NATIONAL INSTITUTE OF TECHNOLOGY DELHI
राष्ट्रीय प्रौद्योगिकी संस्थान दिल्ली

Data Mining

[CSB 352]

Project Report

Cricket Score and Winner Predictor

Submitted to:

Dr. Rishav Singh
(171210033)

Assistant Professor

Submitted By:

Keya Shukla

Srijan Gupta (171210051)

CSE Department

Tanika Jindal (171210056)

[B.Tech CSE, 3rd Year]

Index

No.	Contents	Page No.
1.	Acknowledgment	03
2.	Glimpse: An Introduction	04
3.	In a Nutshell: The Project Abstract	05
4.	Methodology	06
5.	Model Used	09
6.	Our WebApp	10
7.	Related Work	14
8.	Result and discussion	17
9.	Conclusion	18
10.	Future Work	19
11.	References	20

Acknowledgment

In performing our assignment, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. The completion of this assignment gives us much pleasure. We would like to show our gratitude, **Dr. Rishav Singh,** for giving us a good guideline for assignment throughout numerous consultations. We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us in writing this assignment.

Many people, especially our classmates and team members themselves, have made valuable comment suggestions on this proposal which gave us the inspiration to improve our assignment. We thank all the people for their help directly and indirectly to complete our assignment.

Glimpse

(An Introduction)

Cricket is one of the most popular sports in the world, viewed by the majority of the world's population. It is a game played between two teams of eleven players each. With the advent of statistical modelling 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. We focus our research on T20s, the most popular format of the game. With this article, we embark on predicting the outcome of an Indian Premier League (IPL) cricket match. In an IPL season, there may be a minimum of 8 to 10 teams playing and each team play with remaining all teams for a minimum of two times. Matches are held at different venues. 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 the run rate may not yield good results since various factors might affect the score of the innings. We develop a model for T20 format games by mining existing game data which can be available from the Cricinfo website.

In a nutshell

(The Project Abstract)

Cricket matches are known to be tremendously exciting but also, at times, extremely unpredictable. Players are in a constant state of training to emerge triumphant in their matches. To train their teams, coaches use previous performances of their respective teams to target areas where the team needs improvement. This would entail that coaches spend a lot of hours going through video footage trying to analyze what happened and what could have happened had their tactics been different. This wastes precious time and is a major cause of inefficiency in the work-flow. Resolving this would be of tremendous help to coaches as well as their teams and would give them an edge over other teams. This project aims to optimize this process of analyzing cricket matches to change tactics and encourage teams to perform better against certain rival teams through data mining algorithms. The goal is to create a model through the Linear Regression algorithm that predicts the score of an ongoing match by giving ball-to-ball data of previous similar matches (played on the same ground, played against the same team, etc as the ongoing match) and determining the chances of positive outcomes for a particular team.

Data Set Used: https://cricsheet.org/

Methodology

The methodology is a process in which data is selected, transformed, and prepared for the calculations needed to generate useful insights. For this research methodology is SEMMA modeling.

