Prediction of Winning Team using Machine Learning

Yash Ajgaonkar
Dept. of Information Technology Vidyavardhini's College of Engg. And Tech., University of Mumbai Mumbai, India

Kunal Bhoyar
Dept. of Information Technology Vidyavardhini's College of Engg. And Tech, University of Mumbai Mumbai, India

Prof. Anagha Patil
Dept. of Information Technology Vidyavardhini's College of Engg. And Tech, University of Mumbai Mumbai, India

Jenil Shah
Dept. of Information Technology Vidyavardhini's College of Engg. And Tech, University of Mumbai Mumbai, India

***Abstract*—Machine learning (ML) is one of the intelligent techniques. It has shown optimistic results in the field of classification and prediction accuracy. Sports Prediction is one of the expanding areas in good predictive accuracy as it involves huge money in betting. Since football is an interesting area of research, it is regarded as complex and dynamic when compared to other sports. It is the most widely played sport and currently being played in more than 190 countries. In this paper, prediction of winning team in English Premier League (EPL) is implemented using Machine Learning techniques. The objective is to predict the full time result (FTR) of the football match, which decides the winning team. We implement algorithms viz. Support Vector Machines, Random Forest and Naïve Bayes for training the data and the one that gives the maximum and best accuracy will be used for predicting the winning team. The dataset used were gathered from [6] for the past seasons.**

Keywords*—Machine learning, soccer game, winning prediction, accuracy, classification.

I. INTRODUCTION

Accurately predicting season results of soccer event is a billion dollar undertaking more commonly it is a multi-million dollar industry, with gamblers and die-hard fans eager for more accurate prediction and probabilities. Predicting the results of soccer matches has attracted so many people who love and have passion towards soccer. Right from the managerial football team which can analyze the performance of the squad and thus improve their game strategy to the fans that are eager to know the results about their favorite team/club. The analytics used behind the game and the strategy implemented plays as one of the deciding factor for a team to win.

Soccer prediction has become an intrigued research problem because there are many factors which can influence the outcome of matches such as home/away goals , rankings, match types(day/night) , team work , player skills, home team advantage , weather. The main motive behind this project is giving an accurate dataset for soccer matches and predicts the winners in upcoming matches and thus yielding efficient results.

In this paper, we propose a model of soccer match prediction on the basis of FTR which is Full Time Result which would be our class label i.e. Home, Away or Draw.

II. RELATED WORK

A considerable amount of work and surveys are carried out in sports prediction especially winning team prediction domain.

The initial paper i.e. [1] deals with the prediction of winning team in case of NBA matches. The algorithms used were Linear Regression, Maximum Likelihood Classifier and the Multilayer Perceptron (Back Propagation) approach. The results by using these algorithms were : Naive Majority Vote Classifier: 63.98%, Linear Regression (67.89%) and Multilayer Perceptron Method: 68.44%. Here they considered the parameters: win or loss percentage of home team games, win-loss percentage as visitor or home as the respective situation of the teams and the difference in point of home team. Expectation of results was to be highly correlated. The highest accuracy was given by Linear Regression with a prediction rate of winners and losers as 68%. The further improvements in this model would be a comparatively large dataset, feature classification.

In [2] prediction of winning team in an NBA match is put forth. A certain amount of algorithms were used on which model was trained .These algorithms were implemented using the sklearn library of python. The dataset were considered were from past 32 seasons from NBA api and had over 50 features, hence feature classification was one of the most crucial steps carried out for yielding better results. Due to this, only those attributes that were required were taken into consideration i.e. selection of only those features that impart information about the output variable independently or conditionally on other relevant variables [2]. The dataset that was chosen was comparatively more than in [1]. The algorithms implemented were Extra Trees, Gradient Boosting, k-nearest neighbors, Logistic Regression, SVM, Neural Networks (MLP algorithm) and Non-Linear SVM. The highest accuracy was given by SVM i.e. 71% at a certain extent. Some of the given conclusion was: Gradient Boost algorithm had a better classification as compared to Random Forest or SVM although SVM had better accuracy. Logistic Regression has advantages over Linear Regression as it had a good starting estimate. K-nearest neighbors has

the ability to reduce effect of noise during final prediction. Future works proposed were like incorporation of live data and then testing the model also each player characteristics can be considered for the same.

