Project Seminar Report

On

**Improving Duckworth Lewis Method using**

**Machine Learning**

Submitted by

Rayees Ahmed Taj 1604-18-733-091

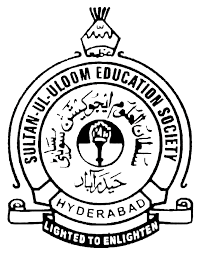
Syed Amir Ahmed Razvi 1604-18-733-092 Mir Rafath Ali 1604-18-733-100

Under the Project Supervision of

**Mr. Ahmed**

Assistant Professor

Depatment of CSE



Computer Science and Engineering Department

Muffakham Jah College of Engineering and Technology

2021-2022

**Index**

|  |  |  |
| --- | --- | --- |
| S.No | Title |  |
|  | Abstract |  |
|  | i |  |
|  | List of Figures |  |
|  |  |  |
|  |  |  |
|  | ii |  |
|  | List of Tables |  |
|  |  |  |
|  |  |  |
|  | iii |  |
|  | 1.Introduction |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | 2. Literature Survey |  |
|  |  |  |
|  |  | .. |
|  | 3. System Analysis |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | a. |  |
|  |  |  |
|  |  |  |
|  |  | .. |
|  | b. |  |
|  |  |  |
|  |  |  |
|  |  | .. |
|  | c. |  |
|  |  |  |
|  |  |  |
|  |  | .. |
|  |  |  |
|  | 4. System Design |  |
|  |  |  |
|  |  |  |

**Abstract**

Technology is continuing to play an integral part in sports. In cricket too, there are many areas where technology can be used. Machine learning will play an important role in Sports Analytics. We believe that we can use Machine Learning to analyze historical cricket games, and use this to continuously improve the Duckworth Lewis (D/L) Method of computing target scores in rain-shortened matches.

The Current D/L method is a statistical method invented by statisticians Frank Duckworth and Tony Lewis. It is designed to calculate the target score (or the PAR score) that the second batting team (in a rain-interrupted match) needs to achieve. Today, there are two D/L models/editions that are available to the cricket community: Standard Edition and Professional Edition. The Standard Edition is a chart-based model used for non-ICC matches and local matches. The Professional Edition is a software-based black box model and is used by the ICC for all official matches.

After Duckworth and Lewis retired, Professor Steven Stern (from the Queensland University of Technology) became the custodian for the method, and the method was renamed as the Duckworth-Lewis-Stern method (or D/L/S method). In many pieces of existing literature, it continues to be referred to as the D/L method.

Duckworth Lewis methodology is used in rain-interrupted matches in cricket to predict par scores. It uses parameters like overs, wickets to predict the target. But, it is not fair if we don’t consider batsmen statistics like his average in the respected format of the game, his Strike Rate, or any bowler’s economy, and some other terms that are not defined by the Standard Duckworth Lewis Method. So, our project intends to improve the existing Duckworth Lewis method including the above attributes and for that, It requires Machine Learning practices and algorithms to complete this project.

Keywords: Cricket Prediction, Supervised Learning, Sports Analysis, Naïve Bayes classification, Random Forest Algorithm, Duckworth Lewis Method, DLS Method, One Day International(ODI), Classification, Regression analysis, Standard Edition of DL Method, Professional Edition of DL Method.

**Introduction**

**Background**

After football, cricket is most loved and watched by many individuals in the world but in India cricket is the most loved sport. In the past few years, lots of research papers are published and lots of work is done which predicts the result of a cricket match by determining the number of overs to be reduced and even the reduced target score if somehow an interruption occurs during a match in any of the innings say 1st or 2nd depending on the pitch conditions and some other factors the estimated playing overs or target is determined by using the Duckworth Lewis method and to this method, we can apply any of the supervised machine learning algorithms to predict the outcome of the match like Linear regression, support vector machines, logistic regression, decision tree, Bayes network, random forest.

So, basically, if a Cricket match is interrupted due to rain or bad pitch condition, or even bad lights in the stadium then we can determine the favorable chances of continuing the game by calculating the resources available to both the teams and setting a respective target for the team batting 2nd or reducing some number of overs for the team who is batting 1st. Back in 1997-2004, a Standard Edition was followed for any Cricket match which is used to predict the targeted score for the team batting 2nd depending on the run-rate, which is really not fair. Then later on Professional Edition(2004-2015) was followed for all the international matches and was also officially declared by the International Cricket Council(ICC) for calculating the target score in rain-interrupted matches which was formulated by two statisticians named Frank Duckworth and Tony Lewis, so it was named as Duckworth Lewis method. Finally, it got replaced by another method called the Duckworth Lewis Stern method and is currently followed by the ICC but the working of this model is not known to people as ICC has not given any access to it other than their management team and it’s confidential.

