MINI PROJECT REPORT

on

**"Score Prediction For Cricket Matches"**

By

Shuchi Mandal – 117A1037

Omkar Krishnapurkar-117A1033

Jayadev Sitaraman-117A1026

UNDER THE GUIDANCE OF

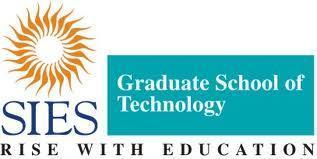
**Prof. Suvarna Chaure**

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF

BACHELOR OF ENGINEERING

In

Computer Engineering



**DEPARTMENT OF COMPUTER ENGINEERING**

**SIES GRADUATE SCHOOL OF TECHNOLOGY**

**NERUL, NAVI MUMBAI – 400706**

ACADEMIC YEAR

2019 – 2020

**CERTIFICATE**

This is to certify that this is a bonafide record of Mini Project titled **“Score Prediction For Cricket Matches”** carried out by the following students of third year in Computer Engineering.

Sr. No. Name Roll No.

1. Shuchi Mandal 117A1037

2. Omkar Krishnapurkar 117A1033

3. Jayadev Sitaraman 117A1026

The report is submitted in partial fulfillment of the degree course of Bachelor of Engineering in Computer Engineering Department, of University of Mumbai during the academic year 2019-20.

**Internal Guide Head of Department Principal**

We have examined this report as per University requirements at SIES Graduate School of Technology, Nerul (E), Navi Mumbai on \_\_\_\_\_\_\_\_\_\_\_\_.

**Name of External Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature with Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Name of Internal Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature with Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| Sr.No. | Topic | Page No. |
| 1. | List of Figures | 3 |
| 2. | Abstract | 4 |
| 3. | - Introduction | 5-6 |
| 4. | Literature Survey | 7-8 |
| 5. | Proposed System | 9-12 |
| 6. | System Design :Experimental/Modeling/ Designing/Programming/data collection etc. | 13-15 |
| 7. | Experimental Results and System Screenshots | 16-21 |
| 8. | Conclusion and Future Scope | 22 |
| 9. | References | 23 |

**List Of Figures**

|  |  |  |
| --- | --- | --- |
| Sr.No. | Topic | Page No. |
| 1. | Figure1:Block Diagram | 12 |
| 2. | Ffffffffffff Figure2:Activity Diagram | 13 |
| 3. | Figure2:Dat Figure3:Data Flow Diagram Level 0 | 14 |
| 4. | Figure2:Dat Figure4:Data Flow Diagram Level 1 | 14 |

**ABSTRACT**:

Applications of machine learning supplemented with data mining techniques has become a hot topic for research worldwide, sports analytics is no exception though. Cricket is one of the most popular sports in Australia, Caribbean, UK and South Asian nations with a net fan base of around 2.5 billion. The game has tremendous spectator support in more than 100 nations and the masses show great interest in predicting the game outcomes. There are lots of pre-game and in-game attributes which decides the outcome of a cricket match. Pre-game attributes like the venue, past track-records, innings(first/second), team strength etc. and the various in-game attributes like toss, run rate, wickets remaining, strike rate etc. influence the result of a match in a predominant manner. In this study, Two Algorithms namely Linear Regression and Random Forest have been used to Predict the score which can be made in a cricket match.The designed tool takes into consideration five features namely Runs,Wickets,Overs,Striker’s score and non-striker’s score for predicting the final score of a given match.

**Chapter 1**

**INTRODUCTION**

Cricket is basically a bat and ball game which is played between 2 teams having 11 players each. Each team comes to bat and has a single inning in which it seeks to score as many runs as possible, while the other team fields. The innings ends when the total quota of deliveries, which depends on game format has turned up, or the 10 batsmen have been dismissed, whichever comes first. The prime objective is to score more runs & thus runs are the decisive factor. Game of cricket is highly unpredictable in nature. Until the very last moment, it is difficult to make accurate predictions about the game. Various natural factors affecting the game output, huge betting market and enormous media coverage have given strong incentives to model this game from the machine learning perspective.

