Turkish Journal of Computer and Mathematics Education Vol.12 No.6 (2021), 5111-5124

Research Article

Predictive Analysis of Cricket

Aman Sahu[a], Devang Kaushik[b] , A. Meena Priyadharsini[c]

[a,b]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
[c] Assistant Professor (O.G), Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
[a] aa9057@srmist.edu.in, [b] dr9003@srmist.edu.in, [c] meenapra@srmist.edu.in

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: In cricket, especially the emergence of the availability of a different format the size of automatic machines for learning by guessing and predicting is important. As cricket is highly regarded and popular, there is no one who can imagine who will win the game until the final over-the-top ball. And there are a number of factors such as individual performance, team performance and environmental factors that need to be considered in planning a game strategy. So, we decided to make a machine learning model to predict the outcome of its games. For this project study, we used a data analysis tool called Google Colab to process data and provide recommendations. Improved models can help decision-makers during cricket games to test a team's strengths against the other and environmental factors. For preprocessing of dataset, we have used label encoder and in compilation we have used random forest classifier algorithm to describe the conditions and recommendations of the solutions

1. Introduction

Machine learning is a department of artificial intelligence which tries to solve real engineering problems. It affords the opportunity to learn without being exclusively programmed and is centered at learning from data. It is used everywhere so much a day that we may not notice it. The use of mathematical models, heuristic learning, information acquisition and decision-making trees is the advantage gained in machine learning. Therefore, it provides control, recognition and resilience.

There are a variety of ways in which cricket is played, namely, T20, One Day International, and Test Matches. IPL (or Indian Premier League) is a 20-20 cricket league recognized for endorsing cricket in India and thus encouraging new and competitive players. The league is held annually. IPL teams are selected by auction to represent different Indian cities. Player auctions are not new to sports. However, in India, it was in IPL the selection of a team from an existing collection of players through a player auction was made for the first time. Due to huge fan following, team spirit and financial involvement, the outcomes of these matches are crucial for the stakeholders. This, in turn, depends on the rules governing the game, the toss winner which depends on the team's luck, the competence of the players and their performance on the match day. Many environmental factors, such as player's past performance data, play an important part in forecasting the outcome of a cricket match.

Team selection can be aided with a system to predict a match's outcome taking place between different teams. However, various parameters involved pose major difficulties in predicting the exact outcome of the match. In addition; the correctness of the forecast relies on the size of the data.

For this project study, we used a data analysis tool called google colab to process data and provide recommendations. Decision makers can be aided by improved models during cricket games to test a team's strengths against the other and environmental factors. We plan to contribute to the proposed project in the areas below based on data acquisition,

- Provide analysis of player statistics based on various factors.
- Predict team performance based on individual player statistics.
- Predict Successful prediction of cricket results.
- Predict Predictably predict environmental factors affecting the cricket league.

2. State Of The Art (Literature Survey)

Bowler Performance Prediction for One-day International Cricket Using Neural Networks

2018

S.Muthuswamy and S.S.Lam (2018) proposed "Bowler Performance for One-Day International Cricket Using Neural Networks," in this paper predicting Indian bowler performance in 7 international teams. The neural network method that uses the Back Propagation Network [BPN] and the Radial Basis Function Network (RBFN) used to predict the performance of the Indian bowler team bowler. The recent performance of the BPN and RBPN model was compared with prediction and classification.

A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket

2012

G. D. I. Barr no-B. S. Kantor (2012) states that "How to Compare and Select Cricketers in Overs Overs Cricket,". In this paper the author outlines the main criteria for comparing the selection of strikers in limited colleges. This paper shows a clear 2D representation of the strike rate on one axis and the probability of exit, i.e., P (exit) on another axis. We are building a strike selection strategy based on this 2D framework that combines scale and strike rate As an example of this application we use this principle in beating the 2003 World Cup performance to show the strong and consistent performance of batsmen playing in Indian and Australia team.

Prediction of athletes performance using neural networks: An application in cricket team selection

2010

S. R. Iyer and R. Sharda (2010) reported on "Predicting Performance of Athletes Using Neural Networks: An Application to Cricket Team Selection," used later to predict future performance of players based on their previous performance in which they place the batsmen and throwers in three different categories “maker”, “balance” and “failure” with the services of cricket professionals. They show how these heuristic scales will be used in selecting batsmen for the World Cup. By choice the callers should get a 1 or 2 "character" rating or a rated rating but not a "failure" rating. The conditions for selecting throwers are exactly the same as hitting, the thrower has “moderate” and “good” values and does not get a failure rating selected from the team.

Predicting the performance of batsmen in test cricket

2015

I.P. Wickramasinghe (2015) explained "Guessing the performance of batsmen in test cricket," in this paper explained how the performance of batsmen in a series of tests can be predicted using long-distance and long-range methods. In this paper they collect sample data from test cricket batsmen who played during the 2006-2010 season in nine overseas teams and show that these sample data shows long and high-level formations of three levels.

A Multivariate Data Mining Approach to Predict Match Outcome in One-Day International Cricke, 2016

In the above paper they have used different mathematical methods of data construction and tried different dividing methods to predict who will win the 50 over match. The winner prediction accuracy was 80%.

Predicting ODI Cricket Result, 2017

Here they predict the results of One Day International matches using the details of ICC match ratings, ICC scorers' scorers, home factor, ICC ratings scores and match results. Logistic Regression was applied to the data and found 74.9% accuracy of the predicted result of the games and in 81% of the games predicted the team that would succeed.

Title	Year	Description	Pros	Cons

Predicting C Cricket Result	2015	Predicted the results of the One Day International game using ICC match ratings data, ICC level scores, home feature, ICC rating differences and ground results	They used Logistic Regression on this data and the accuracy in predicting match results was 74.9%.	Theoretical explanation and no practical approach with explanation
Quantitative Assessment Player Performance and Winner Prediction in ODI Cricket	2017	It is predicted that who will win the One Day International cricket match	They used Logistic Regression, KNN, Random Forest and Decision Trees.	The cross verification process did not happen.
Dynamic Win Prediction Twenty20 Cricket: Based on Relative Team Strengths	2017	They did some research by guessing who would win the game at the end of the over, the player's recent and past performance and other statistics' needed to predict who would win the game were used.	Random Forest classifier is been proposed.	They have arrived 74.1 % accuracy only.
Predicting a T20 cricket result while the match is in progress	2015	Guessing the performance of each player is what is referred to in this paper e.g. each batsman scores runs and each bowler takes wickets.	The use of Naive Bayes, Random Forest Classifier, Multi-class SVM and Decision Tree Classifiers was done to create a model for runs and wickets	Not much detailed explanation about experimental results, tools and algorithm implementation.

Predicting Players Performance in One Day International Cricket Matches Using Machine Learning	2018	The method is described to increase the accuracy of the prediction of the game of cricket by Niravkumar Pandey and Kalpdrum Passi by comparing multi-cast division algorithms including Decision Trees, Naïve Bayes, SVM and Random Forest.	Weka and Dataiku tools used for forecasting statistics. The Random Forest is used to predict runners and wickets taken by the bowler for given data sets.	Only statistical methodology based approach.