In the paper [3], the authors have described their approaches which are able predict the match results in the 2015/16 English Premier League (EPL) with an accuracy of about 67%. The algorithms used were Support Vector Machine (SVM), libsvm and the one which gives the best accuracy is chosen. They trained first 85% of 2015/16 English Premier League match results and kept the rest 15% for testing purposes. But due to this the last games of a season became much more predictable than other games. So they got poor accuracy in testing data large data. Plus the SVM algorithm fails to give good accuracy with huge data. According to the authors, there is scope for further improvement, like more data of the previous season and statistics could be taken into account.

In [4], it deals with a model to predict the results of soccer matches in the Barclays' English Premier League. In this the data-set contains around 65 attributes every season like the away team goals, venue, scores, and home team. They trained the final data-set on three ML classifiers viz. Support Vector Machines (SVM) , XGBoost and Logistic Regression to further improve this research, they could bring in sentiment analysis, features such as individual player metrics , the posts from fans on social media, etc. to increase the accuracy of the model in their future work.

In [5], a model of logistic regression was proposed by the authors for estimating the 2015/16 Barclays Premier League match results having the accuracy of around 69.5%. The considered datasets were from Barclays Premier League and the number of variables considered was just four namely: Away Offense, Home Defense, Home Offense, and Away Defense. Their system predicts who will win the match and the details of it like the odds/probability, coefficients of regression. In spite of the number of variables considered, the system gives strong prediction accuracy.

The comparative survey done for the referred papers:

Table 1: Comparison of papers

III. PROPOSED MODEL

In this paper, we propose a model to predict the outcome of football matches in the English Premier League. We train the dataset of past seasons on various machine learning classifiers. Comparisons amongst the algorithms would be made and the one that turns out to be the most accurate i.e. having the better prediction accuracy will be considered. Then, optimization can be made on that classifier to further enhance the model accuracy in making predictions. The label that would be considered would be Home Win (H), Away Win (A), and Draw (D).

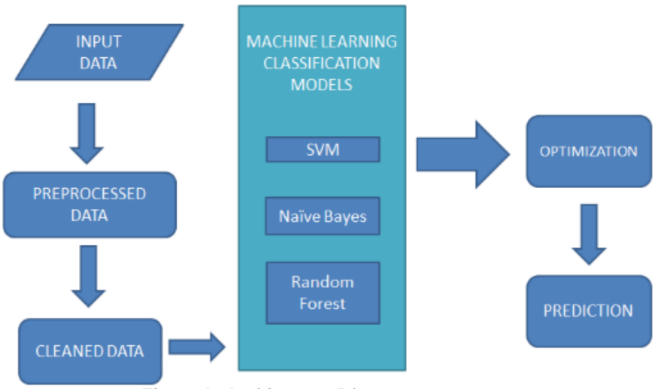

Figure 1: Architecture Diagram

A. Dataset Description

Prediction would be done on the basis of data from past games for recent seasons. We have obtained the data set from

[6] that has tremendous amount of data right from the old games to the ones that are being played. There are about 65 attributes per season like the Home team, Away team, scores, venue to be named few. After having filtered these attributes we get around 8-10 attributes that are actually going to predict the results. The dataset size is 3000.