International Cricket Council (ICC) is the governing body which decides the rules of cricket. There are 3 widely accepted formats of cricket on international level - T20 match, One Day Internationals and Test match. The scheduled duration of the game is the prime difference between these three formats, which directly modifies the number of deliveries each team get to play in their respective innings.

The Board of Control for Cricket in India(BCCI) founded a new league named Indian Premier League in 2008.

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India contested during March or April and May of every year by eight teams representing eight different cities in India. The league was founded by the Board of Control for Cricket in India (BCCI) in 2008. The IPL has an exclusive window in ICC Future Tours Programme.

One Day International i.e. ODI format is of limited overs, where each team faces 300 deliveries(50 overs). Generally, ODI match falls in any of the 2 categories: Day or Day-Night match. T20 is the shortest internationally recognized format of this game, where each team innings consist of 20 overs. This is more of an ”explosive” and more ”athletic” than the other two formats. The study is focused on the popular formats of Cricket, One Day Internationals or the ODIs and the T20s.

The final score outcome of One Day Internationals and T20 is influenced by a varied no. of features . The best attributes or factors that influence the final score outcome need to be found. For this analysis, the factors have been considered, which are proven to have a substantial impact on the final score outcome. The factors considered for analysis include: Runs, Wickets, Overs, Striker’s runs,

Non-Striker’s runs.

**Chapter 2**

**Literature Survey**



The above picture clearly tells us how bad is taking run rate as a single factor to predict the final score in an limited overs cricket match.  In ODI and T-20 cricket, many factors play a key role in deciding what the final score will be.

So, instead of considering only a single factor multiple factors are considered for better accuracy of the score in the long run of the innings whence the match is taking place.

During the past few years lot of work and research papers have been published which measure the player performance and their winning predictions. Two research papers namely

Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks andMovie Success Prediction using Machine Learning Algorithms and their Comparison were referred in order to understand the mechanism of predictions.

**Analysis on Attributes Deciding Cricket Winning** does an analysis on various factors responsible for winning predictability of cricket matches. Swetha & Saravanan.KN**[2]** considered the factors include Pitch, Toss, Team Strength, Past Records, Home Ground Advantage, Current Performance & Weather. Moreover, various pre-game attributes & in-game attributes are also key for prediction of match. The paper states various approaches like multinomial logistic regression, Artificial Neural Network, etc. as models or the algorithms chosen for the better accuracy in the score prediction of cricket match.

**Movie Success Prediction using Machine Learning Algorithms and their Comparison** makes use of Support Vector Machine (SVM) , Random Forest , Ada Boost, Gradient Boost & KNN Classifier Algorithms to predict the success of the movie. Rijul Dhir & Anand Raj**[3]** found that the highest accuracy among these models was shown by Random Forest Algorithm. The IMDB dataset was used for testing & training purposes. A movie success does not only depend on features related to movies. Number of audience also plays very important role for a movie to be successful. Hence, the necessary features have to be taken into consideration for max. precision & accuracy.

Thus, The past methods or research that has been done considered only partial factors of a match resulting in less accuracy of the Outcome(Score) of the Match.

**Chapter 3**

**Proposed System**

We created a system which will consider various factors to decide the outcome of the match.

Whether the match may be IPL ,ODI or T-20 the factors have to be considered to formulate an outcome with higher accuracy.

First, we collected the necessary data ( data collection) and pre-processing step to prepare the dataset and selecting the features for the training purposes. Appropriate standard scalar was applied to maintain the standardization between all the columns of the dataset.

Furthermore, after choosing the algorithms ( Random Forest & Linear Regression) the dataset was trained on 75% of the whole dataset. The Model was then tested based on the remaining 25% of data by choosing both the algorithms on same custom input.R-Square value determines how well the model is fit on the chosen algorithm .Higher it’s value more is the model suited for the system. Custom Accuracy function tests if the predicted score is less than the threshold value specified then the model is fit.Finally, comparative analysis of the two algorithms is done by testing with a custom input in both cases and checking the accuracy by R-Square value & Custom Accuracy function built in the system.