3. Inference From The Literature Survey

The inference derived from related works is that, in the field of predictive analysis, there have been many methods to get the task of prediction done. Also, from very early times, the task of prediction is processed using machine learning algorithms. Different authors propose different machine learning techniques to get this task done until the very concept of neural network was explored in this domain. As proposed by S.Muthuswamy and S.S.Lam (2018), S. R. Iyer and R. Sharda (2010), the authors use various machine learning algorithms for prediction of match outcome, but as proposed by Sasank Viswanadha, Kaustubh Sivalenka, Madan Gopal Jhawar and Vikram Pudi (2017) highlights the better accuracy of using random forest classifier for the task.

Predicting outcome of the game has recognized some fundamental problem. In the existing method, by extensive literature survey many research papers have researched on this topic but most of them have used primitive machine learning algorithms like Naïve Bayes and logistic regression. They have taken into consideration factors like teams, toss winners, winners by runs and winners by wickets. We aim to develop an effective machine learning tool which is needed to predict the outcome of a game, taking humidity and wind speed into consideration. In the current situation, many franchise owners have lost money on the negative prediction results of players.

Also, having gone through the works of different authors and researchers, a keen inter- est in the wide applications of Machine Learning algorithms has been seen. The significant reason for the application of random forest classifier and adaboost, is because these algorithm can take a large amount of input features, train the decision trees and stumps and can give a very accurate prediction.

The old machine learning techniques like Naïve Bayes and Logistic Regression have run out of date and a little change in the application needs rebuilding the whole prediction system, according to the stakeholder's needs. With the advancements of parallel processing and increase in computational capacities the Machine Learning algorithms, tend to outperform the older systems and hence are significant motivation for our work.

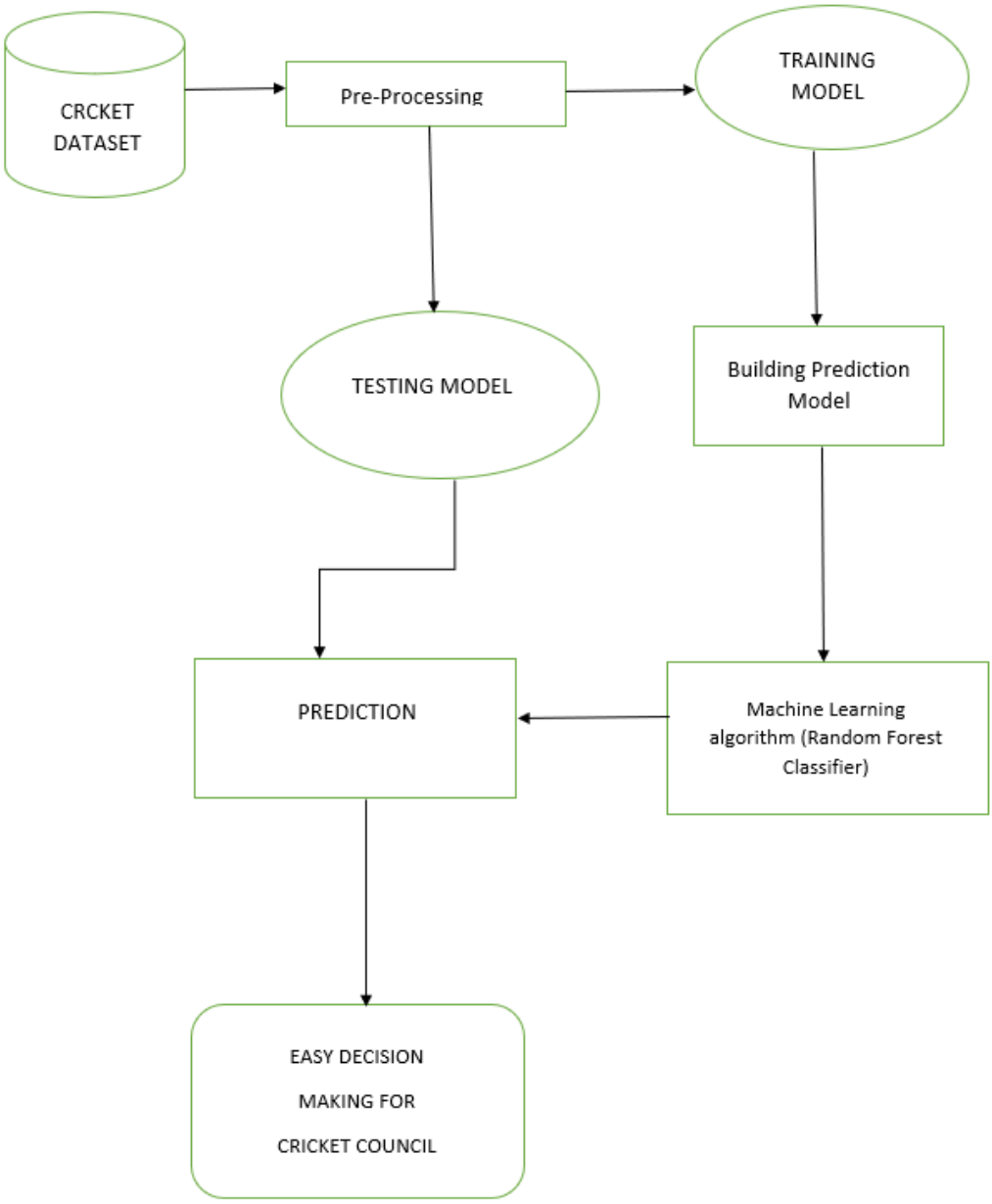
4. System Design

The purpose of this project research is to create data analysis tools to process cricket data with promising results. This project tutorial analyzes all the parameters used in the game namely player statistics, team statistics, environmental features to provide an effective solution to make the game more fun and keep the fun maintained.