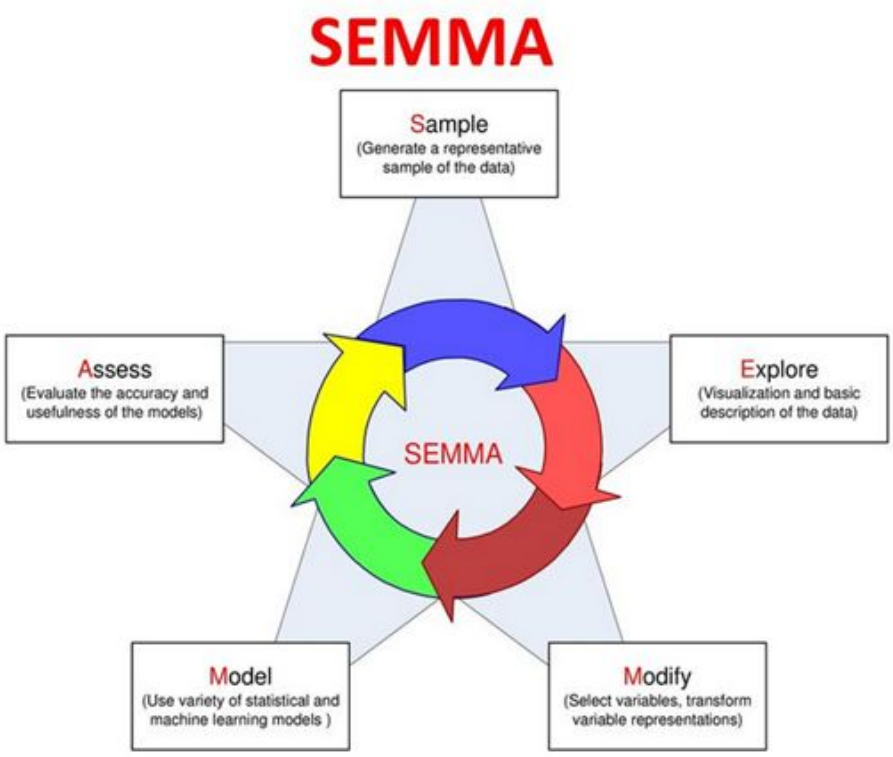


Figure 1: SEMMA Methodology

1. SEMMA:

The SEMMA process was developed by the SAS Institute that considers a cycle with 5 stages for the process: Sample, Explore, Modify, Model, and Assess.

Data mining is the process of discovering predictive information from the analysis of large databases. Python is used for the data mining of the following steps:

- There should be one informational dataset that contains enough information to fulfill the purpose of data mining and should be able to do calculations on it to generate useful insights.

- The target variable on which all the analyses will perform should be there in the dataset. This progression includes the utilization of information planning devices for information import, union, consolidation, filtering, connection, and sifting, just as measurable examining systems.

- Finding patterns between data points by concatenating different options, finding correlations, and relations between attributes. This step includes the exploration of data which includes checking missing values, inconsistencies, exploring variable distributions, techniques for the determination of variables, and finding factors.

- Purification of data includes treating missing values if there is any, removing outliers, and transforming variables for getting the normal distributions of variables. This step is very important in modeling as it's about the modification of data. If the data will not be good, then good results cannot be generated.

- Using Artificial Intelligence techniques to generate useful insights, this includes training of AI models on selected data to generate results in a desirable way. This step includes implementing suitable machine learning models according to the nature of data for the forecasting of values of the target variable.

- The last step is the evaluation of implemented models. Checking the fitness of models whether the model is overfitting or underfit and comparing performances of models by different statistical techniques. If the model, is not appropriate and not giving the best results then try different techniques to make it appropriate.

2. Data Visualizations:

Visualizations are an important part of any research to understand the business and behavior of data in a way that how different attributes are relating to the target variable and what attribute should be the point of focus. Visualizations of data give valuable meaning insights. By the visualizations, every end-user can easily represent the data into an understandable interactive graphic. Cubes will be generated related to different aspects of data. There are various visual analytic tools to create visualizations but as this research has been done in python so visualizations will also be made in python programming using mat plot lib libraries. As the topic of this research is to predict the winner of the match so all the cubes will be related to how different attributes of data are interacting with a match-winner variable.

Model Used

Multiple linear regression (MLR), also known simply as **multiple regression**, is a statistical technique that uses **several** explanatory variables to predict the outcome of a response variable. **Multiple regression** is an extension of **linear** (OLS) **regression** that uses just one explanatory variable.

Here: The attributes- Runs, Overs, Wickets, Striker, and Non-Striker were weighted as opposed to The Total Score.