Cricket is one of the most well-liked sports in the world. Especially the Twenty20 format is very popular as it is a fast-paced form of the game that attracts spectators at the ground and the viewers at home. The Indian Premier League (IPL) is a professional Twenty20 cricket league that is governed by the Board of Control for Cricket in India (BCCI). The Indian Premier League is conducted every year and participating teams represent a city in India. Various natural factors affect the game, the hype has given by the media, and a huge market like fantasy 11 and betting on sites has provided a lot of importance to the model. The rules of the game, the skill of the players, their form, and various other natural factors are very important in the prediction accuracy of the result of a cricket match.

The use of machine learning makes life easier in many aspects. To predict the outcome of a cricket match we are not going to rely on a single machine learning algorithm we are going to use all the machine learning algorithms. In machine learning, there are two types of learning supervised learning and unsupervised learning. In Unsupervised learning, the data is not properly labeled so the machine has to sort the data according to patterns, combinations without any training given. But in supervised learning, the data is labeled with the proper classification so the machine can easily analyze it and produce the correct result. For our application, the unsupervised learning models are not of any use because the data of cricket matches are properly labeled. So we are going to use the supervised learning models. In Supervised learning, there are again two types are classification and regression. Classification is used to classify among categories like red or blue and Regression is used when the output is a real number like rupees or height. In our model, we are going to use regression because the outcome will be the winning percentage and it is of type number. Our main objective is to find the key factors that affect the match outcome and select the best machine learning model that best fits this data and gives the best results. Some works already have been published in this area of predicting the outcome of a cricket match. In some papers, only a few key factors are taken for prediction so the accuracy is less. Whereas in some papers the machine learning model is not appropriate. So it is important to take all the key factors that can affect the match outcome and as well as to select the best model for training and testing the data. This will increase the prediction accuracy drastically.

**Problem Statement**

As we know the entire Duckworth Lewis method works on the resources available to both the teams in any rain-interrupted matches to predict the par score. But we can improve the existing model by considering some factors like batsman’s statistics with respect to a particular team or bowler, his average, strike rate, bowler’s economy, bowler’s average, etc. so that it could be helpful in predicting the par score and chances of winning a team in a cricket match. So, our project intends to improve the Duckworth Lewis method including the above attributes and It requires a Machine Learning algorithm depending on how efficient and accurate that algorithm is we can complete this task.

**Objective**

The main objective behind this project and the problem statement is to work on the existing Duckworth Lewis method and update the Duckworth Lewis method by including the individual statistics of the players, factors such as player ranking, an average of the batsmen, bowler’s economy, an average of any bowler, comparison of any individual batsman with respect to a particular bowler and home and way conditions(if possible), etc. So all these attributes are not included in the standard DL method and only resources available to both are taken into consideration for predicting the target score, but here in our model with the help of all the above-mentioned attributes, we can predict not only the par / target score but also the chances of a team winning the game.

**Organization of Report**

The following section, Basic background, statement, and Objective, is intended to stand alone as a complete synopsis of the report. The arrangement is keyed to understand the importance of the project In the section entitled Literature Survey, the discussion is centered around three major topics that include: the papers that inspired us to work on the project, the performance comparison with respect to important parameters used, and a brief description of technology and tools used throughout the entire project. In the next section, Existing problem in the current solutions, Our proposed solution to the problem, Technical Specifications(Software and tools) required to build the proposed project In the final section, discussion of the Software architecture to be followed, and the estimated project work timeline.

**Literature Survey**

**Literature review on methods that are similar to the Duckworth Lewis Method**

The following are methods that have been used so far in one-day cricket together with a brief description. Most of these do not take account of the stage of the innings at which the overs are lost or of the number of wickets that have fallen.

1. **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 favor of the team batting second.
2. **Most productive overs (MPO):** The target is determined for the overs the team batting second (Team 2) are to receive by totaling 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 favor Team 1.
3. **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 as that method.
4. **Parabola (PARAB):** This method, by a young South African (do Rego8), calculates a table of 'norms' y, for overs of an innings, x, using the parabola y = 7.46x - O.059x2 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 a 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.
5. **World Cup 1996 (WC96):** This is an adaptation of the PARAB method. Each of the norms has been converted into a percentage, of 225 as an approximation for the 50 over the norm and is generally regarded as the mean of first innings scores in one-day international matches.
6. **Clark Curves (CLARK):** This 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 inning, 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 that have fallen. There are discontinuities between the revised target scores at the meeting points of two adjacent types of stoppage.