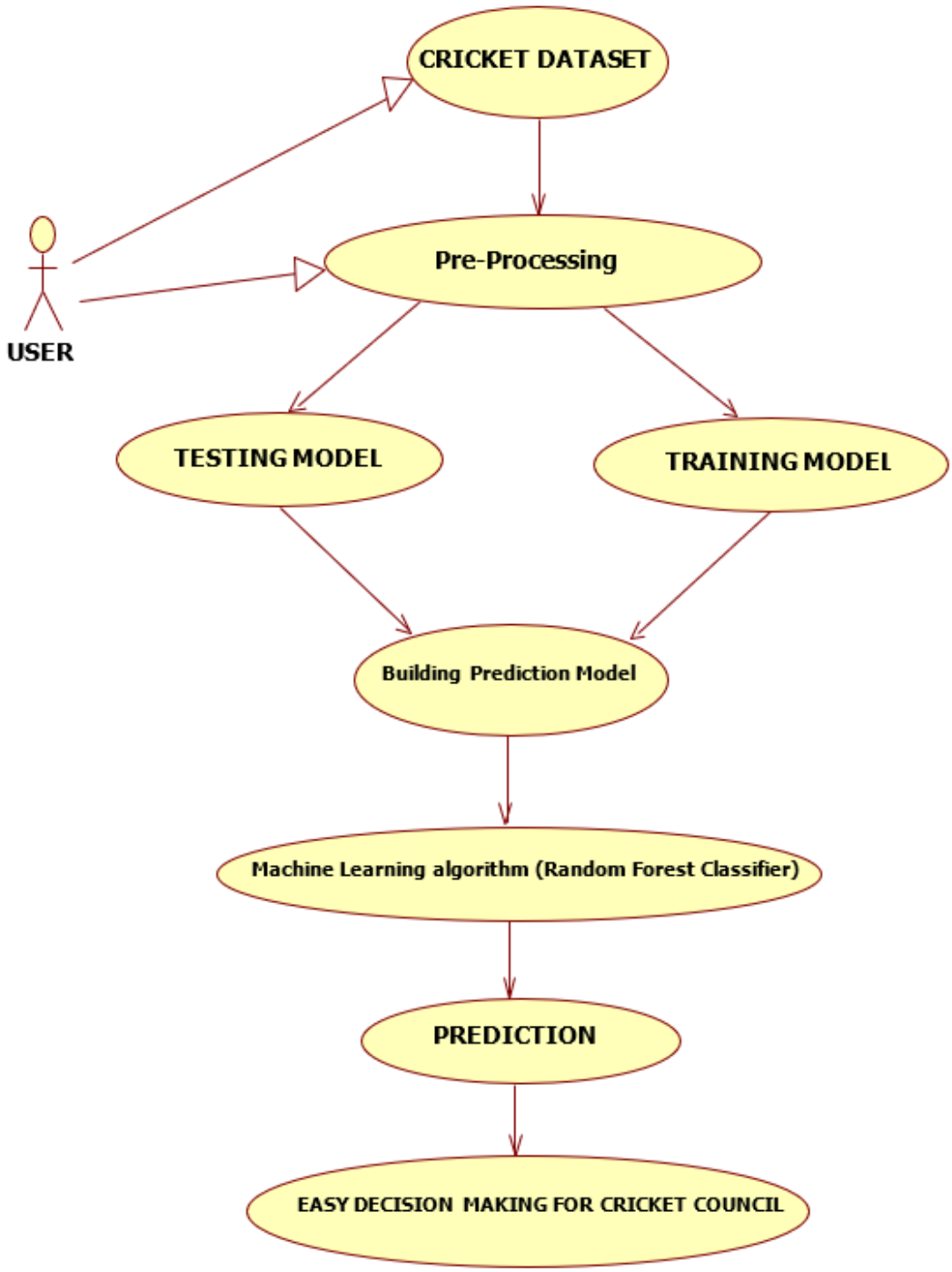
Cricket began in the 16th century in England. Cricket is a game with many formats, different levels of play and different lengths. The Twenty20 is one of three current types of cricket known to the International Cricket Council (ICC). In that format, two teams with one innings each team has more than 20 overs. Due to the short duration and the excitement, it is productive, twenty-two cricket has been so successful. There are many competitions held annually at the local and international level. There is a strong commercial interest in predicting player performance in cricket leagues. This has encouraged a lot of analysis of individual and team performance, as well as predictions of upcoming games, across game formats.

This project research aims to improve the application of accurate prediction in the game of cricket using machine learning about the game, the environment, the players. We have proposed a system that overcomes the major weaknesses of time-consuming and labor-intensive work to keep individual player records and statistics. In the proposed system the user can upload historical data collected from a machine learning tool such as Pycharm or

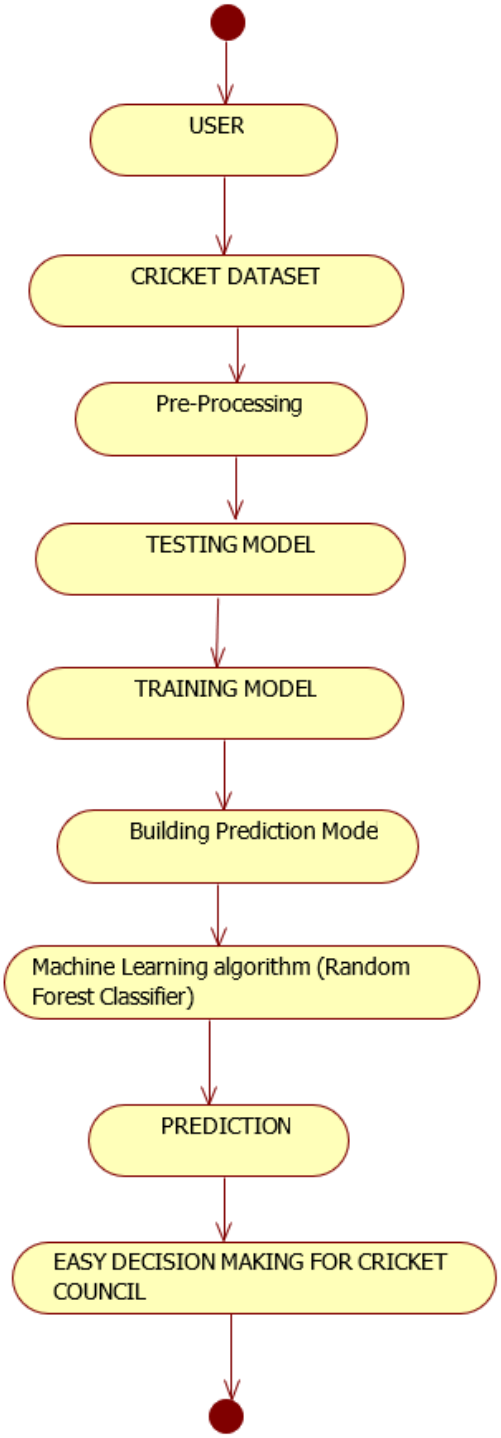
Jupiter or Google Colab and perform data analytics to obtain the result. In data analytics we use the random forest classifier algorithm for building the prediction model defining conditions to process and produce a result.