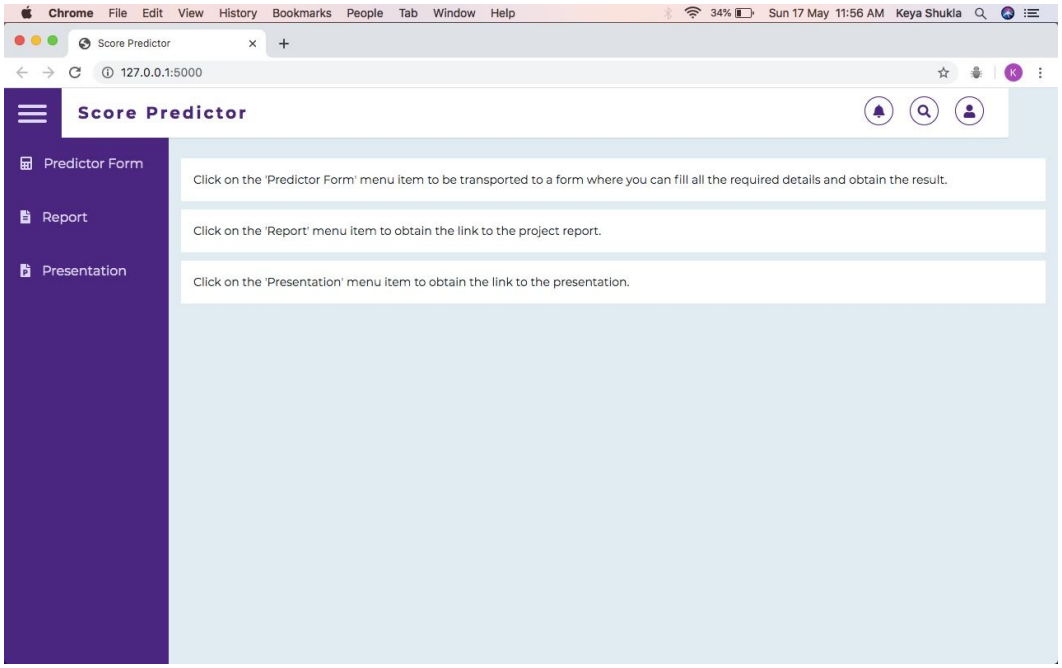
Formally, the model for multiple linear regression, given *n* observations, is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \ldots n.$$

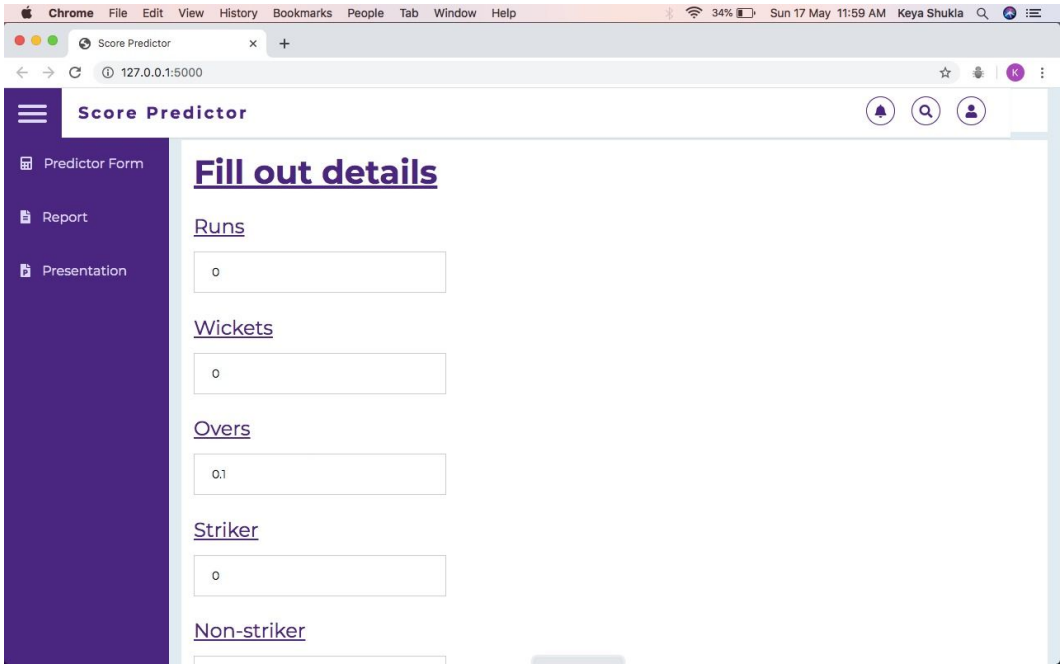
But Since Python provided us with the inbuilt function we used that.