All of these methods have flaws that are easily exploitable: The average run-rate method takes no account of how many wickets the team batting second have lost, but simply reflect 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. • The most productive overs method also takes no account of how many wickets the team batting second has lost and also has the further effect of penalizing the team batting second for good bowling, as their best overs are ignored in setting the revised target.

**Limitations in the Existing Duckworth Lewis method**

The D/L model has come under criticism from players, supporters, journalists, and academics, and as a result papers have been published outlining possible flaws with the model and attempts to improve upon it. Bhattacharya, Gill, and Swartz (2011) question whether the Standard edition of D/L, originally designed to be used in 50 over cricket and created before the invention of 20 over cricket, is suitable for use in T20 cricket. They assert that, due to different rules and the excessive risk-taking encouraged by the short duration of a T20 inning, the runs scoring pattern in T20s may differ from the runs scoring pattern in 50 over cricket and thus D/L, in its current form, may not be suitable. The way the D/L model looks at T20 cricket is it assumes a T20 game is like the last 20 overs of a 50 over a game still with 10 wickets in hand. Therefore the D/L resource table for T20s (appendix 1.3) is just a linear transformation of the 50 over resource table and can be created by dividing the resources percentages for the last 20 overs of a 50 over match by 0.566 (the resources left in a 50 over match with 20 overs remaining and 10 wickets in hand).

**Performance Comparision of Different Algorithms**

Now the core part of the analysis starts. We will be going to use the machine learning algorithm to predict the outcome of the match, to predict the par score, and even try to improve the existing Duckworth Lewis method. The main motive of classification is to classify the unknown input data to the specific output. Machine learning has the most powerful technique to learn the data as we know that there are 2 kinds of learning supervised and unsupervised learning. In our project, we are going to use supervised learning algorithms they are as follows Support vector machine, Decision tree, Random forest, Logistic regression, Multinomial logistic regression, and Neural networks. We have chosen these machine learning algorithms because of the following reasons:

* It set a definite distinction between classifiers.
* It is very specific about the definition of the classifiers. We can create perfect decision boundaries in supervised learning algorithms.
* We can specifically specify the number of classes we want in our research.
* After training the machine we don’t want the training data again in the memory for any future classification. It runs in the logarithmic run time.
* It can give the accountability of the machine learning model using statistical tests.
* It can handle missing data.
* It is very fast at testing times. Machine learning algorithm clearly classifies the data.

As we know that Random Forest can be used for both continuous and categorical target (dependent) variables. In a random forest i.e, the combination of decision trees, the classification model refers to the categorical dependent variable, and the regression model refers to the numeric or continuous dependent variable. So, this algorithm is can be used to complete this project as it gives predictions that can be understood easily and it can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

**Technology and Tools**

As our project is completely based on predictions and training a model with the help of datasets so Machine Learning is the technology that we have used in our project that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. We have used Python as a programming language in our project over R language because there is greater availability of machine learning packages like sklearn in Python, it’s better for generic programming tasks and is more easily preoductionized, plus Python is better for data cleaning and for analysis purposes. We have also used Streamlit for deploying our Machine Learning model, which is an open-source python framework for building web apps for Machine Learning. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code.

**System Analysis**

## Existing Problem:

Duckworth Lewis methodology is used in rain-interrupted matches in cricket to predict par scores. It uses parameters like overs, wickets to predict the target. But, it is not fair if we don’t consider batsmen statistics like his average in the respected format of the game, his Strike Rate, or any bowler’s economy, and some other terms that are not defined by the Standard Duckworth Lewis Method. So, our project intends to improve the existing Duckworth Lewis method including the above attributes and for that,

At First, there was Standard Edition which was being used in 1997. It used the concept of the current run rate to give the par score or revised score. But that was opposed as the current run rate could not only be the factor to predict the par score. Because the batsman plays cautiously by not losing the wickets and the batsman tends to score more at the death overs. So as a result, the professional edition of DLS was proposed and was implemented in the year 2004. This edition added the concept of resource available, based on which a par score would be calculated. This Edition which was introduced was quite stable. It lasted for around 10+ years. And in the year 2015, ICC introduced the concept of the DLS Stern method. This method is confidential and the software is available to only the members of ICC.