Table 2: Input Data

	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
0	Man United	Leicester	2	1	H	1	0	H	8	13	6	4	11	8	2	5	2	1	0	0
1	Bournemouth	Cardiff	2	0	H	1	0	H	12	10	4	1	11	9	7	4	1	1	0	0
2	Fulham	Crystal Palace	0	2	A	0	1	A	15	10	6	9	9	11	5	5	1	2	0	0
3	Huddersfield	Chelsea	0	3	A	0	2	A	6	13	1	4	9	8	2	5	2	1	0	0
4	Newcastle	Tottenham	1	2	A	1	2	A	15	15	2	5	11	12	3	5	2	2	0	0
5	Watford	Brighton	2	0	H	1	0	H	19	6	5	0	10	16	8	2	2	2	0	0
6	Wolves	Everton	2	2	D	1	1	D	11	6	4	5	8	7	3	6	0	1	0	1
7	Arsenal	Man City	0	2	A	0	1	A	9	17	3	8	11	14	2	9	2	2	0	0
8	Liverpool	West Ham	4	0	H	2	0	H	18	5	8	2	14	9	5	4	1	2	0	0
9	Southampton	Burnley	0	0	D	0	0	D	18	16	3	6	10	9	8	5	0	1	0	0
10	Cardiff	Newcastle	0	0	D	0	0	D	12	12	1	6	14	16	5	5	2	2	0	1
11	Chelsea	Arsenal	3	2	H	2	2	D	24	15	11	6	12	9	5	1	0	2	0	0
12	Everton	Southampton	2	1	H	2	0	H	13	15	7	4	8	20	2	5	0	5	0	0
13	Leicester	Wolves	2	0	H	2	0	H	6	11	2	3	10	8	1	9	2	1	1	0
14	Tottenham	Fulham	3	1	H	1	0	H	25	10	11	3	9	5	5	2	0	0	0	0

SR NO	ALGORITHMS	ACCURACY	MERITS	DEMERITS
1	Linear Regression, Maximum Likelihood, Back Propogation	67.89 %	Linear Reg is easy to Implement and training can be done in short duration.	Limited Dataset. Feature Classification wasn't done.
2	Support Vector Machine KNN, Gradient Boosting Neural Network	71%	Dataset had enough instances Feature Classification	Lack of complexity of algo used.
3	Support Vector Machine Libsvm, NN(sklearn implementation)	67%	Risk of over-fitting is less in SVM.	poor- accuracy in testing large data.
4	Support Vector Machine XGBoost, Logistic Regression	68.5%	Xgboost - scalable and accurate	Slower training for Xgboost.
5	Logistic Regression	69%	Easy to implement	Complex algo, Overfitting is more.

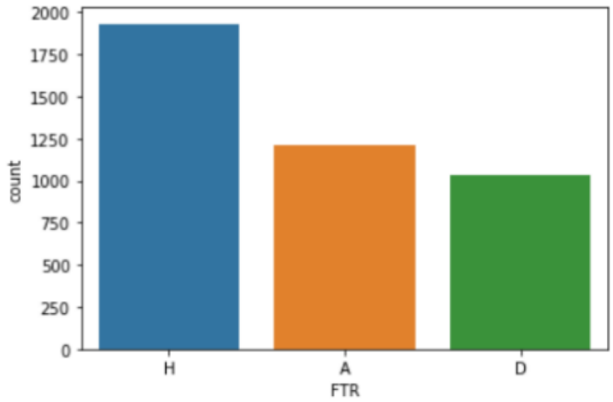

Figure 2: Count Plot

Countplot displays the observation counts in each categorical bins using bars. As our FTR is the dependent variable or the outcome, the countplot has given the count for each of the values namely H – Home, A-Away and D-Draw.

B. Data Preprocessing

The dataset that is obtained from consists of several attributes of each season. Some of those attributes are less important or rather irrelevant for predicting the result .So data cleaning is carried out for retaining only those attributes that are relevant for prediction. Following is the boxplot for the feature scaled attributes.