Our WebApp

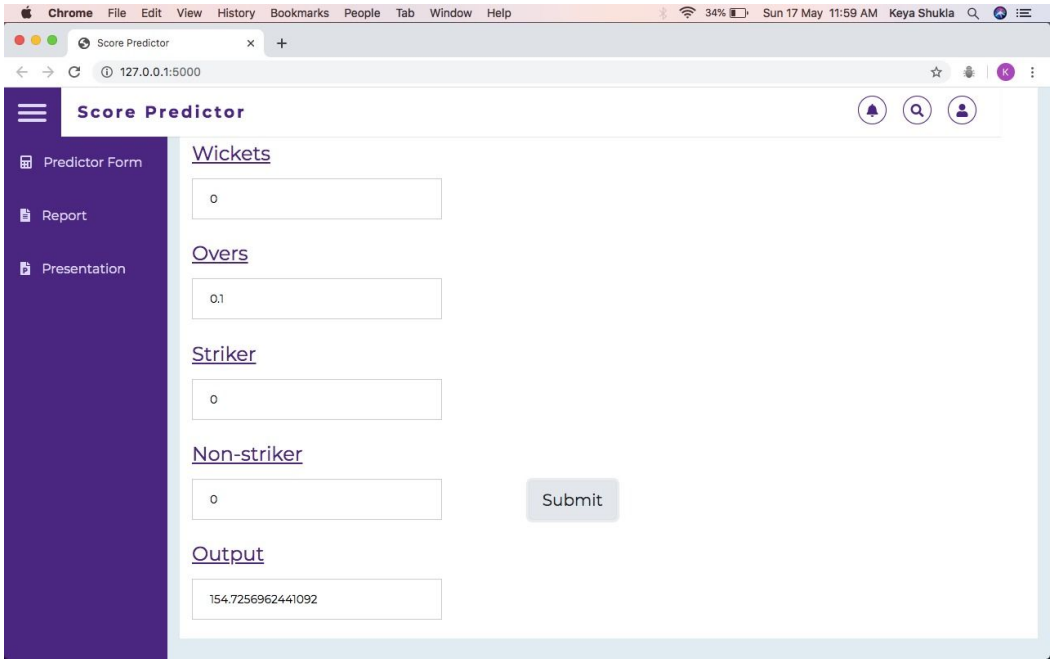
- The Main Page



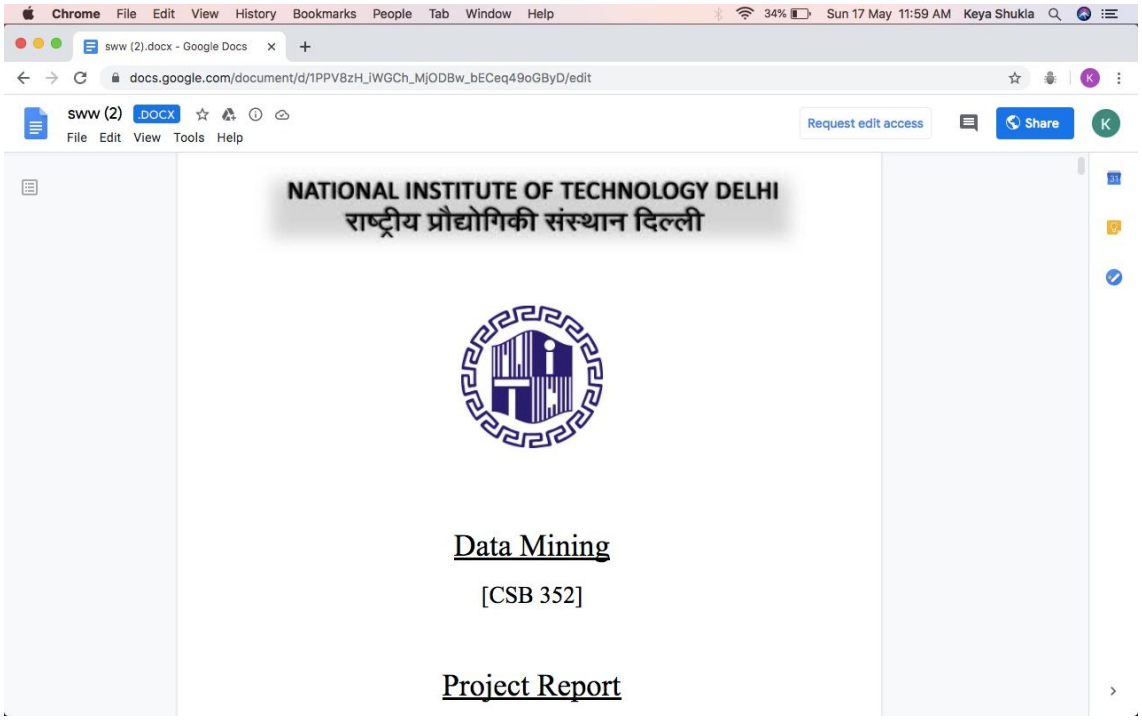
