Samriddhi College

Tribhuvan University

Institute of Science and Technology

A PROJECT REPORT
ON
CRICKET SCORE PREDICTION

Submitted To:

Department of Computer Science and Information Technology

Samriddhi College

In partial fulfillment of the requirement for the Bachelor Degree in Computer Science and Information Technology

Submitted By:

Anish Budhathoki (16137/074)

Sandip Bhujel (16173/074)

Suresh Suwal (16178/074)

June 2, 2022

Samriddhi College

Tribhuvan University

Institute of Science and Technology

A PROJECT REPORT
ON
CRICKET SCORE PREDICTION

Submitted To:

Department of Computer Science and Information Technology

Samriddhi College

In partial fulfillment of the requirement for the Bachelor Degree in Computer Science and Information Technology

Submitted By:

Anish Budhathoki (16137/074)

Sandip Bhujel (16173/074)

Suresh Suwal (16178/074)

Under the supervision of
Mr. Loknath Regmi

June 2, 2022

Samriddhi
Estd. 2013 AD

Samriddhi College
[T.U. Affiliated]
Lokanthali-16, Bhaktapur

Supervisor’s Recommendation

I hereby recommend that this project is prepared under my supervision by **Anish Budhathoki,Sandip Bhujel** and **Suresh Suwal** entitled ”A Project Report On CRICKET SCORE PREDICITON” in partial fulfillment of the requirements for the degree of B.Sc.CSIT Computer Science and Information Technology is processed for the evaluation.

....

Mr. Loknath Regmi
Supervisor
Department of Computer Science and Information Technology
Samriddhi College
Lokanthali,Bhaktapur
Nepal

Samriddhi
Estd. 2013 AD

Samriddhi College
[T.U. Affiliated]
Lokanthali-16, Bhaktapur

Letter of Approval

I hearby recommend that this project is prepared under my supervision by **Anish Budhathoki, Sandip Bhujel** and **Suresh Suwal** entitled ”**Cricket Score Predicition**” in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Information Technology has been well studied. In our opinion it is satisfactory in the scope and quality as a project for the required degree.

....
Mr. Loknath Regmi
(Supervisor)
Samriddhi College
Lokanthali, Bhaktapur

....
Mr. Sandeep Shrestha
(Pricipal)
Samriddhi College
Lokanthali, Bhaktapur

....
Mr. Navaraj Poudel
(External Examiner)
Associate Professor
Central Department of Computer Science
and Information Technology
Tribhuvan University, Nepal

Acknowledgement

We would like to thank **Mr. Loknath Regmi**, our Project Supervisor, who with his continuous support and effort, made the research upon the problem domain possible. Similarly, we cannot stay without expressing our gratitude to **Mr. Sandeep Shrestha**, the Principal of Samriddhi College, for providing us an opportunity to perform our research activities and come up with a working solution that would be beneficial for a large mass of people.The faculty members **Mr. Mohan Bhandari** and **Mr. Bikash Balami** are also thankful for proper guidance on our course to complete the project.It would have never been possible without the support of these amazing personals.

We are thankful to all the individuals who aid us on the project directly or indirectly. The continuous coordination and the dedication toward project has led on time completion of the project. We feel blessed to have such dedicated individuals on our team and it was wonderful working together. Also, we would like to apologize for the mistakes that may have been committed in this project.

Any kind of suggestion or criticism will be highly appreciated and acknowledged.

Anish Budhathoki	16137/074
Sandip Bhujel	16173/074
Suresh Suwal	16178/074

Abstract

Machine learning and data mining in sports analysis,is a new research field in computer science with a lot of challenge. In this research, the goal is to design a cricket score prediction system for a T20 cricket match between two teams while the match is in progress.Currently, in T20 cricket matches,innings score is predicted on the basis of a current run rate which can be calculated as the number of runs scored per the number of overs bowled.Using only current run rate,prediction may not be accurate.It does not include factors like the number of wickets fallen, current score,overs left,last five overs runs,etc.Using these features, Multiple Linear Regression and Random Forest Regression is implemented that predicts the total score of the innings and comparison and analysis of algorithm is studied.At last, it will be implemented in the web UI portal for the users.

Keywords: *Random Forest Regression, Multiple Linear Regression, Score Prediction ,MAE, Decision Regression Tree*

Table Of Contents

List of Figures

List of Tables

List of Abbreviations

CRR	Current Run Rate
ICC	International Cricket Council
MAE	Mean Absolute Error
MSE	Mean squared error
ODIs	One Day Internationals
SVM	Support Vector Machine
T20I	Twenty20 International
UI	User Interface

Chapter 1: Introduction

1.1 Introduction

Cricket is a bat-and-ball game played between two teams of eleven players on a field.Today, it is a most watched game in many countries.It is popular in South Asian countries, England, Australia, New Zealand. It is a game played between two teams of eleven players each. With the advent of statistical modeling in sports, predicting the outcome of a game has been established as a fundamental problem. Cricket is one of the most popular team games in the world. The game of cricket is played in three formats - Test Matches, ODIs and T20s. This project focus research on T20s, the most popular format of the game and embark on predicting the outcome of a T20 cricket match.Initially toss plays as a crucial factor in deciding the winner of the match. Toss winning team can wish to either field or bat. The team batting first will try to pose as many runs as possible in their 20 overs in order to set a target. The team batting second need to chase the target in order to win the game with wickets in hand. For years while watching limited overs cricket, we have seen projected scores at different intervals being displayed on our television screens. Projected scores are completely based on runs scored and looking at different totals at the end of an innings, using various run rates. For example, if a team's score is 100 at the end of 10 overs. There could be four variations of projected scores:

Current run-rate:	200
6 per over:	160
8 per over:	180
12 per over:	220

Considering only run rate may not yield good results since various factors might affect the score of the innings.If a score is predicted through supervised machine learning it could give a mostly accurate score. In this research, Multiple linear regression and random forest will be used and compared so that best score is obtained.A model is developed for T20 format games by mining existing game data which can be available from cricsheet[1] website.

1.2 Problem Statement

Cricket Data-set has a very large domain.In cricket various factors might affect the total score like toss, current run rate, wickets, overs, batsman, past records, bowler and all should be considered in predicting the accurate result.Mostly the cricket score prediction are performed considering only current run rate per over and wickets in hand.This does not provide the accurate result. With the use of machine learning algorithms, it can provide more accurate score with multiple features.

1.3 Objectives

The main objectives of the project is:

- To predict a score of cricket match and compare the accuracy of linear regression and random forest.

1.4 Scope and Limitations

Cricket Score Prediction is a system to analyze the total score(runs).Users can predict the score by giving inputs to the system and the predicted score will be obtained. The system is capable to produce different score accordingly to given inputs. It can be used to predict the score in live match while broadcasting. However, there are some limitations of this system.Some limitations of this application are listed as below:

- It may not be developed as a critical decision making tool.
- Since the prediction is made based on few selected parameters, any parameter outside the selection can effect on the prediction

1.5 Report Organization

The purpose of report organization is to inform about contents and chapters present in the report.The first chapter is ’Introduction’ that includes the basic information about the system.It briefs about the problem statement,objective and scope and limitation of the Cricket Score Prediction system.