Table 3: Preprocessed Data

	HomeTeam	AwayTeam	FTR	HST	AST	HC	AC	HAS	HDS	AAS	ADS
0	Man United	Leicester	H	6	4	2	5	1.067323	0.835341	0.939759	0.896552
1	Bournemouth	Cardiff	H	4	1	7	4	1.067323	0.887550	0.469880	1.024631
2	Fulham	Crystal Palace	A	6	9	5	5	0.725780	1.357430	0.939759	0.939245
3	Huddersfield	Chelsea	A	1	4	2	5	0.256158	1.148594	0.991968	0.853859
4	Newcastle	Tottenham	A	2	5	3	5	0.512315	0.939759	1.566265	0.555008
5	Watford	Brighton	H	5	0	8	2	0.725780	0.939759	0.626506	1.024631
6	Wolves	Everton	D	4	5	3	6	0.853859	0.991968	0.783133	0.768473
7	Arsenal	Man City	A	3	8	2	9	1.195402	0.574297	1.305221	0.384236
8	Liverpool	West Ham	H	8	2	5	4	1.494253	0.365462	0.730924	0.853859
9	Southampton	Burnley	D	3	6	8	5	0.640394	1.096386	0.730924	1.067323
10	Cardiff	Newcastle	D	1	6	5	5	0.640394	1.200803	0.522088	0.683087
11	Chelsea	Arsenal	H	11	6	5	1	1.110016	0.469880	1.305221	1.110016
12	Everton	Southampton	H	7	4	2	5	0.896552	1.096386	0.678715	0.981938
13	Leicester	Wolves	H	2	3	1	9	0.555008	0.678715	0.730924	0.597701
14	Tottenham	Fulham	H	11	3	5	2	1.024631	0.626506	0.417671	1.366174

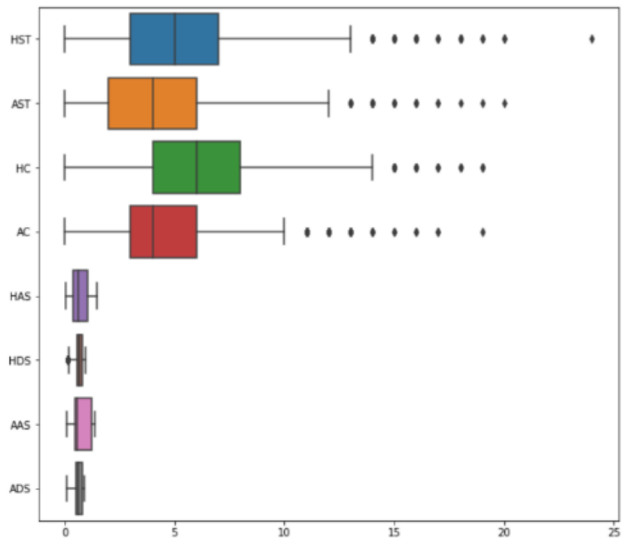

Figure 3: Box Plot

A boxplot is a graph that helps us to understand how the values in the data are spread out. It is a standardized way to display data distribution. We obtain a five number summary viz. "minimum", first quartile (Q1), median, third quartile (Q3) and "maximum" [10].

C. Data Splitting

We have split the data into two parts – training and testing in order to fit the model and get the desired outcome. They have been split on the percentages 80 and 20.

D. Modelling

In this system, we have implemented the following three algorithms: Naïve Bayes, Random Forest, and Support Vector Machine (SVM).

Naïve Bayes Classifier:

A Naïve Bayes Classifier is a probabilistic model which is most commonly used for classification task. It is based on the Bayes Theorem. The outcome can be determined given the predictors or the independent variables. For instance if there are two events A and B, we can find the probability of A happening given that B has already occurred wherein B is the evidence and A is the hypothesis. The assumptions made here is that a particular feature has no effect on the other which is called to be 'naïve'. In the dataset chosen for instance the outcome is FTR and the independent or predictor variables are the parameters [7].

