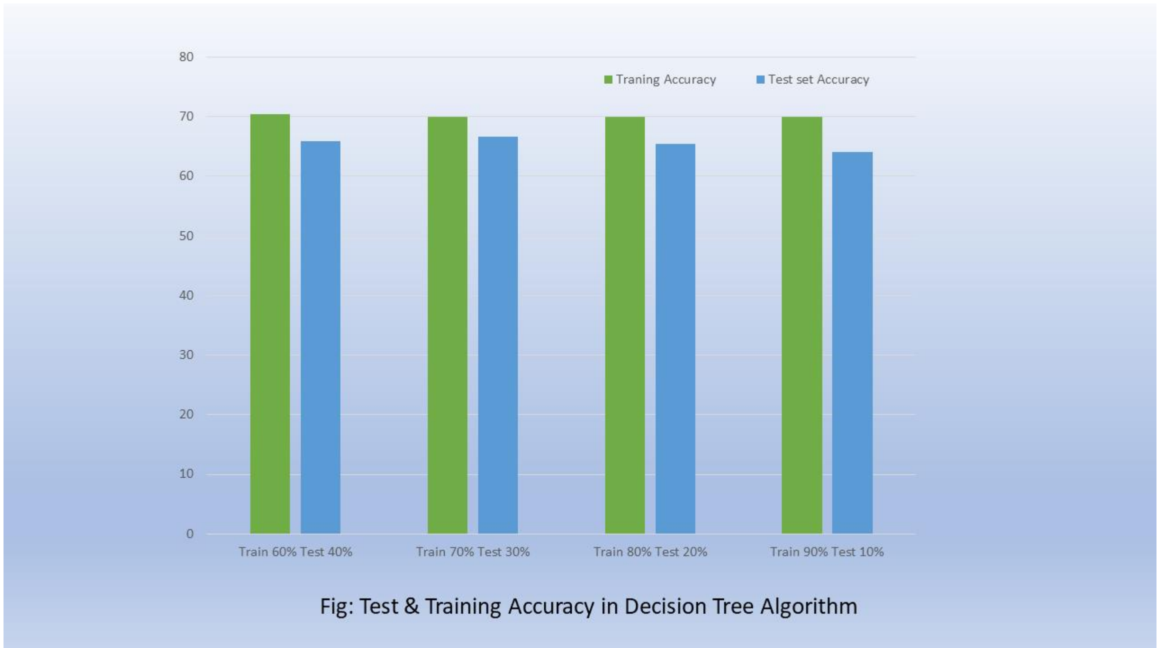


Figure 5.2.1 Decision Tree Algorithm (2019)

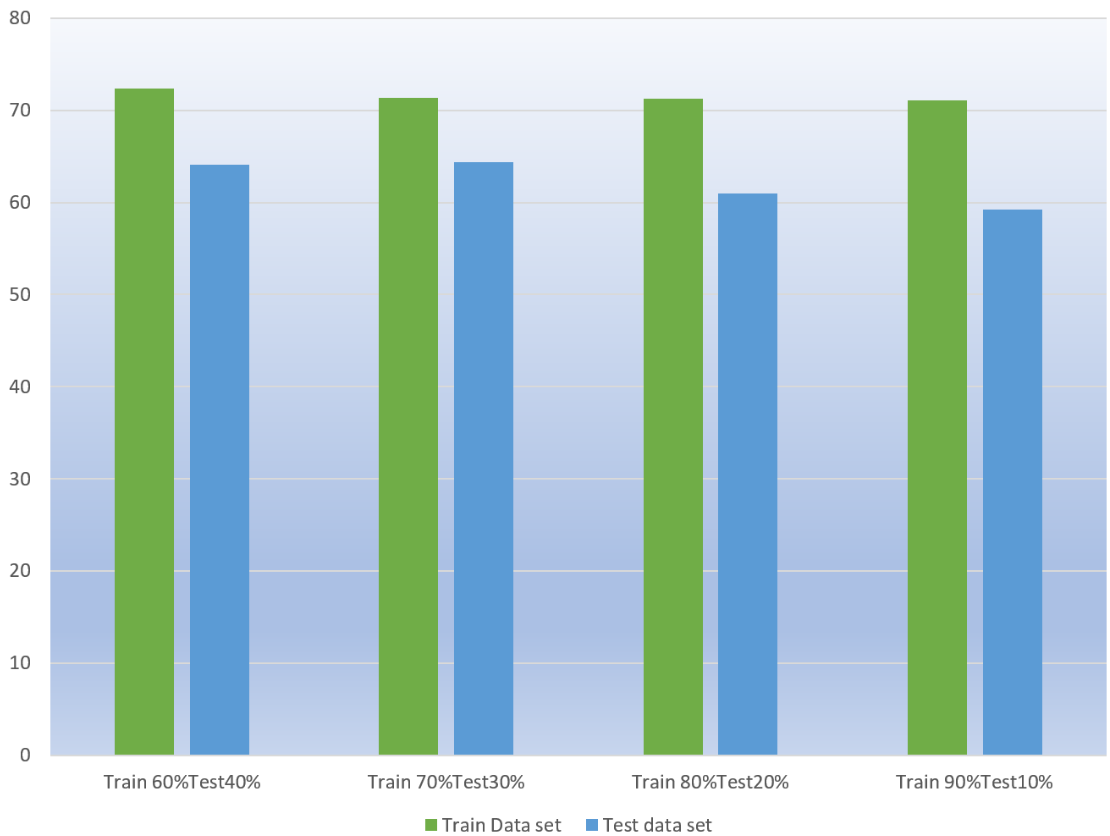


Figure 5.2.2 Decision Tree Algorithm (2015)

