SPORTS PREDICTOR USING MACHINE LEARNING ALGORITHM

Abstract

Machine learning is a branch of Artificial Intelligence and is more often used to predict outcomes. A lucrative area like sport aids from ML as it forecasts players' performance. Sports coaches, fanatics, and enthusiasts get an insight into improving a team's performance. In-game various factors, such as past fixtures, player statistics, and opposition information-related data, are used to form the sports prediction model. This work focuses on the use of SVM to predict a team and a player's performance, namely in hockey and football. This work has achieved R2 0.8780, MSE 0.0028, MSE 0.0530 for football and R2 0.8189, MSE 0.0021, RMSE 0.0464 for hockey using SVM algorithm.

Keywords: Sports, Win Predictor, Football, Hockey, Machine learning

Authors

Ignatius Almeida
Department of Information Technology
St. Francis Institute of Technology
Mumbai, India
ignatiusalmeida1999@student.sfit.ac.in

Gladstone D'souza
Department of Information Technology
St. Francis Institute of Technology
Mumbai, India
glady18dsouza@student.sfit.ac.in

Amit Kumar Yadav
Department of Information Technology
St. Francis Institute of Technology
Mumbai, India
goblinaryan0000@student.sfit.ac.in

Siddhart Sandu
Department of Information Technology
St. Francis Institute of Technology
Mumbai, India
siddhartsandu@student.sfit.ac.in

Aruna Pavate
Department of Information Technology
St. Francis Institute of Technology
Mumbai, India
arunapavate@sfit.ac.in

I. INTRODUCTION

Various sports matches are regulated constantly as the growth of this sector has excelled a lot over the years. Nowadays, technology has contributed enormously to the change in the gameplay of these sports. Sports prediction makes a difference for sports enthusiasts. It gives an idea of which factors have contributed to a certain way the team or player has performed [1] [2]. For sports fans, it gives them a platform to put their skills to the test and judge their sports knowledge.

In this work, sports prediction is viewed as a classification problem, with one outcome (win, lose, or draw) to predict, while there are some researches where the outcome is a numeric value [1]. Vast features can be collected such as the past performance of the teams and data on players, which help to understand the chances of winning or losing forthcoming games.

Machine Learning is to make computers behave like humans and improve their learning over time in an autonomous fashion. It automates analytical model building. It's a subset of AI that helps models with the flexibility to learn automatically and improvise from experiences without being directly programmed. Machine learning algorithms overcome the disadvantages of statistical models by creating data-driven predictions or decisions using a model from sample input. It has strong bonds to mathematical optimization [3] [4]. Plenty of companies are investing heavily in machine learning to predict sports results.

Support Vector Machine (SVM) is a commonly used machine learning algorithm. It's a supervised algorithm that is used for classification and regression problems. Some of its applications are Cusine recommendation systems and recommendation system [5]-[7], The SVM algorithm target is to plot the best line or decision function using a subset of training points. SVM chooses extreme points that help in creating the hyperplane. These extreme points are called support vectors. Hence the algorithm is termed a Support Vector Machine. Objectives of the work:

1. The research focuses on use of machine learning algorithm such as Support Vector Machine. It is used for training both supervised and unsupervised machine learning models. This work concentrated on supervised machine learning model.
2. We are focusing on eliminating the limitations by adding more features and getting a more accurate and realistic outcome.
3. This work offers the users of freedom to choose and work on their favorite sport with any formation of their choice.

The rest of the research layout is as follows: Section 2 introduces the literature survey followed by section 3 the proposed system and implementation, after which section 4 discusses the results and analysis and section 5 is the conclusion.

II. RELATED WORK

This section comprises some of the literature work for Sports Prediction System.

McCabe & Trevathan et al. [8] used a three-layer multi-layer perception (MLP) with nineteen input units, ten hidden units, and one output unit. The feature set calculated values

for every team for every spherical competition. The reason to use the neural network (NNs) is the model's ability to find the link between inputs and outputs. Victimizing this deep learning technique, they achieved a median accuracy of 62.6% across the four different leagues they worked.

Sushant & Rana et al. [9] used different classifiers for result prediction, compared the result, and selected the best classifier for the correct match prediction. The info's first classification was enforced using SVM, XGBoost, and supply Regression. In the end, results conjointly extended the performance and preciseness processes by mistreating some advanced tools. They achieved a mean accuracy of 56.67% across the three machine learning algorithms.