Architecture Diagram of the prediction Model



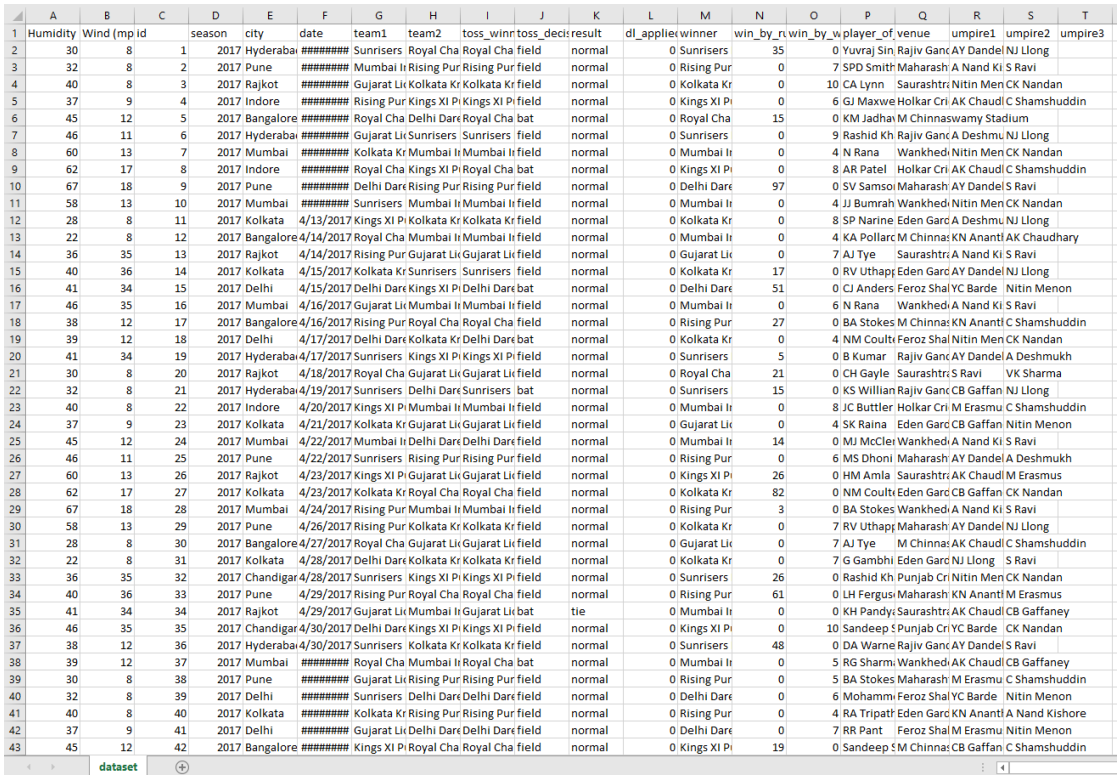
Use Case UML Diagram



Activity UML Diagram

Data Source :

The database used for this project is available in kaggle, a database obtained by the cricket crowd. The database in the database is in csv format. Contains natural value, per historical data. Therefore, this last publicly available site can be used for analysis.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Humidity	Wind (mp	id	season	city	date	team1	team2	toss_winn	toss_decis	result	dl_applie	winner	win_by_r	win_by_w	player_of	venue	umpire1	umpire2	umpire3
2	30	8	1	2017	Hyderaba	########	Sunrisers	Royal Cha	Royal Cha	field	normal	0	Sunrisers	35	0	Yuvraj Sin	Rajiv Gan	AY Dandel	NJ Llong	
3	32	8	2	2017	Pune	########	Mumbai In	Rising Pur	Rising Pur	field	normal	0	Rising Pur	0	7	SPD Smith	Maharash	A Nand Ki	S Ravi	
4	40	8	3	2017	Rajkot	########	Gujarat Li	Kolkata Kr	Kolkata Kr	field	normal	0	Kolkata Kr	0	10	CA Lynn	Saurashtr	Nitin Men	CK Nandan	
5	37	9	4	2017	Indore	########	Rising Pur	Kings XI P	Kings XI P	field	normal	0	Kings XI P	0	6	GJ Maxwe	Holkar Cri	AK Chaud	C Shamshuddin	
6	45	12	5	2017	Bangalore	########	Royal Cha	Delhi Dare	Royal Cha	bat	normal	0	Royal Cha	15	0	KM Jadha	M Chinnaswamy Stadium			
7	46	11	6	2017	Hyderaba	########	Gujarat Li	Sunrisers	Sunrisers	field	normal	0	Sunrisers	0	9	Rashid Kh	Rajiv Gan	A Deshmu	NJ Llong	
8	60	13	7	2017	Mumbai	########	Kolkata Kr	Mumbai In	Mumbai In	field	normal	0	Mumbai In	0	4	N Rana	Wankhed	Nitin Men	CK Nandan	
9	62	17	8	2017	Indore	########	Royal Cha	Kings XI P	Royal Cha	bat	normal	0	Kings XI P	0	8	AR Patel	Holkar Cri	AK Chaud	C Shamshuddin	
10	67	18	9	2017	Pune	########	Delhi Dare	Rising Pur	Rising Pur	field	normal	0	Delhi Dare	97	0	SV Samso	Maharash	AY Dandel	S Ravi	
11	58	13	10	2017	Mumbai	########	Sunrisers	Mumbai In	Mumbai In	field	normal	0	Mumbai In	0	4	JJ Bumrah	Wankhed	Nitin Men	CK Nandan	
12	28	8	11	2017	Kolkata	4/13/2017	Kings XI P	Kolkata Kr	Kolkata Kr	field	normal	0	Kolkata Kr	0	8	SP Narine	Eden Gard	A Deshmu	NJ Llong	
13	22	8	12	2017	Bangalore	4/14/2017	Royal Cha	Mumbai In	Mumbai In	field	normal	0	Mumbai In	0	4	KA Pollarc	M Chinnas	KN Ananth	AK Chaudhary	
14	36	35	13	2017	Rajkot	4/14/2017	Rising Pur	Gujarat Li	Gujarat Li	field	normal	0	Gujarat Li	0	7	AJ Tye	Saurashtr	A Nand Ki	S Ravi	
15	40	36	14	2017	Kolkata	4/15/2017	Kolkata Kr	Sunrisers	Sunrisers	field	normal	0	Kolkata Kr	17	0	RV Uthapp	Eden Gard	AY Dandel	NJ Llong	
16	41	34	15	2017	Delhi	4/15/2017	Delhi Dare	Kings XI P	Delhi Dare	bat	normal	0	Delhi Dare	51	0	CJ Anders	Feroz Shal	YC Barde	Nitin Menon	
17	46	35	16	2017	Mumbai	4/16/2017	Gujarat Li	Mumbai In	Mumbai In	field	normal	0	Mumbai In	0	6	N Rana	Wankhed	A Nand Ki	S Ravi	
18	38	12	17	2017	Bangalore	4/16/2017	Rising Pur	Royal Cha	Royal Cha	field	normal	0	Rising Pur	27	0	BA Stokes	M Chinnas	KN Ananth	C Shamshuddin	
19	39	12	18	2017	Delhi	4/17/2017	Delhi Dare	Kolkata Kr	Delhi Dare	bat	normal	0	Kolkata Kr	0	4	NM Coulte	Feroz Shal	Nitin Men	CK Nandan	
20	41	34	19	2017	Hyderaba	4/17/2017	Sunrisers	Kings XI P	Kings XI P	field	normal	0	Sunrisers	5	0	B Kumar	Rajiv Gan	AY Dandel	A Deshmukh	
21	30	8	20	2017	Rajkot	4/18/2017	Royal Cha	Gujarat Li	Gujarat Li	field	normal	0	Royal Cha	21	0	CH Gayle	Saurashtr	S Ravi	VK Sharma	
22	32	8	21	2017	Hyderaba	4/19/2017	Sunrisers	Delhi Dare	Sunrisers	bat	normal	0	Sunrisers	15	0	KS Willian	Rajiv Gan	CB Gaffan	NJ Llong	
23	40	8	22	2017	Indore	4/20/2017	Kings XI P	Mumbai In	Mumbai In	field	normal	0	Mumbai In	0	8	JC Buttler	Holkar Cri	M Erasmu	C Shamshuddin	
24	37	9	23	2017	Kolkata	4/21/2017	Kolkata Kr	Gujarat Li	Gujarat Li	field	normal	0	Gujarat Li	0	4	SK Raina	Eden Gard	CB Gaffan	Nitin Menon	
25	45	12	24	2017	Mumbai	4/22/2017	Mumbai In	Delhi Dare	Delhi Dare	field	normal	0	Mumbai In	14	0	MJ McCle	Wankhed	A Nand Ki	S Ravi	
26	46	11	25	2017	Pune	4/22/2017	Sunrisers	Rising Pur	Rising Pur	field	normal	0	Rising Pur	0	6	MS Dhoni	Maharash	AY Dandel	A Deshmukh	
27	60	13	26	2017	Rajkot	4/23/2017	Kings XI P	Gujarat Li	Gujarat Li	field	normal	0	Kings XI P	26	0	HM Amla	Saurashtr	AK Chaud	M Erasmus	
28	62	17	27	2017	Kolkata	4/23/2017	Kolkata Kr	Royal Cha	Royal Cha	field	normal	0	Kolkata Kr	82	0	NM Coulte	Eden Gard	CB Gaffan	CK Nandan	
29	67	18	28	2017	Mumbai	4/24/2017	Rising Pur	Mumbai In	Mumbai In	field	normal	0	Rising Pur	3	0	BA Stokes	Wankhed	A Nand Ki	S Ravi	
30	58	13	29	2017	Pune	4/26/2017	Rising Pur	Kolkata Kr	Kolkata Kr	field	normal	0	Kolkata Kr	0	7	RV Uthapp	Maharash	AY Dandel	NJ Llong	
31	28	8	30	2017	Bangalore	4/27/2017	Royal Cha	Gujarat Li	Gujarat Li	field	normal	0	Gujarat Li	0	7	AJ Tye	M Chinnas	AK Chaud	C Shamshuddin	
32	22	8	31	2017	Kolkata	4/28/2017	Delhi Dare	Kolkata Kr	Kolkata Kr	field	normal	0	Kolkata Kr	0	7	G Gambhi	Eden Gard	NJ Llong	S Ravi	
33	36	35	32	2017	Chandigar	4/28/2017	Sunrisers	Kings XI P	Kings XI P	field	normal	0	Sunrisers	26	0	Rashid Kh	Punjab Cri	Nitin Men	CK Nandan	
34	40	36	33	2017	Pune	4/29/2017	Rising Pur	Royal Cha	Royal Cha	field	normal	0	Rising Pur	61	0	LH Fergus	Maharash	KN Ananth	M Erasmus	
35	41	34	34	2017	Rajkot	4/29/2017	Gujarat Li	Mumbai In	Gujarat Li	bat	tie	0	Mumbai In	0	0	KH Pandy	Saurashtr	AK Chaud	CB Gaffaney	
36	46	35	35	2017	Chandigar	4/30/2017	Delhi Dare	Kings XI P	Kings XI P	field	normal	0	Kings XI P	0	10	Sandeep S	Punjab Cri	YC Barde	CK Nandan	
37	38	12	36	2017	Hyderaba	4/30/2017	Sunrisers	Kolkata Kr	Kolkata Kr	field	normal	0	Sunrisers	48	0	DA Warne	Rajiv Gan	AY Dandel	S Ravi	
38	39	12	37	2017	Mumbai	########	Royal Cha	Mumbai In	Royal Cha	bat	normal	0	Mumbai In	0	5	RG Sharm	Wankhed	AK Chaud	CB Gaffaney	
39	30	8	38	2017	Pune	########	Gujarat Li	Rising Pur	Rising Pur	field	normal	0	Rising Pur	0	5	BA Stokes	Maharash	M Erasmu	C Shamshuddin	
40	32	8	39	2017	Delhi	########	Sunrisers	Delhi Dare	Delhi Dare	field	normal	0	Delhi Dare	0	6	Mohamm	Feroz Shal	YC Barde	Nitin Menon	
41	40	8	40	2017	Kolkata	########	Kolkata Kr	Rising Pur	Rising Pur	field	normal	0	Rising Pur	0	4	RA Tripath	Eden Gard	KN Ananth	A Nand Kishore	
42	37	9	41	2017	Delhi	########	Gujarat Li	Delhi Dare	Delhi Dare	field	normal	0	Delhi Dare	0	7	RR Pant	Feroz Shal	M Erasmu	Nitin Menon	
43	45	12	42	2017	Bangalore	########	Kings XI P	Royal Cha	Royal Cha	field	normal	0	Kings XI P	19	0	Sandeep S	M Chinnas	CB Gaffan	C Shamshuddin	