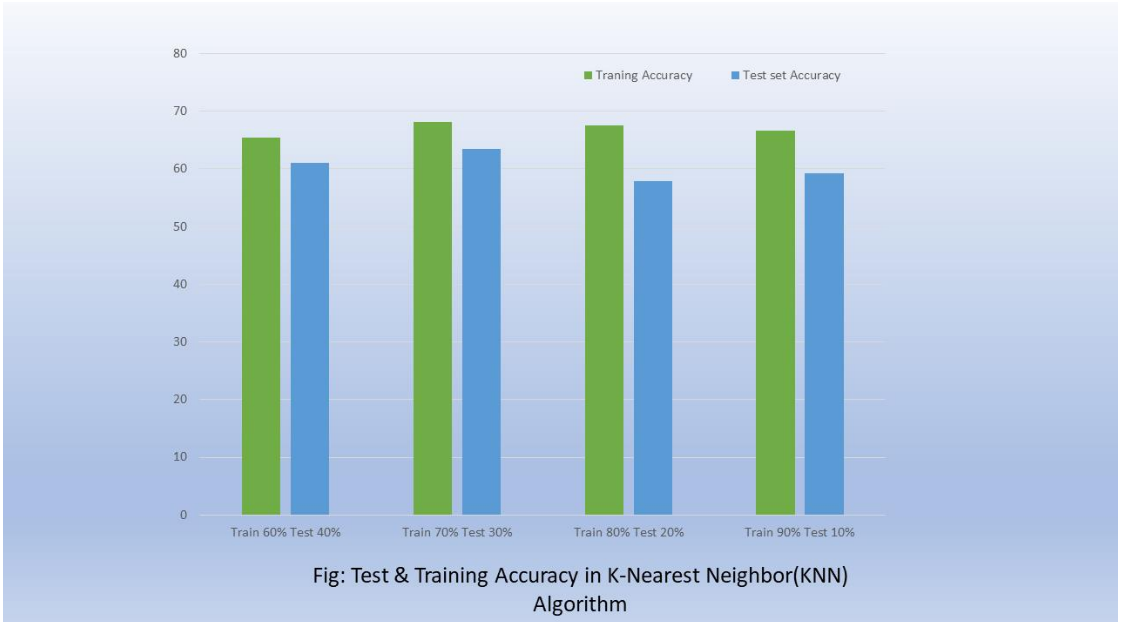


Figure 5.2.3 KNN Algorithm (2019)

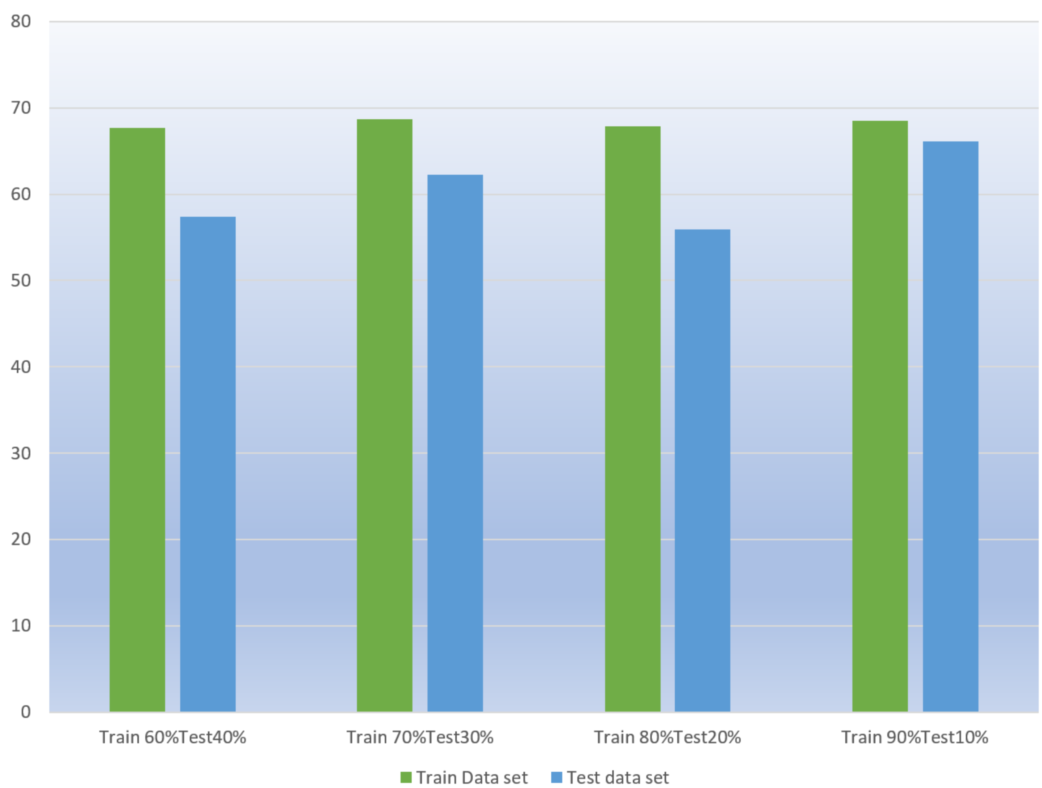


Figure 5.2.4 KNN Algorithm (2015)