Random Forest:

Random Forest comprises of a large number of individual decision trees that operate as an ensemble. An ensemble method can be defined as the usage of multiple learning algorithms that leads to a better predictive performance that any algorithm would do as alone. Each individual tree in random forest splits out a class prediction and the class which has the maximum votes becomes the model prediction [8].

Support Vector Machine (SVM):

Support Vector Machines are Machine Learning models which are useful for regression analysis and classification tasks. It falls under the supervised learning category of Machine Learning. These are widely used in classification tasks. Support Vector Machines are based on the idea of finding the best hyperplane that divides the dataset into two parts [9].

UML Diagram for our model is:

Figure 4: UML Diagram

IV. EXPERIMENT

Experiment is conducted for getting the best accuracy. In this paper, we are using the data from past recent seasons of the English Premier League. It is done to determine whether amount of training data has any impact on prediction accuracy. The data has been split into training and testing datasets which can be seen in table 4 and table 5.

Table 4: Training Dataset

Index	HST	AST	HC	AC	HAS	HDS	AAS	ADS
816	8	1	8	2	1.22692	0.62185	0.760455	0.88
2303	4	4	10	5	0.121282	0.198542	0.505721	0.487949
2405	2	2	5	4	0.296154	0.44953	1.25119	0.634615
1498	4	3	8	6	1.06051	0.670549	1.28491	0.753077
2290	2	7	3	4	0.634615	0.89906	1.25868	0.626154
2639	4	2	10	6	0.761538	0.96649	1.2437	0.789744
2845	8	3	6	5	0.403333	0.400831	0.389593	0.609231
2416	5	5	8	5	1.28897	0.651819	0.325909	0.414615
1872	3	1	3	1	1.23821	0.644326	1.25119	0.634615
2027	7	4	3	2	0.558462	0.595627	0.760455	0.88
2248	6	1	9	5	1.06051	0.670549	0.325909	0.414615
1175	3	2	11	7	1.46667	0.629342	0.902806	0.784103
1915	4	2	2	6	0.301795	0.543182	0.576897	0.767179

The training dataset includes 80 percentage of the whole dataset and has eight attributes which are: Home Attacking Strength (HAS), Home Defensive Strength (HDS), Away Attacking Strength (AAS), Away Defensive Strength (ADS), Home and Away Corners, Home and Away Shots on Target. Here, in this training dataset the shots and corners are included. From this we come to know which team has better attacking strength.

Table 5: Testing Dataset

Index	HST	AST	HC	AC	HAS	HDS	AAS	ADS
2107	7	6	6	7	1.00974	0.842869	0.318417	0.499231
810	6	5	3	4	0.504872	0.741725	1.25119	0.634615
2315	9	2	11	4	1.28897	0.651819	0.603119	0.82359
2235	1	7	4	7	0.0733333	0.209781	1.2437	0.789744
1058	8	2	0	7	0.504872	0.636834	0.576897	0.767179
337	6	11	7	6	0.298974	0.531944	0.902806	0.784103
1533	7	2	6	5	1.28897	0.651819	0.239749	0.437179
979	4	4	2	4	0.366667	0.55442	0.584389	0.744615
684	5	5	5	7	1.00974	0.842869	1.25868	0.626154
620	13	13	3	3	0.55	0.689279	0.0749217	0.112821
1395	11	4	6	3	0.558462	0.595627	0.584389	0.744615
1251	6	3	4	7	0.761538	0.96649	0.27721	0.502051

The testing dataset includes 20 percentage of the whole dataset and has eight attributes which are: Home Attacking Strength (HAS), Home Defensive Strength (HDS), Away Attacking Strength (AAS), and Away Defensive Strength (ADS), Home Shots on Target (HST), Away Shots Total (AST), Home Corners (HC), and Away Corners (AC).

Visualization:

A heatmap is one of the most popular way of visualizing the dataset [12].It helps us to understand which attribute is the most co-related. It uses a system of color-coding to represent different values. The heatmap of the feature scaled attributes in our dataset gives the co-relation between the attributes. Via color-coding we come to know that the most co-related attributes to HST are ADS and AAS respectively followed by HAS and HDS.