Bosch & Bhulai et al. [10] showed how Deep Learning strategies might outdo Machine Learning strategies in predicting NFL-game winners. SVM, supplying Regression and Random Forest, was the machine learning technique utilized by the authors. Artificial Neural Network (ANN) & repeated Neural Network (RNN) models have an associate degree optimum range of 3 layers. Within the given models, machine learning techniques achieved a mean accuracy of 62.95%, and deep learning techniques had a mean accuracy of 63.36%.

Singla et al. [11] used numerous parameters while predicting the sports associated with offensive and defensive properties of the team, such as goals scored, conceded, corner kicks, red, yellow cards, etc., for each season. Models like SVM, Multinomial Naïve Thomas Bayes & supplying Regression were used for classification and compared before and when applying standardization. The category employing a machine learning rule displayed the foremost correct outcome with and while not standardization. From the results achieved, the typical accuracy obtained before standardization was 55.8%; however, after standardization, the accuracy leaped down to 54.7%.

Vistro et al. [12] worked on different machine learning algorithms to predict the IPL match winner before the match even began. The winner of the IPL was foreseen by coaching machine learning models such as Random Forest, SVM, Naive Bayes, supplying Regression, and call Tree on the chosen features. Three classifiers, the Decision tree classifier, random forest classifier & XGBoost classifier, were accustomed to acquiring the result. Victimization of these machine learning algorithms achieved a median accuracy of 89.95%.

Many researchers worked on sports predication system using machine learning algorithms, but each one having their own limitations. Here in this work we have tried to attempt most commonly used machine learning algorithm for sports prediction system.

III. PROPOSED SYSTEM

The proposed work collects the dataset from websites "fbref.com" and "hockey-reference.com." The Premier League has been considered played in England for football. There are 20 teams where each team plays the other twice, so one team plays 38 games per season. For this project, data is collected from 498 players who have played at least one or more games in the Premier League from 2020-to 2021. The attributes of the dataset are mentioned in Table 1 with the description. For Hockey, this work has considered the NHL, within which one regular-season plays 82 games per team. It's divided into two conferences, i.e., Western and Eastern conferences, each conference has two divisions, and every division

has eight teams. The data is considered from 2016-17 to 2021-22. The data attributes are mentioned in Table 2 with the description.

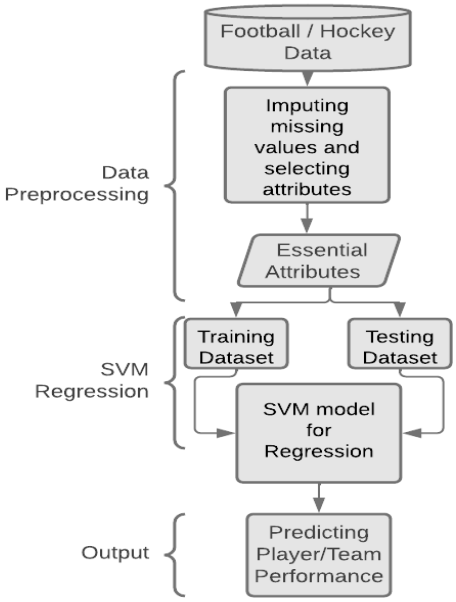

Figure 1: Proposed System for Sports Prediction

Figure 1 shows the proposed model. Support vector regression method works on the principle of SVM with few minor differences. Given data points, it tries to find the curve. As it's a regression algorithm, it uses the curve to find the match between the vector and the position of the curve. It helps to determine the closest match between the data points and the function used to represent them. The regression algorithm known as Support Vector Regression, as its name suggests, enables both linear and non-linear regressions. The Support Vector Machine is the basis for how this approach operates. In contrast to SVM, which is used to predict discrete categorical labels, SVR is a regression that is used to forecast continuous ordered parameters. This is one manner in which SVR and SVM differ from one another.

The goal of basic regression is to reduce error rates, whereas the goal of SVR is to fit the error inside a predetermined threshold. This means that the goal of SVR is to approximate the best value within a predetermined margin known as the "-tube." There are some basic terminologies required that are

1. **Kernel:** Without increasing the processing overhead, a kernel enables us to locate a hyperplane in the higher dimensional space. Usually, as the dimension of the data grows, the computing cost grows as well. When we can't go in a certain dimension because there isn't a dividing hyperplane there, we must move in a higher dimension instead.

2. In SVM, a hyperplane essentially acts as a boundary between two data classes. However, this line will be utilized in Support Vector Regression to forecast the continuous output.