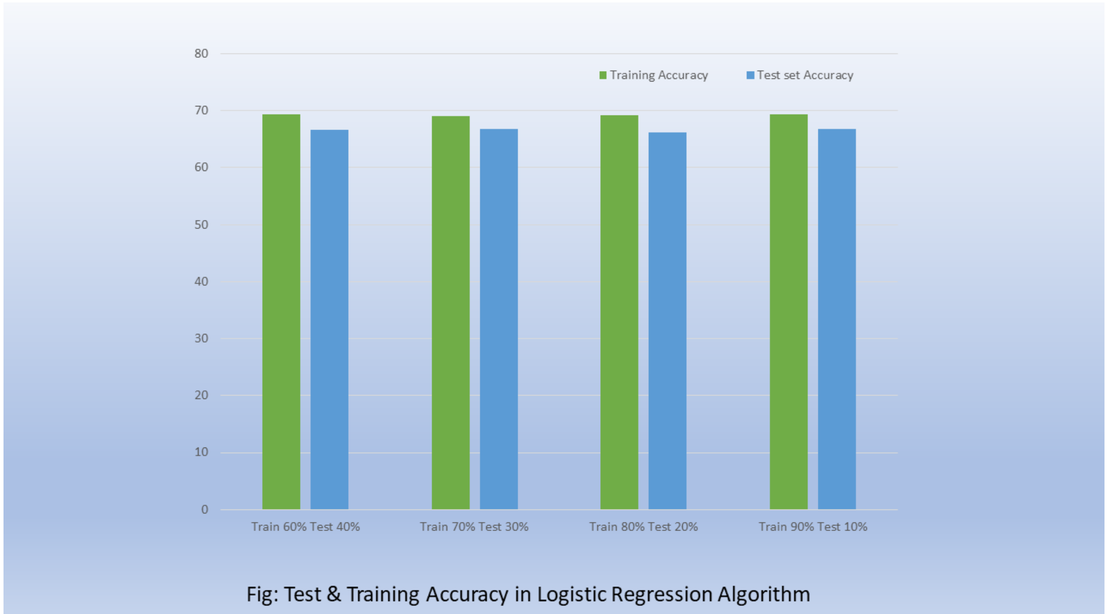


Fig: Test & Training Accuracy in Logistic Regression Algorithm

Figure 5.2.5 Logistic Regression Algorithm (2019)

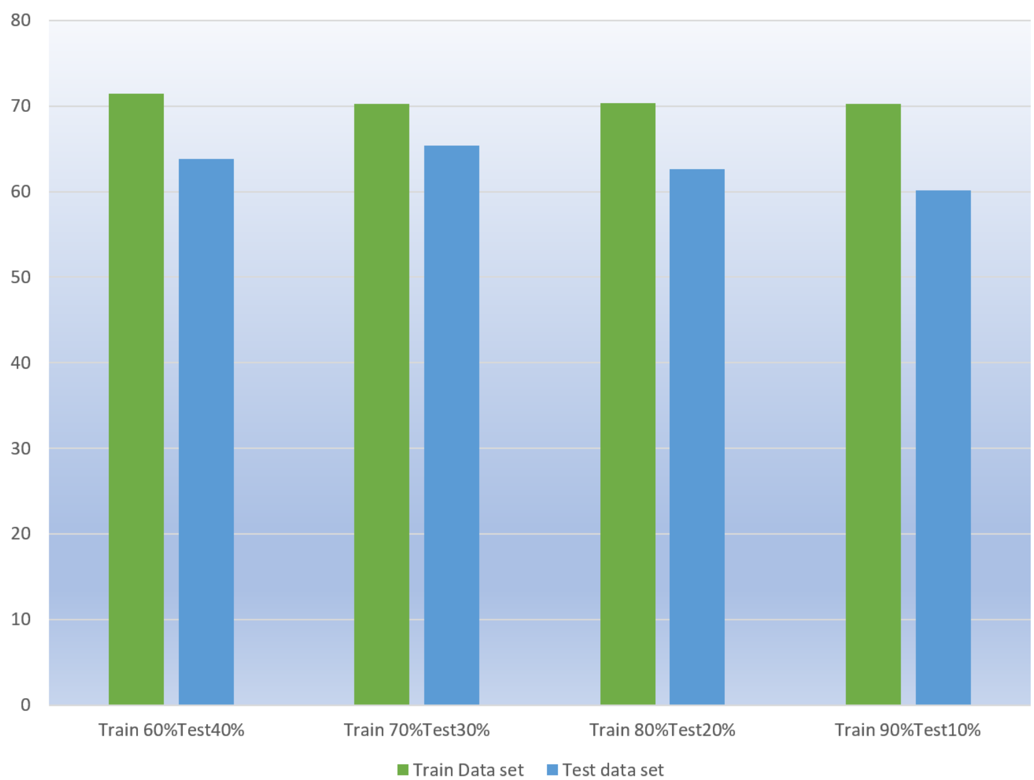


Figure 5.2.6 Logistic Regression Algorithm (2015)

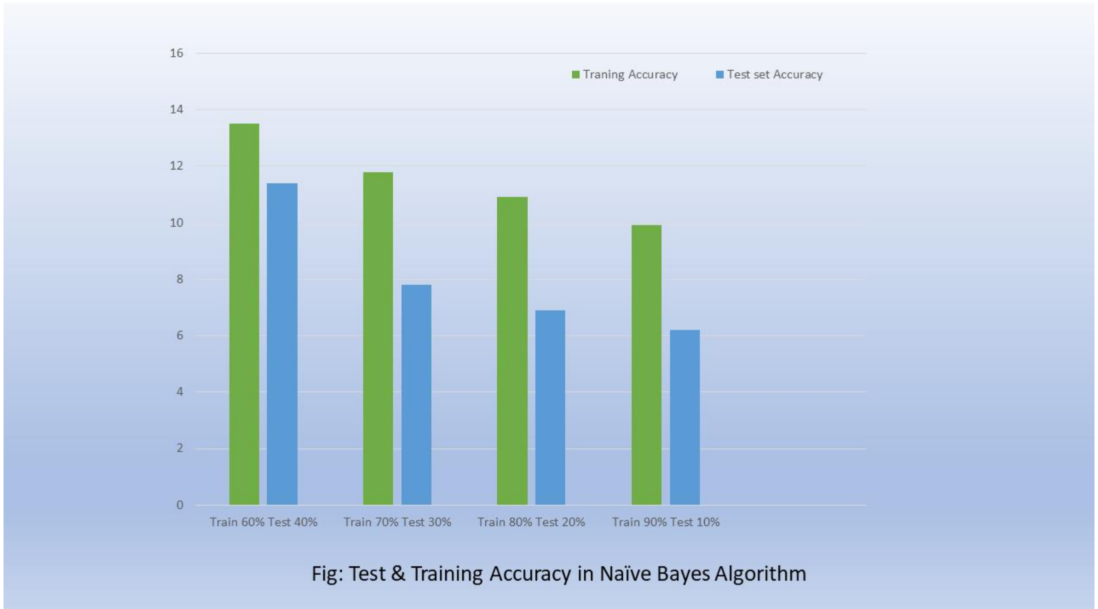


Figure 5.2.7 Naïve Bayes Algorithm (2019)