- On Clicking ‘Predictor Form’



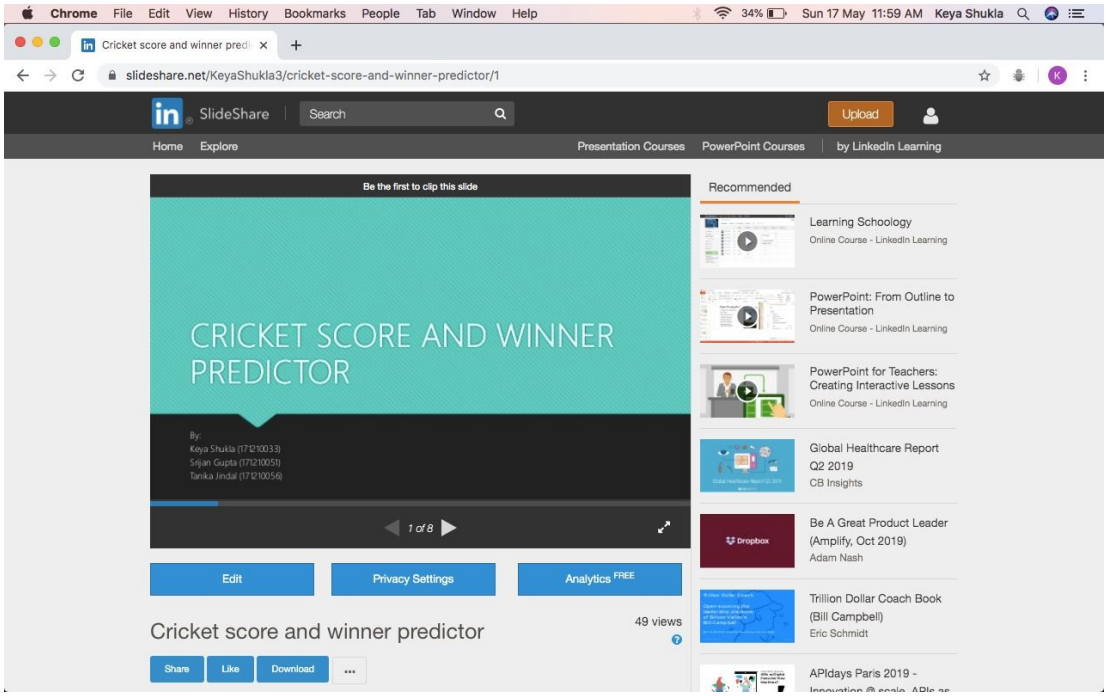
- Filling out details and clicking ‘submit’ button



