Design and Implementation of Statistical Estimation Based Model for Fair Assessment of Rain Interrupted Cricket Matches

M. Tech. Dissertation in MLIS
Department of Computer Science Engineering

Submitted By	**: Praveen D Chougale**
Reg. No.	**: 17ETCS075002**
Supervisors	**: Mr. Divya Kiran**

August – 2019

FACULTY OF ENGINEERING AND TECHNOLOGY

RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Bengaluru -560 054

FACULTY OF ENGINEERING AND TECHNOLOGY

FET

Certificate

*This is to certify that the Project titled “**Design and Implementation of Statistical Estimation Based Model for Fair Assessment of Rain Interrupted Cricket Matches**” is a bonafide record of the project work carried out by **Mr.Praveen D Chougale** bearing Reg. No. **17ETCS075002** Department of Computer Science Engineering, FT-2017 batch in partial fulfilment of requirements for the award of M.Tech Degree of M. S. Ramaiah University of Applied Sciences.*

August – 2019

Mentor

Mr. Divya Kiran

Asst. Professor

Dept. of CS, RUAS

Dr. P.V.R Murthy
HoD – Dept. of CSE, RUAS

Dr. Arulanantham
Dean – FET, RUAS

Declaration

Design and Implementation of Statistical Estimation Based Model for Fair Assessment of Rain Interrupted Cricket Matches

The Dissertation report submitted herewith is a result of our own work and in conformance to the guidelines against plagiarism as laid out in the University Student Handbook. All sections of the text and results which have been obtained from other sources are fully referenced. We understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly.

Signature :

Name of the Student : Praveen D Chougale

Reg. No. : 17ETCS075002

Date :

Acknowledgments

At the very outset, I sincerely thank Almighty for bestowing His blessings upon me, endowing me with capabilities to accomplish my goals throughout my entire life.

I express my heartfelt gratitude to my academic supervisor, **Mr. Divya Kiran**, Assistant Professor, Dept of CSE, RUAS, who directed and motivated me to have better understanding, knowledge and dexterity to undertake this project. Without his valuable guidance, great insights and meticulous corrections, the progress of my dissertation would have been stagnated. I feel really privileged to have availed this opportunity of working under his guidance.

I would like to express my sincere gratitude to **Dr. Hariharan Ramasangu** for his valuable support to persuade me to accomplish the virtuous and meritorious in my academic pursuit. His valuable insights and multifaceted subject knowledge and idea constantly motivated and contributed in enriching my dissertation work.

I take this opportunity to express my deep sense of gratitude to **Dr. P.V.R Murthy,** HOD, Department of CSE, RUAS for his encouragement and support. I express my gratitude to **Dr. Arulanantham**, Dean-FET, RUAS for providing all facilities and resources and to **Dr. Govind R. Kadambi**, Pro-Vice-Chancellor, RUAS for providing me the opportunity of enrolment for this course. I am thankful to the management of **Ramaiah University of Applied Sciences** for providing all the facilities and resources for the successful completion of the course and extend my sincere thanks to all the **staff** and **faculties of CSE department,** Ramaiah University of Applied Sciences for their support and co-operation.

I express my indebtedness and sincere gratitude to my parents **Mr. Dundappa Chougale** and **Mrs. Bharati Chougale** for their constant love and motivation. I express my heartfelt appreciation to my beloved brother **Mr. Pramod D Chougale** and sister **Ms. Laxmi D Chougale** for being always with me, boosting my morale in tough times and helping me to make this work possible with their boundless love and encouragement. I sincerely thank all my friends for their understanding, endless support, encouragement and forbearance.

Abstract

Despite the magnanimous impact of cricket in India, ironically, not much research has been carried out, nor has much of work been concluded. Developing an algorithm for fair assessment of the rain interrupted cricket will reduce the bias toward the chasing team in the second innings. The current form of algorithm being used to recalculate the total when rain plays spoilsport in a match is the Duckworth-Lewis (D/L) method

In rain interrupted matches a decision has to be reached within an allocated time of the game and the game cannot be postponed to another day. In such situations, the target score is currently calculated using the D/L method. It has been reported that the D/L method delivers unrealistic target scores for certain cases exhibiting its unfairness. The proposed algorithm formulated is a better approach that could serve well to reset the target score because of this intrinsic problem of the D/L method.

The formulation of such intrinsic algorithm demanded the processes of data cleaning and structuring on the raw available data, followed by feature extraction. Exploratory analysis and statistical test have been carried out on the independent variables. Developed mathematical functions for both and batting and bowling teams and these functions are trained by neural network to learn the functions. The developed algorithm is trained and validated for all the completed ODI matches as well as for D/L matches. The implemented algorithm can be extended to player selection, modelling using other features (apart from batting and bowling related) to improve the prediction a for the rain interrupted matches implementing a D/L method, to give a fair evaluation of outcomes. Accuracy of the model tested on completed ODI matches and for rain interrupted matches is 57 % and 61 % respe

Contents

List of Tables

Table No. **Title of the Table** **Page. No.**

List of Figures

Figure No. **Title of the figure** **Page. No.**

Abbreviation and Acronyms

ARR	Average Run Rate
CSS	Cascading Style Sheets
DMPO	Discounted Most Productive Overs
D/L	Duckworth Lewis Method
HTML	Hypertext Markup Language
MPO	Most Productive Overs
ODI	One Day International
PARAB	Parabola
SVM	Support Vector Machine
T20	Twenty 20 Match
WC96	World Cup 1996

1.Introduction and Motivation

The chapter enclosed herein is a preface of the envisaged approach and the motivation for its development. The incentives behind the specific subject chosen after a thorough literature survey done, is also explained and the motive behind the proposed algorithm is established with its relevant significance. A preamble and the motives behind the proposed work is showcased.

Current Scenario – International Cricket

In comparison to other sports, limited overs cricket is particularly vulnerable to inclement weather – when it rains, or becomes too dark, cricket becomes too dangerous to play. Consequently, when a One-Day International (ODI) or Twenty-20 International (T20I) match is interrupted by rain or bad light, either or both of the competing teams can often not complete their allotted overs. Incomplete games are unsatisfactory for the players and fans alike and, to some extent negate the purpose of the shorter formats since an abandoned match offers minimal levels of excitement. Furthermore, to enable knockout tournament play, such as the ODI and T20I World Cups, games must reach a positive conclusion. Therefore, the cricket authorities have adopted quantitative methods to adjust scores and reset targets in order to ensure interrupted matches are concluded with positive results.

1.1 Problems due to Interruptions during cricket matches

Generally, interruptions in cricket matches are due to rain and adverse weather conditions. Rain is the major reason for the interruption of International cricket matches. Such unforeseen interruptions cause disappointment among fans as they are not able to experience the game as expected. Interruptions also have a cost and resource implication as the match has to be continued on the next day (reserve day) which may not be as convenient for the fans and organisers. Continuation on another day also implies brand new conditions and environment for the players which could favour a particular team or players. This would give an undue advantage to them and may constitute an unfair advantage.

Currently, ICC has approved and accepted Duckworth Lewis method as the way of evaluating and resetting the scores of interrupted matches. For the D/L to be applied at least 20 overs of the game should have been played in the second innings of the match. Only then, can the target score for the team playing second be reset or the winner of the match can be declared. This proves a huge impediment to the team management and the players as the game strategy has to be drastically altered in the new scenario, which could be favourable to any one team.

The following relevant sections are from the ICC rules:

Section 12.4.2.B.iii states that a minimum of 20 overs in the second innings is required, subject to a result not being achieved earlier.

Section 27.7.2 says that a minimum of 20 overs is required in the second innings before applying the D/L par score. Hence, the D/L score above is irrelevant, and the match is declared a no-result.

1.2 Existing solutions for rain interrupted matches

There several methods adopted in the past for rain interrupted matches.

Average run rate (ARR)

The winning team is decided by the higher average number of runs per over that each team has had the opportunity to receive. It is a simple calculation but the method's major problem is that it very frequently alters the balance of the match, usually in favour of the team batting second.

Most productive overs (MPO)

The target is determined for the overs the team batting second (Team 2) are to receive by totalling the same number of the highest scoring overs of Team 1. The process of determining the target involves substantial bookwork for match officials and the scoring pattern for Team 1 is a criterion in deciding the winner. We believe that it is only Team l's total that should be used in setting the target and not the way by which it was obtained. The method strongly tends to favour Team 1.

Discounted most productive overs (DMPO)

The total from the most productive overs is discounted by 0.5% for each over lost. This reduces slightly the advantage MPO gives to Team 1 but it still has the same intrinsic weaknesses of that method.

Parabola (PARAB)

PARAB method designed by a young South African (do Rego8), calculates a table for overs of an innings, x, using the equation (5) the parabola

$$y = 7.46x - 0.0592x^2 \tag{5}$$

to model, rather inappropriately since it has a turning point (at about 63 overs, the 'diminishing returns' nature of the relationship between average total runs scored and total number of overs available. The method is an improvement upon ARR but takes no account of the stage of the innings at which the overs are lost or of the number of wickets that have fallen.

World Cup 1996 (WC96)

This is an adaptation of the PARAB method. Each of the norms has been converted into a percentage, shown in Table 1, of 225 as an approximation for the 50 over norm and generally regarded as the mean of first innings scores in one-day international matches.

Clark Curves (CLARK)

The CLARK method, fully described on the Internet,9 attempts to correct for the limitations of the PARAB method. It defines six types of stoppage, three for each innings, for stoppages occurring before the innings commences, during the innings, or to terminate the innings. It applies different rules for each type of stoppage some of which, but not all, allow for wickets which have fallen. There are discontinuities between the revised target scores at the meeting points of two adjacent types of stoppage.

Duckworth–Lewis–Stern method (DLS)

The Duckworth–Lewis–Stern method (DLS) is a mathematical formula designed to calculate the target score for the team batting second in a limited overs cricket match interrupted by weather or other circumstances. The method was devised by two English statisticians, Frank Duckworth and Tony Lewis and was formerly known as the Duckworth–Lewis method (D/L). It was introduced in 1997, and adopted officially by the ICC (International Cricket Council) in 1999. After the retirements of Duckworth and Lewis, Professor Steven Stern became the custodian of the method and it was renamed to its current title in November 2014.

When overs are lost, setting an adjusted target for the team batting second is not as simple as reducing the run target proportionally to the loss in overs, because a team with 10 wickets in hand and 25 overs to bat can play more aggressively than if they had 10 wickets and a full 50 overs, for example, and can consequently achieve a higher run rate. The DLS method is an attempt to set a statistically fair target for the second team's innings, which is the same difficulty as the original target. The basic principle is that each team in a limited-overs match has two resources available with which to score runs (overs to play and wickets remaining), and the target is adjusted proportionally to the change in the combination of these two resources.

1.3 Fairness issues in existing solutions

These earlier methods had flaws that meant they produced unfair new target scores that altered the balance of the match, and were easily exploitable:

1. The Average Run Rate method took no account of how many wickets were lost by the team batting second, but simply reflected how quickly they were scoring when the match was interrupted. So, if a team felt a rain stoppage was likely, they could attempt to force the scoring rate without regard for the corresponding highly likely loss of wickets, skewing the comparison with the first team.
2. The Most Productive Overs method also took no account of wickets lost by the team batting second, and effectively penalised the team batting second for good bowling by ignoring their best overs in setting the revised target. Figure 1.2 shows the snapshot of rain interrupted match.

Figure 1. 1 Rain interrupted matches

1.4 Motivation

- **The semi-final in the 1992 World Cup between England and South Africa at Sydney**, where the Most Productive Overs method was used. Rain stopped play for 12 minutes with South Africa needing 22 runs from 13 balls. The revised target left South Africa needing 21 runs from 1 ball, a reduction of only 1 run compared to a reduction of 2 overs, and a virtually impossible target given that the maximum score from 1 ball is generally 6 runs.

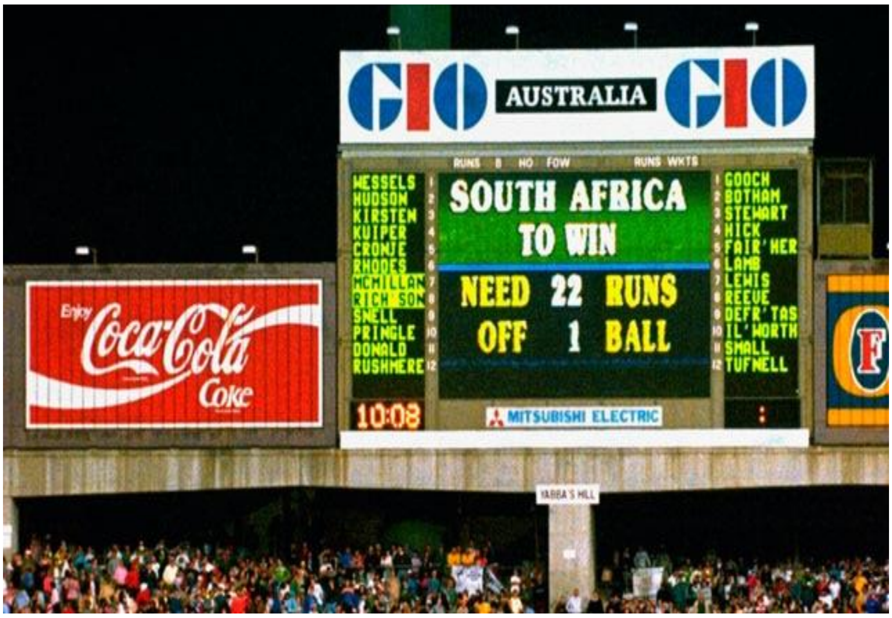

GIO AUSTRALIA GIO
SOUTH AFRICA TO WIN
NEED 22 RUNS OFF 1 BALL
WESSELS HUDSON KIRSTEN KUIPER CRONJE RHODES MCMILLAN RICH'SON SNELL PRINGLE DONALD RUSHMERE
GOOCH BOTHAM STEWART HICK FAIR'HER LAMB LEWIS REEVE DEFR'TAS IL'WORTH SMALL TUFNELL
10:08
MITSUBISHI ELECTRIC
Enjoy Coca-Cola Coke

Figure 1. 2 World Cup 1992 Semi Final Scoreboard between England and South Africa

Duckworth said, "I recall hearing Christopher Martin-Jenkins on radio saying 'surely someone, somewhere could come up with something better' and I soon realised that it was a mathematical problem that required a mathematical solution." The D/L method avoids this flaw: in this match, the revised D/L target would have left South Africa 4 to tie or 5 to win from the final ball. A definitely achievable target!

The D/L method was first used in international cricket on 1 January 1997 in the 2nd match of the Zimbabwe versus England ODI series, which Zimbabwe won by 7
runs. The D/L method was formally adopted by the ICC in 1999 as the standard method of calculating target scores in rain-shortened one-day matches.

- **2003 World Cup - Group Stage: Rain, brain freeze, D/L math pangs - Kingsmead, Durban**

Sri Lanka with help of Marvan Attapattu's century posted 268/9. Herschelle Gibbs scored 73 to get South Africa going. Play was called off after the 45th over with the Proteas placed at 229 for 6. At the end of the 44th over, the dressing room sent a message to Mark Boucher, who was going great guns then, that 229 at the end of the 45th over would ensure them victory as per D/L if they did not lose any more wickets. Boucher and Klusener took 13 runs off the first 5 balls of Muralitharan's over. Boucher pumped his fist in the air after sending the 5th ball flying over the boundary. He defended the last ball under the impression that they had already won. Covers were brought on. Figure 1.3 shows scene of disappointed South African dressing room.

Figure 1. 3 2003 World Cup - Group Stage: Rain, brain freeze

- **2015 Semi-Final: Rain gods land a cruel blow Eden Park, Auckland**

The rain played hide-and-seek for almost an hour-and-a-half. When play resumed, the match was reduced to 43 overs a side. David Miller provided impetus in these final overs to ensure that South Africa posted 282 on the board. New Zealand got a D/L target of 298**.** Brendon McCullum came out all guns blazing and scored 59 off 26 balls to set Kiwis on course for victory. After a few hiccups, and a few reprieves New Zealand romped home with 4 wickets in hand and 1 ball remaining. De Villiers was sporting enough to credit the victory to New Zealand, saying, "The better team won”. The Figure 1.5 shows disappointed Morne Morkel in the 2015 world cup match between New Zealand and south Africa.

Figure 1. 4 2015 Semi-Final Eden Park, Auckland

As cricket lovers and enthusiasts, having loved to both play and watch cricket, the thought of making a significant scientific contribution to it has always been a dream. Moreover, the scope for cricket in any way is far greater in the subcontinent than other countries simply because of the love and popularity of the sport.

Despite the huge impact of the sport in our country, ironically, much educational research has not been carried out, nor has much work been concluded hence this served as a major motivating factor to carry out the proposed technique.