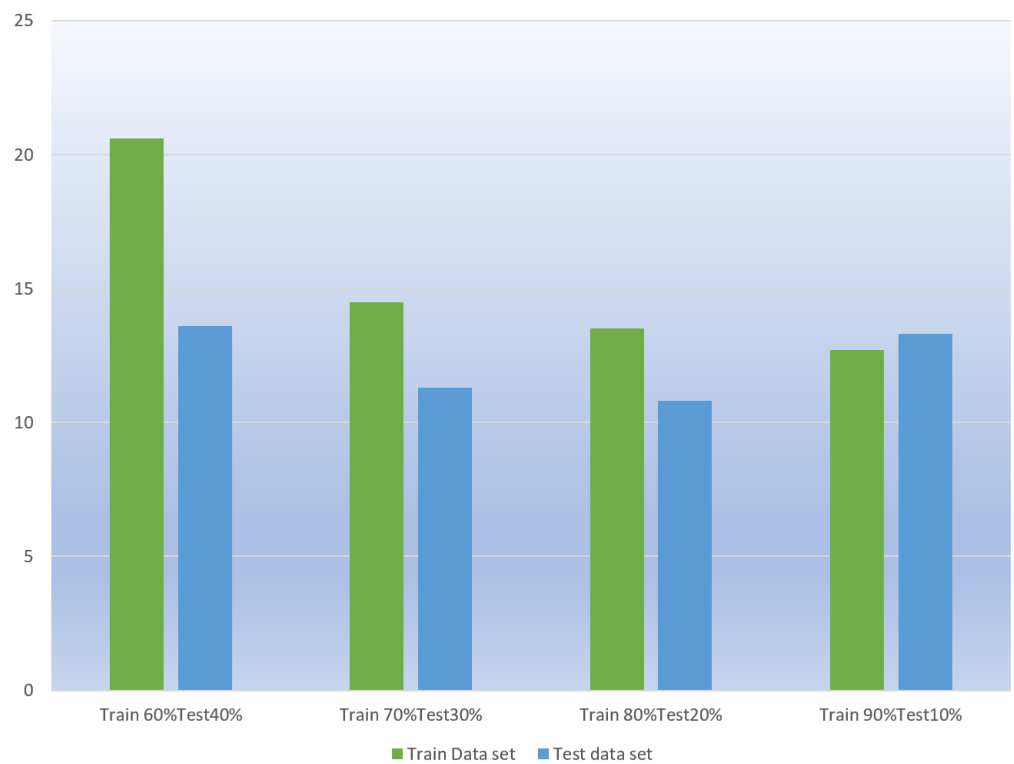


Figure 5.2.8 Naïve Bayes Algorithm (2015)

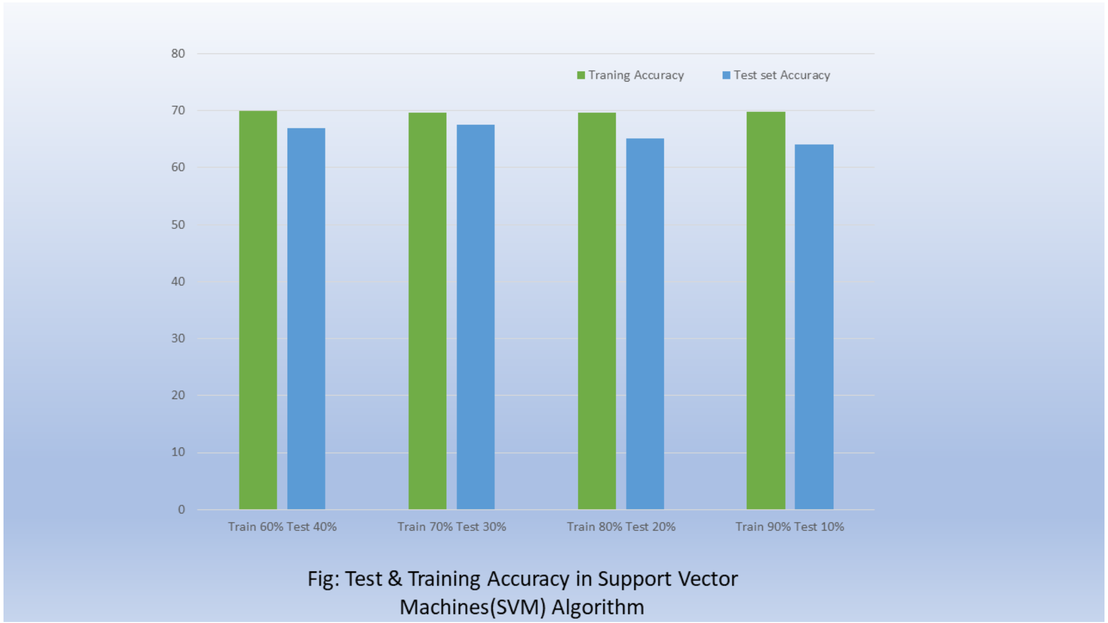


Figure 5.2.9 SVM Algorithm (2019)