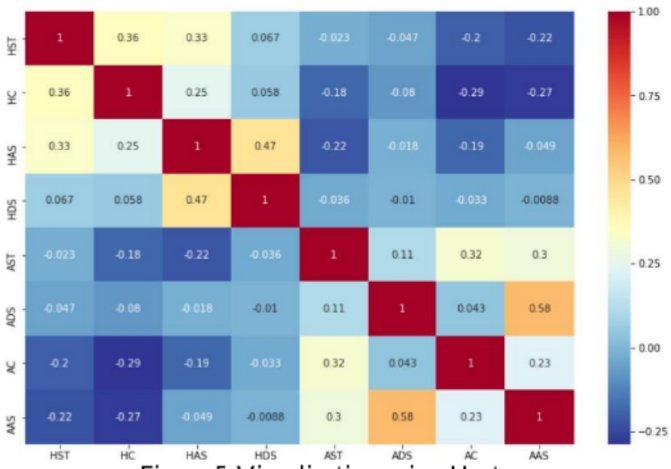

Figure 5:Visualization using Heatmap

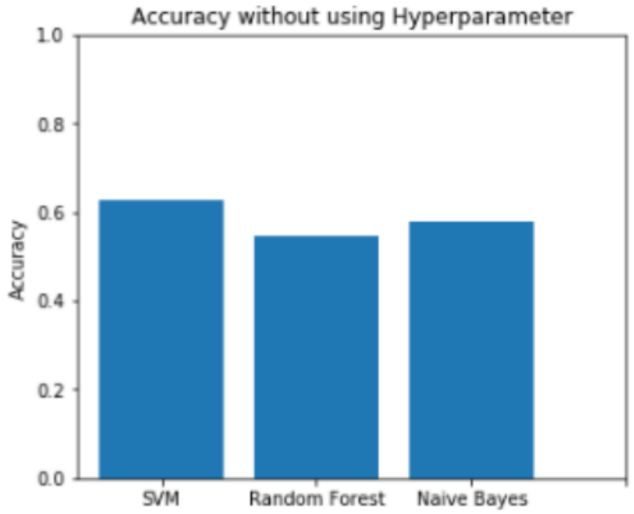

Figure 6: Accuracy Comparison (Without using Hyperparameter)

The accuracy obtained for Support Vector Machine is 63% which was greater than the Random Forest and Naïve Bayes that were obtained as 55% and 57% respectively. From the results obtained,we can say that SVM has better accuracy.

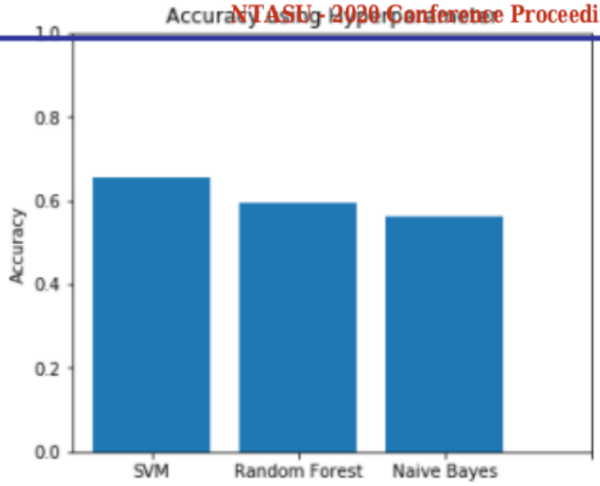

Figure 7: Accuracy Comparison (Using Hyperparameter)

The accuracy obtained for Support Vector Machine is 67% which is comparatively greater than Random Forest and Naïve Bayes obtained as 60% and 56% respectively. We used GridSearchCV function [13] from sklearn toolkit to build the models in order to tune the parameters and which led to increase in accuracy. From this we can say that SVM has better accuracy.