Next chapter is ’Literature Review’ .A ’Literature Review’ is a source and evaluation of the available literature in the chosen subject area.A ’Literature Review’ includes about the details areas of the article and journal from where the system information is gathered.

’Methodology’ is included in chapter 3.It includes the system block diagram and details of dataset and algorithms used in project.It contain details of Multiple linear and Random Forest algorithm.It also includes feasibility analysis and requirements analysis of system.Requirements analysis defines a functional and non-functional requirements.feasibility analysis defines economic,technical, operational and scheduling details.

Next chapter is ’System Design’ that in detail explains about the system architecture,sequence diagram,class diagram, activity diagram.sequence diagram are interaction diagram that tells how operation are carried out.They capture the interaction between the objects in the context of a collaboration.Activity diagram defines the flow from one activity to another activity. Class diagram describes the structure of a system by showing the system’s classes, their attribute operation and relationship among objects.

’Implementation and Result Analysis’ chapter describes about tools used in the system development and model result analysis.Result analysis is done with the MSE and R-squared method.

The last chapter included in the report is ’Conclusion and Recommendation’.Conclusion is the last part of the report which contains the system conclusion.The recommendation contains the future enhancement of the system and to overcome the limitations of the current system.

Chapter 2: Literature Review

Cricket itself is a game of uncertainty. Many types of research on Cricket Score Prediction have been proposed for Score and result prediction with help of supervised machine learning approaches.During the research,chance to learn about various perspectives and approaches that already been developed for Cricket Score Prediction was achieved. The contributions of other's in the field of the Score prediction and the innovations and development activities that are to be done in this filed was obtained and various research papers,books,journals were studied during this time.Some of the research papers that were effective for our project are given below.

In [2], Cricket match prediction was done specific in One Day International match.This paper compares various supervised machine learning algorithms that can be used to predict the match result of ODIs match. This paper briefs about the key factors that affect the result of the cricket match and the regression model that best fits the data and gives the best predictions.

In [3],two models was developed one to calculate a total score in a innings and second to calculate a wining percentage of both teams.Naive Thomas Bayes, Random Forest, multi-class SVM, and call Tree classifiers were used for the prediction. Random Forest classifier was found to be the foremost correct for prediction.

In [4], Rabindra Lamsal and Ayesha Choudhary has used a different machine learning model to predict the score and accuracy of the models.Among algorithms, multi-layer perceptron algorithm has the highest 71.33% accuracy rate.The Multi-layer perceptron classifier outperformed other classifiers by correctly predicting 43 out of 60, 2018 Indian Premier League matches. Based on the classification accuracy, the Multi-layer perceptron classifier was followed by Logistic Regression, Random Forests and Support Vector Machine classifiers. However, Naive Bayes and Extreme Gradient Boosting classifiers performed poorly in predicting the outcomes of 2018 IPL matches.

In [5], Apurva Devnath has used the Random forest Regression algorithm for predicting the final score. Random forest is then trained with the training set and the testing set is on the trained model to find out the accuracy of the accuracy of the Random Forest

Regression model.

In [6],Prasad Thorat, Vighnesh Buddhivant, Yash Sahane have implemented the Linear Regression ,Random Forest Regression and Lasso Regression to predict the score of the match.Linear Regression achieved less Mean Absolute Error than other algorithms.They have done a dynamic feature selection on data.From the results, it is concluded that the Linear Regression algorithm has the highest accuracy of the prediction.

Above research papers and journals have helped a lot in choosing our algorithm. It was seen that mainly the linear regression and random forest performs the best in regression tasks.Random forest will have the decision tree regression as a base model. The Naive Bayes, support vector machine classifiers were used to classify the winning teams of the match in the above research paper.Feature selection was most important for the better accuracy of the models.The MAE and R-squared method was used for the result analysis and performance of the models.

Linear regression and Random Forest regression will be implemented for project to predict the total score of T20I first innings match.The linear and random forest models will be compared and analysed for the best outcome of the score.

Chapter 3: Methodology

3.1 System Block Diagram

The method for developing the system consists of mainly three main steps.Firstly data is collected and sorted for relevancy from various sources.Secondly,analysis is carried out on the collected data by examining the current-score,balls-left,wickets-left,crr,last-five-overs-runs and total-runs.At last, a multiple linear and random forest regression is designed and implemented to predict the final score.

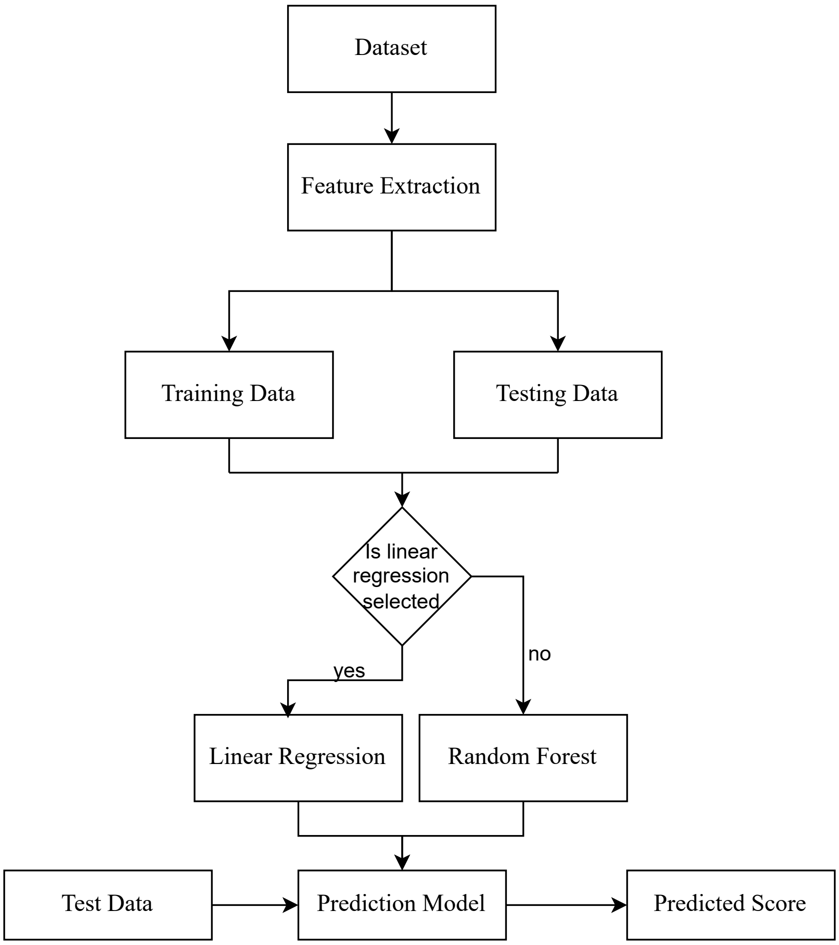


Figure 3.1.1: System Block Diagram of Cricket Score Prediction

The proposed algorithm contains of linear regression and random forest.