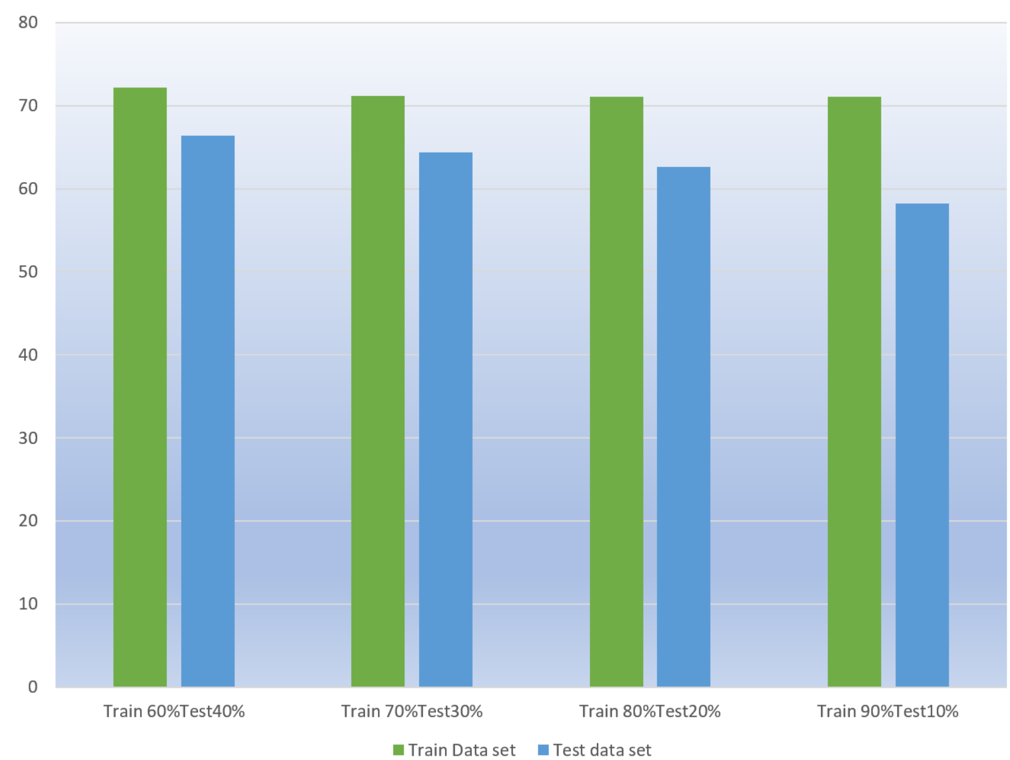


Figure 5.2.10 SVM Algorithm (2015)

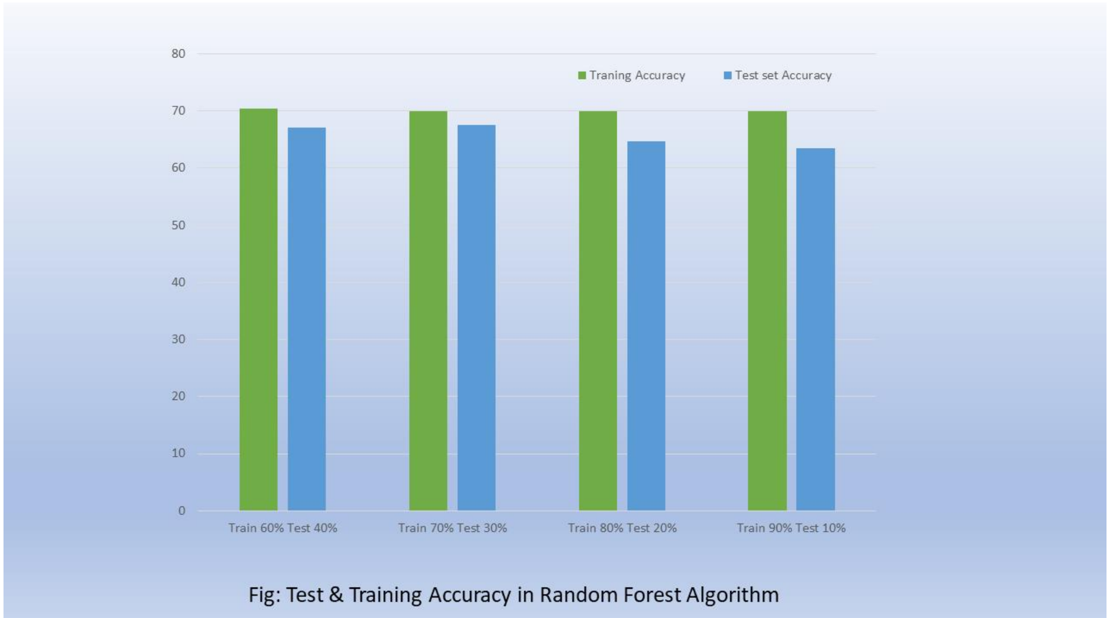


Fig: Test & Training Accuracy in Random Forest Algorithm

Figure 5.2.11 Random Forest Algorithm (2019)