Scanty evident research work is found in the field of sports and cricket pertaining to one; hence the innovative idea was to come up with our own stipulation as to how cricket can be more technologically enhanced.

During the primary research, it was found that the Duckworth Lewis method has its own limitations and such could be enhanced to be fairer to both teams. This dissertation is thus motivated by the keenness to contribute to the game and its outcomes.

1.5 Organisation of the thesis

The overall goal is to develop an alternative algorithm to the Duckworth-Lewis method. Although many researches and development are being done to overcome the limitation of D/L, no other method is as successful as D/L.

In this thesis, we begin with exploring the different available solutions for interrupted cricket matches. As D/L is widely accepted and popular method, the major focus revolves around evaluating it and finding out the limitations of this method. An algorithm is proposed and evaluated to overcome these gaps.

In Chapter 2 comprises of literature survey and background study of the algorithms used and history, implementation of D/L method.

In Chapter 3, we arrive at aims/objectives of thesis from the literature surveys.

In Chapter 4, Modelling and problem solving of 4 experiments. This involves coming up with a mathematical function for batting and bowling and training these functions using neural network, so that any point of the game will be able to declare the winner based on functions values.

In Chapter 5, we discuss Result and discussion of all experiments

And Conclusion and future works are mentioned in chapter 6.

2. Background Theory and Literature Survey

The chapter enclosed herein is a prologue of the detailed literature survey carried out and the background study of the techniques implemented. The algorithms used in the methodology are elucidated and justified as well. The details of the web page application built with a User Interface developed is also elaborated.

2.1 Literature Survey

An extensive literature survey revealed that research and analysis done by far in the cricketing field is quite confined. The secondary research carried out revealed the research available is limited to very few studies in machine learning domain being carried out in the sports domain especially cricket. There have been not many successful attempts at improvising the existing D/L method hence no alternative is being implicated by the ICC.

2.1.1 Papers /Journals related to D/L methods

Various research publications were confined to just domestic cricket or only player performances or only one format of the game and also had results with a generalized accuracy of around 50-55% approximately. The researchers in [1] Bayesian Inference is applied to build a resource table which overcomes the non-monotonicity problem of the current D/L resource table to show that it gives better prediction for teams in first innings score and hence it is more suitable for using in rain affected matches. For each match they have defined R(u,w(u)) as the run scored from the stage in first innings where u overs are available and w(u) wickets are lost until the end of the first innings. They have also calculated R(u,w(u)) for all values of u that occurred in the first innings. The estimated resource percentage table is then calculated by averaging R(u,w(u)) over all matches

where w(u) = w and dividing by the average of R(50, 0) (which is the average first innings score) over all matches. Just like D/L table, this non-parametric resource table suffers from the lack of monotonicity. Authors of [2] have used isotonic regression method to overcome this issue, whereas in [1] they have taken a parametric Bayesian approach. Non-parametric model based resources decay as overs remaining decrease for different wickets. Throughout this paper in resource decay plots index w ∈ W indicates loss of w wickets. Instead of throwing out those columns or rows that have missing entries, [1] have used the Bayesian inferential framework that provides a natural way of imputing the missing entries using the posterior predictive distribution once a full hierarchical model is specified. Adopting the following nonlinear regression model:

$$\boldsymbol{R}(\boldsymbol{u},\boldsymbol{w})\sim\boldsymbol{N}\left(\boldsymbol{m}(\boldsymbol{u},\boldsymbol{w};\boldsymbol{\theta}),\frac{\boldsymbol{\sigma}}{\boldsymbol{nuw}}\right), \in \mathbf{u},\mathbf{w}\ \in \mathbf{W} \qquad \textbf{(1)}$$

From equation (1) ,where R¯(u,w) is the sample average of runs scored by a team among the total number of matches considered in the data set and m(u,w; θ) is the corresponding modelled population average of runs scored by a team when a large number of games are taken into consideration and θ denotes a vector of parameters to be specified later in our model. As R(u,w) is not observed for each of the match (in the sample), the average is taken over all those matches, denoted by nuw, over which the sample mean ¯R(u,w) is calculated. If there is no observation for R(u,w) across all of the matches sampled.

Residual Sum of Squares than the D/L method specially when the match is interrupted in situations where there are lots of overs left is shown. Under the MAR assumption, the proposed Bayesian model provides a natural method to carry out imputations using the posterior predictive distributions which is an advantage over many existing methods (e.g., compared to the non-parametric method). This method is broadly applicable in the sense that it is not restricted to only 50-overs cricket match interruption problem and can be applied many similar sports events. Moreover, the model can be used to estimate the nonlinear mean function of two variables under bi-monotonicity constraint. One future direction for research can be to develop a nonparametric approach for modelling such constrained bivariate functions that is not necessarily based on an exponential decay model. Another alternative method to calculate the revised target in interrupted 50 overs ODI matches is found in [3]. Existing D/L method and its modified versions only take available batting resources of the batting team into account and ignore the individual player's excellence to calculate the revised target. Here, it is worth mentioning that individual player's excellence varies in reality, and therefore quality of the available resources may affect the revised target significantly. Furthermore, in D/L method the revised target calculation depends only on the available batting resources of the batting team and does not consider the available bowling resources of the fielding team. Their method overcomes these two shortcomings by taking individual player's excellence and available bowling resources of the fielding team into account. Individual player's excellence has been determined by Data Envelopment Analysis (DEA), a well-known non parametric mathematical programming technique.

Analysing the D/L method using graphical and mathematical methods and find out the root cause of this unfairness is done in [4]. Here, the reason for the unfairness of D/L method using graphical methods and chi-square tests is shown. It has been shown that this is due to the inherent nature of the D/L method that use graphs produced using past statistics of all teams of the world. The expected performance specified by those graphs deviate considerably from the actual performance of the teams participating in the game resulting in the unfairness. In each innings, there were at most 50 data points. As such, the degrees of freedom was less than or equal to 49. For this value of the degrees of

2.1.2 Papers /Journals related to other variants.

To facilitate the comparison, the absolute values of the differences between the two tables was imputed, and a heat map is produced. The darker shades of the heat map indicate the greatest disagreement between the two tables. On investigating these areas of disagreement, it is observed that the greatest absolute differences occur in three regions. First, large differences occur in the top-right hand corner and bottom-left hand corner of the table. These are precisely the regions where very little or no data are available. These regions are not viewed as too important as the resetting of targets would rarely use these entries. It is interesting however that the non-parametric approach provides more resources in these regions than the D/L approach.

The more interesting discrepancy occurs in the 'middle' of an innings (8–13 overs available with 3–6 wickets lost). In this stage of an innings, the non-parametric approach based on Gibbs sampling suggests that there is up to 5% fewer resources remaining than provided by the D/L method. In 1-day cricket, a team needs to protect its wickets over a longer period of overs. Consequently, up until the middle stage, more resources are conserved in the 1-day game than in Twenty20. They remark that a difference of 5% resources may be very meaningful as a target of 240 runs diminished by 5% gives 228 runs. As more Twenty20 matches become available, authors of [5] endorse a review of the use of D/L in Twenty20 and the estimation techniques used in the construction of the associated resource table. The method is based on a simple model involving a two-factor relationship giving the number of runs which can be scored on average in the remainder of an innings as a function of the number of overs remaining and the number of wickets fallen [6]. It is shown how the relationship enables the target score in an interrupted match to be recalculated to reflect the relative run scoring resources available to the two teams, that is overs and wickets in combination. The method was used in several international and domestic one-day competitions and tournaments in 1997.

Therefore, need a two-factor relationship between the proportion of the total runs which may be scored and the two resources, overs to be faced and wickets in hand. To obtain this it is necessary to establish a suitable mathematical expression for the relationship and then to use relevant data to estimate its parameters

The basis of this method is that it recognises that the batting side has two resources at its disposal from which to make its total score; it has overs to face and it has wickets in hand. The number of runs that may be scored from any position depends on both of these resources in combination. Clearly, a team with 20 overs to bat with all ten wickets in hand has a greater run scoring potential than a team that has lost, say, eight wickets. The former team have more run scoring resources remaining than have the latter team although both have the same number of overs left to face. The mechanisms of other methods used for resetting target scores in interrupted one-day cricket matches is explained in [7]. Each of these methods yields a fair target in some situations. None has proved satisfactory in deriving a fair target under all circumstances. We have presented a method which gives a fair revised target score under all circumstances.

This is based on the recognition that teams have two resources, overs to be faced and wickets in hand, to enable them to make as many runs as they can or need. They have derived a two-factor relationship which gives the average number of runs which may be scored from any combination of these two resources and hence have derived a table of proportions of an innings for any such combination. This enables the proportion of the resources of the innings of which the batting team are deprived when overs are lost as a result of a stoppage in the play to be calculated simply and hence a fair correction to the target score to be made.

Though the examples given, both hypothetical and real, it is shown that this method gives sensible and fair targets in all situations. They include the circumstances where overs are lost at the start of the innings, part way through, or at the end of an innings and where the game is abandoned requiring a winner to be decided if Team 2's innings is terminated. The examples have shown the importance of taking into account the wickets that have been lost at the time of the interruption and the stage of the innings at which the overs are lost. Our method was adopted by the England and Wales Cricket Board for the 1997 domestic and Texaco one-day international competitions and the International Cricket Council has used it for several international one-day competitions.

2.2 Duckworth-Lewis method

The Duckworth Lewis Method:

The Duckworth Lewis method is a statistical method devised by two British statisticians, Frank Duckworth and Tony Lewis. It aims to reset a target score if play in a one-day match is interrupted. **(DLM)** is a mathematical formulation designed to calculate the target score for the team batting second in a limited overs cricket match interrupted by weather or other circumstances. It was introduced in 1997 and adopted officially by the ICC in 1999.

When overs are lost, setting an adjusted target for the team batting second is not as simple as reducing the run target proportionally to the loss in overs, because a team with ten wickets in hand and 25 overs to bat can play more aggressively than if they had ten wickets and a full 50 overs, for example, and can consequently achieve a higher run rate.

The DLS method is an attempt to set a statistically fair target for the second team's innings, which is the same difficulty as the original target. The basic principle is that each team in a limited-overs match has two resources available with which to score runs (overs to play and wickets remaining), and the target is adjusted proportionally to the change in the combination of these two resources.

The essence of the D/L method is 'resources'. Each team is taken to have two 'resources' to use to score as many runs as possible: the number of overs they have to receive; and the number of wickets they have in hand. At any point in any innings, a team's ability to score more runs depends on the combination of these two resources they have left. Looking at historical scores, there is a very close correspondence between the availability of these resources and a team's final score, a correspondence which D/L exploits.[9]

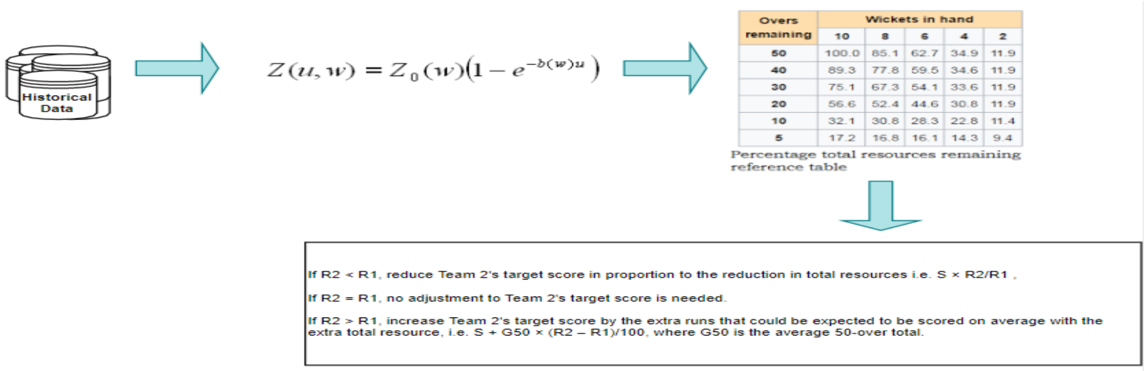
The D/L method converts all possible combinations of overs (or, more accurately, balls) and wickets left into a combined resources remaining percentage figure (with 50 overs and 10 wickets = 100%), and these are all stored in a published table or computer. The target score for the team batting second ('Team 2') can be adjusted up or down from the total the team batting first ('Team 1') achieved using these resource percentages, to reflect the loss of resources to one or both teams when a match is shortened one or more times.

In the version of D/L most commonly in use in international and first-class matches (the 'Professional Edition'), the target for Team 2 is adjusted simply in proportion to the two teams' resources, i.e.

If, as usually occurs, this 'par score' is a non-integer number of runs, then Team 2's target to win is this number rounded up to the next integer, and the score to tie (also called the par score), is this number rounded down to the preceding integer. If Team 2 reaches or passes the target score, then they have won the match. If the match ends when Team 2 has exactly met (but not passed) the par score then the match is a tie. If Team 2 fail to reach the par score then they have lost.

For example, if a rain delay means that Team 2 only has 90% of resources available, and Team 1 scored 254 with 100% of resources available, then 254 × 90% / 100% = 228.6, so Team 2's target is 229, and the score to tie is 228. The actual resource values used in the Professional Edition are not publicly available, so a computer which has this software loaded must be used.

If it is a 50-over match and Team 1 completed its innings uninterrupted, then they had 100% resource available to them.Figure 2.1 shows the D/L work flow.



$$Z(u, w) = Z_0(w)\left(1 - e^{-b(w)u}\right)$$

Overs remaining	Wickets in hand				
	10	8	6	4	2
50	100.0	85.1	62.7	34.9	11.9
40	89.3	77.8	59.5	34.6	11.9
30	75.1	67.3	54.1	33.6	11.9
20	56.6	52.4	44.6	30.8	11.9
10	32.1	30.8	28.3	22.8	11.4
5	17.2	16.8	16.1	14.3	9.4

Percentage total resources remaining reference table

If R2 < R1, reduce Team 2's target score in proportion to the reduction in total resources i.e. S × R2/R1 ,

If R2 = R1, no adjustment to Team 2's target score is needed.

If R2 > R1, increase Team 2's target score by the extra runs that could be expected to be scored on average with the extra total resource, i.e. S + G50 × (R2 – R1)/100, where G50 is the average 50-over total.

Figure 2. 1 Duckworth-Lewis workflow

2.3 Algorithms Explained:

Numerous algorithms were used for this research to be concluded. The algorithms and techniques used are explained as follows:

1. Confusion Matrix:

A confusion matrix is a table that is commonly used to describe the performance of a classification model, on a set of test data for which the original values are known.(Kevin Markham, 2014) The confusion matrix is relatively simple to understand. The detailed explanation of a confusion matrix are as follows:

- **True Positives (TP):** These are the cases in which the predicted outcome is yes/true. Example: a hypothesis, where a person has a disease and the prediction also says yes, they do have the disease, such a case will be considered as a TP.
- **True Negatives (TN):** Where prediction is false/no, and in reality, they don't have the disease.
- **False Positives (FP):** Where the predicted result is a yes/true, but in actual they don't have the disease. This is commonly known as a "Type I error".
- **False Negatives (FN):** Where prediction is false/no, but they actually do have the disease. This is also known as a "Type II error".

From the above parameters of the confusion matrix the accuracy is therefore calculated using the equation (6) :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

2.4 Justification for choosing specific algorithms:

The stages of the diabetic retinopathy are to be predicted, for which a supervised machine learning classification algorithm is required.

Though there are numerous techniques available to help classify, CNN is best suited for Image classification problems, hence CNN model is implemented in the design flow and tested to check the performance of the CNN classifier

2.5 Background research for Web Application and User Interface developed:

Subsequently after the model validation, a visualization using a user interface (UI) is done in order to provide a platform to showcase the eminent results and to enhance the project to a product level.

The visualization is done on a User Interface (UI), for which the specific routes were formed, and static HTML pages were designed. To support the routing and the design Python Tool is used in the backend. The research carried out in order to implement the set objective of developing a UI is elaborately explained as follows:

2.5.1 Website development:

A website makes the first impression about the company or product directly impacts on the global audience. The world is at the fingertips now as everything is done online now, So, the requirement for a website to be developed along with the project carried out is justified. It adds as a valuable investment and benefits one by giving business exposure.

The prime requirements for web development include a user interface and user experience. Web development has significantly evolved in the recent years, particularly with the apparition of web frameworks. A complete website using Flask web development framework is created.

Flask working and development:

The Flask is a python based micro web development-based framework. It actually acts like a glue that sticks together the Jinja2 and Werkzeug frameworks(Dwyer, 277AD), responsible for answering requests and presenting the output (HTML).A Flask tries to find the HTML file in the templates folder, the same folder in which the specific script is present.

The Application folder contains:

- Hello.py
- Templates

The term **'web templating system'** refers to designing an HTML script in which the variable data can be inserted dynamically.(Pocoo, 2015)

The **route ()** decorator in Flask is used to bind the URL to a function.