1. Importing the dataset from cricsheet.
2. Clean the dataset by only keeping columns like currrent_score, wickets_left, balls_left,

last_five_runs and total_runs.

3. Partition the clean data-set into 80% training and 20% testing.
4. Train the model using Random forest and multiple linear regression for calculating inning score.
5. Testing the module by providing some input and checking accuracy using Mean Absolute Error(MAE), R-squared method.
6. Deployment of system so prediction model can be accessed by everyone.

3.1.1 Data-source

The data-set will be obtained from crick-sheet[1].It will be processed and cleaned for further processing.The data-set contains ball by ball coverage of T20 matches.
Each data set consists of following columns (features):
mid: Each match is given a unique number
date: When the match happened
city: Stadium where match is being played
batting_team: Batting team name
bowling_team: Bowling team name
current_score: Total runs scored by team at that instance
wickets_left: Total wickets fallen at that instance
balls_left: Total overs bowled at that instance
last_five_runs: Total runs scored in last 5 overs
runs_x: Total runs scored by batting team after first innings.

3.1.2 Data Set

This project attempts to predict the total score of the cricket game with respect to the parameters input to the model. It requires historic data of cricket match between teams for data mining.for this project,the data are collected from the circsheet website of the T20s international match between countries.

Table 3.1.1: Data-set

match_id	batting_team	bowling_team	city	current_score
2	Australia	Sri lanka	Melbourne	43
79	India	England	Cardiff	80
1167	Pakistan	Australia	Dubai	50
900	Australia	India	Sydney	72
1136	Pakistan	Bangladesh	Mirpur	49

balls_left	wickets_left	crr	last_five	runs_x
90	10	8.6	43	168
45	6	6.4	41	148
82	9	7.89	37	147
71	8	8.81	47	197
70	10	5.88	41	141

3.1.3 Feature Extraction

In feature extraction,the raw data sets are reduced and the required data like total runs,wickets_left, current-run-rate(crr),last_five runs will be extracted and the data will be split-ed and implemented into training data and testing data.further,Multiple linear and random forest regression will be applied for the score prediction.The data-set is given below:

Table 3.1.2: Featured Data-set

current_score	balls_left	wickets_left	crr	last_five	runs_x
43	90	10	8.6	43	168
80	45	6	6.4	41	148
50	82	9	7.89	37	147
72	71	8	8.81	47	197
49	70	10	5.88	41	141

3.1.4 Data Normalization

The data is normalized for faster processing of data by model.The input vectors of the training and testing data are normalized such that all the features are zero-mean and unit variance.The target values are normalized using min max function such that all the values are converted into the values within the range of 0 to 1. The minimum value is represented by 0 and maximum value is represented by 1.

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

After normalizing the dataset with the min max algorithm, the dataset will look like follows:

Table 3.1.3: Normalized Data-set

current_score	balls_left	wickets_left	crr	last_five	runs_x
0.137255	0.918367	1.0	0.466667	0.432099	0.543269
0.141176	0.908163	1.0	0.461075	0.444444	0.543269
0.145098	0.897959	1.0	0.455833	0.456790	0.543269
0.145098	0.887755	1.0	0.438788	0.444444	0.543269
0.145098	0.877551	1.0	0.422745	0.419753	0.543269

The normalized dataset will be divided into 80% for training and 20% for testing of the models.The training dataset is used to train the model for cricket prediction model,and the model is then tested with the test dataset.

3.2 Algorithm Details

3.2.1 Multiple Linear Regression

Multiple Linear regression is used for the predictive modeling purposes. Regression is an inherently statistical technique used regularly in data mining. In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and nature of regression line is linear. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The Multiple linear regression equations is as follows:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ... + \theta_5 x_5 \tag{3.2}$$

Where, y is the total runs which is the predicted or expected value of the dependent variable.

x1 is current_score,x2 is balls_left,x3 is wickets_left, x4 is crr x5 is last_five_runs are five distinct independent features.θ_0 is the value of y when all of the independent variables/features (x1 through x5) are equal to zero

θ_1 through θ_5 is the estimated regression co-efficients. At last total runs is calculated by the dot product of features(x1,x2,...,x5) and theta(θ_0,θ_1,...,θ_5).

3.2.2 Cost Function

Cost function is the error between actual values and predicted values in the multiple regression models.Mean Squared Error is used to calculate the cost.If there is a lower the cost value then there is a better accuracy in regression and vice versa.

^

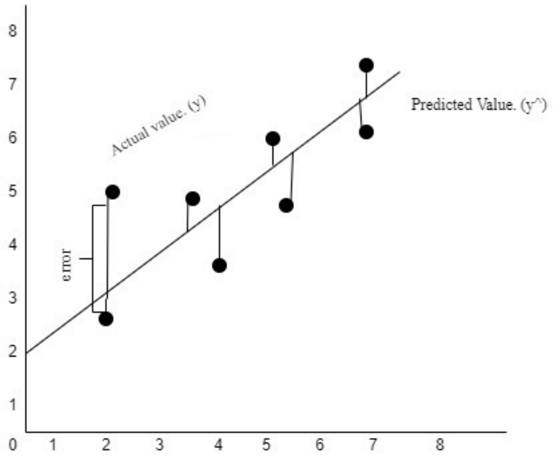


Figure 3.2.1: Cost Function.

The mathematical equation is as:

$$CostFunction(MSE) = \frac{1}{m}\sum_{i=1}^{m}(y - \hat{y})^2 \tag{3.3}$$

Where,m is a size of training set,predicted value($\hat{y}$) is predicted line of the total_runs, actual value(y) is the actual total_runs in the data-set. Here,cost function(MAE) is an error value in the prediction

3.2.3 Gradient Descent Function

Gradient descent function is used to minimize a cost function so that there will be better accuracy.It finds the best-fit line for a given training data-set in a smaller number of iterations.

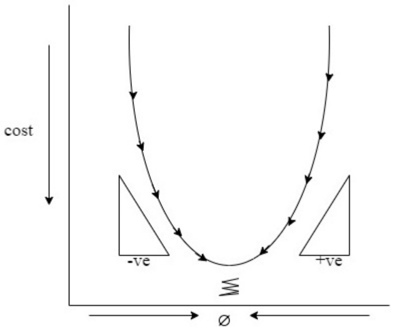


Figure 3.2.2: Gradient Descent.

The value of theta is calculated through the iteration by calculating the cost function and getting the local minimal value of the cost function.The value of θ can be obtained by:

$$\theta_{new} = \theta_{old} - \alpha * \frac{\partial cost}{\partial theta} \tag{3.4}$$

α is the learning rate of the model and $\frac{\partial cost}{\partial theta}$ is the partial derivative of cost with respect to theta which is the slop of tangent.

$$\hat{y} = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ... + \theta_n x_n \tag{3.5}$$

$$\hat{y} = \sum_{i=0}^{n} \theta_i x_i \tag{3.6}$$

The predicted value ŷ (total_runs) is obtained by dot product of theta(θ_0,...,θ_5)(coefficients) and (x1,...,x5)features.