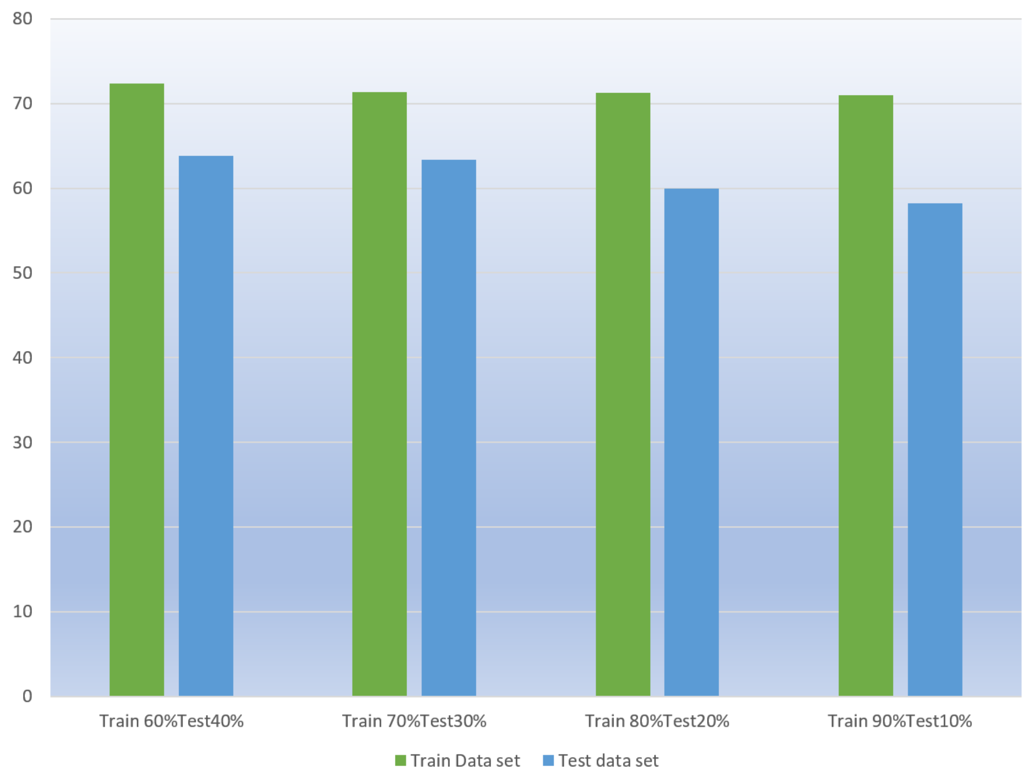


Figure 5.2.12 Random Forest Algorithm (2015)

Figure 5.2.13 Training accuracy of different Algorithm (2019)

Figure 5.2.14 Training accuracy of different Algorithm (2015)

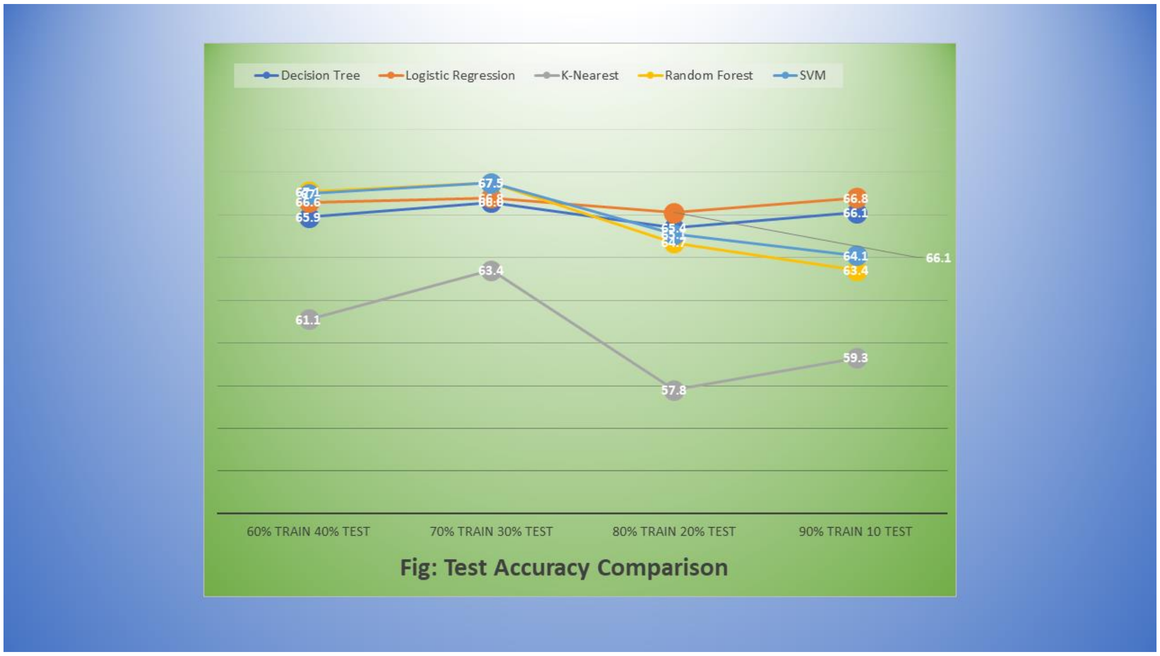


Figure 5.2.15 Test accuracy of different Algorithm (2019)

Figure 5.2.15 Test accuracy of different Algorithm (2015)

TABLE 1: RESULTS OF DIFFERENT ALGORITHMS FOR 2019 DATA

Classifier		Accuracy			
		60% train 40% test	70% train 30% test	80% train 20% test	90% train 10% test
Decision Tree Algorithm	Training set Accuracy	70.4	70	69.9	69.9
	Test set Accuracy	65.9	66.6	65.4	66.1
Logistic Regression Algorithm	Training set Accuracy	69.4	69.1	69.2	69.4
	Test set Accuracy	66.6	66.8	66.1	66.8
K-Nearest Neighbor	Training set Accuracy	65.4	68.2	67.6	66.6
	Test set Accuracy	61.1	63.4	57.8	59.3
Random Forest	Training set Accuracy	70.4	70	69.9	69.9
	Test set Accuracy	67.1	67.5	64.7	63.4
Support Vector Machines	Training set Accuracy	70	69.7	69.7	69.8
	Test set Accuracy	67	67.5	65.1	64.1
Naïve Bayes Algorithm	Training set Accuracy	13.5	11.8	10.9	9.9
	Test set Accuracy	11.4	7.8	6.9	6.2

We can see the accuracy of 6 different algorithms. All the algorithms are showing a good accuracy. But Decision Tree and Random Forest these two algorithms are showing 70.4% accuracy. So if we compare all 6 algorithms we can say that Decision Tree and Random Forest can predict or detect Cricket match wining prediction most accurately.