V. RESULT ANALYSIS

Our analysis is to predict the outcome of the match and when incorporated with training data, SVM algorithm predicted the highest accuracy. So, we have normalized the dataset during preprocessing stage. Normalization is done in order to bring the features of the training dataset to the same scale. The goal of normalization is to alter values of numeric columns in the dataset to a common scale, without distorting the differences in the ranges of values [11]. Also, hyperparameter tuning was carried for SVM as it had shown maximum accuracy prior to tuning. The hyperparameter tuning eventually lead to increase in accuracy obtained for the dataset for SVM which is 67%.

VI. CONCLUSION

In this paper, we built a classification model to predict the outcome of English Premier League (EPL) matches. From visualizing we find that the significant variables are Attacking Strength and Defensive Strength of Home and Away team, but the prediction cannot be done by including only these four attributes. It was learned that data from recent seasons is more relevant than the data from past seasons. Additionally, adding more featured attributes like corners and shots on target bring more value to the predicted accuracy.

VII. FUTURE WORK:

The currently devised model is purely based on past statistical results which do help to predict the winning team based on the chosen parameters. In order to increase the accuracy of the model, sentiment analysis like trending twitter hashtags on a regular basis like during or before the

match can be studied and worked upon. Also managers have an important role to decide the strategies and tactics, hence manager's past record can also be considered as a criteria as to whether the team can perform better in its next game.

REFERENCES

[1] Renator Amorim Torres "Prediction of nba games based on machine learning methods". University of Wisconsin, Madison, 2013.
[2] Weronika Swiechowicz , Jacob Perricone, Ian Shaw "Sports Data Mining: Predicting Results for Professional Basketball Games", Stanford University,CA,CS229 Autumn 2016.
[3] Steffen Smolka ,"Beating the bookies :Predicting the outcome of soccer games", Stanford University,CA,CS229 Autumn 2017.
[4] 1Anand Ganesan, 2Harini M , 1Student, 2Assistant Professor, "ENGLISH FOOTBALL PREDICTION USING MACHINE LEARNING CLASSIFIERS ", International Journal of Pure and Applied Mathematics, Volume 118 No. 22 2018, 533-536,SRM UNIVERSITY 2018
[5] Darwin Prasetio , Dra Harlili Predicting football match results with logistic regression, International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) 2016.
[6] football-data . (2019). data | football-data, [online] Available at: http://www.football-data.co.uk/ [Accessed on 7 Aug. 2019].
[7] Rahman, M. M., Faruque Shamim, M. O., & Ismail, S. An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach. International Conference on Innovations in Science, Engineering and Technology (ICISET) 2018.
[8] Oughali, M. S., Bahloul, M., & El Rahman, S. A. Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models. International Conference on Computer and InformationSciences(ICCIS) 2019.
[9] Anik, A. I., Yeaser, S., Hossain, A. G. M. I., & Chakrabarty, A. Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms. 4th International Conference on Electrical Engineering and Information & Communication Technology(iCEEiCT) 2018.
[10] Thirumalai, C., Kanimozhi, R., & Vaishnavi, B. Data analysis using box plot on electricity consumption. International Conference of Electronics, Communication and Aerospace Technology(ICECA) 2017.
[11] statisticshowto.datasciencecentral. (2015). Normalized | statisticshowto.datasciencecentral. [online] Available at: https://www.statisticshowto.datasciencecentral.com/normalized/ [Accessed on 10 Jan. 2020].
[12] likegeeks. (2019). Seaborn-heatmap-tutorial | likegeeks. [online] Available at: https://likegeeks.com/seaborn-heatmap-tutorial/ [Accessed on 15 Jan 2020].
[13] towardsdatascience. (2019). Hyperparameter-Tuning-c5619e7e6624 | towardsdatascience. [online] Available at: https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624 [Accessed on 13 Jan 2020].