3. **Decision Boundary:** To simplify, think of a decision boundary as a line where the positive examples are on one side and the negative examples are on the other. The

instances can be classified as either positive or negative along this exact line. Support Vector Regression will also use this same SVM concept.

The three kernels that SVM most frequently use are: Linear kernel: Given two specified observations, the dot product, kernel of a polynomial Curved lines are thus permitted in the input space and the third RBF: Radial Basis Function Complex areas are produced inside the feature space. This work concentrated on use of RBF kernel function. The procedure for SVR is discussed below:

- Gathering the training materials
- Choosing a kernel, its parameters, and any necessary regularization.
- Correlation matrix creation
- Get the contraction coefficients
- Utilize the coefficients to create an estimator.

First, read the dataset. We have to take care that the features of the training dataset ought to fulfil the domain that we tend to expect as the SVR will solely build on the knowledge within the training dataset. These datasets which are used are real-world datasets due to which the features may vary. Feature scaling helps to normalize this variation in the data which helps the outcome.

To fit SVR to the dataset. We have to assign a kernel and set all the other parameters of the SVR algorithm if required. We have used Gaussian Radial Basis Function (RBF) as its for non-linear data and helps to make proper separation when there is no prior knowledge of data as shown in equation 1.

$$f(x, xj) = \exp(-\Gamma * \|x - xj\|^{\wedge 2}) \quad \text{---------- (1)}$$

Once the steps mentioned above are completed then train the model and test it by predicting results. To evaluate the performance of the model different metrics used as MSE, RMSE, and R2 score. The data is visualized to see the best fit and correlation of parameters with the output.

Table 1: Football Data Set Attribute List

xG	Expected Goals
npxG	Non-Penalty Expected Goals
xA	Expected assists
npxG+xA	Non-Penalty Expected Goals + Expected assists
Per_xG	Expected Goals per 90
Per_xA	Expected assists per 90
Per_xG+xA	Expected Goals + Expected assists
Per npxG	Non-Penalty Expected Goals per 90
Per_npxG+xA	Non-Penalty Expected Goals + Expected assists per 90
SoT%	Shot on Target Percentage
Sh/90	Shots per 90

SoT/90	Shot on Target per 90
G/Sh	Goals per shot
G/SoT	Goals per shot on target
Dist	Average distance(yards) from the goal of all shots taken
npxG/Sh	Non-Penalty Expected Goals per Shot
DribSucc%	Percentage of Dribbles Completed successfully
PassRec%	Percentage of Passes Received Successfully
PassCmp%	Percentage of Passes Completed Successfully
SCA90	Shot Creating Actions per 90
GCA90	Goal Creating Actions per 90
SuccTkl%	Successful Tackles Percentage
Press%	Successful Pressures Percentage
Per_Gls	Goals per Game

Table 2: Hockey Data Set Attribute List

SRS	Simple Rating System
SOS	Strength of Schedule
GF/G	Goals for per game
GA/G	Goals against per game
PP%	Power Play Percentage
PK%	Penalty Kill Percentage
PIM/G	Penalty in minutes per game
oPIM/G	Opponent penalty in minutes per game
S%	Shooting Percentage
SV%	Save Percentage
PDO	Shooting plus save percentage
CF%	Corsi for percentage
FF%	Fenwick for percentage
xGF	Expected Goals For
xGA	Expected Goals Against
SCF%	Scoring Chances For Percent
HDF%	High Danger Scoring Chances For Percent
HDC%	High Danger Scoring Chances For Converted Percent
HDCO%	High Danger Scoring Chances Against Percent
PTS%	Points Percentage

IV. RESULT AND ANALYSIS

In order to determine the characteristics of the SVR model that best fits the data, we fi rst displayed it. So, using these two variables, we have produced a scatter plot. The data visualization for the best fit is represented from figure 2 to figure 9. The plot figure 7,8 and 9 shows a linear relationship between the two variables due to this the linear SVR to model applied to this data.