3.2.4 Random Forest Algorithm

A Random Forest[7] is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. Here,the trained dataset is D of mxn matrix where m is rows of data and n is the number of features(x1,x2,x3,x4,x5). D‘,D“...,$D‘n$ is the bootstrapped dataset and DT1,...DTn is the different decision tree from different bootstrapped dataset. m1,m2,m3..mn are the output from different decision trees

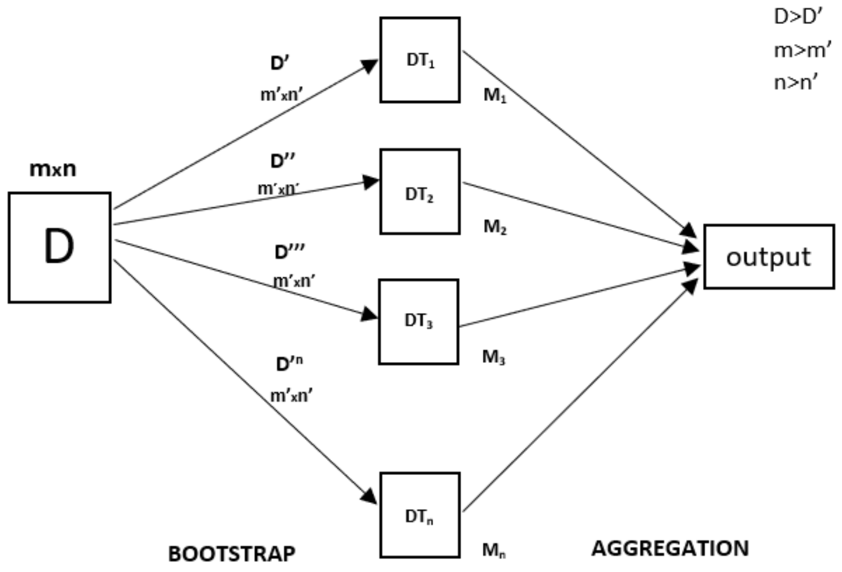


Figure 3.2.3: Random Forest Algorithm

[7]

3.2.5 Bootstrapping

Bootstrapping is the process where the row sampling and feature sampling is done to form a bootstrapped dataset for every decision tree. A bootstrapped data could contains the repeated number of rows and randomness so that there will be random number of decision tree.Row sampling contains the unique and repetitive number of data from the dataset and feature sampling/ random subspace contains the randomly selected features to form the decision trees.

3.2.6 Aggregation

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output does not depend on one decision tree but multiple decision trees. The process of combining the decision tree outputs to give a perfect output is known as aggregation. Here ,m1,m2...mn are the different outputs(total runs) and r is the mean output of different decision tree outputs. The output (r) is the total runs.

3.2.7 Decision Regression Tree

A regression tree is basically a decision tree that is used for the task of regression which can be used to predict the continuous valued output.in our data-set it contains input and output(runs_x) as a continuous valued data.

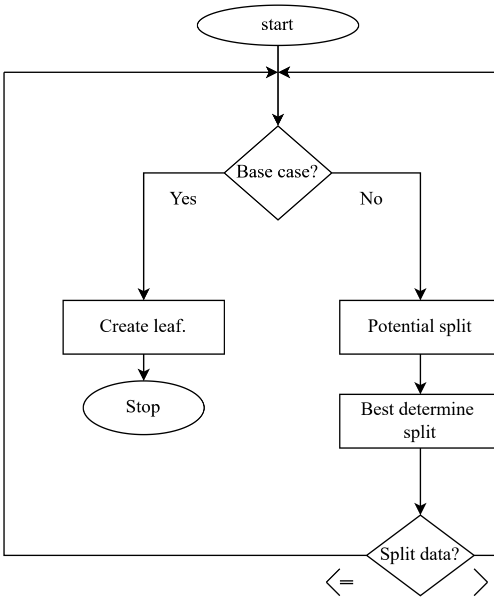


Figure 3.2.4: Decision Tree Algorithm

1. **Base case**

   Base case function is used to create a leaf in the decision tree.Here, a leaf is created if the stopping criteria is satisfied.if the samples is smaller than specified samples and a depth of the tree is smaller than the specified depth of tree. a leaf of tree is formed.

2. **Potential splits**

   This function is used to get the potential splits in the input data-set.It return the unique values form the each data set input columns.In a dataset,it contains current_score,crr,last_five_runs, wickets_left,balls_left,etc

3. **Determine best split**

   This function is used to get the best split column and best split value from the data set.It calculates the overall Mean Squared Error(MSE) and if overall MSE is smaller than given threshold value of MSE it return the best_split_column,best_split_value

4. **Split data**

   The data is split-ted into data_below and data_above according to the given best_split_column and best_split_value.This iterative process will be terminated after the base case is fulfilled.The stopping criteria is minimum samples and a minimum depth. A decision tree should not exceed the threshold value of minimum samples and depth of the tree.

3.3 Requirement Analysis

(i) **Functional Requirements**

Functional requirements are the functions or features that must be performed by a system to satisfy the needs and be acceptable to the users.The functional requirements that the system must require are as follows:

(a) The system should be able to produce total score.

(b) The system should be able to maintain higher accuracy.

The admin will collect the required training and testing data.The user will give the testing data according to input fields and the total score will be predicted.In our system admin are the system developers and users are the clients who gets the total score with respective to the given input data.

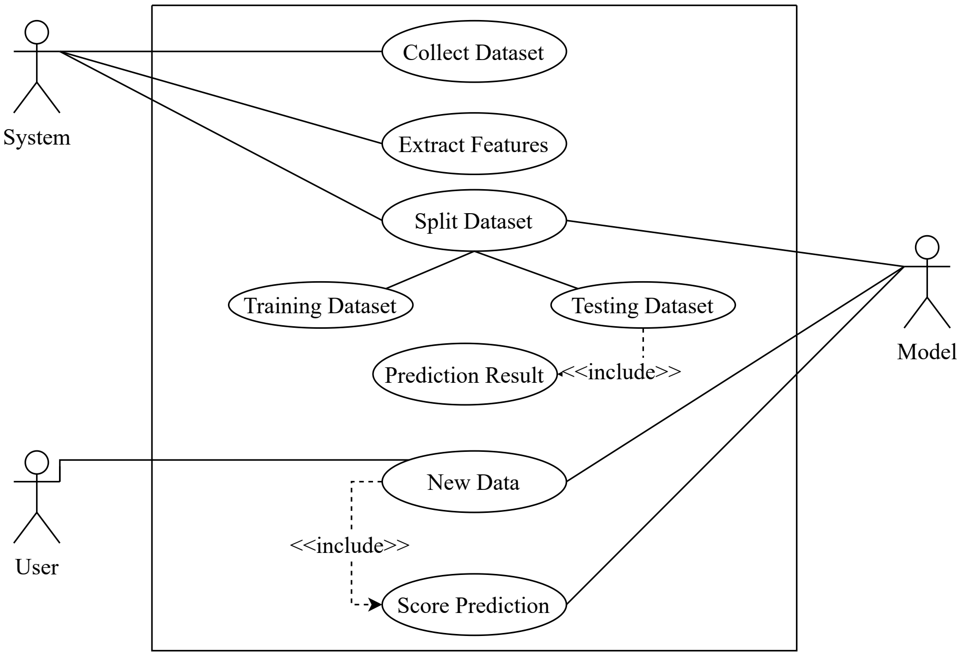


Figure 3.3.1: Use case diagram of Cricket Score System

(ii) **Non-functional Requirements**

Non-functional requirement is a description of features,characteristics and attribute of the system as well as any constraints that may limit the boundaries of the proposed system. Non-functional requirements covers all the remaining requirements that are not covered by functional requirements.The non-functional requirement of the system are:

- **Security:**

  Security requirement are needed in system to ensure unauthorized access to system and to ensure the integrity of the system from accidental or malicious damage.