Flask uses **jinga2** template engine. A web template contains HTML syntax interspersed place holders for variables and expressions (in this case Python expressions) which are replaced values when the template is rendered.

The following code is saved as **hello.html** in the templates folder.

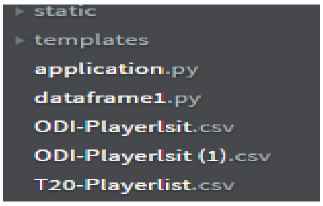
The usage of the developed application contains the following features:

- The layout page provides basic information about the cricket format.
- The web application provides a unique URL for each Cricket format.
- Once the navigation to the specific Cricket format is made, the web application provides a unique URL for each Teams.
- After navigating to the specific Cricket Team, web application provides an information about each player in a team.

Flask implementation:

To implement a Flask application that has features required for the specified web application, the following steps are to be taken:

- Before implementing the features make an application folder.
- In application folder make template subfolder and static subfolder. In template subfolder the required html files should be placed.
- CSS, multimedia and static contents needed for designing the application will be placed in static subfolder.
- Python application server file is also placed in the application folder.



```
▸ static
▸ templates
  application.py
  dataframe1.py
  ODI-Playerlsit.csv
  ODI-Playerlsit (1).csv
  T20-Playerlist.csv
```

Figure 2. 2 Web application setup for the prediction model output

2.5 Research Gaps

The D/L implemented and the background research on other variants of D/L yielded the research gap for the study.

1. The D/L method has been criticized on the grounds that **wickets are a much more heavily weighted** resource than overs, leading to the suggestion that if teams are chasing big targets, and there is the prospect of rain, a winning strategy could be to not lose wickets and score at what would seem to be a "losing" rate (e.g. if the required rate was 6.1, it could be enough to score at 4.75 an over for the first 20–25 overs). The 2015 update to DLS recognised this weakness and changed the rate at which teams needed to score at the start of the second innings in response to a large first innings.

2. Another criticism is that the D/L method does not account for changes in proportion of the innings for which **field restrictions** are in place compared to a completed match.
3. Most common informal criticism from cricket fans and journalists of the D/L method is that it is **unduly complex** and can be misunderstood. For example: in a one day match against England on 20th Mar 2019, the West Indies coach (John Dyson) called his players in for bad light, believing that his team would win by one run under the D/L method, but not realizing that the loss of a wicket with the last ball had altered the Duckworth–Lewis score. In fact, Javagal Srinath (the match referee), confirmed that the West Indies were two runs short of their target, giving the victory to England.
4. D/L does not consider the quality of batsmen yet to bat, pitch condition, number of overs that are to be bowled by different bowlers, quality of the opposition team.
5. Concerns have been raised as to its **suitability for Twenty20** matches, where a high scoring over can drastically alter the situation of the game and variability of the run-rate is higher over matches with a shorter number of overs.

2.6 Summary

After thorough examination of the mentioned publications, primary research has been carried out. After processing a careful and explicit literature survey, the premier insight was that, numerous slacks were found in terms of algorithms used, the objectives, methodologies and the accuracy.

The premier research fulfilled herein, answered the most specific issues and constraints faced in the above publications. In contrast to the known publications, this study ventures into an amalgamation of what is done and what more could be done in these studies.

The ideas generated through this literature survey formulated the problem statement, main objectives and aim of this analysis. Consequently, an algorithm that overpowers the previous ones effectively and efficiently has been proposed and processed. The methodology executed is novel in terms of data collection and preparation, model accuracy and performance and the provision of a UI which is unique in its own terms.

3. Aim and Objectives

The chapter enclosed herein is a foreword of the title suggested with aims and objectives formulated for the proposed design. The elucidation of each specific aim and objective is also mentioned. A brief utterance of the methods and methodologies used along with the resources utilized per objective is tabulated.

3.1 Title:

The Title of the Project is: **Design and Implementation of Statistical Estimation Based Model for Fair Assessment of Rain Interrupted Cricket Matches**

3.2 Aim:

- To implement Duck-worth Lewis method, analyze and identify the pitfalls
- To arrive at a statistical estimation method to assess the winner of the match at any given point of time in the match

3.3 Problem statement:

The primary objective of the research is to predict the winner of the match among the two teams in a truncated match. The total overs in the match may be truncated due to either rain or any other unforeseen interruption to the game. In such cases, when play has been cut short, the goal is to:

- Fair assessment of the winner in the match based on batting and bowling related parameters of the two teams

3.4 Objectives and Methodologies

The Table 3.1 shows the statement of objectives ,methodology and resources required to meet those objectives. In future chapters all the objectives are described in details with respect to the proposed block diagram.

Table 3. 1 Objective and Methodologies

Objective No.	Statement of Objective	Method / Methodology	Resources Required
1	To mathematically analyse the existing method of Duckworth-Lewis and its variants	Referring peer journals, conference papers	Google scholar, researchgate
2	To perform a comparative analysis in order to deduce the pitfalls of the Duckworth-Lewis and its variants	Data collection and preparation	Cricksheet, R tool
3	To formulate a statistical estimation technique based on the resource availability at any instance in the game.	Design and Implementation of formula (using wickets and runs)	R tool
4	To develop Machine learning algorithm by applying the formulated statistical estimation technique to reset the target	Naïve Bayes, support vector machine, Logistic Regression	R tool
5	To validate /test the implemented model on the past D/L applied rain interrupted ODI matches and	Leave out one cross validation, ROC curve, confusion matrix	R tool , Python , HTML

	develop an user interface to visualize the results		

4.Design of Statistical Model

The chapter enclosed herein is an epilogue of the statistical methodology designed and implemented in the proposed technique. The detailed description of each stage that is data collection and data preparation, data pre-processing, data modelling are explained with their relevant figures and tabulations.

4.1 Implementation of Duckworth -Lewis method:

The first step is to apply the current data is
Duckworth and Lewis provided a mathematical definition of the production functions:

$$\boldsymbol{Z(u,w) = Z_0(w)(1 - e^{-b(w)*u})} \qquad \textbf{(2)}$$

From equation 2, production function ' $\boldsymbol{Z(u,w)}$' and they say that this production function tells us the average total score. Each function will vary with the number of wickets lost (which they abbreviate to w) and the number of overs remaining (which is abbreviated to u).

$\boldsymbol{Z_0}$ represents the upper bound of runs that we'd expect the team to score; even if they weren't limited by overs. This value decreases as the number of wickets lost increases. Therefore, $\boldsymbol{Z_0}$ is a function of w

This theoretical upper bound is then multiplied by the value inside the brackets. The number of overs remaining turns up again. As does the rate at which the team approaches the theoretical upper bound (this is abbreviated as b).

Taking an example of the last wicket partnership, Duckworth and Lewis estimate this partnership to have a theoretical upper bound of 7.6 runs, on average. In their tables they indicate that this theoretical upper bound is actually reached after about 12 overs. It means that an average last wicket partnership could be used out on the pitch for 12 overs or fifty overs or even a hundred overs and they'd still only score 7.6 runs on average and this is because they'd almost certainly lose that last wicket before they run out of overs.

So, the basic message of the DWL equation is that there's a theoretical maximum which a team can score given the number of wickets they have left and then lower this maximum by the actual number of overs they've got to face.

The next step to getting to the resources remaining tables is to get this average total score into a percentage. The above equation can be re-written to represent the start of a 50 over match by saying that:

$$\boldsymbol{Z(50, u) = Z_0(0)(1 - e^{-b(0)50}}) \qquad \textbf{(3)}$$

From equation 3, $\boldsymbol{u}$ =50 because there are 50 overs remaining and w=0 because no wickets lost.

The percentage of runs left at any point in the innings is given by:

$$\boldsymbol{P(u, w) = Z(u, w)/(Z(50, 0)} \qquad \textbf{(4)}$$

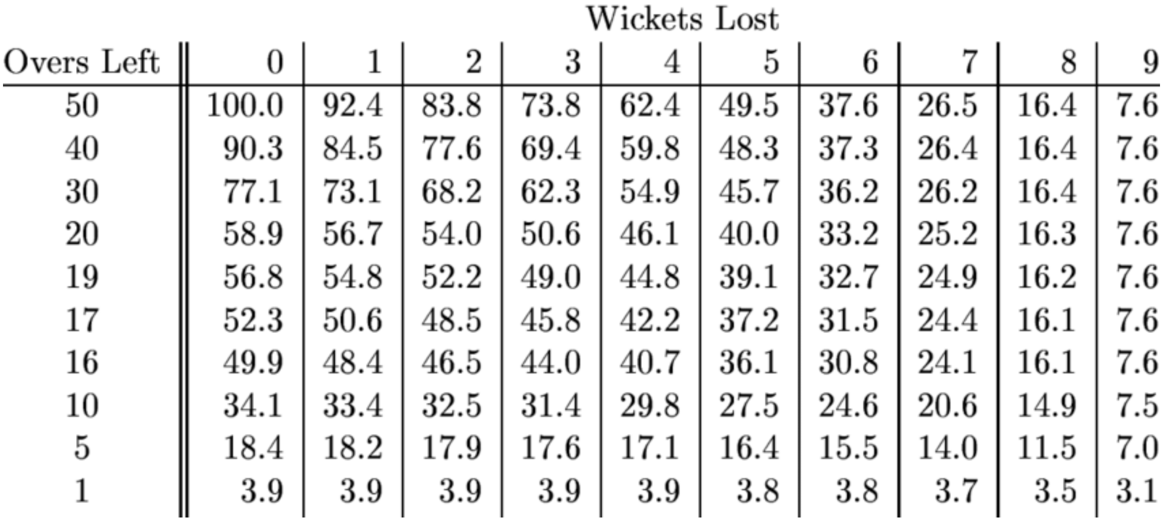
From equation 4,That is, at any point in the match (described by there being u overs remaining and w wickets lost), the percentage of runs remaining is simply the average total score expected (given by $\boldsymbol{Z(u, w)}$) divided by the average total score expected in a full fifty over match.

4.2 Estimating the parameters in the production function

The run production has the following hypothesis

- More wickets in hand mean more resources
- More balls remaining mean more resources
- Earlier batter’s worth more than later batters
- There is some upper bound beyond which extra overs have less chance to score.

With the data, the next step is to calculate average runs remaining for each combination of wickets lost and overs remaining. Figure 4.1 shows the D/L resource table.



	Wickets Lost									
Overs Left	0	1	2	3	4	5	6	7	8	9
50	100.0	92.4	83.8	73.8	62.4	49.5	37.6	26.5	16.4	7.6
40	90.3	84.5	77.6	69.4	59.8	48.3	37.3	26.4	16.4	7.6
30	77.1	73.1	68.2	62.3	54.9	45.7	36.2	26.2	16.4	7.6
20	58.9	56.7	54.0	50.6	46.1	40.0	33.2	25.2	16.3	7.6
19	56.8	54.8	52.2	49.0	44.8	39.1	32.7	24.9	16.2	7.6
17	52.3	50.6	48.5	45.8	42.2	37.2	31.5	24.4	16.1	7.6
16	49.9	48.4	46.5	44.0	40.7	36.1	30.8	24.1	16.1	7.6
10	34.1	33.4	32.5	31.4	29.8	27.5	24.6	20.6	14.9	7.5
5	18.4	18.2	17.9	17.6	17.1	16.4	15.5	14.0	11.5	7.0
1	3.9	3.9	3.9	3.9	3.9	3.8	3.8	3.7	3.5	3.1

Figure 4. 1 Resource table

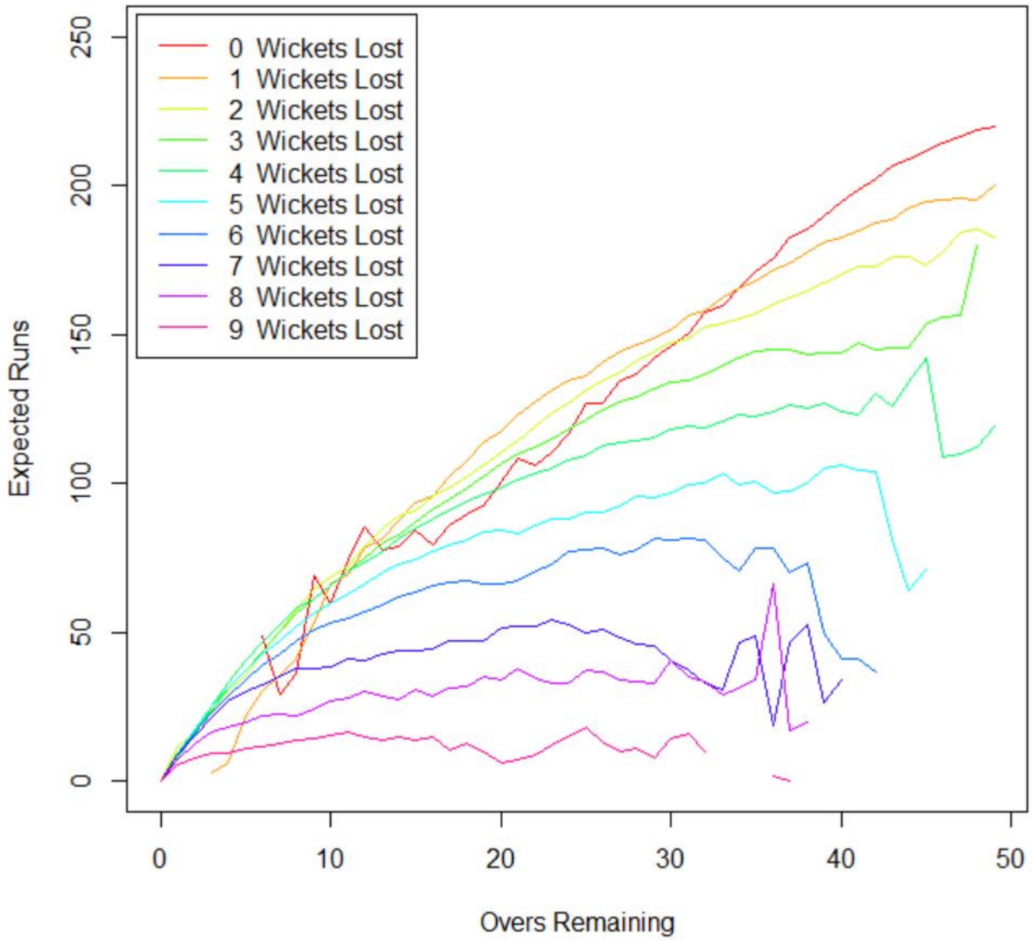


Figure 4. 2 Plot showing Overs remaining versus Expected runs

The Figure 4.2 shows the graph of overs remaining versus expected runs versus scored ,which shows the predicted runs scored if game would have continued after interruption.

The following observations for the graph:

- More wickets in hand mean more resources:

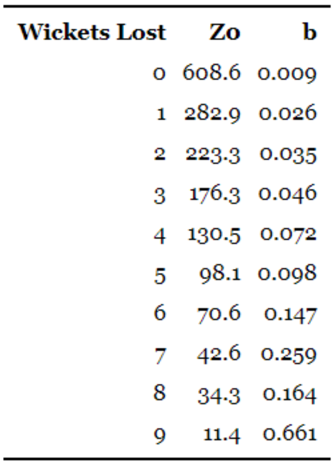
  The 1 wickets lost line is generally above the 2 wickets lost line which is generally above the 3 wickets lost line.
- More balls remaining should mean more resources:

  Each line increases as we go from the left to the right
- Earlier batters add more value than later batters:

  The drop from 2 wickets lost to 3 wickets lost is bigger than from 7-8 wickets lost
- There is some upper limit beyond which extra overs aren't adding much to the score:

  Although the lines are increasing from left to right they tend to flatten off near the end.