**Preparing The Dataset:**

We have downloaded the dataset from cricsheet. The site gives us ball by ball details of matches.

The dataset contains ball by ball coverage of:

* 1188 ODI matches: data/odi.csv
* 1474 T-20 matches: data/t20.csv
* 617 IPL matches: data/ipl.csv

Each dataset consists of following columns(features):

* mid: Each match is given a unique number
* date: When the match happened
* venue: Stadium where match is being played
* bat\_team: Batting team name
* bowl\_team: Bowling team name
* batsman: Batsman name who faced that ball
* bowler: Bowler who bowled that ball
* runs: Total runs scored by team at that instance
* wickets: Total wickets fallen at that instance
* overs: Total overs bowled at that instance
* runs\_last\_5: Total runs scored in last 5 overs
* wickets\_last\_5: Total wickets that fell in last 5 overs
* striker: max(runs scored by striker, runs scored by non-striker)
* non-striker: min(runs scored by striker, runs scored by non-striker)
* total: Total runs scored by batting team after first innings.

The features used in our system are as follows:

**Features Used:**

* runs
* wickets
* overs
* striker
* non-striker

While experimenting, all the other features didn’t make much difference in the results. The other features being as follows:

* How much the team had scored in last 5 overs?
* How much the team had lost wickets in last 5 overs?
* The nature of the pitch
* How strong is the batting and bowling team?

**Label Used:** Total

**Splitting the data into train data and test data:**

We are using 75% of the dataset for training purpose and the rest 25% for testing purpose.

**Feature Scaling the Data:**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values . So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

**Training the dataset:**

1.Using Linear Regression:

A linear approach for modelling the relationship between dependent variable(Total in our case) with the independent variables(Five features namely Runs, Wickets, Overs, Striker’s runs,

Non-Striker’s runs which are used in our case).

2.Random Forest:

An ensemble learning method that operate by constructing a multitude of decision trees at training time and determining the final output rather than relying on individual decision trees.

**Testing the dataset on trained Model:**

R-squared value:

R-squared is a statistic that will give some information about the goodness of fit of a model. The R-squared coefficient of determination is a statistical measure of how well the predictions approximate the real data points. An R-squared value of 1 indicates that the predictions perfectly fit the data.

Custom Accuracy:

We have defined our own function to measure accuracy of the model. Custom Accuracy is defined on the basis of difference between the predicted score and actual score. If this difference falls below a particular threshold(20 taken in our case), we count it as a correct prediction.

**Testing with a custom input:**

Here, we give custom input in the form of features which we have used for prediction.For eg, we have taken a custom input([100,0,13,50,50]) i.e 100 runs with 0 wickets lost in 13 overs with both the striker and non-striker batting on 50 runs each and now our model will predict the final score which can we made based on the current input given.