TABLE 2: RESULTS OF DIFFERENT ALGORITHMS FOR 2015 DATA

Classifier		Accuracy			
		Test 60% Train 40%	Test 70% Train 30%	Test 80% Train 20%	Test 90% Train 10%
Decision Tree Algorithm	Training set Accuracy	72.4	71.4	71.3	71.1
	Test set Accuracy	64.1	64.4	61	59.2
Liner Regression Algorithm	Training set Accuracy	71.5	70.3	70.4	70.3
	Test set Accuracy	63.8	65.4	62.6	60.2
K Nearest Neighbor Algorithm	Training set Accuracy	67.6	68.7	67.9	68.5
	Test set Accuracy	57.4	62.3	55.9	54.1
Random Forest Algorithm	Training set Accuracy	72.4	71.4	71.3	71
	Test set Accuracy	63.8	63.4	60	58.2
Support Vector Machine Algorithm	Training set Accuracy	72.2	71.2	71.1	71.1
	Test set Accuracy	64.4	64.4	62.6	59.2
Navies Bayes Algorithm	Training set Accuracy	20.6	14.5	13.5	12.7
	Test set Accuracy	13.6	11.3	10.8	13.3

We can see the accuracy of 6 different algorithms. All the algorithms are showing a good accuracy. But Decision Tree and Random Forest these two algorithms are showing 72.4% accuracy. So if we compare all 6 algorithms we can say that Decision Tree and Random Forest can predict or detect Cricket match wining prediction most accurately.

There some other figures we can include to give some more idea about our research work.

```
In [27]: #group matches
         predictions = rf.predict(pred_set)
         for i in range(fixtures.shape[0]):
             print(backup_pred_set.iloc[i, 1] + " and " + backup_pred_set.iloc[i, 0])
             if predictions[i] == 1:
                 print("Winner: " + backup_pred_set.iloc[i, 1])

             else:
                 print("Winner: " + backup_pred_set.iloc[i, 0])
             print("")

         South Africa and England
         Winner: England

         West Indies and Pakistan
         Winner: Pakistan

         Sri Lanka and New Zealand
         Winner: New Zealand

         Afghanistan and Australia
         Winner: Australia

         Bangladesh and South Africa
         Winner: South Africa

         Pakistan and England
         Winner: England

         Afghanistan and Sri Lanka
```

Figure 5.2.16: Predict winner of Group matches in world cup 2019

```
In [41]: clean_and_predict(semi, ranking, final, rf)

         New Zealand and India
         Winner: India

         South Africa and England
         Winner: England
```

Figure 5.2.17: Predicting the semifinal (2019)

```
In [23]: # Finals
         finals = [('India', 'England')]

In [24]: clean_and_predict(finals, ranking, final, rf)

         India and England
         Winner: England

In [ ]:
```

Figure 5.2.18: Final winner Prediction (2019)

```
In [38]: results.head()
         print(results.to_string())
```

```
          date      Team_1        Team_2      Winner     Margin        Ground
0      4-Jan-10  Bangladesh     Sri Lanka   Sri Lanka  7 wickets         Dhaka
1      5-Jan-10       India     Sri Lanka   Sri Lanka  5 wickets         Dhaka
2      7-Jan-10  Bangladesh         India       India  6 wickets         Dhaka
3      8-Jan-10  Bangladesh     Sri Lanka   Sri Lanka  9 wickets         Dhaka
4     10-Jan-10       India     Sri Lanka       India  8 wickets         Dhaka
5     11-Jan-10  Bangladesh         India       India  6 wickets         Dhaka
6     13-Jan-10       India     Sri Lanka   Sri Lanka  4 wickets         Dhaka
7     22-Jan-10   Australia      Pakistan   Australia  5 wickets      Brisbane
8     24-Jan-10   Australia      Pakistan   Australia   140 runs        Sydney
9     26-Jan-10   Australia      Pakistan   Australia    40 runs      Adelaide
10    29-Jan-10   Australia      Pakistan   Australia   135 runs         Perth
11    31-Jan-10   Australia      Pakistan   Australia  2 wickets         Perth
12     5-Feb-10 New Zealand    Bangladesh New Zealand   146 runs        Napier
13     7-Feb-10   Australia   West Indies   Australia   113 runs     Melbourne
14     8-Feb-10 New Zealand    Bangladesh New Zealand  5 wickets       Dunedin
15     9-Feb-10   Australia   West Indies   Australia  8 wickets      Adelaide
16    11-Feb-10 New Zealand    Bangladesh New Zealand  3 wickets  Christchurch
17    12-Feb-10   Australia   West Indies   no result        NaN        Sydney
```

Figure 5.2.19: Showing data from 2010 to 2015 before world cup

```
In [65]: # Create new columns with ranking position of each team
         fixtures.insert(1, 'first_position', fixtures['Team_1'].map(ranking.set_index('Team')['Position']))
         fixtures.insert(2, 'second_position', fixtures['Team_2'].map(ranking.set_index('Team')['Position']))
         # We only need the group stage games, so we have to slice the dataset
         fixtures = fixtures.iloc[:45, :]
         print(fixtures.to_string())
```

```
    Round Number  first_position  second_position       date      Location        Team_1        Team_2 Group  Result
0              1               4                5  14-Feb-15  Christchurch   New Zealand     Sri Lanka     A     NaN
1              1               3                8  14-Feb-15     Melbourne     Australia       England     A     NaN
2              1               2               10  15-Feb-15      Hamilton  South Africa      Zimbabwe     B     NaN
3              1               1                6  15-Feb-15      Adelaide         India      Pakistan     B     NaN
4              1              12                9  16-Feb-15        Nelson       Ireland   West Indies     B     NaN
5              1               4               13  17-Feb-15       Dunedin   New Zealand      Scotland     A     NaN
6              1              11                7  18-Feb-15      Canberra   Afghanistan    Bangladesh     A     NaN
7              1              14               10  19-Feb-15        Nelson        U.A.E.      Zimbabwe     B     NaN
8              1               4                8  20-Feb-15    Wellington   New Zealand       England     A     NaN
9              1               3                7  21-Feb-15      Brisbane     Australia    Bangladesh     A     NaN
10             1               6                9  21-Feb-15  Christchurch      Pakistan   West Indies     B     NaN
11             1              11                5  22-Feb-15       Dunedin   Afghanistan     Sri Lanka     A     NaN
12             1               1                2  22-Feb-15     Melbourne         India  South Africa     B     NaN
13             1               8               13  23-Feb-15  Christchurch       England      Scotland     A     NaN
14             1               9               10  24-Feb-15      Canberra   West Indies      Zimbabwe     B     NaN
15             1              12               14  25-Feb-15      Brisbane       Ireland        U.A.E.     B     NaN
16             1              11               13  26-Feb-15       Dunedin   Afghanistan      Scotland     A     NaN
17             1               7                5  26-Feb-15     Melbourne    Bangladesh     Sri Lanka     A     NaN
18             1               2                9  27-Feb-15        Sydney  South Africa   West Indies     B     NaN
19             1               4                3  28-Feb-15      Auckland   New Zealand     Australia     A     NaN
20             1               1               14  28-Feb-15         Perth         India        U.A.E.     B     NaN
```