Using non-linear least squares function and calculate Z0 and b, the



Wickets Lost	Zo	b
0	608.6	0.009
1	282.9	0.026
2	223.3	0.035
3	176.3	0.046
4	130.5	0.072
5	98.1	0.098
6	70.6	0.147
7	42.6	0.259
8	34.3	0.164
9	11.4	0.661

Figure 4. 3 Parameter Estimates based on raw data 2005-2017

Figure 4.3 shows estimated parameters for D/L. Graph of the production functions generated from the parameter estimates helps understand the underlying statistics

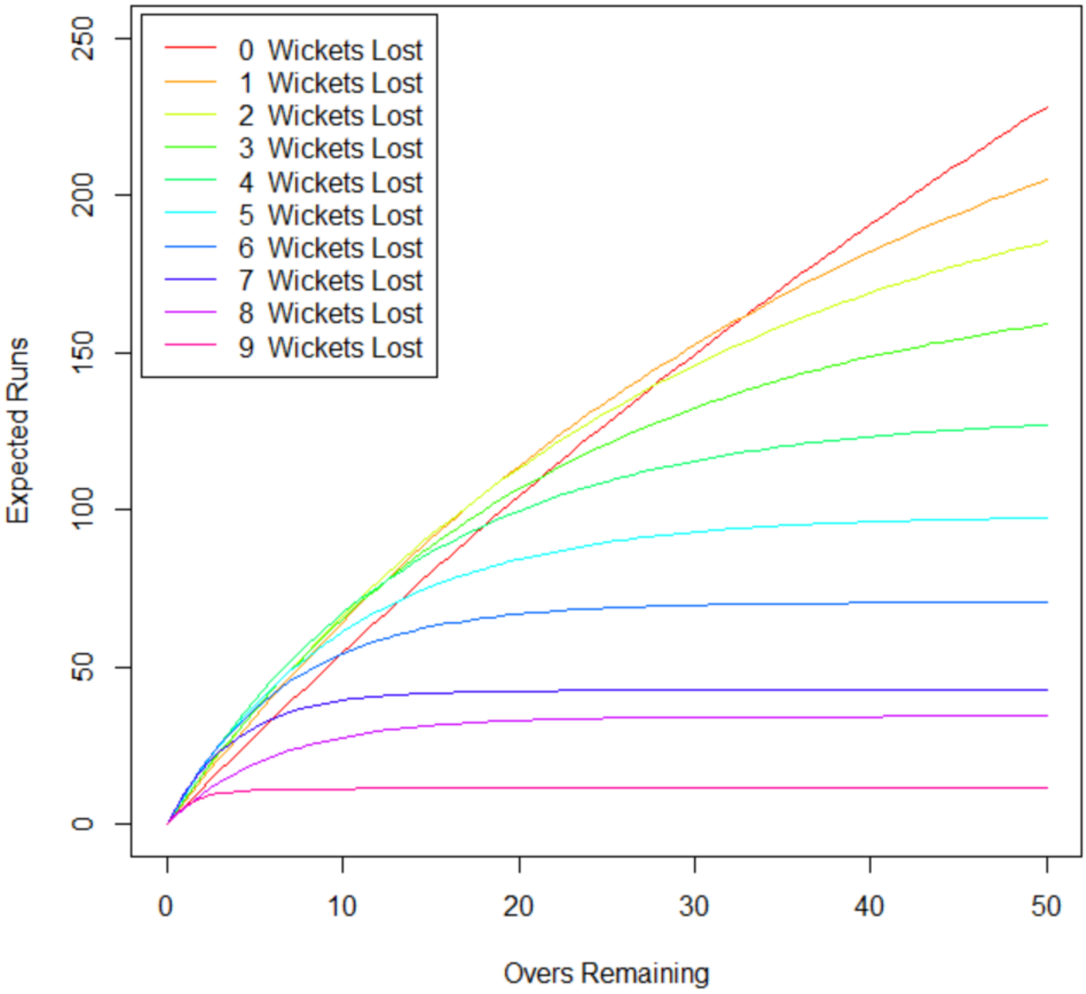


Figure 4. 4 Graph for Fitted Production functions based on real data 2005-2017

Figure 4.4 shows graphs of fitted production functions . So, this is good but not the perfect one. In the functions we see, that they cross each other, which means that sometimes the players batting 2nd and 3rd are more valuable than the player batting first. Also, the production function for 0 wickets lost doesn’t appear to flatten out.

In an idealised world, the best batters come out first, the worst last and all those in between are on a steadily sliding scale. While in the real world it seems like 8th wicket partnerships actually perform better than others when there are only a few overs left to go. Maybe this is because 8th position on the batting order is normally given to a player who can go out there and slog a bunch of balls to the boundary in a few overs.

Figure 5.5 shows the total runs scored pattern for all the ODI matches

Changes in the ODI runs pattern

Innings totals in one day internationals were starting to become quite large, quite rapidly and a change had to be made. The change in frequency of large scores in the graph below shows how rapidly things were changing

In response, Duckworth and Lewis issued a revised theory, a revised paper and a revised set of estimates. This led to the "Professional" method, still in use in ODIs, and the "Standard" method, for use in lower levels of cricket

Resetting Scores

Interruption to team 2's innings

Lets say team one bat out their innings and so set a target score for team two. Team two's innings are then interrupted due to something. This interruption to team two's innings simply deprives them of a certain number of overs. We know that these overs represent a certain number of resources to the team. Calculate the percentage of total resources that these lost overs represent and multiply it by team one's score then this needs to be taken off the target for team two.

For example, team one scores 200 runs in their innings. Team two come into bat and at the end of the 15th over are 3 out for 50 runs. Team two therefore need to score 151 more runs to win the match, a run rate of 4.31 runs an over. Rain starts to fall and the players head back to the pavilion. The rain clears but 10 overs have been lost. Team two are still 3 out for 50 runs but have lost 10 overs of their resources. The Duckworth Lewis tables can be used to figure out what percentage of their total resources these ten overs represent. From the 2002 Resource table we get:

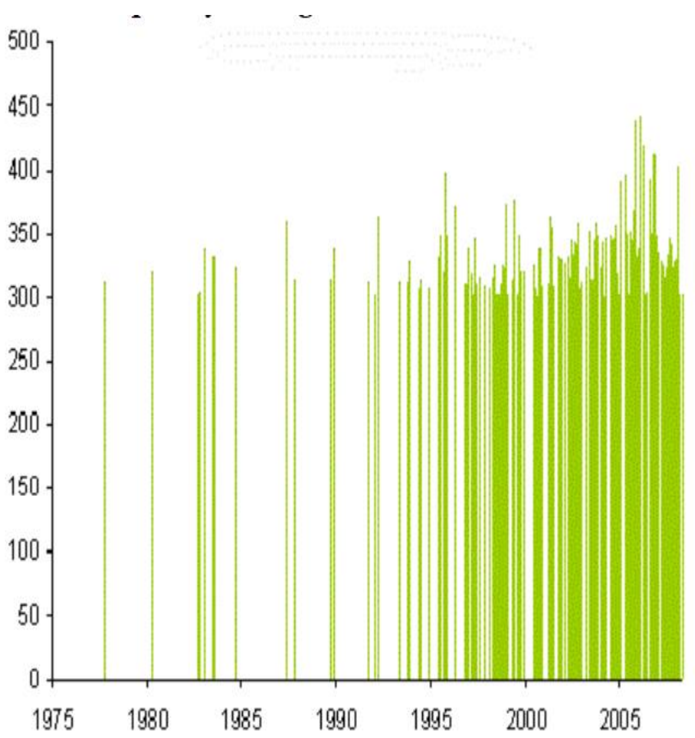


Figure 4. 5 Total runs pattern for all the ODI Matches

- At the interruption there were 3 wickets lost and 35 overs remaining. The table indicates that the team had 66% of their resources remaining.
- After the interruption there were 3 wickets lost and 25 overs remaining. The table indicates that the team had 56% of their resources remaining.
- Team two therefore lost 66% - 56% = 10% of their resources.

It is fair to reduce team one's score by 10%, giving a new target score of 180. Team two then need to score 181 runs to win the match. This means they need 131 more runs to win, a run rate of 5.24 runs an over.

Interruption to team one's innings

Interruptions to team one's innings are complicated by the fact that lost overs are normally spread out over the two innings. For example, if there are 10 overs lost halfway through team one's innings, then it is often the case that team one will be given a total of 45 overs to face and team two will also be given 45 overs to face. However, team one has lost 5 overs from the middle of its innings while team two has lost five overs from the start of its innings. These two groups of five overs are likely to be worth different amounts.

As before, calculate the amount of resources both teams have lost. If the resources both teams have lost are equal then there's no need to adjust anyone's scores.

Another case might be where team one loses less resources than team two. Then simply scale down the target score by the difference in available resources. This is the same thing as done for the interruption to team two's innings.

The final case is if team one loses more resources than team two. Most of the time this will be the result of an interruption to team one's innings. This case presents some problems as team one may have been scoring at an unsustainable pace before the interruption. To overcome this Duckworth and Lewis suggest that the reset score for team two should be team ones score plus the difference in resources available multiplied by the average score in one day matches

Some of the cases that might occur are:

- **Case 1 - Both teams lose the same amount of resources**

  Team 1 is at bat in the first innings. They are 5 out for 220 at the end of the 40th over. Rain interrupts their innings. The rain continues into the start of team two's innings, removing 21 overs. According to the 2002 table, team one lost just over 26% of their resources. Team two also lost just over 26% of their resources. As both teams were equally disadvantage, team two's target score should be be left unchanged at 220.

- **Case 2 - Team 1 loses less resources**

  Let's use the same scenario as before but let's say that team two lose the first 30 overs of their innings. In this case, team one has lost 26.1% of their resources while team two have lost 43.4%. We should therefore reduce team one's score by 43.3%-26.1% = 17.3%. This gives a target score of 181.94. Team two then have to score 182 runs to win.

- **Case 3 - Team 1 loses more resources**

  Same scenario again but let's say that team two only lose the first 10 overs of their innings. Team one have again lost 26.1% of their resources while team two have lost 10.7%. Team 1 has therefore lost 15.4 percentage points more resources than team two. According to Duckworth and Lewis' method we should take team one's score of 220 and add the difference in resources multiplied by the average score in an uninterrupted match. If the average score in an uninterrupted match is 250 we get 15.4% * 250 = 38.5. This gives a target score of 258.5, or 260 runs to win.

4.3 Pitfalls in Duckworth Lewis Methods and its variants :

There are some shortcomings of DWL method and its variants. The following are some the fundamental problems that this method might be facing:

- DWL method is premised on the idea that teams use wickets and balls to produce runs according to some production function and that we can estimate this function from the data.
- The functions are fitted to data from the first innings only. This means that, for the method to work well, the second innings needs to look the same. Sadly, second innings data looks a bit different to first innings data.
- Revised scores for the second Innings will be based on historical data of previous matches.
- The DWL method, drawing its data from innings one, has put too much emphasis on wickets lost rather than balls remaining.

4.4 Proposed process flow diagram :

After extensive evaluation and research of the existing models the proposed algorithm has been developed.

The Figure 4.6 represents the overview of the proposed algorithm. These steps cover end-to-end development, implementation and evaluation of the model from raw data to end user application. The phases of this process are:

- Data Collection: Collection of unstructured raw data for 1348 international cricket matches.
- Data Preparation & Data Cleansing: Raw data to be cleaned and structured for standardisation and analysis
- Exploratory Analysis / Statistical tests: Breakdown and analysis of data by plotting and statistical testing.
- Design Mathematical functions: Different mathematical functions for Batting and Bowling were appraised and most appropriate were chosen.
- Data Modelling: The chosen mathematical functions were trained using neural network.
- Model Validation & Testing: This model was tested and validated for all the matches.
- Development of UI: Developed a website to showcase the ultimate results for better user interface.

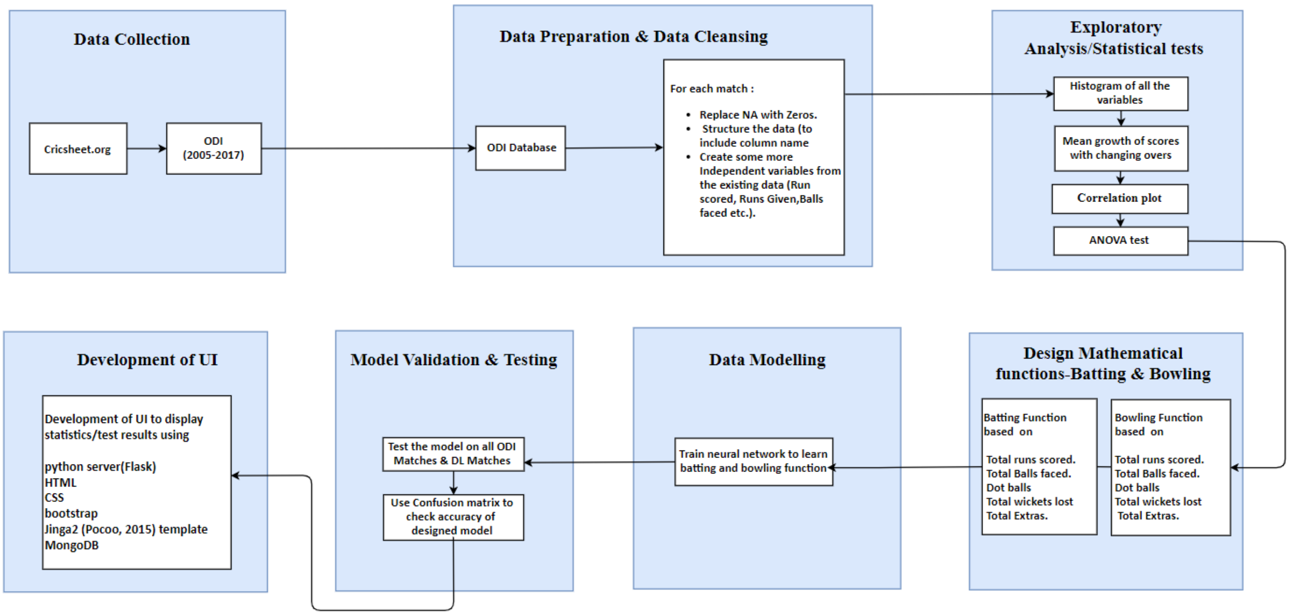
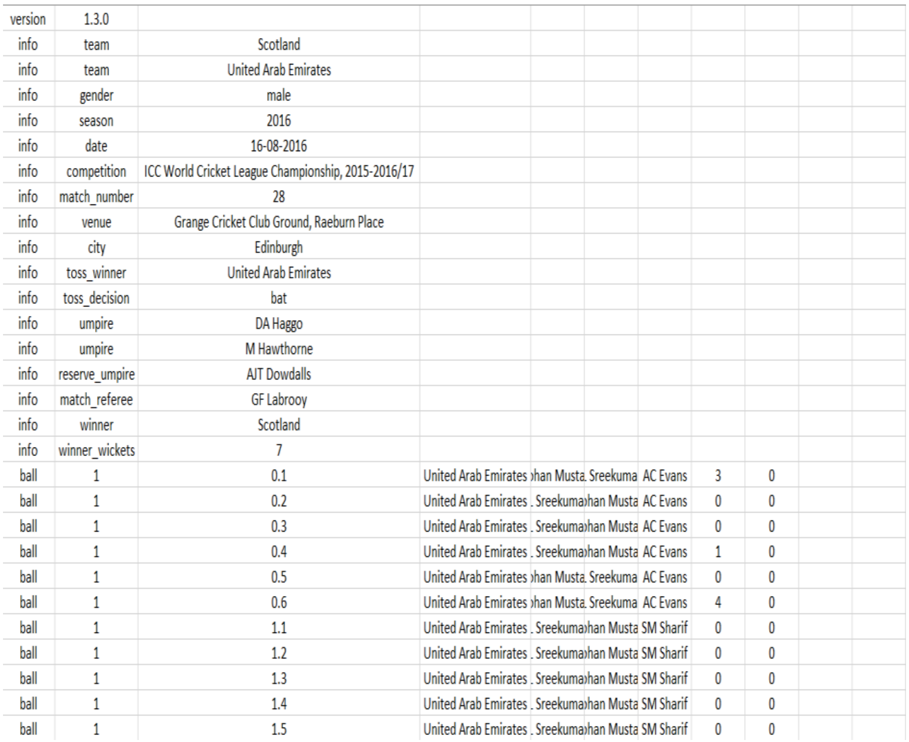

Data Collection
Cricsheet.org
ODI
(2005-2017)
Data Preparation & Data Cleansing
ODI Database
For each match :
Replace NA with Zeros.
Structure the data (to include column name
Create some more Independent variables from the existing data (Run scored, Runs Given,Balls faced etc.).
Exploratory Analysis/Statistical tests
Histogram of all the variables
Mean growth of scores with changing overs
Correlation plot
ANOVA test
Design Mathematical functions-Batting & Bowling
Batting Function based on
Total runs scored.
Total Balls faced.
Dot balls
Total wickets lost
Total Extras.
Bowling Function based on
Total runs scored.
Total Balls faced.
Dot balls
Total wickets lost
Total Extras.
Data Modelling
Train neural network to learn batting and bowling function
Model Validation & Testing
Test the model on all ODI Matches & DL Matches
Use Confusion matrix to check accuracy of designed model
Development of UI
Development of UI to display statistics/test results using
python server(Flask)
HTML
CSS
bootstrap
Jinga2 (Pocoo, 2015) template
MongoDB

Figure 4. 6 Process flow of proposed algorithm

4.4.1 Data collection and data preparation:

The raw data was collected from **cricsheet** for 1348 ODI matches. The raw data was in an unstructured format which required to be first deciphered and structured.