dataset

Dataset acquired from Kaggle with addition of humidity and wind speed

Pre-Processing :

Manual Encoding – Manual Encoding is done to convert categorical values into numerical values which are machine readable.

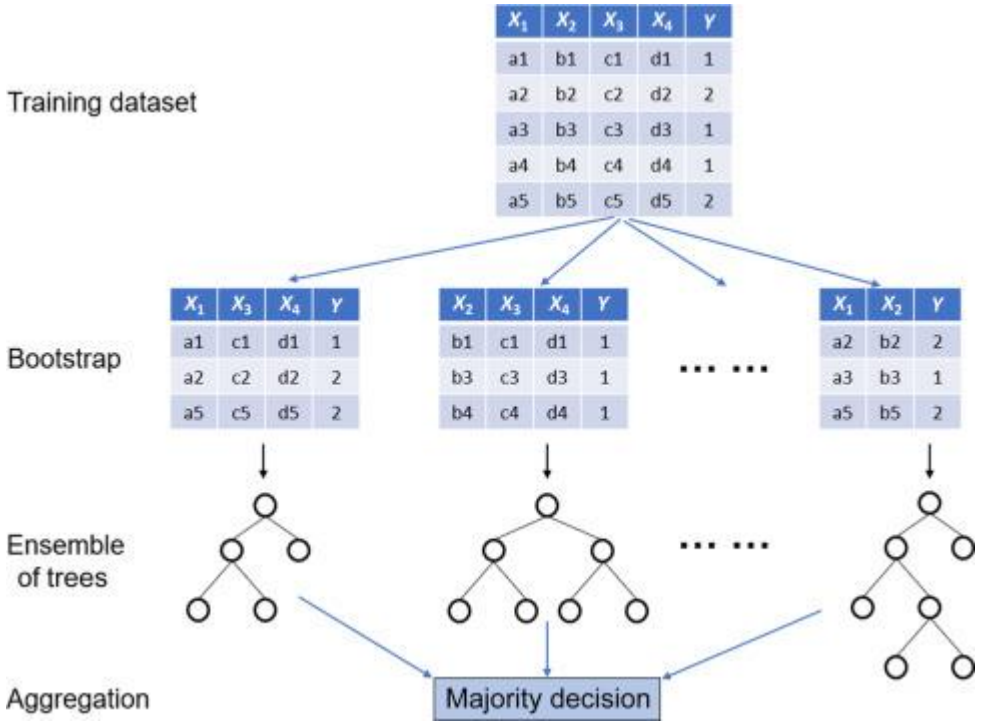
Label Encoder – Label Encoder is a help category to help undo labels that contain values only between 0 and n_class-1. Labeling means converting labels into numbers to convert them into machine-readable form. Machine learning algorithms can then determine the best way to use those labels. It is an important pre-processing step for systematic databases in supervised reading.