- **Reliability:**

  Reliability is the ability of a system to perform its required function under stated condition for a specific period of time

- **Performance:**

  It concerns with the speed of operation of a system like giving the possible accurate score.

- **Maintainability:**

  This is the ability to change the system to deal with new technology or to fix the defects in the system.

- **Usability:**

  Usability is the case with which user can learn to operate,prepare inputs for and interpret output of the system.User-manuals,Help facilities are some of the components that are included in usability.

3.4 Feasibility Analysis

(i) **Economic Feasibility**

This study is undertaken in order to analyze the benefit that we achieved from the cost incurred form the project or system is refereed to as cost benefit analysis. This is vital to understand how feasible the project is economically. The total cost of the project includes both the monetary and non-monetary cost. Monetary cost are incurred upon implementation and the throughout the life of the project.This include start-up fees for WiFi,power supply ,training and testing and source of data-set which are open-source in cricsheet website.

(ii) **Technical Feasibility**

This study is undertaken in order to analyze the technical requirements in order to develop the system.Technical feasibility provides relevant context to the different aspects of project and determine whether the work for the project can be done with the existing equipment, software technology, and available resources. The proposed system is possible and technically feasible as all the technical requirements for the system are realized. The required manpower , software , hardware and equipment are available and system can be build without the technical barriers and objective can be achieved. Data-sets that are required to predict score are obtained through website and extraction and cleansing is done. Thus,System is technically feasible if the proposed technology is easily available to the clients at fewer expenses.

(iii) **Operational Feasibility**

The project is operationally feasible as system is based on web platform where a user can input the data information and obtain the predicted score.

3.5 Schedule

Schedule feasibility is done to analyze the project whether a task is completed in right time.It keeps tracks to accomplish a goal.The schedule is as follows:

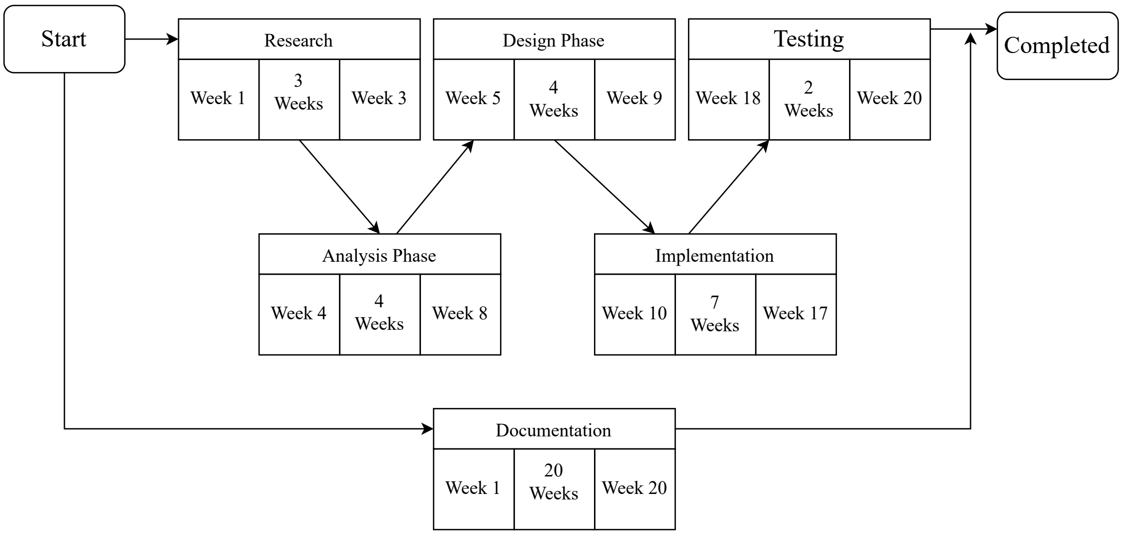


Figure 3.5.1: Schedule of the project

Chapter 4: System Design

4.1 Class Diagram

There are User, Model, admin classes used in the system.The ’admin’ class consists of adddataset(),normalizedataset() methods. Next ’Model’ class consist of the the models used in linear and Random forest regression.It process the training data and tests the models with testing data. ’User’ class consists the new set of testing data as string with methods testData() and getPrediction(). testData() is used to give the input to the model and getPrediction() will give the predicted score to user by the model.

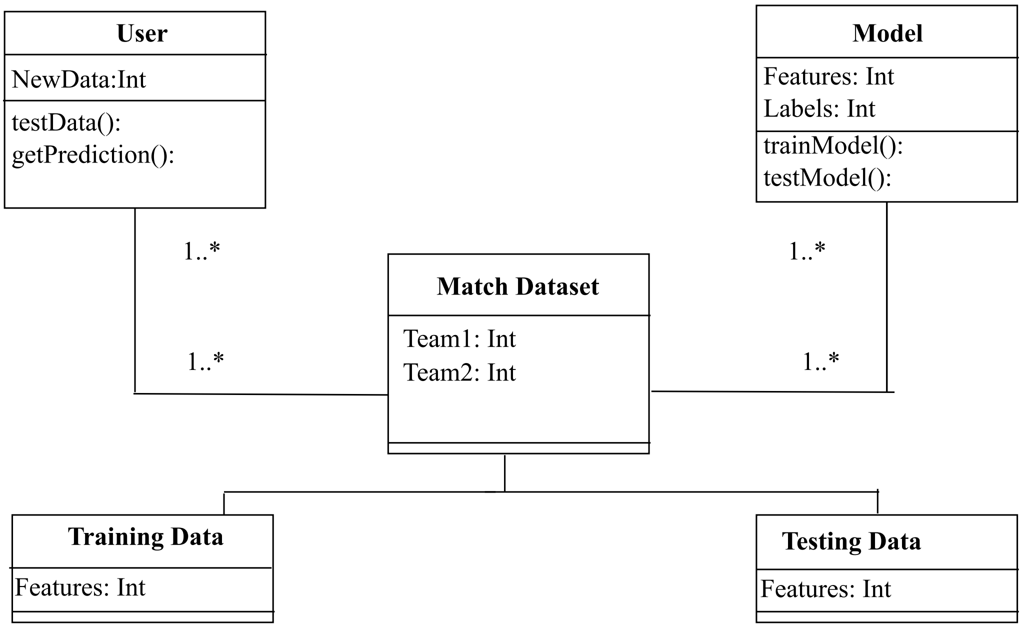


Figure 4.1.1: Class Diagram of cricket score prediction

4.2 Sequence Diagram

A sequence diagram is an interactive diagram that shows how object operate with one another and in what order.Diagram below shows the sequence of action performed one after another under certain process. There are different processes in this system which involves input,model and system and result(predicted score).The sequence of activity which undergoes during this process is given below:

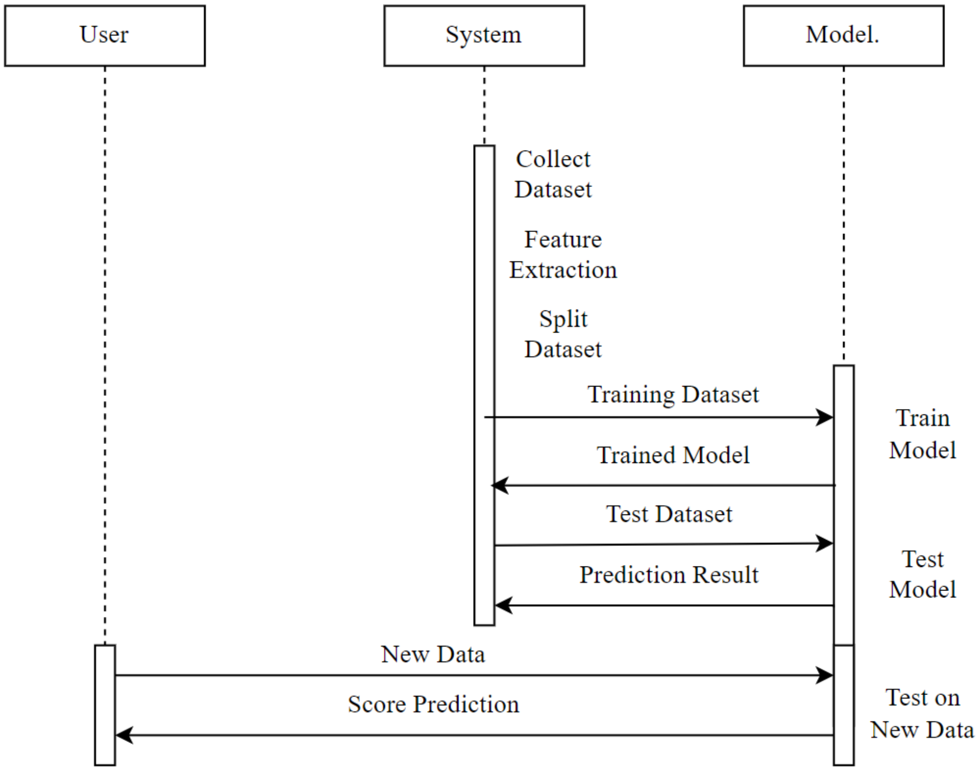


Figure 4.2.1: Sequence Diagram of cricket score prediction

4.3 Activity Diagram

The activity diagram of the cricket score prediction system is given below:

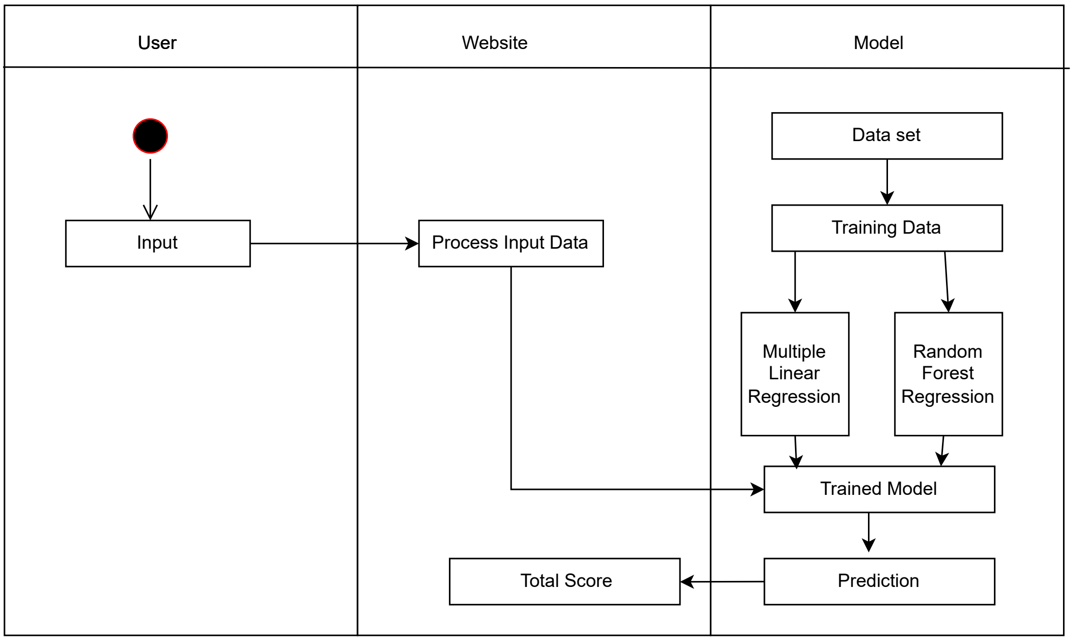


Figure 4.3.1: Activity Diagram of cricket score prediction

User access the website and provides the required input parameters for the system and chooses the required algorithm . The website then gets those required parameters and will be processed by the trained model and generate a predicted total score to the end user’s screen.

The trained model contains the Multiple Linear regression and Random forest regression which will predict the total score of the match and sends response to the user end.

Chapter 5: Implementation and Result Analysis

5.1 Implementation

5.1.1 Tools Used

- Jupyter Notebook
  - The Jupyter Notebook was used for python code and the analysis purpose of the models.It was used for training and testing of the models and result analysis.
- Python libraries
  - Pandas: Pandas was used because the collected data was in yamal and it was changed to a .csv format for creation of dataset.Pandas made this job easier with its modules.
  - Numpy : Numpy was used to make splitting of input data easier. The use of Numpy simplifer the task of array manipulation and multiplication which is widely implemented in th designed models.
  - Scikit-learn : Scikit-learn was used to calculate the MSE,accuracy of the models and for its analysis.
- Python Framework
  - Django: Django was used to build the website system for a clients which takes input parameter and produces the total score of the match.
- Draw.io
  - Draw.io is free diagram creating software.All diagrams are created using draw.io.