The Figure 4.7 shows the prepared data. All the images that were in the same folder

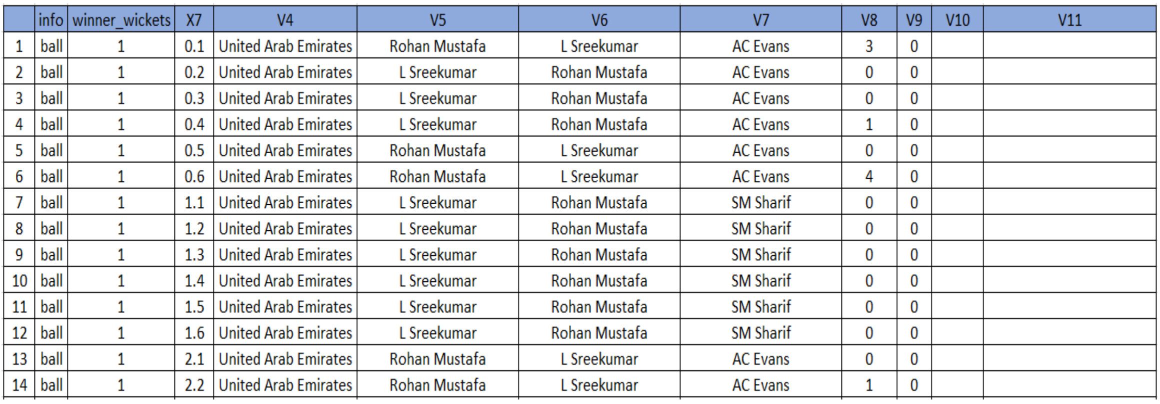


version	1.3.0									
info	team	Scotland								
info	team	United Arab Emirates								
info	gender	male								
info	season	2016								
info	date	16-08-2016								
info	competition	ICC World Cricket League Championship, 2015-2016/17								
info	match_number	28								
info	venue	Grange Cricket Club Ground, Raeburn Place								
info	city	Edinburgh								
info	toss_winner	United Arab Emirates								
info	toss_decision	bat								
info	umpire	DA Haggo								
info	umpire	M Hawthorne								
info	reserve_umpire	AJT Dowdalls								
info	match_referee	GF Labrooy								
info	winner	Scotland								
info	winner_wickets	7								
ball	1	0.1	United Arab Emirates	han Musta	. Sreekuma	AC Evans	3	0		
ball	1	0.2	United Arab Emirates	. Sreekuma	han Musta	AC Evans	0	0		
ball	1	0.3	United Arab Emirates	. Sreekuma	han Musta	AC Evans	0	0		
ball	1	0.4	United Arab Emirates	. Sreekuma	han Musta	AC Evans	1	0		
ball	1	0.5	United Arab Emirates	han Musta	. Sreekuma	AC Evans	0	0		
ball	1	0.6	United Arab Emirates	han Musta	. Sreekuma	AC Evans	4	0		
ball	1	1.1	United Arab Emirates	. Sreekuma	han Musta	SM Sharif	0	0		
ball	1	1.2	United Arab Emirates	. Sreekuma	han Musta	SM Sharif	0	0		
ball	1	1.3	United Arab Emirates	. Sreekuma	han Musta	SM Sharif	0	0		
ball	1	1.4	United Arab Emirates	. Sreekuma	han Musta	SM Sharif	0	0		
ball	1	1.5	United Arab Emirates	. Sreekuma	han Musta	SM Sharif	0	0		

Figure 4. 7 Snippet of the raw data

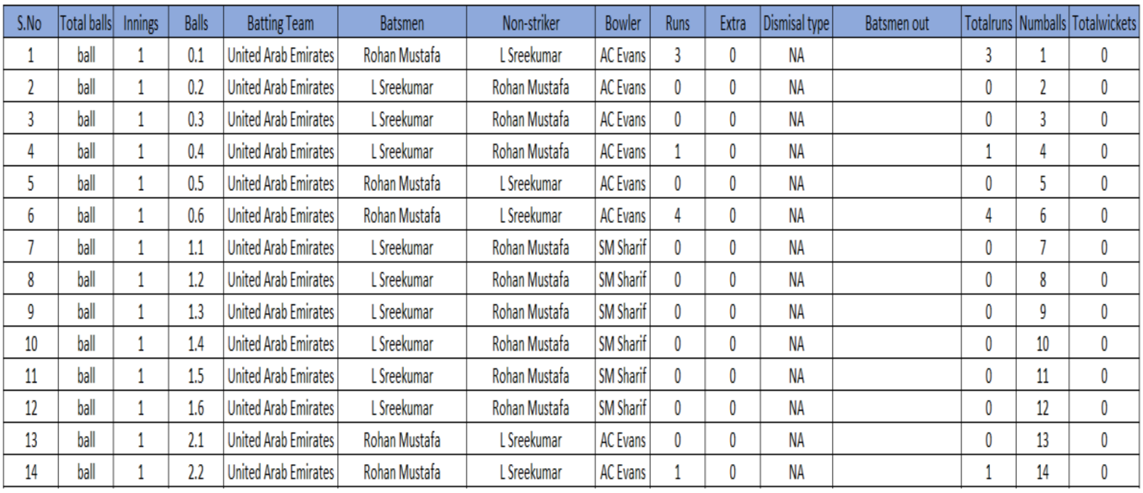
4.4.2 Data Cleansing and Preparation

Raw data requires extensive cleansing and preparation to suit the modelling aspects. Hence the Column names are renamed and structured and additional columns are derived to give a proper structure to the data.



	info	winner_wickets	X7	V4	V5	V6	V7	V8	V9	V10	V11
1	ball	1	0.1	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	3	0		
2	ball	1	0.2	United Arab Emirates	L Sreekumar	Rohan Mustafa	AC Evans	0	0		
3	ball	1	0.3	United Arab Emirates	L Sreekumar	Rohan Mustafa	AC Evans	0	0		
4	ball	1	0.4	United Arab Emirates	L Sreekumar	Rohan Mustafa	AC Evans	1	0		
5	ball	1	0.5	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	0	0		
6	ball	1	0.6	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	4	0		
7	ball	1	1.1	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0		
8	ball	1	1.2	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0		
9	ball	1	1.3	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0		
10	ball	1	1.4	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0		
11	ball	1	1.5	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0		
12	ball	1	1.6	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0		
13	ball	1	2.1	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	0	0		
14	ball	1	2.2	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	1	0		

Figure 4. 8 Snippet of structured data



S.No	Total balls	Innings	Balls	Batting Team	Batsmen	Non-striker	Bowler	Runs	Extra	Dismisal type	Batsmen out	Totalruns	Numballs	Totalwickets
1	ball	1	0.1	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	3	0	NA		3	1	0
2	ball	1	0.2	United Arab Emirates	L Sreekumar	Rohan Mustafa	AC Evans	0	0	NA		0	2	0
3	ball	1	0.3	United Arab Emirates	L Sreekumar	Rohan Mustafa	AC Evans	0	0	NA		0	3	0
4	ball	1	0.4	United Arab Emirates	L Sreekumar	Rohan Mustafa	AC Evans	1	0	NA		1	4	0
5	ball	1	0.5	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	0	0	NA		0	5	0
6	ball	1	0.6	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	4	0	NA		4	6	0
7	ball	1	1.1	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0	NA		0	7	0
8	ball	1	1.2	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0	NA		0	8	0
9	ball	1	1.3	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0	NA		0	9	0
10	ball	1	1.4	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0	NA		0	10	0
11	ball	1	1.5	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0	NA		0	11	0
12	ball	1	1.6	United Arab Emirates	L Sreekumar	Rohan Mustafa	SM Sharif	0	0	NA		0	12	0
13	ball	1	2.1	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	0	0	NA		0	13	0
14	ball	1	2.2	United Arab Emirates	Rohan Mustafa	L Sreekumar	AC Evans	1	0	NA		1	14	0

Figure 4. 9 Snippet of cleaned data

In the Figure 4.8 and 4.9, there are two operations that has been performed on the data. The first is extracting the raw unstructured data from the site and second step is to clean the data and convert it into a structured format with appropriate column names. There are also few additional derived columns used.

4..4.3 Exploratory Analysis

- Statistical tests are done in order to make inferences about the data and to understand which of the independent variable is affecting the dependent variables. The statistical tests carried out in this work is: ANOVA.
- Dependent variable – Innings 1 / Innings 2

  Independent variables – Total runs, total wickets, zeros, extras.

The plot of Overs versus Runs shows the comparative scoring pattern of both the teams,which is shown in Table 4.1.