Figure 2: Parameters Ranked Based on Influence on Goals Per Game. (Football Data Set)

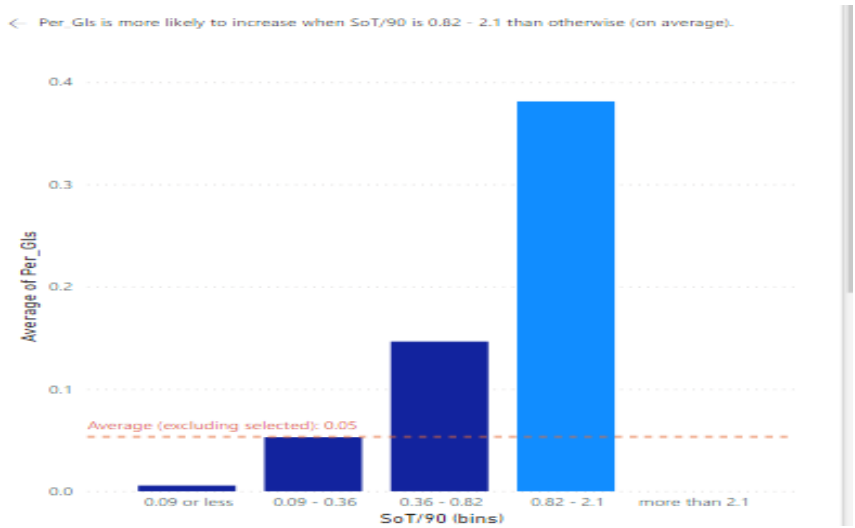

Figure 3: Goals per game vs Shots on target per 90. (Football Data set)

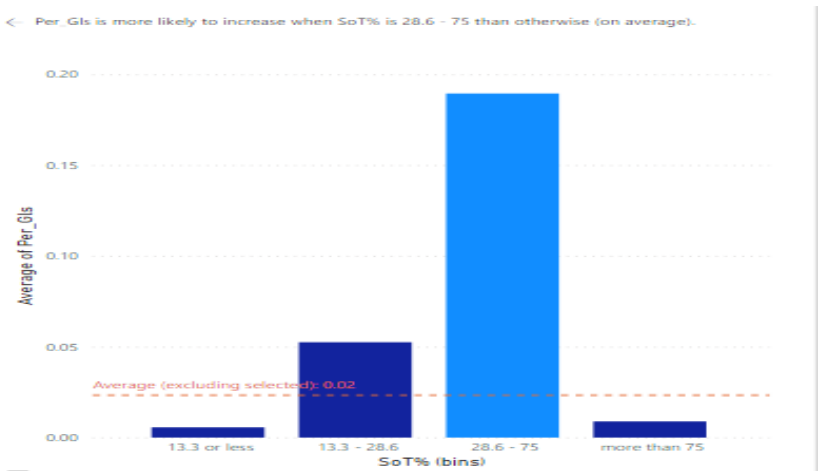

Figure 4: Goals per game vs Shots on target percent. (Football Data set)

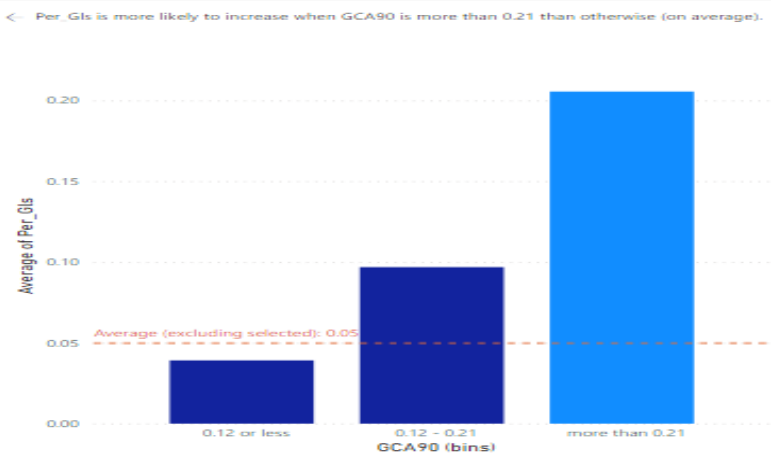

Figure 5: Goals Per Game Vs Goal Creating Actions Per90. (Football Data set)

Figure 6: Parameters Ranked Based on Influence on Points Percent. (Hockey Data Set)