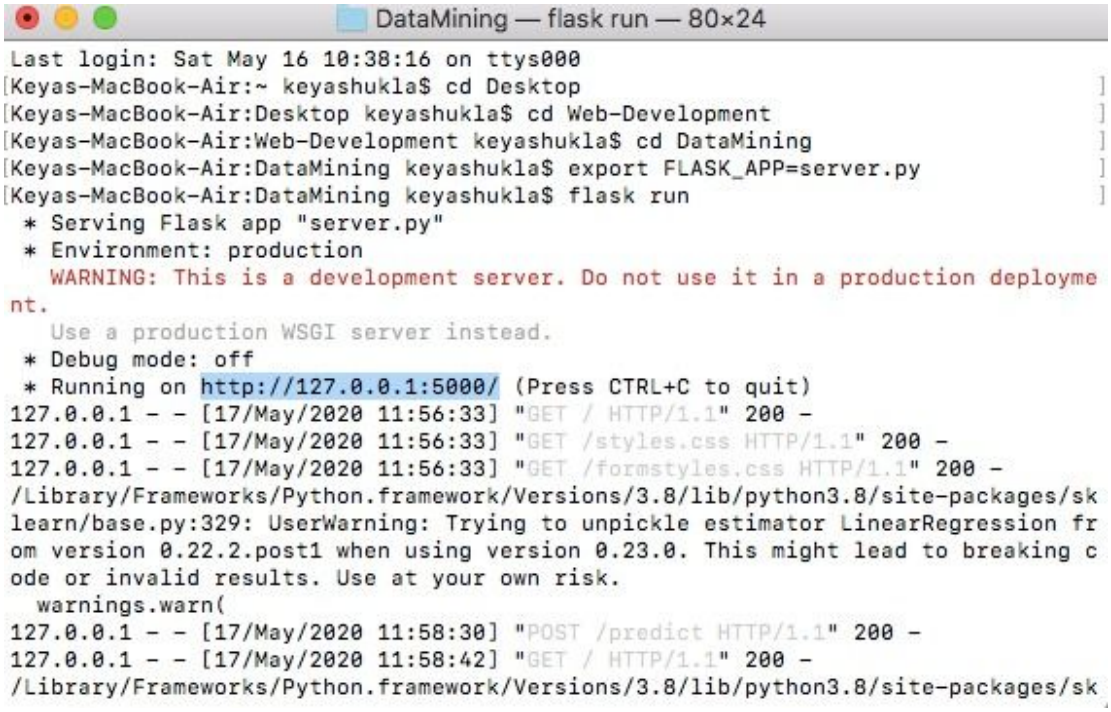
- On clicking ‘Report’ button



- On clicking ‘Presentation’ button



- Commands Used



- Languages and Frameworks Used
  - ❏ HTML
  - ❏ CSS
  - ❏ VANILLA JAVASCRIPT
  - ❏ FLASK
  - ❏ BOOTSTRAP
- **index.html** is inside ‘templates’ folder
- **Css files** are in ‘static’ folder