Table 4. 1 Statistical test : Anova (Analysis of variance) for Innings 1 and Innings 2

```
Anova-Innings-1                       Df    Sum Sq   Mean Sq F value
Pr(>F)
Team1$Totalruns      1 8.500e-32 8.482e-32   1.734 0.1889
Team1$Totalwickets   1 2.090e-31 2.094e-31   4.281 0.0394 *
Team1$zeros          1 0.000e+00 3.100e-34   0.006 0.9362
Team1$Extra          1 2.300e-32 2.331e-32   0.477 0.4905
Residuals          305 1.492e-29 4.891e-32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Anova-Innings-2                       Df    Sum Sq   Mean Sq F value
Pr(>F)
Team2$Totalruns      1 1.000e-29 1.049e-29   0.409 0.5232
Team2$Totalwickets   1 1.900e-29 1.924e-29   0.750 0.3874
Team2$zeros          1 1.720e-28 1.724e-28   6.720 0.0101 *
Team2$Extra          1 0.000e+00 3.000e-32   0.001 0.9710
Residuals          229 5.875e-27 2.565e-29
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In the Table 4.1 & 4.2 the stars(***) represent the significance level. The higher the number of stars the greater is the relevance of the variable. For example there is one star(*) for Team1$Totalwickets which has a p-value of 0.0395. Since this value is less than 0.05 it means that this feature is more relevant for the model to use.

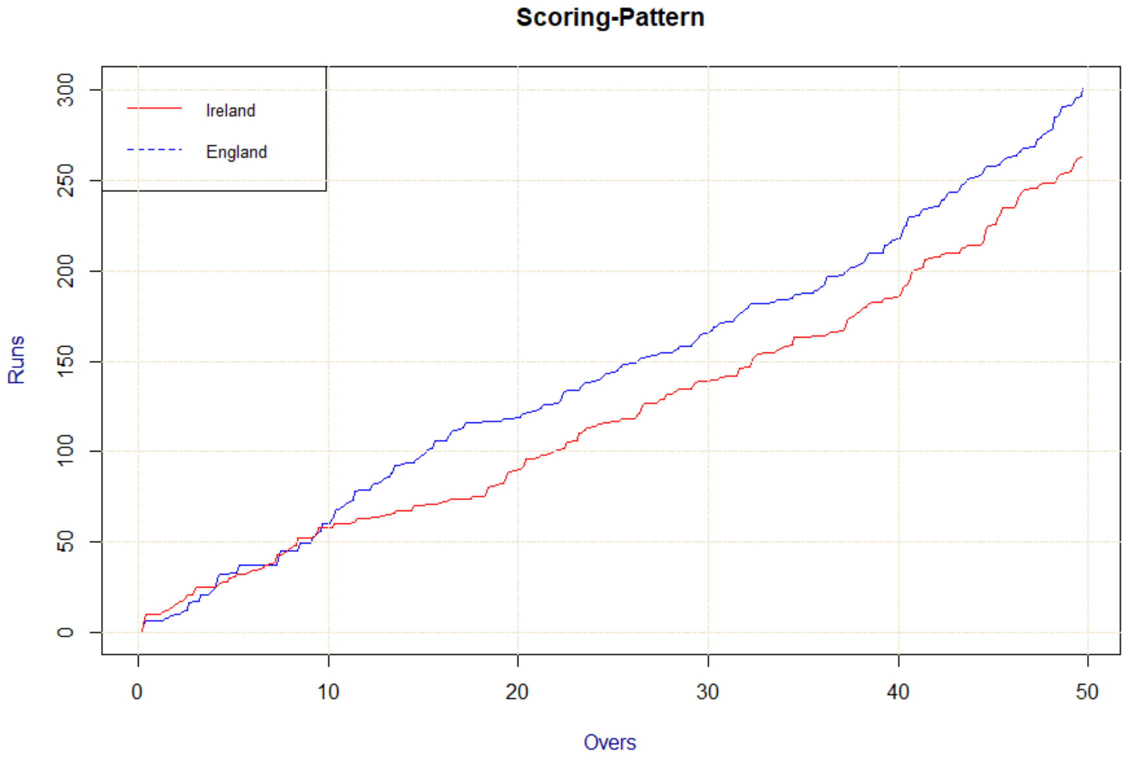


Figure 4. 10 Comparative scoring pattern of both the teams while batting

The figure 4.10 shows the comparative scoring pattern of team 1 and team 2 batting trends. At any given point in the match the comparative performance of each team can be known and analysed at any specific over. This trend analysis of the two teams shows the exact point in the match/ over where the team 2 (Blue – England) overtook the batting scores of the other team (Red – Ireland).

Table 4. 2 Basic statistics of Innings 1 and Innings 2 dataset

Innings 1		Innings 2	
Name	Value	Name	Value
Rows	416885	Rows	353075
Columns	12	Columns	12
Discrete columns	0	Discrete columns	0
Continuous columns	12	Continuous columns	12
All missing columns	0	All missing columns	0
Missing observations	0	Missing observations	0
Complete Rows	416885	Complete Rows	353075
Total observations	5002620	Total observations	4236900
Memory allocation	20.7 Mb	Memory allocation	17.5 Mb

Table 4.2 depicts the summary of the raw data informing about the number of rows and columns, data type of the variables, total number of missing observations and memory allocations for both innings. Innings 1 shows to have more number of data points.

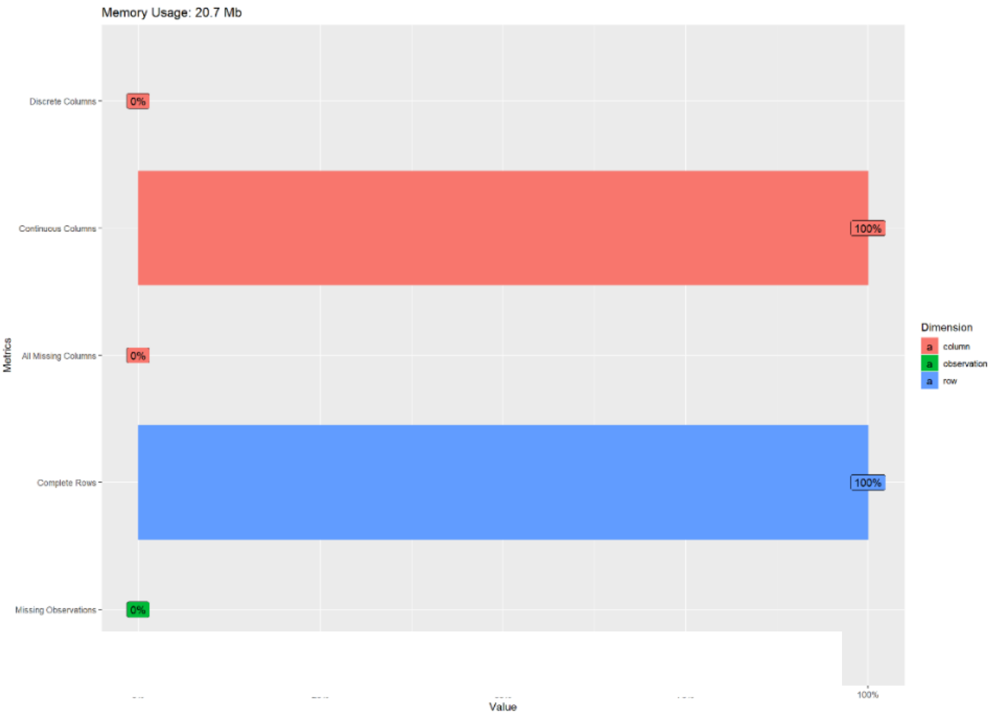


Figure 4. 11 Distribution of data type and missing data -Innings

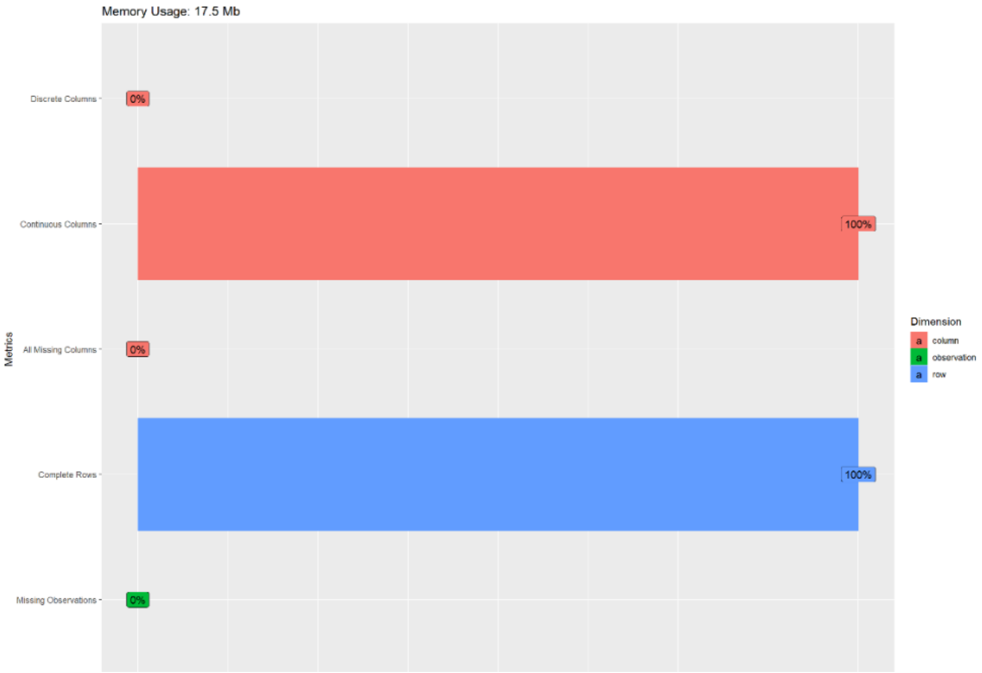
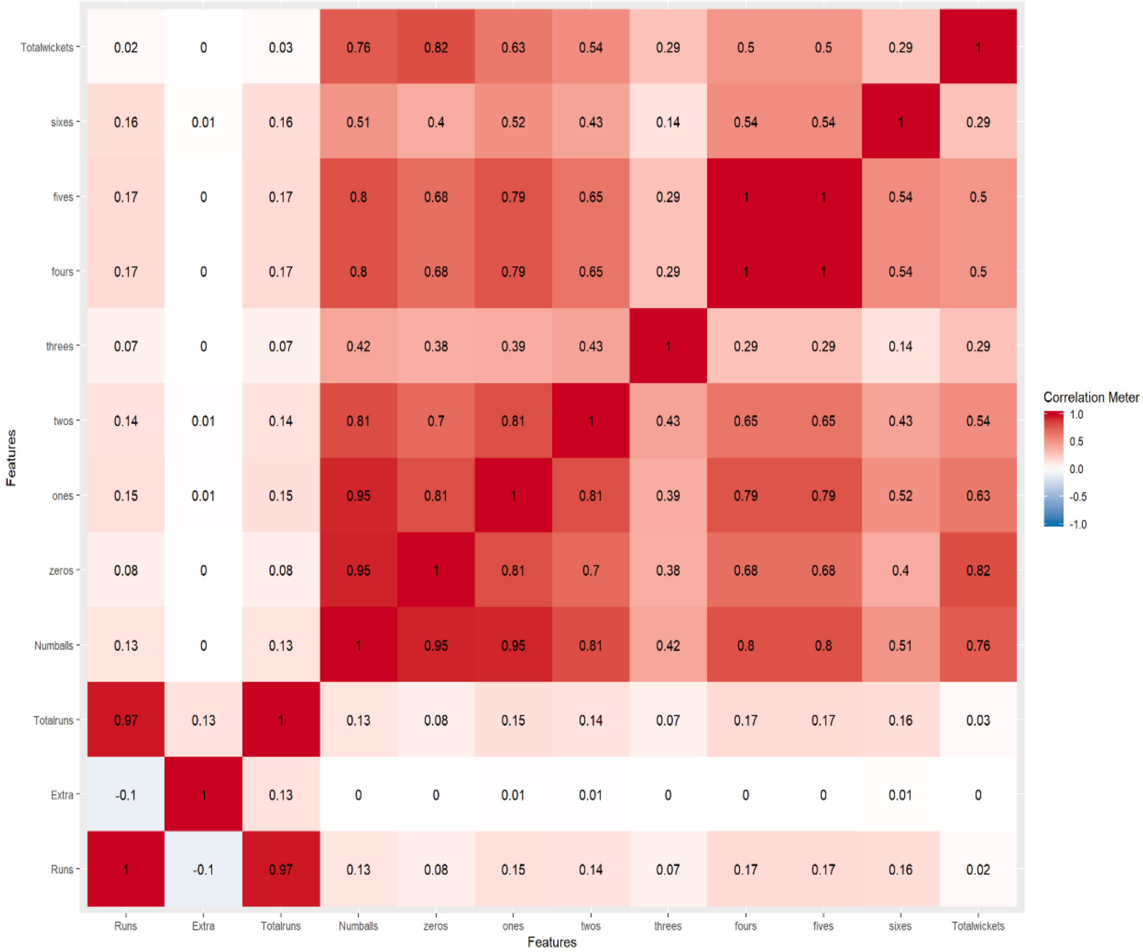


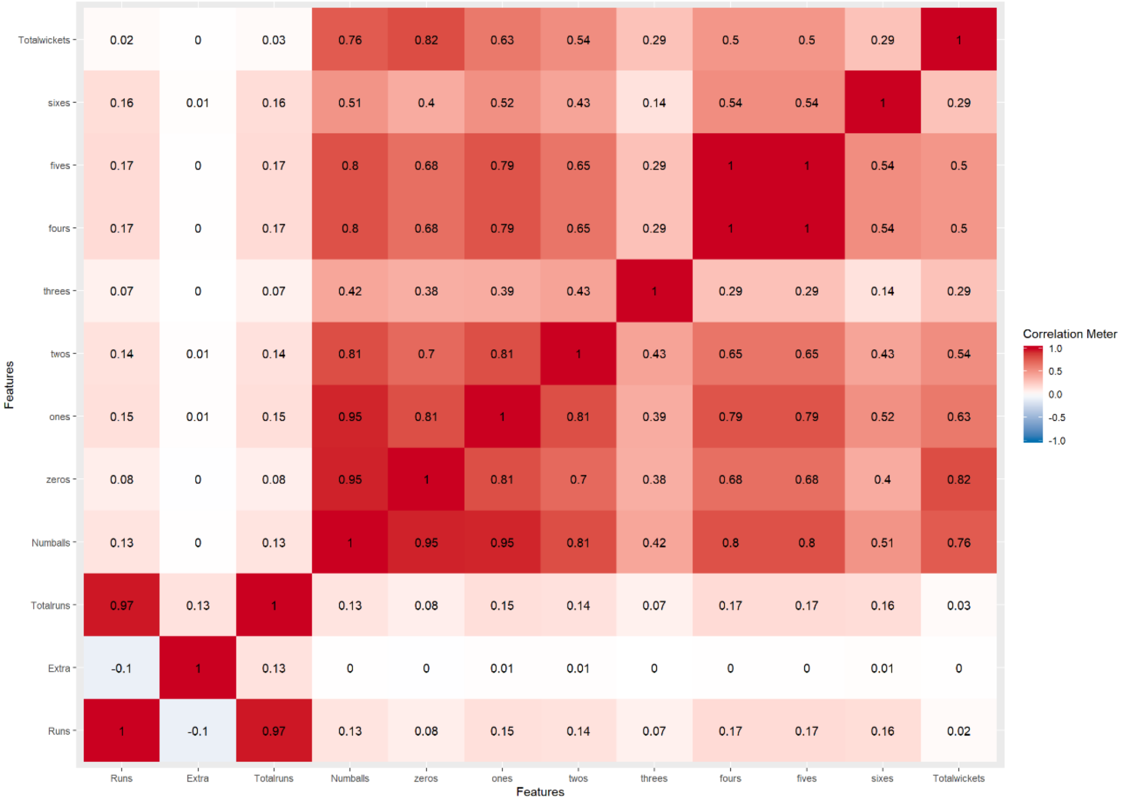
Figure 4. 12 Distribution of data type and missing data -Innings 2

Figure 4.11 and 4.12 depict the total percentage of variables are either discrete or continuous in the dataset, and the percentage of missing values for both innings. It is visible that in both the innings all the variables in the dataset are continuous and there are no missing values.



	Runs	Extra	Totalruns	Numballs	zeros	ones	twos	threes	fours	fives	sixes	Totalwickets
Totalwickets	0.02	0	0.03	0.76	0.82	0.63	0.54	0.29	0.5	0.5	0.29	1
sixes	0.16	0.01	0.16	0.51	0.4	0.52	0.43	0.14	0.54	0.54	1	0.29
fives	0.17	0	0.17	0.8	0.68	0.79	0.65	0.29	1	1	0.54	0.5
fours	0.17	0	0.17	0.8	0.68	0.79	0.65	0.29	1	1	0.54	0.5
threes	0.07	0	0.07	0.42	0.38	0.39	0.43	1	0.29	0.29	0.14	0.29
twos	0.14	0.01	0.14	0.81	0.7	0.81	1	0.43	0.65	0.65	0.43	0.54
ones	0.15	0.01	0.15	0.95	0.81	1	0.81	0.39	0.79	0.79	0.52	0.63
zeros	0.08	0	0.08	0.95	1	0.81	0.7	0.38	0.68	0.68	0.4	0.82
Numballs	0.13	0	0.13	1	0.95	0.95	0.81	0.42	0.8	0.8	0.51	0.76
Totalruns	0.97	0.13	1	0.13	0.08	0.15	0.14	0.07	0.17	0.17	0.16	0.03
Extra	-0.1	1	0.13	0	0	0.01	0.01	0	0	0	0.01	0
Runs	1	-0.1	0.97	0.13	0.08	0.15	0.14	0.07	0.17	0.17	0.16	0.02

Figure 4. 13 Correlation plot -Innings 1



Features	Runs	Extra	Totalruns	Numballs	zeros	ones	twos	threes	fours	fives	sixes	Totalwickets
Totalwickets	0.02	0	0.03	0.76	0.82	0.63	0.54	0.29	0.5	0.5	0.29	1
sixes	0.16	0.01	0.16	0.51	0.4	0.52	0.43	0.14	0.54	0.54	1	0.29
fives	0.17	0	0.17	0.8	0.68	0.79	0.65	0.29	1	1	0.54	0.5
fours	0.17	0	0.17	0.8	0.68	0.79	0.65	0.29	1	1	0.54	0.5
threes	0.07	0	0.07	0.42	0.38	0.39	0.43	1	0.29	0.29	0.14	0.29
twos	0.14	0.01	0.14	0.81	0.7	0.81	1	0.43	0.65	0.65	0.43	0.54
ones	0.15	0.01	0.15	0.95	0.81	1	0.81	0.39	0.79	0.79	0.52	0.63
zeros	0.08	0	0.08	0.95	1	0.81	0.7	0.38	0.68	0.68	0.4	0.82
Numballs	0.13	0	0.13	1	0.95	0.95	0.81	0.42	0.8	0.8	0.51	0.76
Totalruns	0.97	0.13	1	0.13	0.08	0.15	0.14	0.07	0.17	0.17	0.16	0.03
Extra	-0.1	1	0.13	0	0	0.01	0.01	0	0	0	0.01	0
Runs	1	-0.1	0.97	0.13	0.08	0.15	0.14	0.07	0.17	0.17	0.16	0.02

Figure 4. 14 Correlation plot -Innings 2

Tables 4.13 and 4.14 shows the correlation plots between all the variables for the both innings. It can be interpreted that there is high correlation existing between variables:

- Total wickets and number of ball faced – 76%
- Zeros and total wickets - 82%

There is also low correlation between variables:

- Extras and number of balls – 0%
- Sixes and number of balls – 51%
- Wickets and runs – 2%
- Wickets and extras – 0%

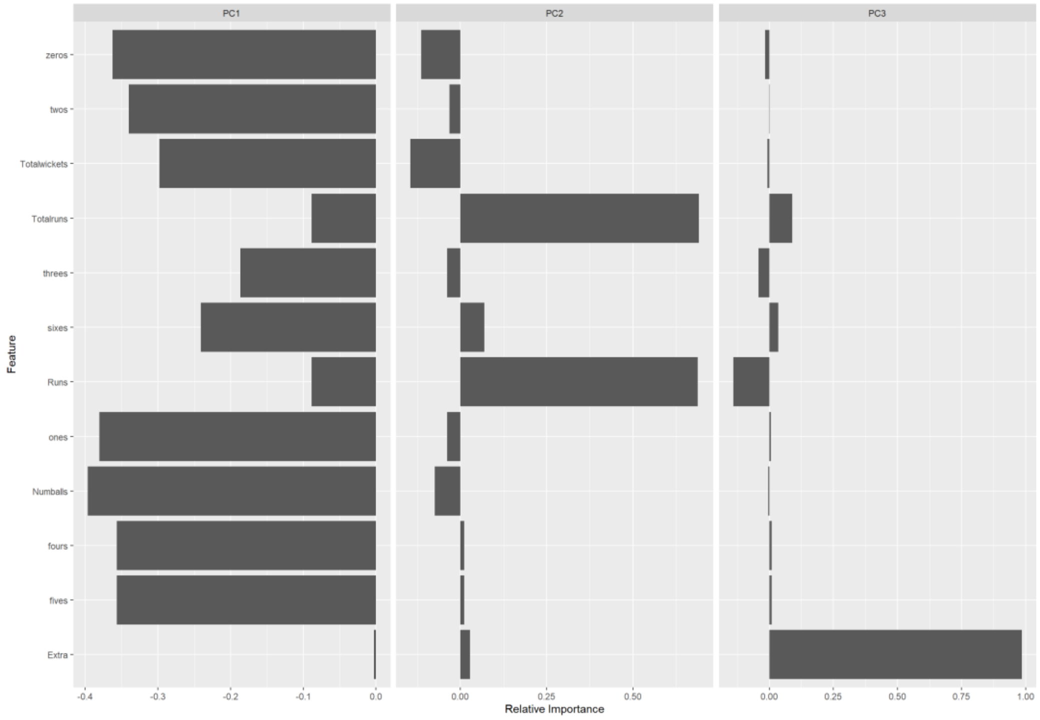


Figure 4. 15 Relative Importance of all variables -Innings 1

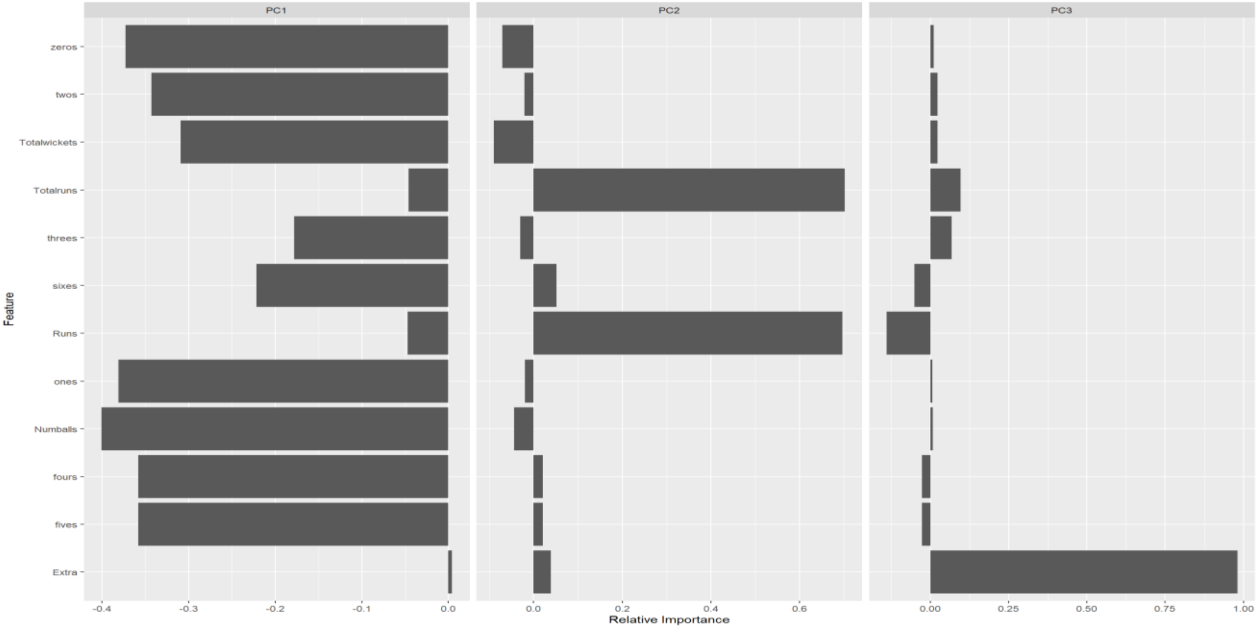


Figure 4. 16 Relative Importance of all variables -Innings 2

Figures 4.16 and 4.17 show the importance of variables in both the innings. For both innings the high importance variables are:

- Number of balls
- Runs scored
- Zeros
- Fours
- Twos

The rest of the variables hardly contribute to the explanation of the variance of dataset.

4.4.5 Design Mathematical Functions (Batting & Bowling)

Various functions of different combinations of variables were tried and tested and only the equations 1 and 2 were found to be aptly suitable.

Functions for both batting and bowling are designed using the following four variables:

- Number of Runs scored
- Number of Extras
- Number of Dot balls
- Number of wickets lost

$$F(Runs, wickets, dot\ balls, extras) = a1 * \frac{Total\ runs}{Total\ Balls\ faced} + a2 * \frac{Total\ extras}{Total\ Balls\ faced} + a3 * \frac{Total\ wickets}{Total\ Balls\ faced} + a4 * \frac{Total\ number\ of\ dot\ balls}{Total\ Balls\ faced}$$

$$F(Runs, wickets, dot\ balls, extras) = b1 * \frac{Total\ runs}{Total\ Balls\ faced} + \mathrm{b2} * \frac{Total\ extras}{Total\ Balls\ faced} + \mathrm{b3} * \frac{Total\ wickets}{Total\ Balls\ faced} + \mathrm{b4} * \frac{Total\ number\ of\ dot\ balls}{Total\ Balls\ faced}$$

Where,

$a1$, $a2$, $a3$, $a4$, $b1$, b2, b3 , b4 are the constants and should be calculated for the data.

- Subsequently a neural network is trained to learn both batting and bowling related functions.

Algorithm designed for choosing the winner are elaborated in the following steps:

- ✓ **Step 1**: Compute the batting and bowling related function values for both the Innings.
- ✓ **Step 2**: If the match stops in between during the second Innings, calculate the total function value for both Innings of the teams.

Table 4. 3 Abbreviations used for Mathematical function for both innings

Innings 1	***Innings 2***
F1 (Batting function-Team 1)	F3 (Batting function-Team 2)
F2 (Bowling function-Team 2)	F4 (Bowling function-Team 1)

Table 4.3 depicts the abbreviations used for both batting and bowling of teams 1 and 2.

- Next is to calculate the total function value of both the Innings.

Total function value for Team 1 = F1 + F4

Total function value for Team 2 = F2 + F3

- ✓ **Step 3:** Winner can be decided based on any one the measures
- Comparing total function value for Team 1 and Team 2 .
- Based on the function value calculate frequency of instances where Team 1 and Team2 are the winners.

- Finally, to Train neural network to learn both batting and bowling related functions.

4.4.6 Train Neural Network

A simple ANN consisting of one input layer two hidden layers and one output layer is used to learn the proposed batting and bowling functions. The structure of the ANN used is shown below in Figure 5 and Figure 6. The four input nodes takes input of total runs, total wickets, extras and dot balls respectively. At a time the output node gives either the batting function value or the bowling function value.

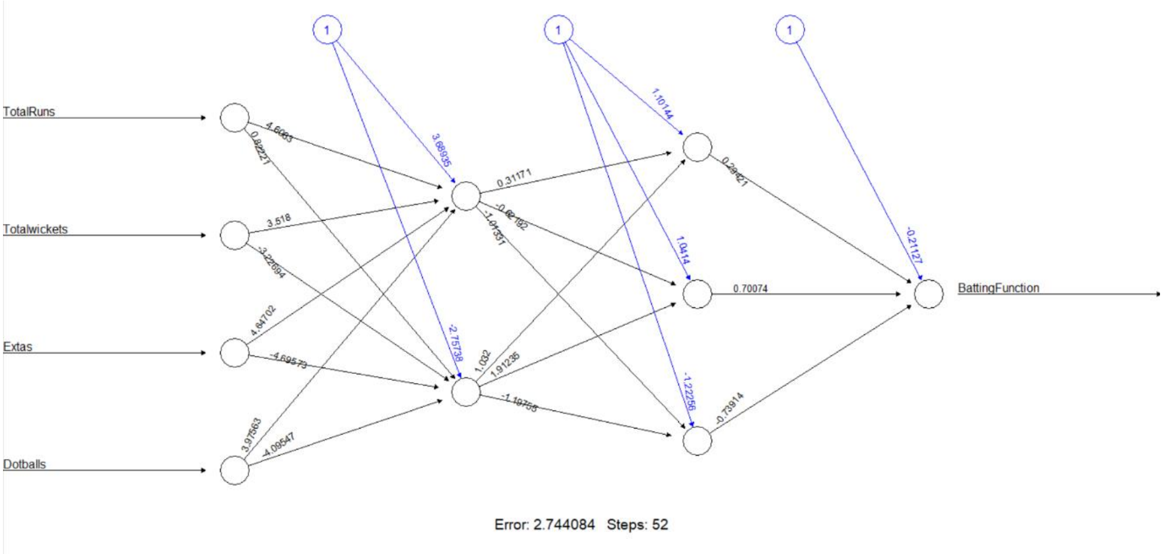


Figure 4. 17 Neural network architecture of trained batting function

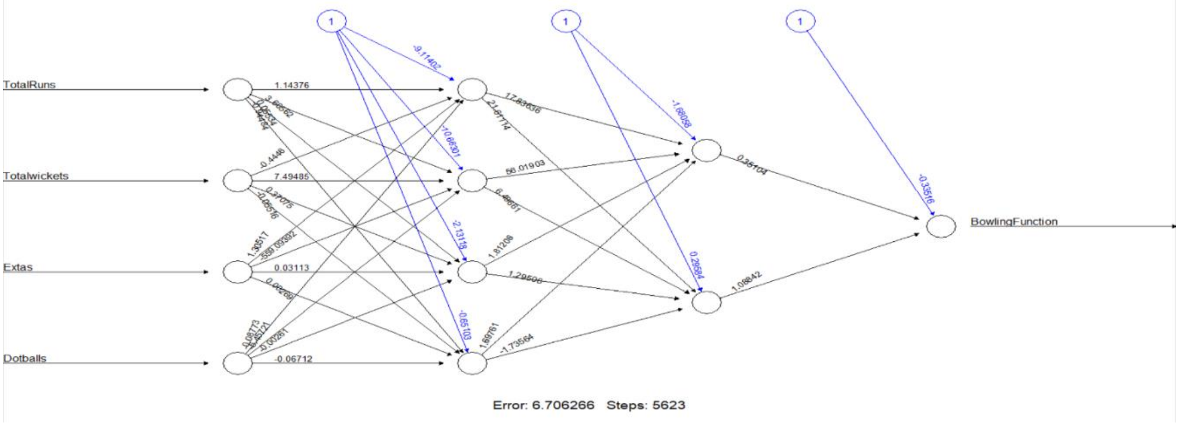


Figure 4. 18 Neural network architecture of trained bowling function

Figures 4.17 and 4.18 show the architecture of the neural networks trained for batting and bowling functions. Several combinations of hidden layers with multiple combinations of neurons are tried but the architecture which gives the least MSE has been chosen.

Table 4. 4 Comparison of training parameters used for batting and bowling functions

Parameters	Batting Function	Bowling Function
Inputs	Total runs,Total wickets,Dot balls,Extras	Total runs,Total wickets,Dot balls,Extras
Ouput	Batting function value	Bowling function value
Hidden Layers	2 Hidden layers (2-3)	2 Hidden layers (4-2)
Steps	52	5623
Mean square error	2.7	6.7

Table 4.4 summarizes the parameters used by both batting and bowling related function. The mean square error is optimal with a value of 2.7 at the 52nd iteration for the batting function and the mean square error is optimal with a value of 6.7 at the 5623rd iteration

for the bowling function. The architecture of batting and bowling function is 4-2-3-1, and 4-4-2-1.

5. Test Results

The chapter enclosed herein is a prologue of the implemented models results and their validation. The main objectives and aim of the proposed algorithm are depicted here. Accuracy of the model is provided and overall performance of the model is also tabulated.

5.1 Algorithm output for each match

The results of the proposed algorithm are standardising both the “Batting” and “Bowling” performances of the 2 teams and evaluating these performances on a standardised common scale. The mathematical model compares varied different factors/ variables of the teams, such as:

- Runs scored
- Wickets lost
- Extras
- Dot balls

The batting team is trying to maximise “Runs scored” & “Extras” and minimise the “Wickets lost” & “Dot balls” and vice versa for the bowling team. The team “Batting” in the first innings will be “Bowling” in the second innings and vice versa for the other team too.

The designed mathematical function captures all these parameters and tries to incorporate these into the model. The mathematical function value for both the “Batting” and “Bowling” teams lies between 0 and 1. This is depicted in Figure 5.1 and 5.2.

From these standardised values, at any point in the game the performance of either team in both the innings can be evaluated. This can be seen in Figure 5.3. As discussed in the proposed algorithm, to bring fairness in the evaluation, the function values are added for both teams (Innings 1 & Innings 2). The team with the greater function value at a point in time is the winner. i.e.at every ball a winner can be declared !!

These instances are converted into percentage as depicted in Figure 5.4.

To give a graphical depiction of the test results, the following sample is chosen of a match between Scotland and UAE. The innings 1 – where Scotland is “Batting” and UAE is “Bowling” is shown in Figure 5.1.

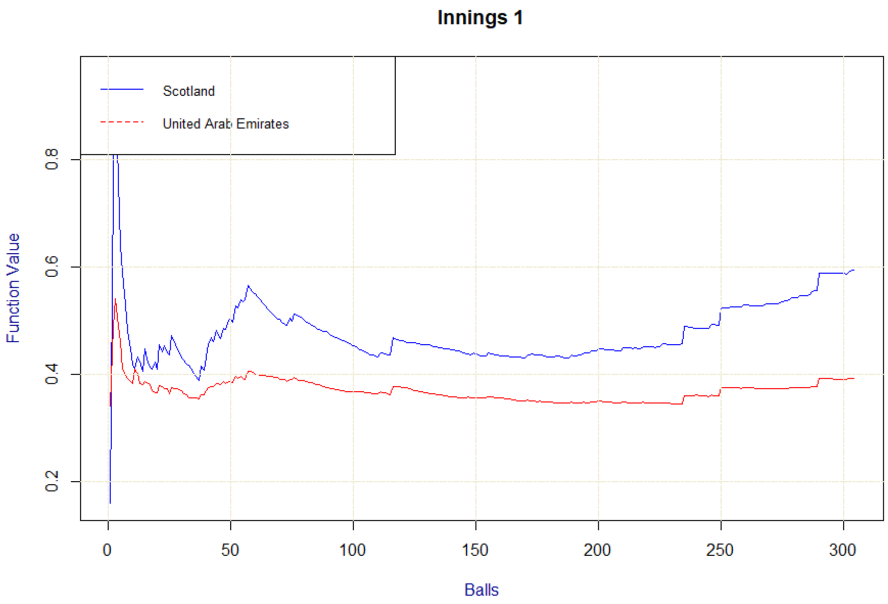


Figure 5. 1 Function value - Plot for Innings 1

Figure 5.1 graphs the plots for Innings 1 between Scotland and UAE

Interpretation of this graph: It can be noticed that Scotland has shown a tremendous spike in performance till the 15th ball pf the match. Post this their performance has been steady and slightly above the performance of UAE.

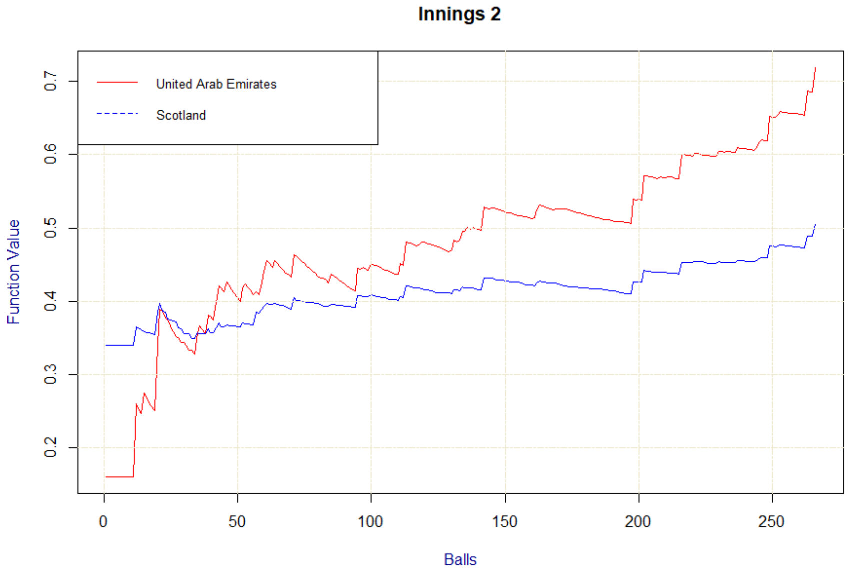


Figure 5. 2 Function value plot for Innings 2

The innings 2 – where Scotland is “Bowling” and UAE is “Batting” is shown in Figure 5.2. Here too both teams performances are being standardised and plotted to show comparative performance at any given ball in the match.

Interpretation of this graph: There is drastic out performance by UAE from the 35th ball of the match. From there onwards, UAE has outperformed Scotland.

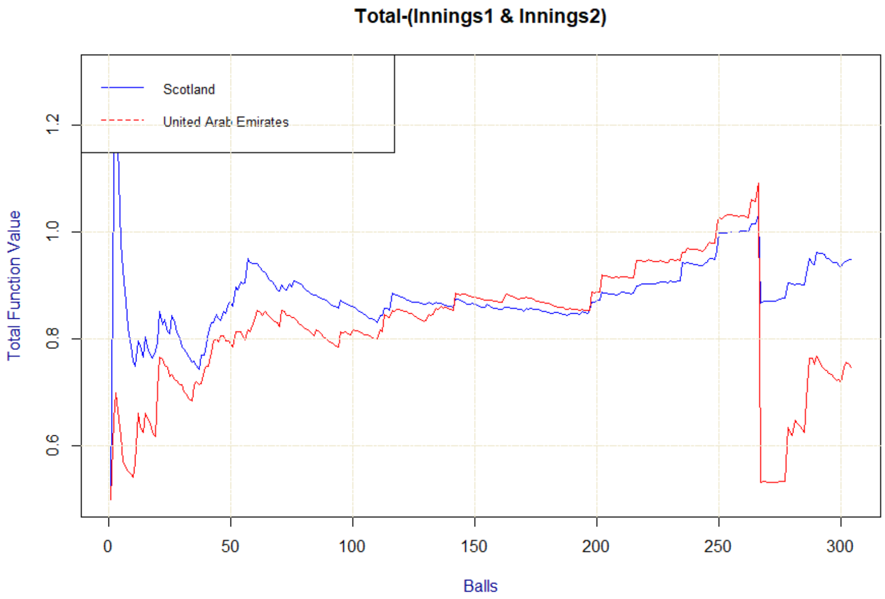
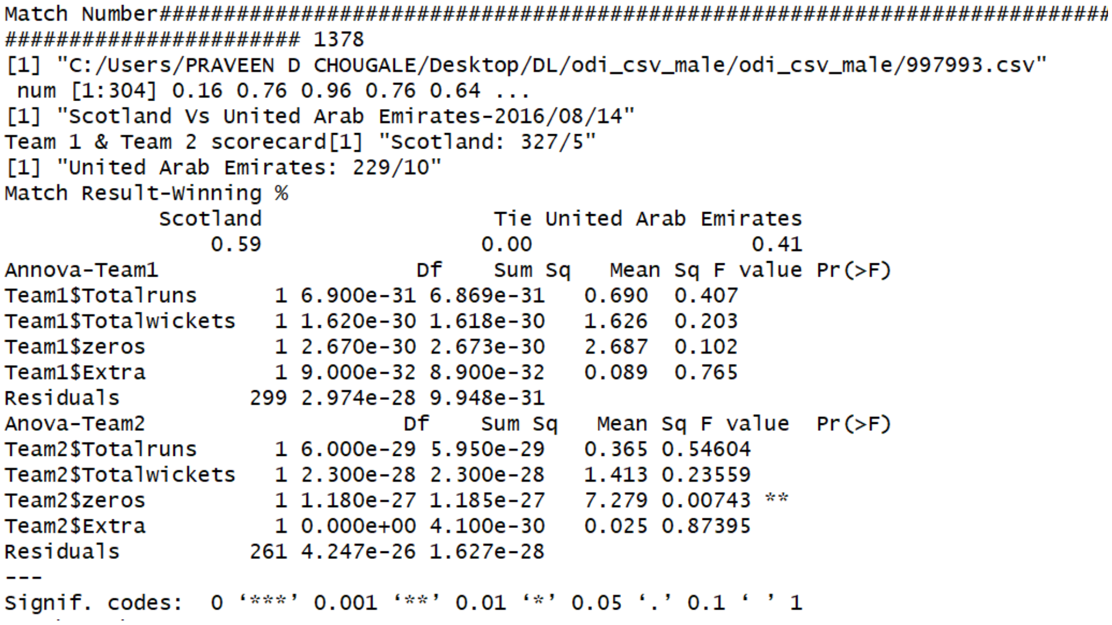


Figure 5. 3 Function value: Plot for combination of Innings 1 and Innings 2

Figure 5.3 shows the comparison of total function values for both the teams. At any point of the game the winner can be chosen based on converting these instances into a percentage.

Interpretation of this graph: Till the 20th ball Scotland displays pumped up performance as shown by the spike in the graph. From the 100th to the 250th ball both teams are equally matched, but the last 50 balls of the match decided the winner – Scotland ! UAE's performance and ability to handle the pressures of the tightly fought match cost them the match.