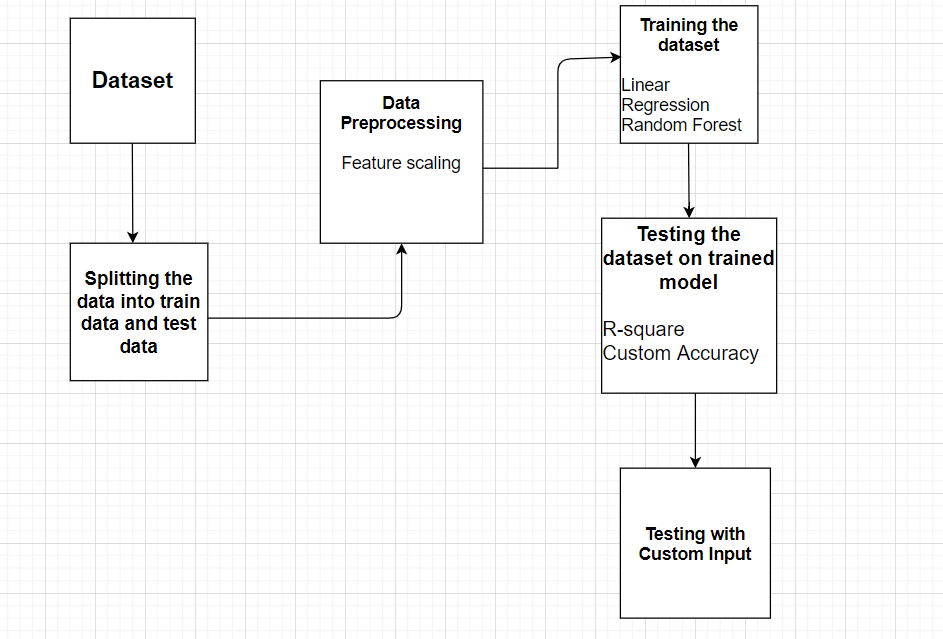


Figure1. Block Diagram

**Chapter 4**

**SystemDesign:Experimental/Modeling/Designing/Programming/data collection**

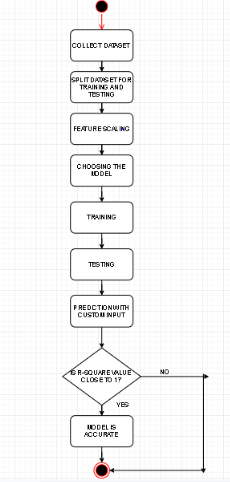
****

Figure 2. Activity Diagram

The above diagram explains the flow of the entire activity. It starts with collection of dataset required for the project followed by the training and testing of the same. Feature scaling is a crucial part of any project implemented using Machine Learning which is done after training & testing. Appropriate models are chosen for the same and the dataset is trained & tested based on those models. In this project, Random Forest Model & Linear Regression Model are used. The next step is the prediction of score using a custom input which is given by the user and finally the R-Squared value decides the accuracy of model. If R-Squared value is closer to 1(100%) then the model is fit for the system i.e. more accurate.

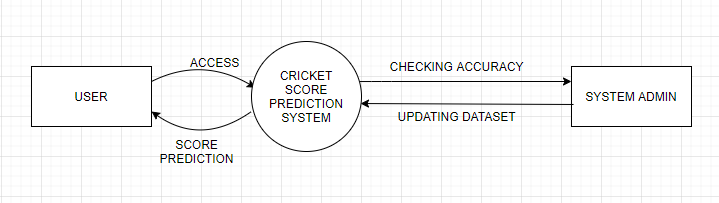
****

Figure3.Data Flow Diagram Level 0

Level 0 DFD showing the system as a single process with its relationship to the User of the system as well as the System Admin. User can access the system using its features to get the prediction of the score. System Admin can update dataset as well as check accuracy of the models being implemented in the system.