5.2 Result Analysis

Linear regression and Random forest model is tested by the mean aboslute error(mae) and r^2-squared value. Mean absolute error(mae) and r^2-squared. R2 is a measure of the goodness of fit of a model. In regression, the R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data.R squared value closer to 1,indicates that the model is well fitted to the data. In evaluation metrics, the R squared value is used to measure the accuracy of the model. The Mae and r^2-squared details is given below for number of iterations and learning rate value:

Table 5.2.1: MAE and R^2 squared of Multiple linear regression

Sn.	Iteration	Alpha	MAE	R^2	Accuracy
1	40000	0.1	6.67	0.67	93.33
2	40000	0.01	6.87	0.65	93.13
3	40000	0.001	7.99	0.57	92.01
4	50000	0.1	6.67	0.66	93.33
5	50000	0.0001	9.47	0.60	90.47

In above table 5.2.1, the iteration no 1 with 0.1 alpha has a 0.67 r^2-squared. It tells that our model explains 66 percent of variation within the data with the 93.33% accuracy rate on test data-set. The MAE and r^2-squared details is given below for number of n_trees,n_features and max_depth:

Table 5.2.2: MAE and R^2 squared of Random forest regression

Sn.	max_depth	features	MAE	R^2	Accuracy
1	10	5	6.3	0.69	93.6
2	2	5	8	0.69	92
3	3	2	9.5	0.67	90.5
4	6	5	6.44	0.68	93.5

In table 5.2.2, the Sn. 1 with max_depth 10 and features with 5 has 0.69 r^2-squared.

It tells that our model explains 69 percent of variation within the data with the 93.6% accuracy rate on testing data-set. From linear and Random forest model it is seen that the random forest model has the higher accuracy rate than the multiple linear regression and random forest has the highest R^2-squared rate than the linear regression. From this study, it is seen that the random forest is better than the linear regression in accuracy.

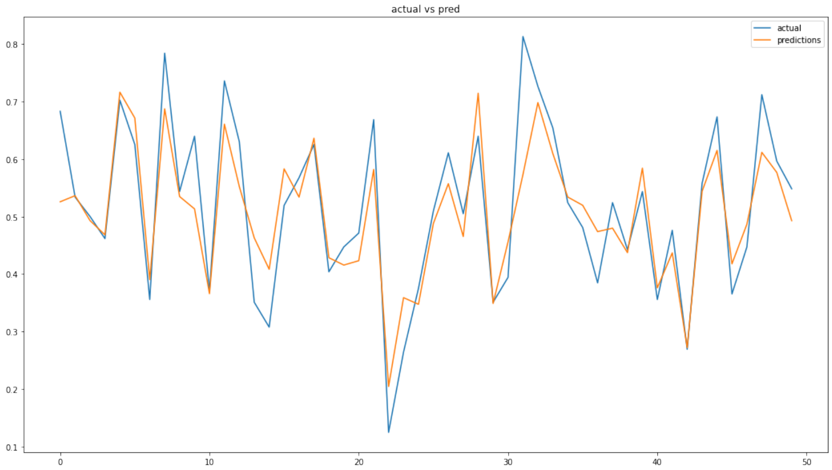


Figure 5.2.1: Actual vs Predicted graph using Multiple linear algorithm

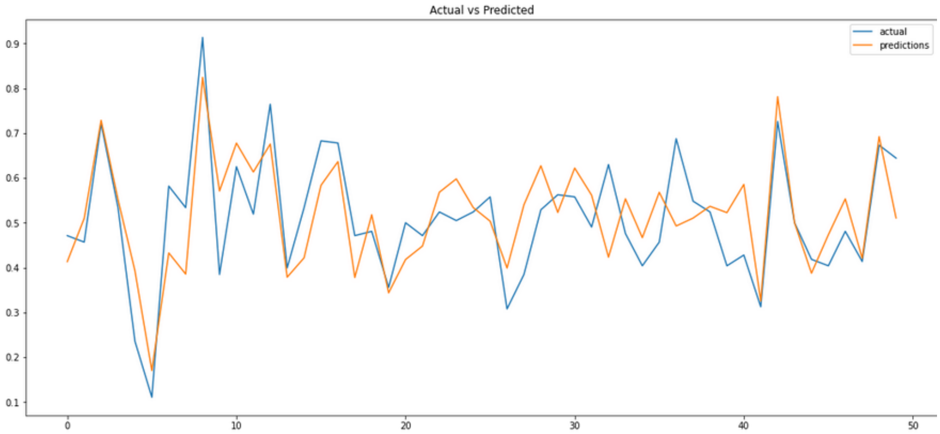


Figure 5.2.2: Actual vs Predicted graph using Random forest algorithm

The figure 5.2.1 and figure 5.2.2 denotes the actual and the predicted score analysis. The blue line denotes the actual total score and the orange line denotes the predicted total score of the match. We are considering the actual and the predicted total scores of 50 matches.The actual and predicted score is quite accurate in both of the models.

Table 5.2.3: Actual Score Versus Predicted Score by Multiple Linear Regression

S.N.	Actual Score	Predicted Score	Difference(%)
1	197	164	16.75
2	166	167	0.60
3	159	158	0.62
4	151	152	0.66
5	201	204	1.49
6	185	195	5.40
7	129	136	5.42
8	218	198	9.17
9	168	166	1.19
10	188	162	13.82

Table 5.2.4: Actual Score Versus Predicted Score by Random Forest Regression

S.N.	Actual Score	Predicted Score	Difference(%)
1	149	145	2.68
2	118	122	3.38
3	179	181	1.11
4	142	147	3.52
5	148	143	3.37
6	135	148	9.62
7	166	169	1.8
8	171	171	0

The table 5.2.3 and table 5.2.4 shows the actual, predicted score and its difference by the Multiple Linear regression and Random Forest Regression respectively in test data set. Random Forest was seen more accurate than Multiple Linear Model in the score prediction

Chapter 6: Conclusion and Future Recommendations

6.1 Conclusion

The multiple linear regression and random forest algorithm was implemented to predict the total score of the cricket match. The initial analysis shows significant correlation between different input parameter.The result obtained in all the cases was fairly accurate.The accuracy of multiple linear regression and random forest is nearly in between 86-93% and 88-94% respectively in majority of the cases.This shows that both models provides similar accuracy for the dataset. After the phase of prediction and analysis,the project will be implemented for users in the website UI portal.

6.2 Future Recommendations

For future enhancement, Potential improvement can be made to our data collection and analysis method. This model predicts a first innings total score but can be made to predict the both innings score.This model can be extended to predict the wining teams.

Here the algorithm was used for continuous data but can be used for categorical classification in future.Future research can be done with possible improvement such as more refined data and more accurate algorithms.

References

[1] A. Zaltzman. (2022) T20 international dataset. [Online]. Available: https://cricsheet.org/

[2] P. Tekade, K. Markad, A. Amage, and B. Natekar, “Cricket match outcome prediction using machine learning,” *INTERNATIONAL JOURNAL*, vol. 5, no. 7, 2020.

[3] R. Kamble *et al.*, “Cricket score prediction using machine learning,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 1S, pp. 23–28, 2021.