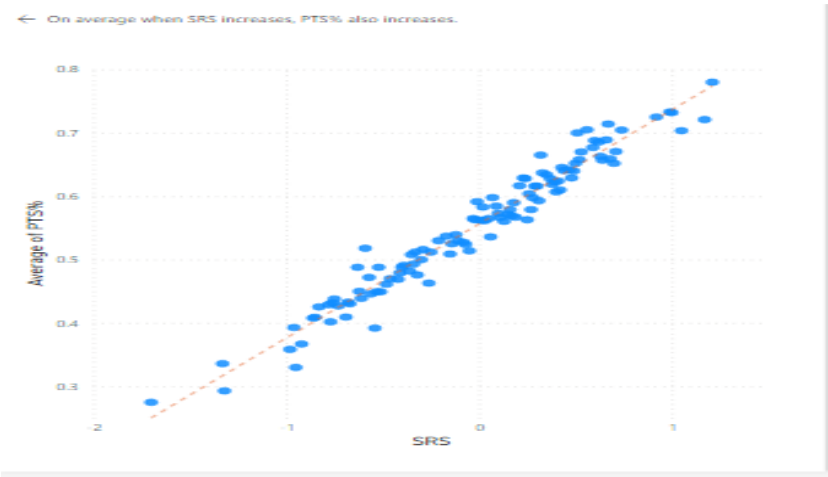

Figure 7: Points percent vs SRS. (Hockey Data set)

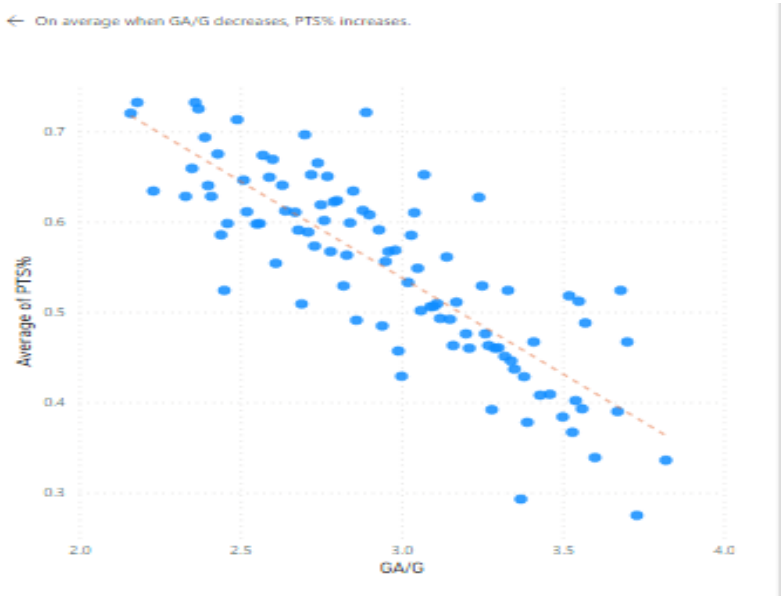

Figure 8: Points percent vs GA/G. (Hockey Data set)

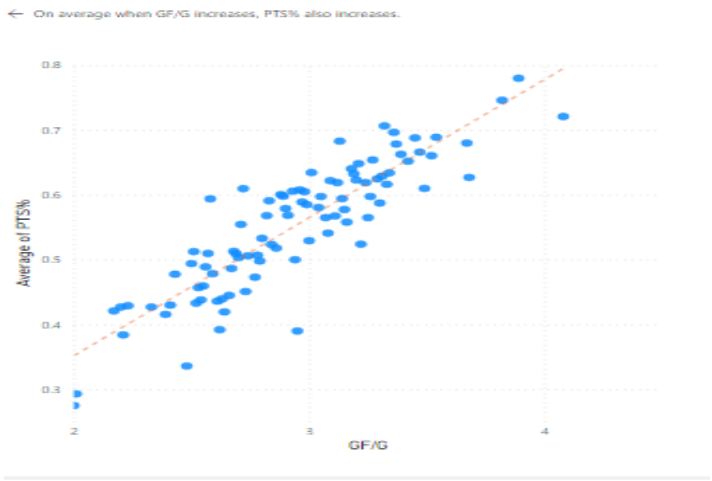

Figure 9: Points percent vs GF/G. (Hockey Data set)

Table 3 represents the results using SVR model applied on football and hockey datasets. The measuring parameters are: R2 score, Mean Squared Error and Root Mean Squared Error.

The R2 score ranges from 0% to 100%. Although not the same, it is closely connected to the MSE. The percentage of the dependent variable's variance that can be predicted from the independent variable is the R2 score.

The average of the squares of the errors is called mean square error (MSE). The uncer tainty increases with increasing number of errors. The discrepancy between the observed valu es y1, y2, y3,...yn and the projected values is what is meant by error in this context.