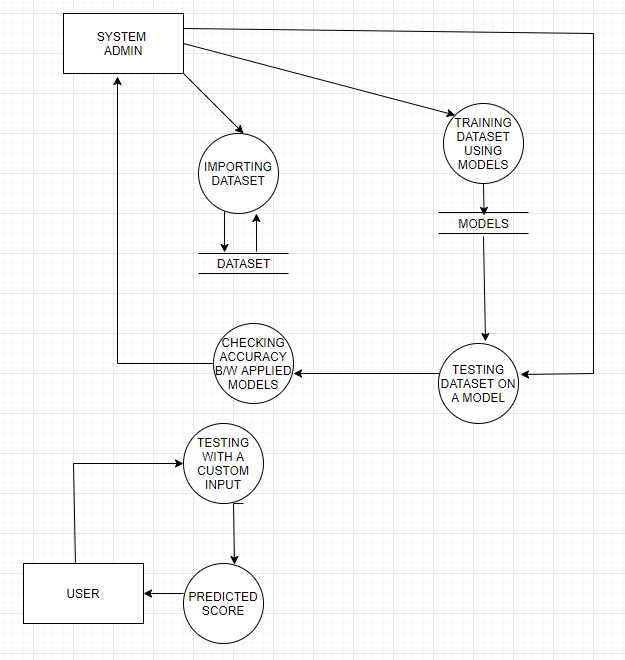
****

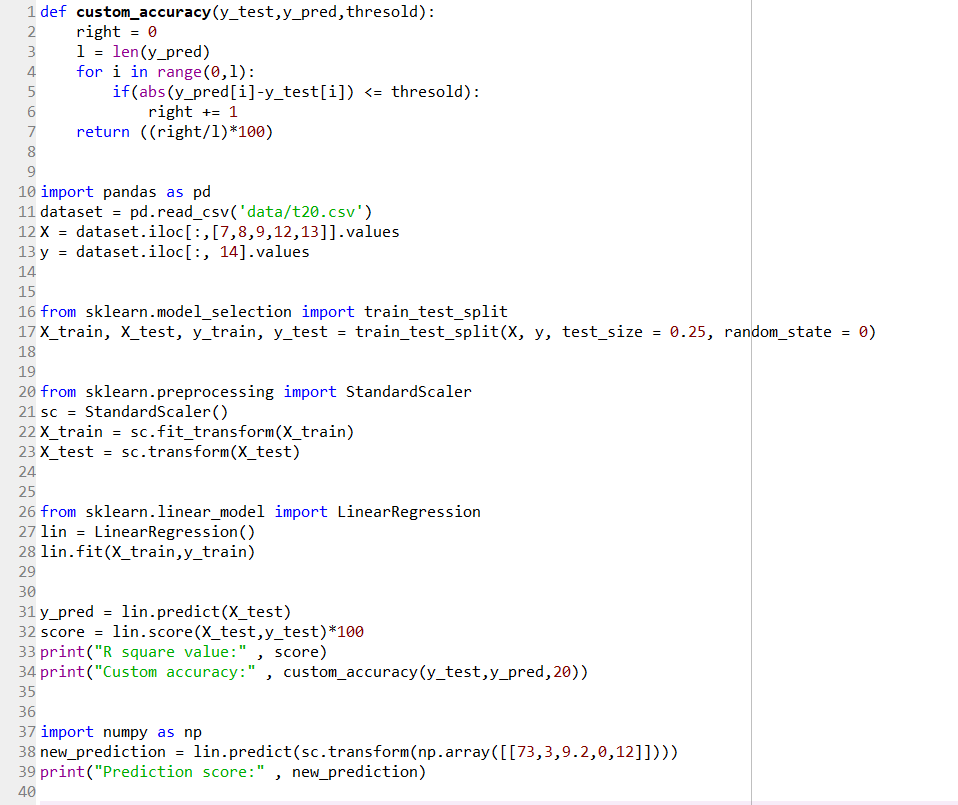
Figure4. Data Flow Diagram Level 1

In Level 1 DFD, the main functions of the system are shown which are breakdown into several processes. Tasks like importing the dataset & Training Dataset using models are handled by the System Admin. Also, the Accuracy between the applied models are handledby the Admin. The Dataset & Models are taken from the Datastores shown above. User can use the system by providing his own custom input and getting the predicted score as output.