```
Match Number###########################################################################
####################### 1378
[1] "C:/Users/PRAVEEN D CHOUGALE/Desktop/DL/odi_csv_male/odi_csv_male/997993.csv"
 num [1:304] 0.16 0.76 0.96 0.76 0.64 ...
[1] "Scotland Vs United Arab Emirates-2016/08/14"
Team 1 & Team 2 scorecard[1] "Scotland: 327/5"
[1] "United Arab Emirates: 229/10"
Match Result-Winning %
            Scotland                  Tie United Arab Emirates
                0.59                 0.00                 0.41
Annova-Team1                    Df    Sum Sq   Mean Sq F value Pr(>F)
Team1$Totalruns      1 6.900e-31 6.869e-31   0.690  0.407
Team1$Totalwickets   1 1.620e-30 1.618e-30   1.626  0.203
Team1$zeros          1 2.670e-30 2.673e-30   2.687  0.102
Team1$Extra          1 9.000e-32 8.900e-32   0.089  0.765
Residuals          299 2.974e-28 9.948e-31
Annova-Team2                   Df    Sum Sq   Mean Sq F value  Pr(>F)
Team2$Totalruns      1 6.000e-29 5.950e-29   0.365 0.54604
Team2$Totalwickets   1 2.300e-28 2.300e-28   1.413 0.23559
Team2$zeros          1 1.180e-27 1.185e-27   7.279 0.00743 **
Team2$Extra          1 0.000e+00 4.100e-30   0.025 0.87395
Residuals          261 4.247e-26 1.627e-28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.4 Result of the match using proposed algorithm

Figure 5.4 shows the scorecard of both the teams, the winning percentage and ANOVA table. This shows the entire match summary.