Related Work

With the evolution of Cricket, it became a very hot topic for sports analysts. A lot of research has been made on cricket but due to inconsistent and complicated data sets, they could not get a breakthrough in predicting match-winner accurately. There are many techniques that have been used in predicting match-winner like KNN, Logistic Regression, SVM, Naïve Bayes but nobody has achieved the accuracy. According to Ahmed & Nazir, they implemented different statistical approaches for the formation of datasets and tried various classification techniques to predict the winner of the One Day Cricket (50 over) match. He has predicted the winner with 80 % accuracy. Shah predicted One Day International match results by using data of ICC match ratings, ICC ranking points for batsmen and bowlers, home factor, ICC rating differences, and ground effects on the match. They implemented Logistic Regression on this data and achieved accuracy in predicting the results of matches 74.9% and in 81% matches they predicted the winning team correctly. Jhawar predicted 71% accuracy in predicting the winner of the One Day International cricket match. He used binary classification models such as Logistic Regression, KNN, Random Forest, and Decision trees. The cross-validation procedure was not carried out. Jhawar has done research on predicting the winner of the match at end of the over, player's

performance recent and past performance, and other statistics' which are necessary for predicting the winner of the match has been used.

The first challenge is to estimate the score that the first team will score at the end of the first innings. Features combination to predict the match outcome is the relative strength of Team B divided by the relative strength of Team A is successful in measuring and comparing the strength of the playing teams. By Random Forest classifier R.F.C. accuracy of 84% has been achieved. Jhanwar analyzed the performances of the One Day International matches played from 2006 and 2016 and accuracy stated that 86% is achieved that top 3 positions of batsman are hot for the man of the match award which is better to previous search and models Random Forests, Decision Trees, KNN and Logistic Regression are the techniques used to predict player performances in a match. Yasir predicted the outcome of a cricket match and for the winner prediction techniques, he proposed a method for predicting the team results and elaborated the working of method which is by using properties of a dynamic team for the winner's prediction like player's history, weather conditions, ground history, and winning percentage. He applied this technique on 100 matches and got an 85% prediction.

1. **Factors to Anticipate Cricket winner**:

Winning a cricket match depends on multiple factors like batting, bowling, fielding, team performances, and player performances. To predict the winner of a cricket match is never an easy task. But there are always some kind of unique aspects or match conditions that may favor some team and sometimes do not such as home advantage, Key Players, Pitch Conditions, and weather conditions.

2. **Cricket Winner Prediction Models**:

Machine learning has become a vast field that is consists of many domain statistics such as artificial intelligence, information technology, and others. Many problems can be solved by the Machine learning model. In the advanced era of today, machines can now work like a human brain because machine learning has been so much evolved. It is learning of computers by creating algorithms that tell the computer how to learn which includes finding the patterns using statistical approaches or similarities in the data. Machine learning algorithms have proved prediction very easy by using classification function to relate the values of attributes in the dataset.

2.1 Naïve Bayes:

Naïve Bayes works on the Bayes probability theorem with the assumption that all the features are independent of the class label (predicted variable) which may be a wrong assumption. Naive Bayes model used in conjunction with recursive feature elimination.

2.2 Decision Tree Regressor Decision:

Tree Regressor has been used to check the overfit by learning from the noise of data using a tree node system. If the max depth of the tree is high, the decision tree regressor takes details from the training data's noise. Decision Trees classification works on tree node principal in which instances are sorted into a tree node system. By this hierarchy, a complex decision-making system is break-down into smaller simpler decisions that provide a simple solution that is easy to implement.

2.3 Random Forest Classifier:

Random Forest classifier is a method used for regression and classification techniques. In the Random Forest Classifiers, to classify a new instance, there is a number of trees in working randomly in a forest putting input vector down and duty of every tree is to give a class label or target variable as a vote for the class. And which node has the highest votes will be chosen by Random Forest Classifier. To increase the accuracy predicted and to control the over-fitting, Random forest uses estimation and averaging approach on the sub-samples of the dataset that is done by fitting a various number of decision tree classifiers. The sub-samples took for this remain equal to the original input size. Random forest is a versatile mechanism enough to deal with both supervised classification and regression tasks. For the datasets under experimentation, approach achieves an accuracy of that of the original Random Forest in a smaller number of trees, and the reduction in size achieved is in the range of 52% to 87%.