Feature Selection :

Adding a lot of features tends to decrease the generalization capability of our model and may also reduce the classifier's overall accuracy. Feature selection aids the prediction model by choosing the best set of features suitable for predicting the match outcome with highest possible accuracy.

Random Forest Classifier :

Random Forest is a variety of targeted learning algorithm based on one of the learning methods used to go back and divide a task. The algorithm in the random forest combines a different algorithm together of the same type which is a multi-choice tree that leads to a forest of trees which is why it is called the "Random Forest". Often random forest provision has a more practical and more accurate result compared to other classification algorithms. This algorithm generates a certain number of decision-making trees by creating a jungle from input cricket databases and produces a professional model for comparison and providing accurate predictions.

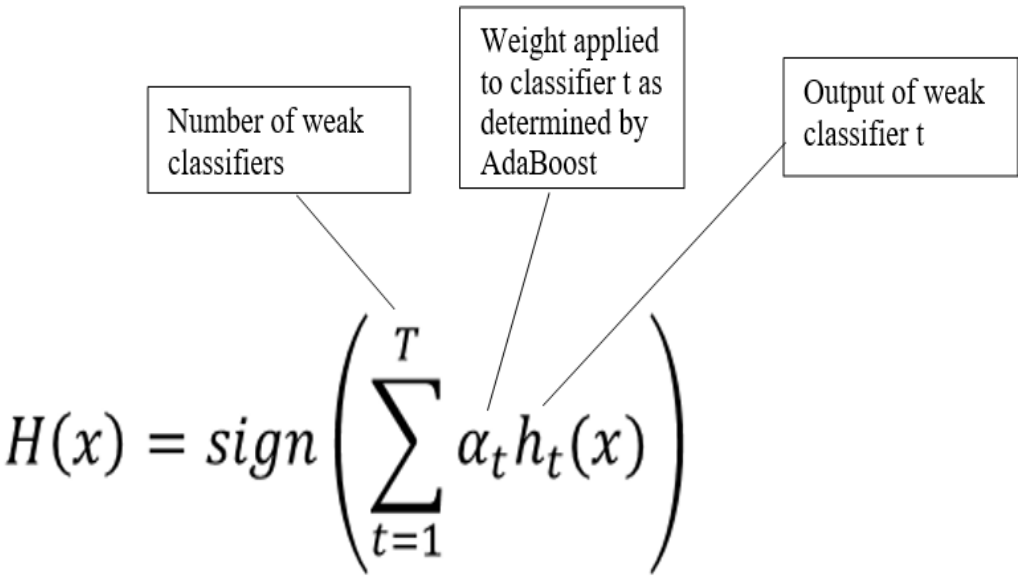


Working of Random Forest Classifier algorithm

AdaBoost :

AdaBoost or Adaptive Boosting was introduced by Robert Schapire and Yoav Freund for combining multiple other learning algorithms or weak classifiers into a single strong learning algorithm or strong classifier. The final output of the boosted algorithm is a weighted sum of the other weaker algorithms. It tweaks the weaker learning algorithms to support those input instances which were classified incorrectly by the previous classifiers.

It is less prone to the overfitting problem than other algorithms. AdaBoost uses the weak learners as its decision trees with only one depth, called decision stumps. It puts more weight on those input instances which are difficult to classify than those already classified.



5. Implementation

The proposed system is implemented by applying the following modules:-

Data Cleaning : Data cleaning is done to identify and correct any errors in the dataset obtained from Kaggle that may impact our predicting model negatively.

In our model data cleaning is accomplished by replacing null values in our dataset:-

Replacing Null Values :-

By default, NaN, n/a and blank cells are all treated as null values. Then among these null values, if a numerical data is missing then it is replaced with either the mean or median of the respective feature column and if a categorical data is missing then it is replaced with mode of the respective feature column.

Data Analysis : For data analysis we encoded the data into numerical format which is machine readable and also selected the best set of feature columns for accurate outcome prediction using Feature selection technique called P-Value Testing.

Encoding :-

In order to analyze the data in our dataset we have to convert the data into a format that is readable by a machine. This task is accomplished by encoding the data.

Manual Encoding :-

In manual encoding we replaced string values with numerical values in teams, winner, and toss winner columns, for example, Mumbai Indians or MI becomes 1.

Label Encoding :-

In label encoding we replaced string data type with integer data type in city and toss_decision columns using an instance of LabelEncoder() and fit_transform() function in loop imported from sklearn.preprocessing.

Feature Selection :

Feature selection is done to select the best set of features to be used to predict the outcome of a match. This task is accomplished here by using P Value Testing.

P-Value Testing :

P-Value testing allows us to verify whether we will get a particular result when a chosen hypothesis or test statistic, say S0 is assumed to be correct. Since we have implemented backward elimination, if the P-Value is greater than 0.5 then we reject the chosen test statistic claiming it to be strange and incorrect or a one-sided hypothesis. P-Value is also called “observed significance level”.

Data Modeling :

In this module we created data models which show the associations formed between different data objects using Logistic regression, Random Forest Classifier and AdaBoost algorithms.

Multinomial Logistic Regression :-

Though the accuracy of our prediction model is less when we tried to implement Multinomial Logistic Regression for classification, we have implemented this module to compare its level of accuracy with that of Random Forest Classifier and AdaBoost algorithms.