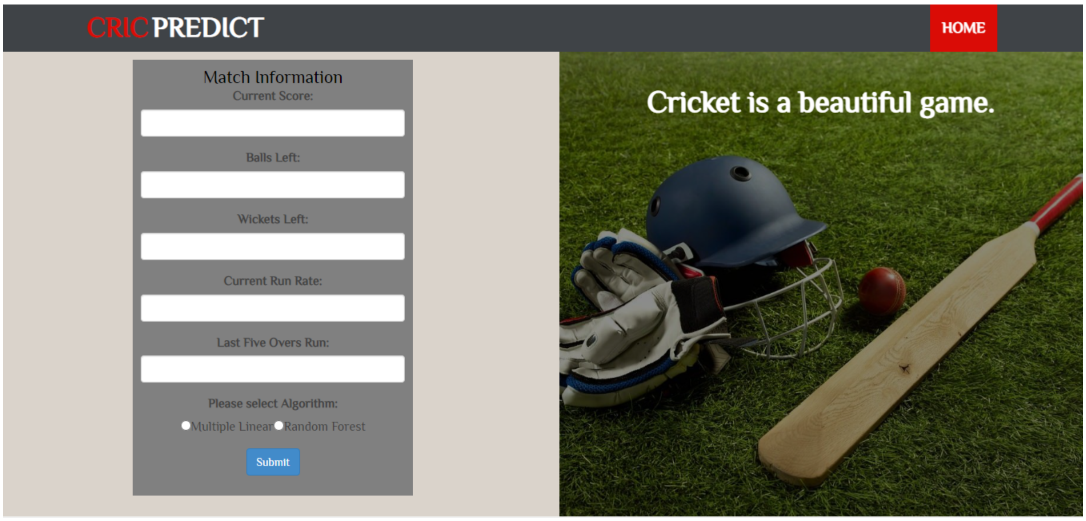
[4] R. Lamsal and A. Choudhary, “Predicting outcome of indian premier league (ipl) matches using machine learning,” *arXiv preprint arXiv:1809.09813*, 2018.

[5] A. Devnath, “Cricket score prediction using machine learning,” 2020. [Online]. Available: https://medium.com/@debnathapurba1/cricket-score-prediction-using-machine-learning-2022b70941ca

[6] P. T. V. B. Y. Sahane, “Cricket score prediction,” 5 2021. [Online]. Available: https://www.ijcrt.org/papers/IJCRT2105677.pdf

[7] “Random forest regression,” Nov 2021. [Online]. Available: https://www.geeksforgeeks.org/random-forest-regression-in-python/

Appendix



Match Information

Current Score:

Balls Left:

Wickets Left:

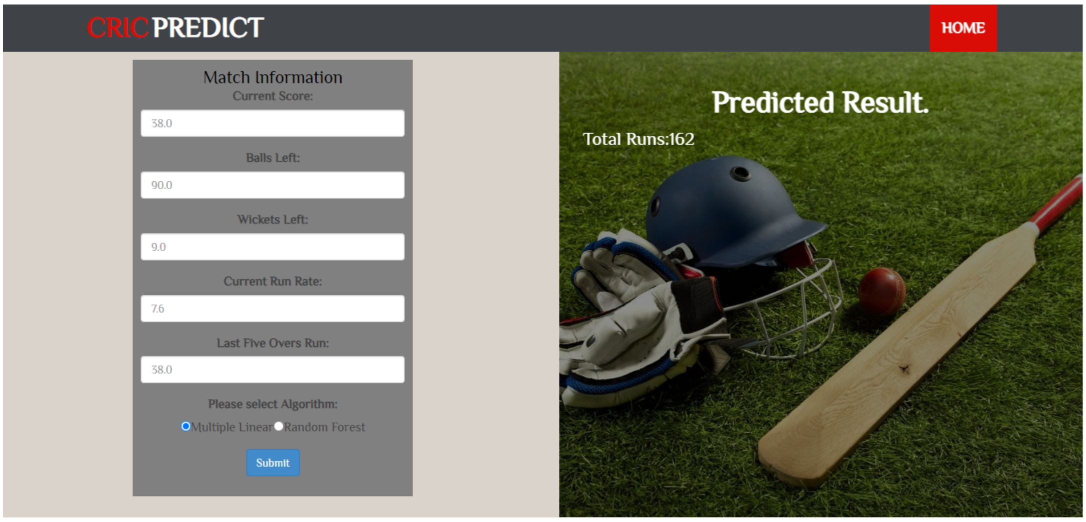
Current Run Rate:

Last Five Overs Run:

Please select Algorithm:

○ Multiple Linear ○ Random Forest

Figure 6.2.1: User Interface



Total Runs:162

Match Information

Current Score: 38.0

Balls Left: 90.0

Wickets Left: 9.0

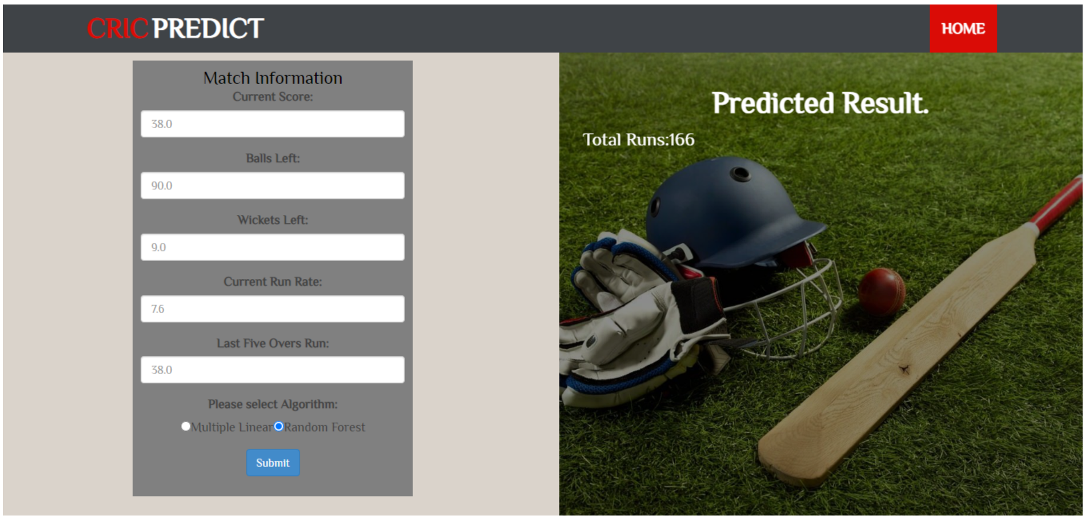
Current Run Rate: 7.6

Last Five Overs Run: 38.0

Please select Algorithm:

● Multiple Linear ○ Random Forest

Figure 6.2.2: Linear regression predicted score



Match Information

Current Score: 38.0

Balls Left: 90.0

Wickets Left: 9.0

Current Run Rate: 7.6

Last Five Overs Run: 38.0

Please select Algorithm:

○ Multiple Linear ● Random Forest

Predicted Result.

Total Runs:166

Figure 6.2.3: Random forest regression predicted score