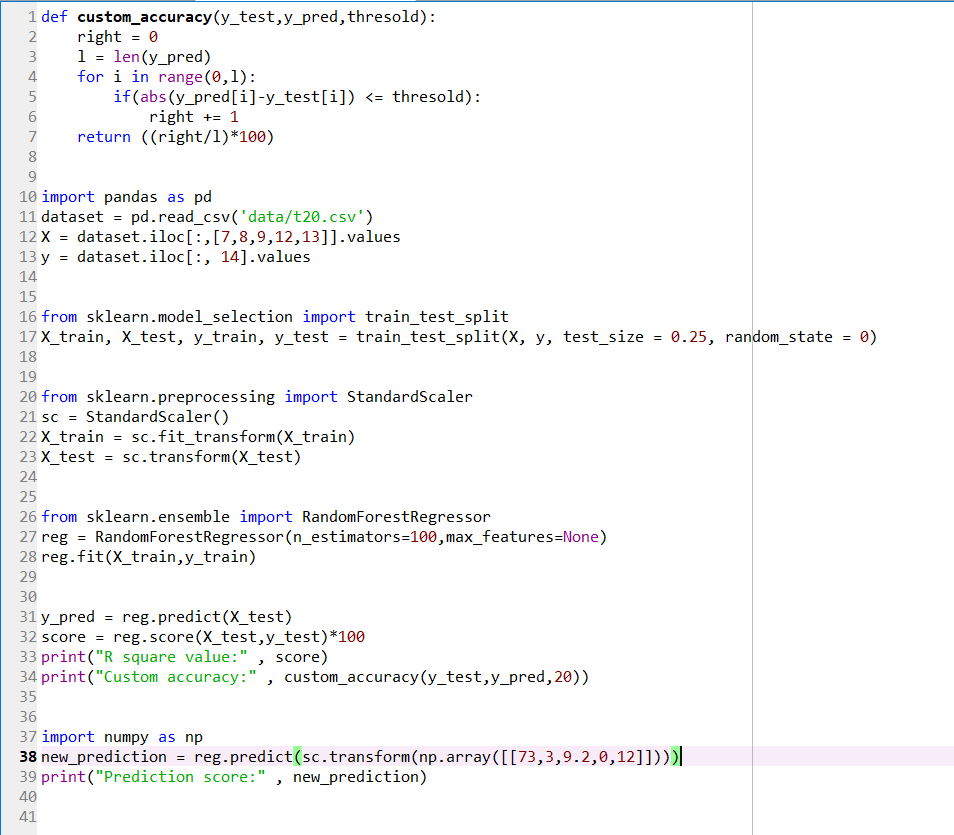
**Chapter 5**

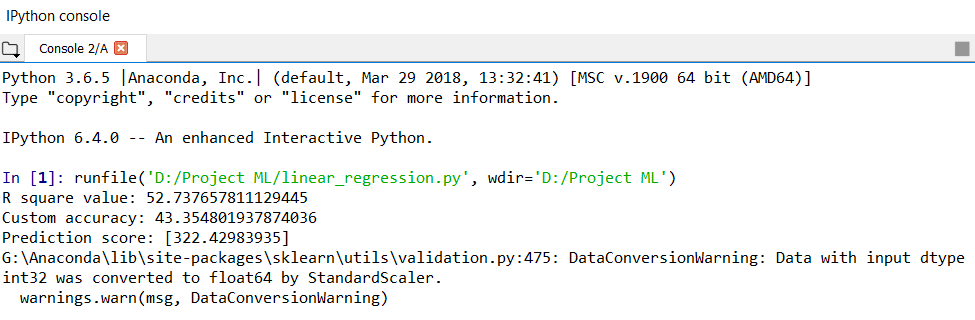
**Experimental Results and System Screenshots**