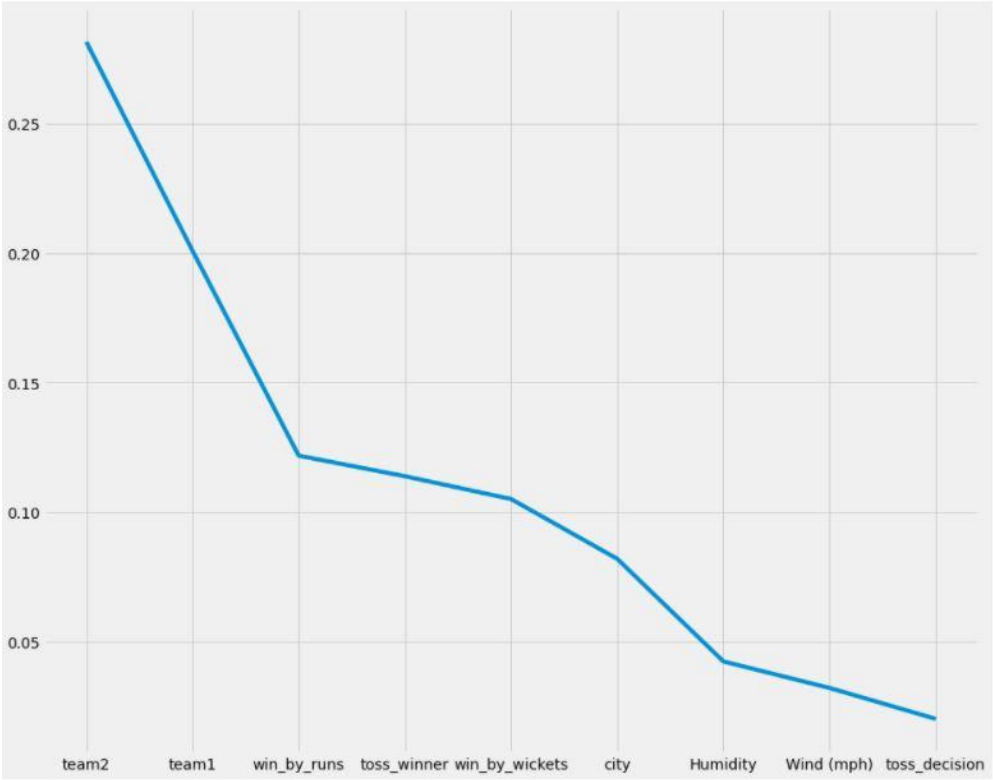
Random Forest Classifier :-

Random Forest Classifier algorithm creates a number of specified (n_estimators) decision trees which are fed data by row and feature sampling making them experts of the training dataset. We created an instance of random forest classifier and used fit() function to feed data and create a bootstrap and get output with the help of a target variable.

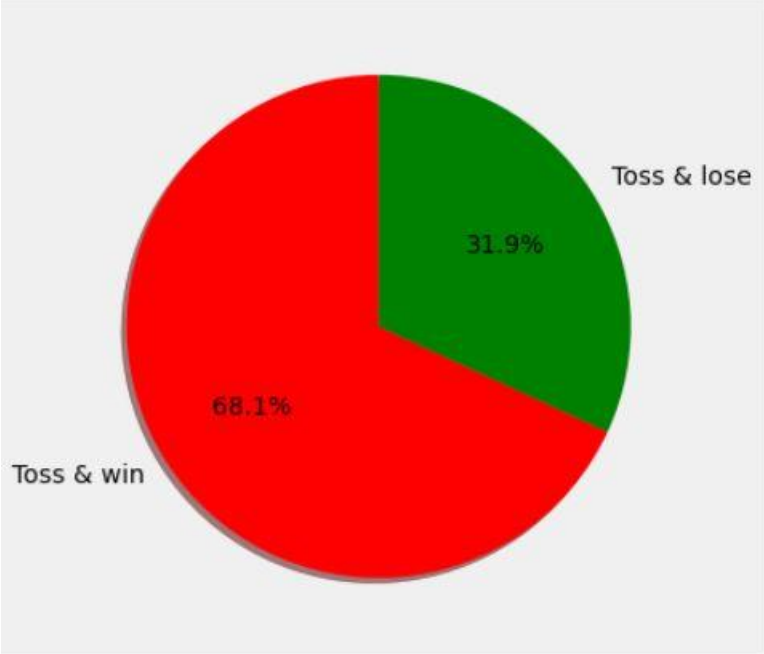
AdaBoost :-

Since it is less prone to overfitting AdaBoost has been implemented for the purpose of increasing the accuracy further. AdaBoost will create stumps and assign weights to the records which will be fed into the first stump. The records for which the first stump predicts the output incorrectly will get passed on to the next stump with an updated weight. This process will keep on iterating until the entire dataset passes through all the sequential decision stumps and the prediction output has a very low error rate. It combines the weak learners (decision stumps) to create a strong learner whose output is more accurate than that of the individual stumps.

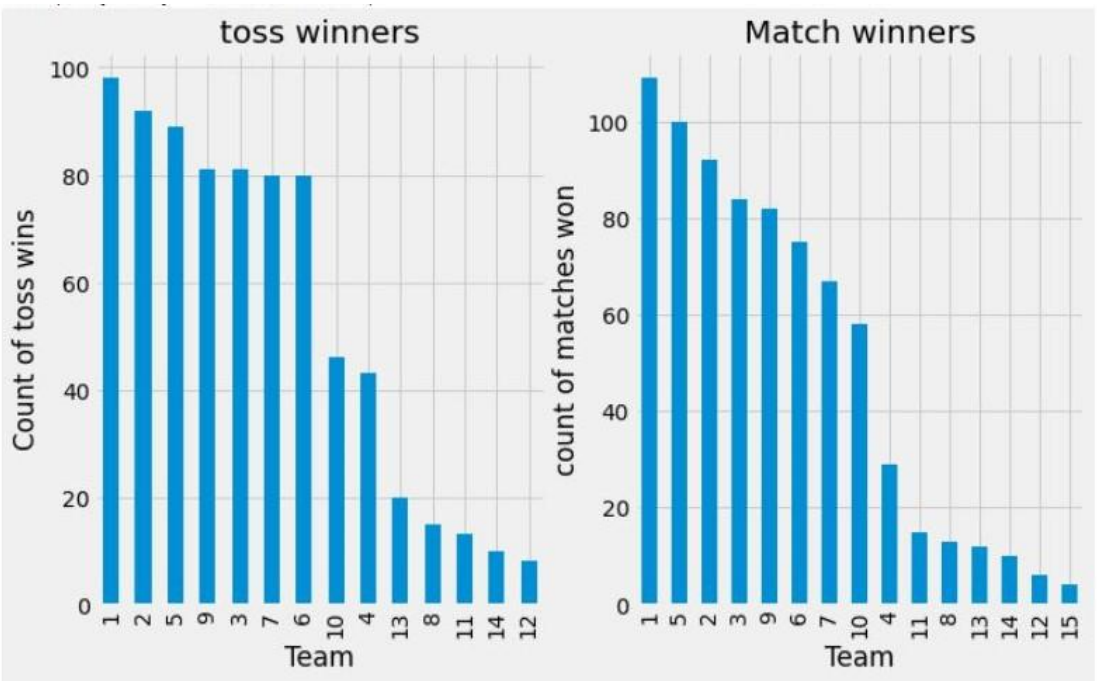
Data Visualization :



Feature Importance Graph



Pie Plot depicting the times a team that has won the toss has won the match as well and the times a team that has won the toss has lost the match



Graphical representation of number of Toss and Match won by each team

6. Results And Discussions

The Software Environment:

Colab is a free cloud-based writing platform. Colab can access electronic libraries that can be downloaded to your notebook. It is ready for everything from improving your Python coding skills to working with deep libraries, such as PyTorch, Cameras, TensorFlow, and OpenCV.

• Zero configuration is required

• Free access to GPUs

• Easy sharing

Google has been fast paced in AI research. TensorFlow, which is an AI framework and Colaboratory, which is a development tool were made by google. TensorFlow is an open-source software and Colaboratory is meant for public use for free and is otherwise called Google Colab or just Colab.

One more feature offered by Google is that developers can make use of the GPU. To make its software a standard in machine learning and data science and also to build a broad customer base for Google Cloud APIs could be some of the reasons why google made its software free to the public

For no apparent reason, there seems to be a reduction in the learning as well as the development of machine learning applications after colab was introduced.