Suppose there are two scenarios. In scenario 1, Rohit Sharma and Virat Kohli are at the crease when it rained, and in scenario 2 Jasprit Bumrah and Mohammed Shami are at the crease and then it rained. Suppose In both scenarios the score was 200/3 after 40 overs. If it wouldn't rain, in scenario-1 the batsman would take the team total to around 270-280 and in scenario-2 the batsman could have taken the total to around 230-240. But the Duck worth method has to be implemented as it rained. And the method would give the revised target of around 260 irrespective of the batsmen playing. But would that be fair in calculating the target irrespective of the batsman playing? So let's take a look at the proposed solution.

## Proposed Solution/Work :

Now we got to know that the Duck worth lewis method must be improved. Including the attributes like player-statistics would make the DLS method effective which covers the problem mentioned above. The Professional Duck worth lewis method can be made with the help of Python and Machine Learning. And for considering the attribute of player statistics we need to first find the data set of player statistics and clean the data set if needed and make the dataset to a required format which is to be used for calculating the target. The important step in it finding the required data set. The data set may not be available in the format which we want. So it must be modified to the format we need. After finding the data set, it is used in the project we intend to make. We include the attributes like strike rate, the average of the batsman, and also the average and strike rate of the bowler to be included too.

The use of machine learning makes life easier in many aspects. To predict the outcome of a cricket match we are not going to rely on a single machine learning algorithm we are going to use all the machine learning algorithms. In machine learning, there are two types of learning supervised learning and unsupervised learning. In Unsupervised learning, the data is not properly labeled so the machine has to sort the data according to patterns, combinations without any training given. But in supervised learning, the data is labeled with the proper classification so the machine can easily analyze it and produce the correct result. For our application, the unsupervised learning models are not of any use because the data of cricket matches are properly labeled. So we are going to use the supervised learning models. In Supervised learning, there are again two types are classification and regression. Classification is used to classify among categories like red or blue and Regression is used when the output is a real number like rupees or height. In our model, we are going to use regression because the outcome will be the winning percentage and it is of type number. Our main objective is to find the key factors that affect the match outcome and select the best machine learning model that best fits this data and gives the best results. Some works already have been published in this area of predicting the outcome of a cricket match. In some papers, only a few key factors are taken for prediction so the accuracy is less. Whereas in some papers the machine learning model is not appropriate. So it is important to take all the key factors that can affect the match outcome and as well as to select the best model for training and testing the data. This will increase the prediction accuracy drastically.

Few algorithms like Random forest, Decision tree, Support vector machine, logistic regression, multinomial logistic regression, Neural network to predict the outcome of the match. After finding the revised score we find the accuracy of the algorithms to be used. From the name of the algorithms, one can say that the most accurate algorithm would be Random forest. And the visualization of the software can be shown up by Streamlit and for the backend, Python is used and for the data sets, machine learning must be used.

One of the most accurate algorithms is the random forest algorithm. This is a supervised learning algorithm is an ensemble learning method. A random forest algorithm is built with a multitude of decision trees to become a forest.

# Technical Specifications

Sofware tools required: Anaconda Prompt, VS code, Streamlit.

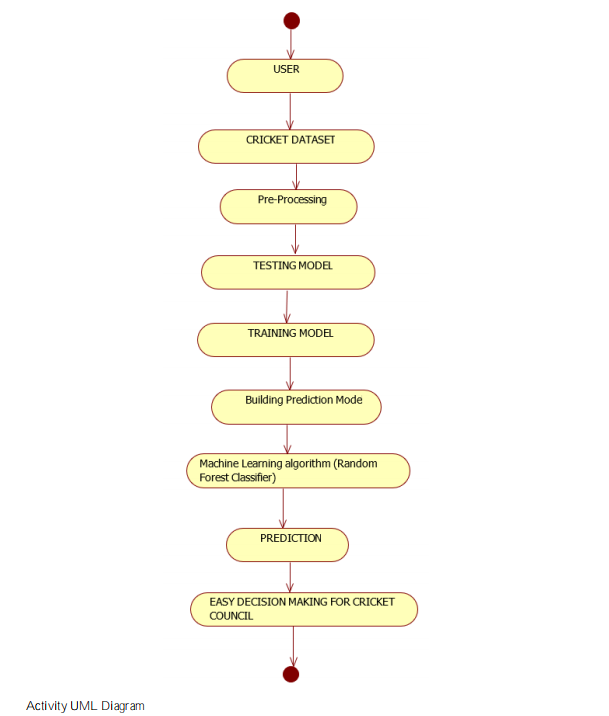
Technologies required: Machine Learning and Python.