**Code for Linear Regression:**

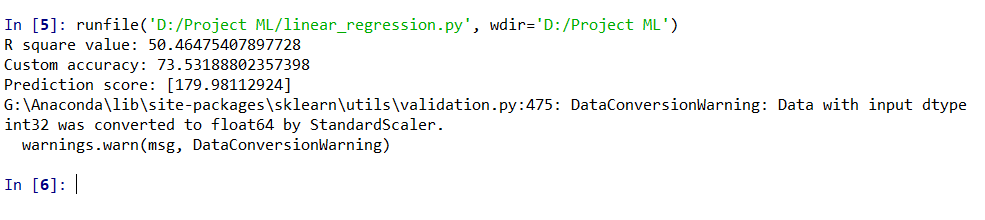
****

**Code for Random Forest:**

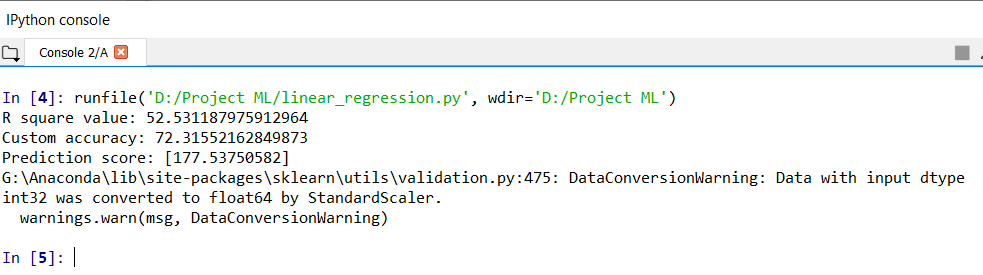
****



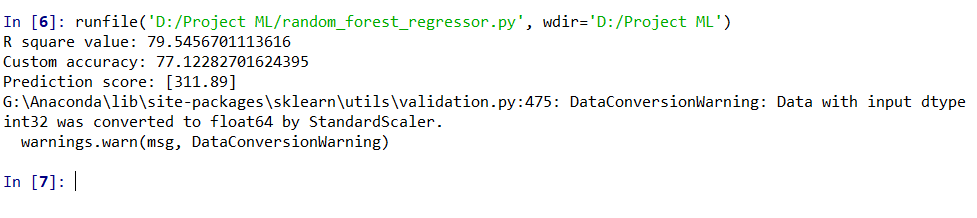
Here, We are using linear regression with the custom input [100,0,13,50,50] to predict the final score for the ODI match and the corresponding R-square value which we got is 52%.



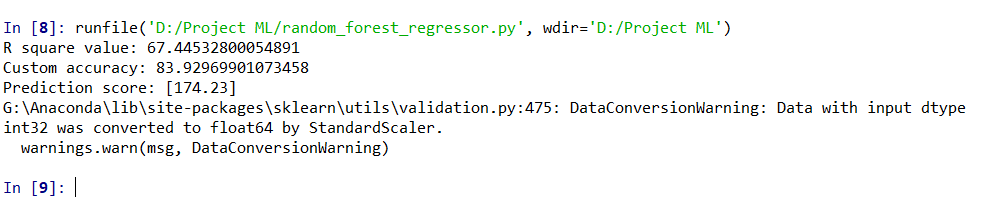
Here, We are using linear regression with the custom input [100,0,13,50,50] to predict the final score for the IPL match and the corresponding R-square value which we got is 50%.



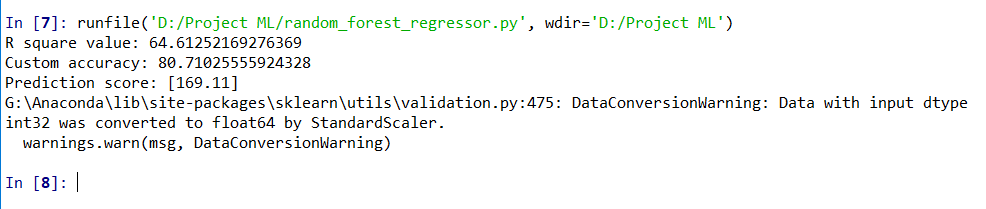
Here, We are using linear regression with the custom input [100,0,13,50,50] to predict the final score for the T20 match and the corresponding R-square value which we got is 52%.



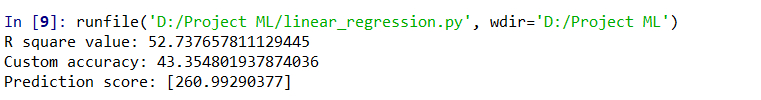
Here, We are using Random Forest with the custom input [100,0,13,50,50] to predict the final score for the ODI match and the corresponding R-square value which we got is 79%.



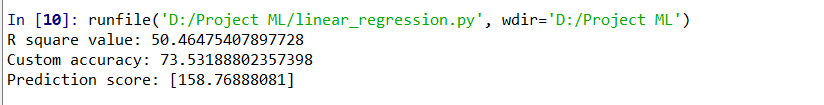
Here, We are using the Random Forest Algorithm with the custom input [100,0,13,50,50] to predict the final score for the IPL match and the corresponding R-square value which we got is 67%%.