Table 3: Result Analysis using R2, MSE and RMSE

Parameters			
Football	R2 0.8780	MSE 0.0028	RMSE 0.0530
Hockey	R2 0.8189	MSE 0.0021	RMSE 0.0464

	Predicted value	Real Value
0	-0.016980	0.00
1	0.080166	0.07
2	0.073966	0.06
3	-0.020187	0.00
4	0.041209	0.04
...	...	...
95	0.092779	0.10
96	0.006908	0.00
97	0.016986	0.00
98	0.167949	0.20
99	0.154086	0.17

Figure 10: Football dataset test results.

	Predicted value	Real Value
0	0.589683	0.598
1	0.560197	0.563
2	0.623376	0.665
3	0.686474	0.714
4	0.582077	0.610
5	0.677320	0.657
6	0.553229	0.541
7	0.583043	0.567
8	0.597062	0.579
9	0.577875	0.664
10	0.650724	0.680
11	0.629284	0.628
12	0.458643	0.427
13	0.504308	0.530
14	0.456232	0.482
15	0.631785	0.634
16	0.659761	0.688
17	0.417290	0.275

Figure 11: Hockey dataset test result

Figure 10 and 11 shows the predicted results. For MSE, there is no ideal value. In other words, the lower the value, the better, and 0 denotes a perfect model. Since there is no right or wrong response, the MSE's primary benefit is in helping us choose one prediction model over another. In a similar vein, there is no ideal value for R2, either. 100% denotes an ideal correlation.

V. CONCLUSION

This work provides an in-depth review of sports prediction using machine learning techniques to predict results in team sports. We have analysed various characteristics of past studies, including the researchers focused on the different machine learning algorithms. The researchers also used many features for an experiment. However, it lacked accuracy and

realism due to the limitations of the elements in the data set. In addition to the limitations we figured out in the research, our research focused on eliminating the restrictions by adding more features and getting a more accurate and realistic outcome. Our analysis also allows the user to choose and work on their favorite sport with any formation.

REFERENCES

[1] Rory P. Bunker, Fadi Thabtah, "A machine learning framework for sport result prediction,Applied Computing and Informatics", Volume 15, Issue 1, 2019, Pages 27-33, ISSN 2210-8327

[2] Zhongbo Bai, Xiaomei Bai, "Sports Big Data: Management, Analysis, Applications, and Challenges", Complexity, vol. 2021, Article ID 6676297, 11 pages, 2021. https://doi.org/10.1155/2021/6676297

[3] Pavate, A.A., Bansode, R. (2021). Performance Evaluation of Adversarial Examples on Deep Neural Network Architectures. In: Balas, V.E., Semwal, V.B., Khandare, A., Patil, M. (eds) Intelligent Computing and Networking. Lecture Notes in Networks and Systems, vol 146. Springer, Singapore. https://doi.org/10.1007/978-981-15-7421-4_22

[4] Aruna Pavate, Rajesh Bansode, "An Analysis of Derivative based Optimizers on Deep Neural Network Models", Book chapter Data Science and Data Analytics Opportunities and Challenges ISBN 9780367628826, 1st Edition, Pages-15, September, 202

[5] Aruna Pavate,Veda waikul, onkar Raghvan, "Restaurant Review Classification and Recommendation System using SVM", International Conference on Innovation and Advance Technologies in Engineering, ISSN (e): 2250-3021, ISSN (p): 2278-8719, PP 49-52, November, 2019

[6] Urvesh rathod, Aruna Pavate, Vaibhav Patil, "Product Rank Based Search Engine for E-Commerce Unification of E-Commerce", 2018 3rd International Conference for Convergence in Technology (I2CT), Electronic ISBN:978-1-5386-4273-3, pp 1-5, April, 2018

[7] A. Pavate, A. Chaudhary, P. Nerurkar, P. Mishra and M. Shah, "Cuisine Recommendation, Classification and Review Analysis using Supervised Learning," 2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW), 2020, pp. 1-6, doi: 10.1109/ICCDW45521.2020.9318646.

[8] Alan Mccabe & Jarod Trevathan , "Artificial Intelligence in Sports Prediction", 1194-1197. 10.1109/ITNG.2008.203.

[9] Rana, D and Amol Vasudeva. "Premier League Match Result Prediction using Machine Learning." (2019).

[10] Bosch, P., & Bhulai, S., "Predicting the winner of NFL-games using Machine and Deep Learning", 2018.

[11] Ragini Singla,Dr.Amardeep Singh, "Sports Prediction Using Machine Learning", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.7, Issue 10, page no.2759-2765, October-2020

[12] Vistro, Daniel Mago, Faizan Rasheed and Leo Gertrude David. "The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics." International Journal of Scientific & Technology Research 8 (2019): 985-990.