Interpretation: It is seen that the designed algorithm interprets Scotland winning percentage is 59 and UAE is 41. The actual match result also shows that Scotland won the match. In addition, this model also shows which factor influenced the performance in both the innings. In Innings 2 – Dot balls influenced the match (** in the Annova Table) with 99.9% confidence.

The proposed statistical model has been run for all the raw data for 1379 ODI matches. A ball by ball interpretation of the performance between two teams has been evaluated for all these matches. Also, the factor that played a crucial role in the winning teams performance can also be attributed for these matches.

5.2 Validation:

The designed algorithm after implementation is validated to check its performance accuracy. Such validation is done for 1311 completed ODI matches and also 120 D/L matches.

S.NO	Team1	Team2	Date	Actual result	Predicted result
1	Scotland	United Arab Emirates	16-08-2016	Scotland	Scotland
2	Scotland	United Arab Emirates	14-08-2016	Scotland	Scotland
3	Ireland	Afghanistan	19-07-2016	Ireland	Afghanistan
4	Ireland	Afghanistan	17-07-2016	Afghanistan	Afghanistan
5	Ireland	Afghanistan	12-07-2016	Afghanistan	Afghanistan
6	Sri Lanka	Australia	01-09-2016	Australia	Australia
7	Sri Lanka	Australia	31-08-2016	Australia	Australia
8	Sri Lanka	Australia	28-08-2016	Australia	Australia
9	Sri Lanka	Australia	24-08-2016	Sri Lanka	Australia
10	Sri Lanka	Australia	21-08-2016	Australia	Australia
11	Afghanistan	Zimbabwe	06-01-2016	Afghanistan	Afghanistan
12	Afghanistan	Zimbabwe	04-01-2016	Zimbabwe	Afghanistan
13	Afghanistan	Zimbabwe	02-01-2016	Zimbabwe	Afghanistan
14	Afghanistan	Zimbabwe	29-12-2015	Afghanistan	Afghanistan
15	Afghanistan	Zimbabwe	25-12-2015	Afghanistan	Afghanistan
16	South Africa	Sri Lanka	10-02-2017	South Africa	South Africa
17	South Africa	Sri Lanka	07-02-2017	South Africa	South Africa
18	South Africa	Sri Lanka	04-02-2017	South Africa	South Africa
19	South Africa	Sri Lanka	01-02-2017	South Africa	South Africa
20	South Africa	Sri Lanka	28-01-2017	South Africa	South Africa

Table 5. 1 Sample of validation results for 20 completed ODI matches

Table 5.1 is a snippet of the results of actuals results versus the predicted results by the designed algorithm for all the completed ODI matches.

Table 5. 2 Confusion matrix for completed ODI matches result against proposed model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 255 306
         1 266 484

               Accuracy : 0.5637
                 95% CI : (0.5363,
0.5907)
    No Information Rate : 0.6026
    P-Value [Acc > NIR] : 0.9981

                  Kappa : 0.1008
 Mcnemar's Test P-Value : 0.1030

            Sensitivity : 0.4894
            Specificity : 0.6127
         Pos Pred Value : 0.4545
         Neg Pred Value : 0.6453
             Prevalence : 0.3974
         Detection Rate : 0.1945
   Detection Prevalence : 0.4279
      Balanced Accuracy : 0.5511

       'Positive' Class : 0
```

Table 5.2 is the confusion matrix which shows the accuracy of the model.

Observations :

- The overall accuracy rate is computed along with a 95 percent confidence interval is 56.37 %
- A p-value from McNemar's test is 0.10 ,which statistically significant .
- The sensitivity and specificity of the model are 48.94 % and 61.27 % respectively , from which it can interpreted that the algorithm is biased towards team batting in innings2.

Table 5. 3 Sample of validation results for 20 ODI matches where D/L was applied

S.No	Team1	Team2	Date	Actual result	Predicted result	Method
1	Zimbabwe	Pakistan	03-10-2015	Zimbabwe	Pakistan	D/L
2	Sri Lanka	West Indies	04-11-2015	Sri Lanka	Sri Lanka	D/L
3	Sri Lanka	West Indies	01-11-2015	Sri Lanka	Sri Lanka	D/L
4	New Zealand	Pakistan	31-01-2016	New Zealand	New Zealand	D/L
5	England	Sri Lanka	29-06-2016	England	England	D/L
6	Bangladesh	India	21-06-2015	Bangladesh	Bangladesh	D/L
7	Sri Lanka	England	03-12-2014	England	England	D/L
8	Sri Lanka	Pakistan	30-08-2014	Sri Lanka	Pakistan	D/L
9	Sri Lanka	Pakistan	23-08-2014	Pakistan	Pakistan	D/L
10	Bangladesh	India	17-06-2014	India	Bangladesh	D/L
11	Bangladesh	India	15-06-2014	India	Bangladesh	D/L
12	Ireland	Australia	27-08-2015	Australia	Australia	D/L
13	England	New Zealand	20-06-2015	England	England	D/L
14	England	New Zealand	12-06-2015	New Zealand	England	D/L
15	Scotland	England	09-05-2014	England	England	D/L
16	Sri Lanka	New Zealand	16-11-2013	Sri Lanka	New Zealand	D/L
17	Sri Lanka	New Zealand	12-11-2013	New Zealand	New Zealand	D/L

18	Bangladesh	New Zealand	29-10-2013	Bangladesh	Bangladesh	D/L
19	England	Sri Lanka	22-05-2014	England	England	D/L
20	England	India	30-08-2014	India	England	D/L

Table 5.3 is a snippet of the results of actuals results versus the predicted results only for D/L matches by the designed algorithm. From the entire dataset only the D/L matches are chosen, a total of 120 matches.

Table 5. 4 Confusion matrix for matches where D/L was applied against proposed model

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 29 16
         1 33 42

               Accuracy : 0.5917
                 95% CI : (0.4982,
0.6805)
    No Information Rate : 0.5167
    P-Value [Acc > NIR] : 0.05989

                  Kappa : 0.1901
 Mcnemar's Test P-Value : 0.02227

            Sensitivity : 0.4677
            Specificity : 0.7241
         Pos Pred Value : 0.6444
         Neg Pred Value : 0.5600
             Prevalence : 0.5167
         Detection Rate : 0.2417
   Detection Prevalence : 0.3750
```

Balanced Accuracy : 0.5959
'Positive' Class : 0

Table 5.4 is the confusion matrix which shows the accuracy to be 59.17% . in this case also specificity is higher i.e. 72.41% from which it can interpreted that the completed ODI matches and D/L matches the batting in second innings is more difficult in D/L.

Observations :

- The overall accuracy rate is computed along with a 95 percent confidence interval is 59.17 %
- A p-value from McNemar’s test is 0.05 ,which statistically significant (Reject the null hypothesis and accept the alternate hypothesis)
- The sensitivity and specificity of the model are 46.77 % and 72.41 % respectively , from which it can interpreted that the algorithm is biased towards team batting in innings 2 .

Conclusion:

The designed mathematical functions for both “Batting” and “Bowling” were plotted for both the Innings and a thorough detailed analysis of all factors and variables was done. The statistical model is tested and validated on both 1311 completed ODI matches and 120 D/L matches. The winners and winning factors for all these matches were identified. The accuracy of the model for both completed ODI matches and D/L matched is 57 % and 61% respectively. The comparison of accuracy, sensitivity, specificity and kappa values shows the proposed model performs better on D/L matches as compared to completed ODI matches. Also, from the sensitivity and specificity values it can observed that the model is biased towards the team batting second.

6. Conclusion

The chapter bound herein gives the conclusive judgement of the analysis of proposed research implemented. The multiple factors pertaining to the analysis results, the model performances, the visualization, UI, and the scope for future work is discussed.

7.1 Conclusion:

The proposed design had set aims and objectives that are successfully fulfilled at each stage. The conclusion drawn from the entire project executed is as follows:

D/L method is implemented and analysed to identify the pitfalls. To mathematically analyze the existing method of Duckworth-Lewis and its variants for resetting the target in rain interrupted One-day Internationals, data collection, preparation, cleansing, structuring and feature extraction processes needed to be executed so as to model the processed data. Ball by ball data for all the ODI matches from 2005-2017 was collected, cleaned and analysed. After which structuring of data in a specific format and replacing of missing data, suitable for modelling was done. Feature extraction performed based on the mean and variance of batting and bowling features are extracted. Next to perform a comparative analysis in order to deduce the pitfalls of the Duckworth-Lewis and its variants by implementing them on historical data, statistical estimation functions related to both batting and bowling are developed and implemented based on the resource availability at any instance in the game.

A Machine learning algorithm is then developed by applying the formulated statistical estimation technique to reset the target/decide the match outcome. The implemented model is then validated /tested on the past D/L applied rain interrupted ODI matches and develop an user interface to visualize the results. The results of the proposed model are validated for all the ODI matches and also for D/L matches. A web application with a UI was designed and implemented using HTML and Python backend support. The proposed design implemented gave an accuracy of 66.6 % on all the D/L matches.

All in all, the proposed design implemented, surpassed previously implemented researches and depicted higher results in terms of model performance accuracy when compared to the already implemented and published algorithms which showed approximately about 50% to 55% on an average.

7.2 Future scope

Various factors can be considered as the future scope of this project. Certain facets that could be considered for the further enhancement of this technique would be as follows:

1. **More robust model** - The present methodology is only confined to batting and bowling related features. This algorithm can be extended to include other factors like:
    - Pitch conditions
    - Toss win
    - Opposition
    - Team composition
    - Player’s form

2. **Team Management Tool** - An optimization algorithm can be used to choose the batting order at any point of the game
3. **All Formats** - The evaluation of model is done only for ODI ,which can be extended for T20 matches as well.
4. **In Game Recokner -** The UI could be made a live running model which could give immediate and instant results at any time of the game.
5. **Better accuracy** - The proposed model uses statistical estimation technique to predict the match result, however there is a scope of improvement of this model by using linear /integer programming.

6. The evaluation of model is done only on ten test samples (matches) for both ODI and T20, which can be extended to numerous samples to check the model to be unbiased and is not overfitting.
7. The data collected was only confined to ODI format and not on T20 format for the reasons of time constraint since it is a tedious process, hence the algorithm developed could be progressed and analysed on Test format data for similar or more features.
8. The Web Application developed will also follow the implementation of other formats data analysis results to be showcased.
9. The UI could be made a live running model which could give immediate and instant results.

- Duckworth, FC & Lewis, AJ, A fair method of resetting the target in interrupted one- day cricket matches, Journal of the Operational Research Society, vol. 49, no. 3, pp. 220-227, 1998.
- Bhogle, S, Is Jayadevan's proposed method better than the Duckworth/Lewis method?
- http://www.rediff.com/cricket/2001/may/21srini.htm, 2001, 2001.
- Bandulasiri, A, Predicting the Winner in One Day International Cricket, Journal of Mathematical Sciences and Mathematics Education, vol.3, no. 1, pp. 6-17, 2008.
- Jayadevan, V, A New Method for the Computation of Target Scores in Interrupted Limited-Over Cricket Matches, Current Science, vol. 83, no. 5, pp. 577–586, 2002.
- de Silva, R, A Fair Target Score Calculation Method for Reduced-Over One day and T20 International Cricket Matches, Journal of Mathematical Sciences & Mathematics Education vol. 8, no. 2, pp. 6-19, 2013.
- Duckworth, FC & Lewis, AJ, Duckworth/Lewis Method of Re-calculating the Target Score in an Interrupted Match, http://static.icc-cricket.yahoo.net/ugc/documents/DOC_1F1135280 40177329F40FE47C77AE2_1254317757686_695.pdf.
- Swartz, TB, Gill, PS, Beaudoin, D & de Silva, MB, Optimal batting orders in one-day cricket, Computers & Operations Research, vol. 33, no. 7, pp. 1939–1950, 2006.
- Norman, JM & Clarke, SR, Optimal batting orders in cricket, Journal of the Operational Research Society, vol. 61, pp. 980-986, 2010.

- de Silva, MB, Pond, GR & Swartz, TB, Estimation of the magnitude of victory in one-day cricket, Australian and New Zealand Journal of Statistics, vol. 43, pp. 259-268, 2001.
- Lemmer, HH, The single match approach to strike rate adjustments in batting performance measures in cricket, Journal of Sports Science and Medicine, vol. 10, no. 4, pp. 630-634, 2010.
- Pocoo (2015) *Flask documentation*. Available at: http://flask.pocoo.org/docs/0.10/. [Accessed 02 October 2015].
- S., G., R., H.-W. and H.P., S. (2010) 'A support vector machine for decision support in melanoma recognition', *Experimental Dermatology*, 19(9), pp. 830–835. doi: 10.1111/j.1600-0625.2010.01112.x.
- Science, C. (2015) 'Predicting a T20 cricket match result while the match is in progress Authors Name Fahad Munir (11201014) Md . Kamrul Hasan (11201032) Sakib Ahmed (11201009) Sultan Md . Quraish (11201017) Supervisor Rubel Biswas Moin Mostakim A thesis presented fo', (11201014).
- Science, C. and Jhanwar, M. G. (2017) 'Quantitative Assessment of Player Performance and Winner Prediction in ODI Cricket', (July).
- Ul Mustafa, R. *et al.* (2017) 'Predicting the Cricket match outcome using crowd opinions on social networks: A comparative study of machine learning methods', *Malaysian Journal of Computer Science*, 30(1), pp. 63–76. doi: 10.22452/mjcs.vol30no1.5.
- Zhang, H. (2006) 'The Optimality of Naive Bayes', *Pattern Recognition Letters*, 27(7), pp. 830–837. doi: 10.1016/j.patrec.2005.12.001.

- [2] M. N. Ashraf, Z. Habib and M. Hussain, "Texture Feature Analysis of Digital Fundus Images for Early Detection of Diabetic Retinopathy," 2014 11th International Conference on Computer Graphics, Imaging and Visualization, Singapore, pp. 57-62, 2014.
- S. Mohammadian, A. Karsaz and Y. M. Roshan, "A comparative analysis of classification algorithms in diabetic retinopathy screening," 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, pp. 84-89, 2017.

Personal experience

The proposed study executed was an elaborate process stretching up to three months. The data collection in itself was a challenge to be able to collect in detail. After which the data preparation and structuring was a tactful task giving us insights to be able to handle various formats of data. There were multiple challenges faced at each step of the process flow starting right from the problem statement to the results.
At each stage a well-coordinated team work was required. An even bifurcation of work based on the strengths of every individual to get the best out of each of us, was planned and delivered. However, the entire project has been in sync with every individual objective even though it was segregated, the mistakes made during each task were rectified, learnt from and updated which aided in all decisions made.

Albeit the various hardships, the team was able to exceed the self-expectations and worked as a team at every stage. At every decisive stage the team was constantly guided and mentored by my guide. Being mentored by my guide helped me understand my limitations and overcome them.

All in all, it was a delightful experience of working extensively and synchronously to achieve such results. This gave a glimpse of the real work world scenario and prepared me to be able handle diversity.

The experience boosted my morale as well as confidence to solve the real-world problems using machine learning algorithms.