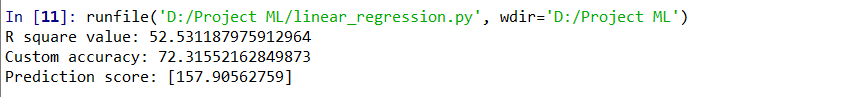
Here, We are using Random Forest Algorithm with the custom input [100,0,13,50,50] to predict the final score for the T20 match and the corresponding R-square value which we got is 64%.



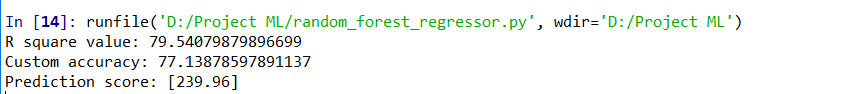
Here, We are using linear regression with the custom input [73,3,9.2,0,12] to predict the final score for the ODI match and the corresponding R-square value which we got is 52%.



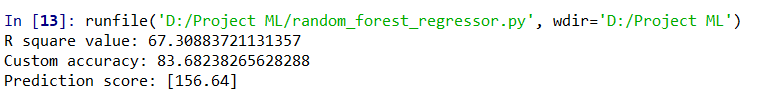
Here, We are using linear regression with the custom input [73,3,9.2,0,12] to predict the final score for the IPL match and the corresponding R-square value which we got is 50%.



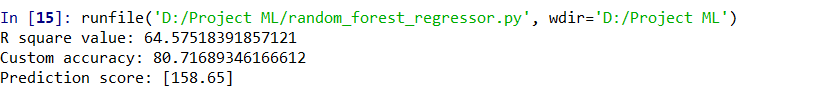
Here, We are using linear regression with the custom input [73,3,9.2,0,12] to predict the final score for the T20 match and the corresponding R-square value which we got is 52%.



Here, We are using the Random Forest Algorithm with the custom input [73,3,9.2,0,12] to predict the final score for the ODI match and the corresponding R-square value which we got is 79%.



Here, We are using the Random Forest Algorithm with the custom input [73,3,9.2,0,12] to predict the final score for the IPL match and the corresponding R-square value which we got is 67%.



Here, We are using the Random Forest Algorithm with the custom input [73,3,9.2,0,12] to predict the final score for the T20 match and the corresponding R-square value which we got is 64%.

Since,the R-square for Random Forest Algorithm is better than the R-square value for Linear Regression Algorithm for both the custom inputs [100,0,13,50,50] and [73,3,9.2,0,12] which shows that the Random Forest Algorithm is indeed more accurate than Linear Regression Technique as more the R-square value the better the prediction would be.

**Chapter 6**

**Conclusion and Future Scope**

In this study, a comparative analysis of the predictions generated by 2 different supervised classification models (algorithms) was performed for the same input dataset.The proposed approaches are better than the statisticalapproach as unlike statistics which uses mathematical equations to formalize the relationships between variables, these approaches require no prior assumptions regarding the data variables and their underlying relationships.In our proposed system, we found out that Random Forest Algorithm was more accurate in predicting the score of the match than the Linear Regression Technique.Moreover, by considering multiple factors the prediction is also spot-on for future scope of the project.

**References:**

[1] Jalaz Kumar, Rajeev Kumar, Pushpender Kumar “Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks” in 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

[2] Swetha, Saravanan.KN “Analysis on Attributes Deciding Cricket Winning” in International Research Journal of Engineering and Technology (IRJET) Volume:04 Issue:03.

[3] Rijul Dhir, Anand Raj “Movie Success Prediction using Machine Learning Algorithms and their Comparison” in 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

[4] Kalpdrum Passi and Niravkumar Pandey ,”Increased Prediction Accuracy in the Game of Cricket using Machine Learning” in International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018.

[5]Documentation sci-kit learn : machine learning in python sci-kit learn 0.20.1,Scikit-learn.org.

Available:https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[6] Documentation sci-kit learn : machine learning in python sci-kit learn 0.20.1,Scikit-learn.org.

Available:https://scikitlearn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html

Regards,

Ms. UjwalaRavale,

T. E. Mini - Project Coordinator