Colab is Jupyter's free note-taking site that works perfectly in the cloud. You do not need setup and your created notebooks are editable simultaneously by your team members in the same way documents are edited in Google Docs.

Python is a translated, high-quality programming language, with a common purpose. Python makes sure that code is readable with its remarkable utilization of white spaces. Its object oriented and language building method tries to assist editors in writing clear, logical code for small and large projects.

Python is a programming language for multiple paradigms. Object-oriented programs and systematic programs, functional programming and interactive programs (including metaprogramming and metaobjects) are supported. Contract programming and logic programming are also supported by the use of extensions.

Powerful typing and a combination of reference counts and a garbage collector that takes care of memory management are used. It also incorporates dynamic word flexibility (late binding), which includes method and word changes during application.

Python attempts to acquire syntax and simple, low-computer programming language while giving developers the opportunity to choose their own writing style. Python adopted "there must be one way — and probably the only one — to make it clear" philosophy for design.

The developers of Python avoid premature execution, and exclude layers in less important components of CPython reference that can provide a small increase in speed at a clear cost. With speed as the crucial factor, the Python program developer can submit timebound tasks by adding modules that are written in C-language languages, or by making use of PyPy, a timely compiler. Cython can be used to convert the Python script into C and direct C-level API calls to Python interpreter.

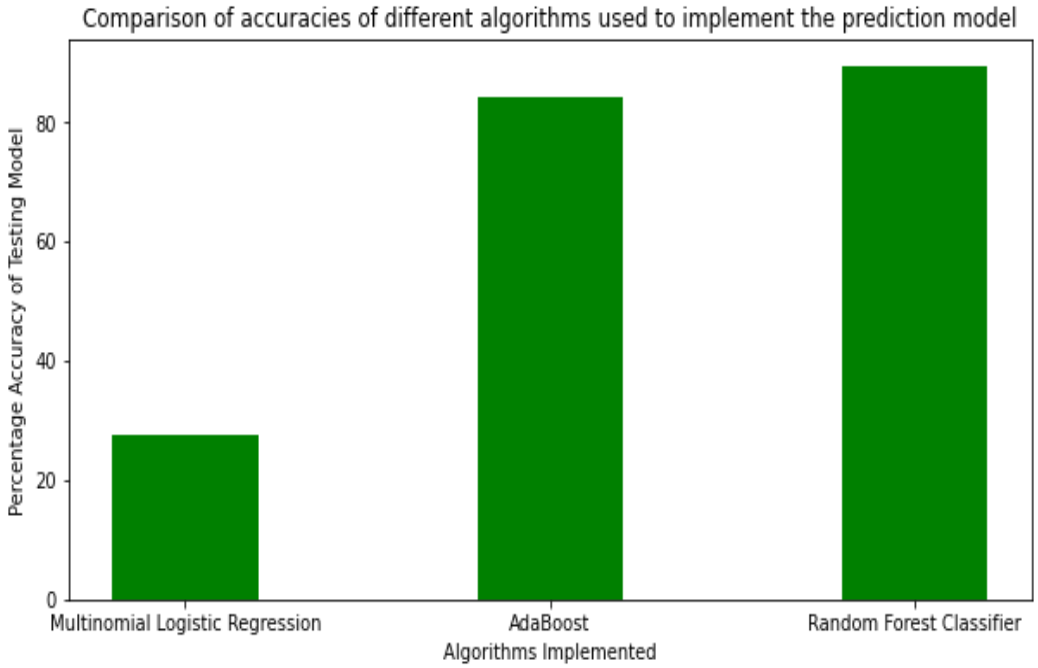
Results:

With the application of Random Forest Classifier algorithm in our prediction model we have achieved an accuracy of 98.14% for the training model and an accuracy of 89.47% for the testing model.

In contrast to Random Forest, when we applied Multinomial Logistic Regression for predicting the match outcome, the accuracy of training model was found to be 29.62% and that of the testing model was found to be 27.63%.

Lastly, when we applied AdaBoost or Adaptive Boosting to our prediction model, we were able to achieve an accuracy score of 1.0 for the training model and accuracy percentage 84.21% for the testing model which are yet less than those of Random Forest Classifier algorithm.

Comparison between the accuracy values has been done by plotting a bar graph shown in figure below.



Bar Graph representing the comparison of accuracies of the Testing Models

7. Conclusion

For building our Prediction Model we have chosen Random Forest Classifier Algorithm and implemented it to receive a good accuracy score as it can be seen in the Results. And, to provide a proof of concept, we have taken into account two more algorithm. First one is Multinomial Logistic Regression, which is one of the mostly used algorithms in the previous prediction models and the other one is AdaBoost which is less prone to overfitting and is considered as a good fit in such prediction model but has not been used in previous cricket match prediction models. As it can be seen from the results, Multinomial Logistic Regression is not at all suited for our model that tries to predict the outcome of IPL matches and AdaBoost is a good fit but not the best. Therefore, as cricket is a very unpredictable game and its outcome depends on number of factors hence, every percent increase in the model's accuracy will be counted as very important and also, we aimed to build a model that outperforms the past iterations of such model. Therefore, we went ahead with the Random Forest Classifier Prediction Model.

8. Future Scope

There still are some scopes to improve upon. We are currently using the IPL game database to demonstrate the functionality of the future method we can create a separate database to predict world-class players.

We can extend the system to manage flexible match data where the manager can add data after each game to make the analysis more dynamic

References

1. S. Muthuswamy and S. S. Lam, "Bowler Performance Prediction for One-day International Cricket Using Neural Networks," in Industrial Engineering Research Conference, 2008.
2. G. D. I. Barr and B. S. Kantor, "A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket," Operational Research Society, vol. 55, no. 12, pp. 1266-1274, December 2004.
3. S.. R. Iyer and R. Sharda, "Prediction of athletes performance using neural networks: An application in cricket team selection," Expert Systems with Applications, vol. 36, pp. 5510-5522, April 2009.
4. I.P.Wickramasinghe, "Predicting the performance of batsmen in test cricket," Journal of Human Sport & Excercise, vol. 9, no. 4, pp. 744-751, May 2014.
5. Lemmer, H. H. (2008). Analysis of players' performances in the first cricket twenty 20 world cup series. South African Journal for Research in Sport, 30(2), pp.71-77.
6. Lewis, A. J. (2005). Towards fairer measures of player performance in one-day cricket. Journal of the Operational Research Society, 56, pp.804-815.
7. Saikia , H., Bhattacharjee, D., & Bhattacharjee, A. (2012). Is IPL Responsible for Cricketers Performance in Twenty20 World Cup? International Journal of Sports Science and Engineering, 6(2), pp.96-110.
8. Brooks, R. , Bussie`re, L. F., Jennions, M. D., & Hunt, J. (2003). Sinister strategies succeed at the cricket World Cup. Proceedings of the Royal Society.
9. "Free Download web scraping tool -web scraper | ParseHub," parsehub,[Online].Available: https://www.parsehub.com.
10. "Data extracted from the web," Import.io, [Online].Available : https://www.import.io. [11] Tim Kam Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 8, pp. 832-844, August 1998