Algorithms required: Random forest, Decision tree, Support vector machine, logistic regression, multinomial logistic regression, Neural networks, etc.

## Planning of Work:

The project is divided into blocks of different blocks of works. First to create a front end of the project. The front end of the project is made so as to get a visual idea of what is to be done in the project. The front end of the project is done is using streamlit. Then the resource table for the project is to be found out and through this dataset, backend calculations are done to find the revised target. The backend calculations are to be done using python and machine learning. Then to include the attributes of player statistics, we find the required dataset. The data set which is found out must be clean i.e it must not have any missing values etc. Now the attributes of player statistics are to be taken into the backend part. Including the attributes, the revised target is to be calculated. After this, if time permits we intend to include the Match Predictor part using different machine learning algorithms mentioned above.

The Project which we as a team are intended to do requires a lot of research.The DLS methodology must be analysed clearly because the analysis of previous methodology makes it easy to upgrade the methodology.As of now, the research on the different methodologies, the front end part of the project, finding the data set for the project is done. The reaming part of the project is in progress.



## Fig.UML Activity Diagram

## References:

Bandulasiri, A. (2008). Predicting the winner in one day international cricket, *Journal*

*of Mathematical Sciences & Mathematics Education* **3**(1): 6–17.

Beaudoin, D. and Swartz, T. (n.d.). One-day cricket.

Hucaljuk, J. and Rakipovi´c, A. (2011). Predicting football scores using machine learning

techniques, *MIPRO, 2011 Proceedings of the 34th International Convention*, IEEE,

pp. 1623–1627.

Joseph, A., Fenton, N. E. and Neil, M. (2006). Predicting football results using bayesian

nets and other machine learning techniques, *Knowledge-Based Systems* **19**(7): 544–553.

Kaluarachchi, A. and Aparna, S. V. (2010). Cricai: A classifification based tool to predict

the outcome in odi cricket, *Information and Automation for Sustainability (ICIAFs),*

*2010 5th International Conference on*, IEEE, pp. 250–255.

Longadge, R. and Dongre, S. (2013). Class imbalance problem in data mining review,

*arXiv preprint arXiv:1305.1707* .Martins, R. G., Martins, A. S., Neves, L. A., Lima, L. V., Flores, E. L. and do Nasci

mento, M. Z. (2017). Exploring polynomial classififier to predict match results in football

championships, *Expert Systems with Applications* **83**: 79–93.

McHale, I. G. and Asif, M. (2013). A modifified duckworth–lewis method for adjusting

targets in interrupted limited overs cricket, *European Journal of Operational Research*

**225**(2): 353–362.

Me, E., Unold, O. et al. (2011). Machine learning approach to model sport training,

*Computers in human behavior* **27**(5): 1499–1506.

Pathak, N. and Wadhwa, H. (2016). Applications of modern classifification techniques to

predict the outcome of odi cricket, *Procedia Computer Science* **87**: 55–60.

Phanse, V. and Deorah, S. (2011). Evaluation and extension to the duckworth lewis

method: A dual application of data mining techniques, *Data Mining Workshops*

*(ICDMW), 2011 IEEE 11th International Conference on*, IEEE, pp. 763–770.

Ramanayaka, A., Semansinghe, W., Ohani, P. and Wehigaldeniya, W. (2016). Application

of econometrics in sport: A probability estimation of getting outin odi cricket, Sri Lanka

Forum of University Economists (SLFUE), Department of Economics, Faculty of Social

Sciences, University of Kelaniya.

Sankaranarayanan, V. V., Sattar, J. and Lakshmanan, L. V. (2014). Auto-play: A data

mining approach to odi cricket simulation and prediction, *Proceedings of the 2014 SIAM*

*International Conference on Data Mining*, SIAM, pp. 1064–1072.

Singh, T., Singla, V. and Bhatia, P. (2015). Score and winning prediction in cricket

through data mining, *Soft Computing Techniques and Implementations (ICSCTI), 2015*

*International Conference on*, IEEE, pp. 60–66.

Xu, B. (2012). Prediction of sports performance based on genetic algorithm and artifificial

neural network, *International Journal of Digital Content Technology and its Applica*

*tions* **6**(22): 141.