Figure 5.2.20: Showing pool A and B fixtures of 2015 world cup

```
In [321]: #group matches
predictions = rf.predict(pred_set)
for i in range(fixtures.shape[0]):
    print(backup_pred_set.iloc[i, 1] + " and " + backup_pred_set.iloc[i, 0])
    if predictions[i] == 1:
        print("Winner: " + backup_pred_set.iloc[i, 1])

    else:
        print("Winner: " + backup_pred_set.iloc[i, 0])
    print("")
```

```
Sri Lanka and New Zealand
Winner: New Zealand

England and Australia
Winner: Australia

Zimbabwe and South Africa
Winner: South Africa

Pakistan and India
Winner: India

Ireland and West Indies
Winner: West Indies

Scotland and New Zealand
Winner: New Zealand
```

Figure 5.2.21: Showing group stage results (2015)

```
# List of tuples before
knockout = [('Australia', 'West Indies'),
            ('Sri Lanka', 'South Africa'),
            ('New Zealand', 'Pakistan'),
            ('Bangladesh', ' India')]
```

```
Australia and West Indies
Winner: Australia

Sri Lanka and South Africa
Winner: South Africa

New Zealand and Pakistan
Winner: New Zealand

Bangladesh and India
Winner: India
```

Figure 5.2.22: Knockout stage results (2015)

```
# List of tuples before
Semi = [('Australia', 'South Africa'),
            ('New Zealand', ' India')]
```

```
Australia and South Africa
Winner: Australia

New Zealand and India
Winner: India
```

Figure 5.2.23: Semi Final Results (2015)

```
# Finals
finals = [('Australia', ' India')]
```

```
clean_and_predict(finals, ranking, final, rf)
```

```
Australia and India
Winner: Australia
```

Figure 5.2.24: Final Results (2015)

CHAPTER 6

Conclusion and Future Scope

6.1 Conclusion:

We mention our judgment and conclusion in this section. Here we examine the scope and possibilities of the future. In general, researchers observed that prototypes on cricket match predictions have been successfully developed using machine learning. But more and more avenues are still available to improve this problem. Different sectors and angles are still available. We are focusing on this.

6.2 Future Scope:

1. We will upgrade the algorithm so that Bangladesh or any team can select their best eleven depending on the all stats
2. Currently we are only working on world cups. In near future we will work on every single ODI.
3. It is possible to predict 2023 or later tournament just upgrading the datasets.
4. We will work on T20s too
5. It plays an important role if it is possible to select best players for a match and results in securing victory. An accurate prediction of most likely score by a batsman and best team selection helps a team to secure a victory. In this paper, we worked regarding to batting and bowling. We have not included other effects such as performances, nature of pitch and weather. In near future it will also be added.

References

[1] J. Kumar, R. Kumar and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 343-347, doi: 10.1109/ICSCCC.2018.8703301.

[2] A. N. Wickramasinghe and R. D. Yapa, "Cricket Match Outcome Prediction Using Tweets and Prediction of the Man of the Match using Social Network Analysis: Case Study Using IPL Data," 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2018, pp. 1-1, doi: 10.1109/ICTER.2018.8615563..

[3] H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier League," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9153972.

[4] T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), Faridabad, India, 2015, pp. 60-66, doi: 10.1109/ICSCTI.2015.7489605.

[5] M. M. Hatharasinghe and G. Poravi, "Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033698..

[6] Akhil Nimmagadda, Nidamanuri Venkata Kalyan, Manigandla Venkatesh, Nuthi Naga Sai Teja, Chavali Gopi Raju. **Cricket score and winning prediction using data mining**, International Journal of Advance Research, Ideas and Innovations in Technology, www.IJARnD.com.

[7] H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier League," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9153972.

[8] A. N. Wickramasinghe and R. D. Yapa Man of the Match using Social Network Analysis: Case Study Using IPL Data," 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2018, pp. 1-1, doi: 10.1109/ICTER.2018.8615563

Cricket Match Winning Prediction

ORIGINALITY REPORT

18%	14%	5%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	8%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.datacamp.com Internet Source	2%
4	Jalaz Kumar, Rajeev Kumar, Pushpender Kumar. "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks", 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018 Publication	2%
5	www.researchgate.net Internet Source	1%
6	Submitted to Rochester Institute of Technology Student Paper	1%
7	Harshit Barot, Arya Kothari, Pramod Bide, Bhavya Ahir, Romit Kankaria. "Analysis and	<1%

Prediction for the Indian Premier League", 2020 International Conference for Emerging Technology (INCET), 2020
Publication

8	www.ijarnd.com Internet Source	<1 %
9	"Forecasting the Outcome of the Next ODI Cricket Matches to Be Played", International Journal of Recent Technology and Engineering, 2019 Publication	<1 %
10	"Novel Method for Cricket Match Outcome Prediction using Data Mining Techniques", International Journal of Engineering and Advanced Technology, 2019 Publication	<1 %
11	Submitted to University of Westminster Student Paper	<1 %
12	Submitted to Liverpool John Moores University Student Paper	<1 %
13	Submitted to University of Edinburgh Student Paper	<1 %
14	ethesis.nitrkl.ac.in Internet Source	<1 %
15	scholarbank.nus.edu.sg Internet Source	<1 %

16	www.en8848.com.cn Internet Source		<1 %
17	zerobio.cs.dm.u-tokai.ac.jp Internet Source		<1 %