2.4 Accuracy Score:

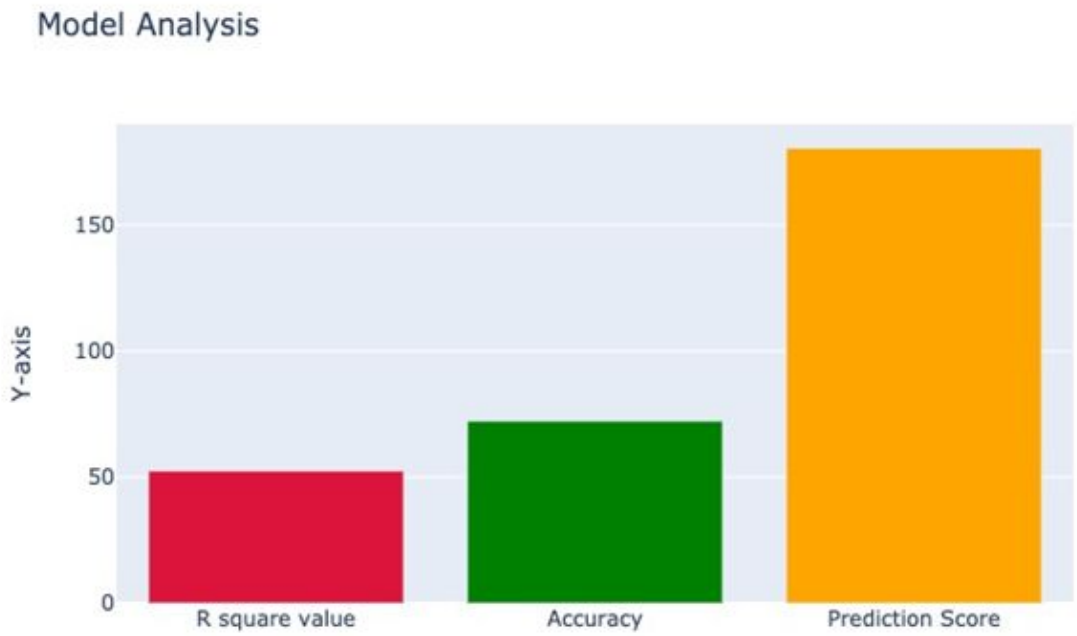
To optimize a model's performance, it should be ensuring that a proper selection of features is under the training of the generative classifiers. To calculate the model's performance or model's accuracy confusion matrix is a matrix that gives the comparison between the predicted class and the actual class into a classification report.

Result and Discussion

For calculation of the accuracy difference between the predicted score and the actual score was calculated. If this difference falls below a particular threshold, we count it as a correct prediction.

We used the Linear regression inbuilt function for our model.

This method gave us an accuracy of 72.34%



Conclusion

This report has discussed successful ventures in the field of cricket score predictor. This was undertaken by us through the application of the concept of multiple linear regressions.

The future work as mentioned hereafter will refine whatever implementation we've had in this project.

Future Work

· As we know Machine Learning and Data Mining are developing at a rapid pace with several new techniques being developed and old techniques being modified to enhance performance, keeping this in mind our work can be expanded to incorporate new methods of classification for outcome prediction.

· More features could be added along with the ones currently considered.

· Although our study is done for ODI matches only, the however similar approach could be applied to predict outcomes in other versions of Cricket matches as well.

· Classification techniques can be applied to other sports such as baseball, football as well, although the method of implementation might differ from one sport to another.

References

1. https://ieeexplore.ieee.org/abstract/document/748960
2. Proceedings of the 2014 SIAM International Conference on Data Mining
3. https://www.sciencedirect.com/science/article/pii/S1877050916304653
4. https://ieeexplore.ieee.org/abstract/document/5715668
5. https://icfhr2018.org/usr/local/pub/GraduateProjects/2161/spm5218/Report.pdf
6. https://www.schrodinger.com/kb/1842