International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

# Football Team Predictor and Analyzer using Machine Learning

Jay Parekh[1], Vedant Kale[2], Ila Savant[3], Gaurang Kulkarni[4], Swarali Bhide[5]
[1, 2, 3, 4, 5]UG Student, Department of Computer Engineering, Marathwada Mitramandal's College of Engineering. Pune, India.

***Abstract: Dilemma is to decide whether a player should be in-team for a certain match or not. Without knowing the abilities of opponent it is hard to decide the team lineup. Selecting the perfect lineup and tactics is the main concern of the team manager. In order to create a perfect team, considering all the factors like player's form, speed, agility, hit ratio, goal ratio, defense, and history and team's win ratio, player relationships, and many more. Building a team so that the current team has a better chance of winning using the system. This system helps predict a better team formation according to the geographical, performance factors.***

***Keywords: ML, AI, KDD***

## I. INTRODUCTION

Football's scope has been limited to studying certain aspects such as physiological factors of team and players. Recently, it has been suggested that researchers should focus upon the performance measures. Performance indicators are defined as the variables that elucidate some aspects of performance. They indicate a profile of ideal performance that help achieve and compare success as well as behavior. Despite pre-existing constructions in team sports such as basketball, cricket and rugby, there has been little research for football. Existing analysis literature  in football suggests that there is paucity of research on team performance factors. Moreover, there is no such known system that predicts the team lineup and performance of the team's players according to the opponents for every match. There are studies and systems that predict the winning probability of the team analyzing previous data but not for individual players. Mostly the creation of lineup and formation is done manually by the coach and team. Considering various factors like player's individual performance measures and team's compatibility with considering geographical factors including venue and attendance.

Each player has fixed position where they play from like forward, mid-field, defender or goalkeeper. Certain players perform better against some specific opponents. The factors to be considered for this system are classified according to their positions which are shown in Figure A using UML diagram and some are universal which are common for all positions.

The aim of this paper is to develop a system using ML  algorithms such as linear regression and gradient boosting to provide perfect lineup for the next match comparing opponent's lineup and team performance. An idea to assign a rank to each player and team as a whole as well makes the analysis easier to understand. Also the features such as analysis of teams, players, and league which helps the management and coach.

### *A. Linear Regression*

Collect football match data (Data Gathering).

Attribute selection

Linear regression equation is $Y = m1x1 + m2x2 \ldots + mnxn + c$

Where,

m1…mn = slope

x1…xn = independent variables

c = intercept

Find out the values of m1,m2…mn and c.

Insert the values of m1,m2…mn and c in linear regression equation.

Find out the value of  Y.

XGBoost

XGBoost (eXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm. It has both linear model solver and tree learning algorithms. So, what makes it fast is its capacity to do parallel computation on a single machine.

Steps

1) *Step 1:* Load all the libraries.
2) *Step 2:* Load the dataset.
3) *Step 3:* Data Cleaning & Feature Engineering.
4) *Step 4:* Tune and Run the model.
5) *Step 5:* Score the Test Population.

### *B. Random Forest*

Random forest is a classification and regression technique which is applied by constructing multiple decision trees. At the time of training the data these multiple decision trees are constructed and output is generated using classifying or predicting the results of these constructed multiple decision trees.

Normally tress are grown very deep which makes them to learn highly irregular patterns that may result into overfitting. Random forest is a way of averaging multiple decision trees and these are trained on different parts of same training dataset.

The training of dataset in random forest applies a general technique of bootstrap aggregating or bagging. Given a training set S=s1,s2...sn and with target as T= t1,t2...tn which baggs repeatedly( a maximum of N times. Then selects a random sample with replacement of training set and fits the trees into the samples:

For n=1,2,...,N

1) A sample with replacement of n training examples form S,T
2) Then train a classification or regression tree on Sb and Tb.

After training the data to trees prediction for undefined samples can be made by taking average of prediction from all individual regression trees.

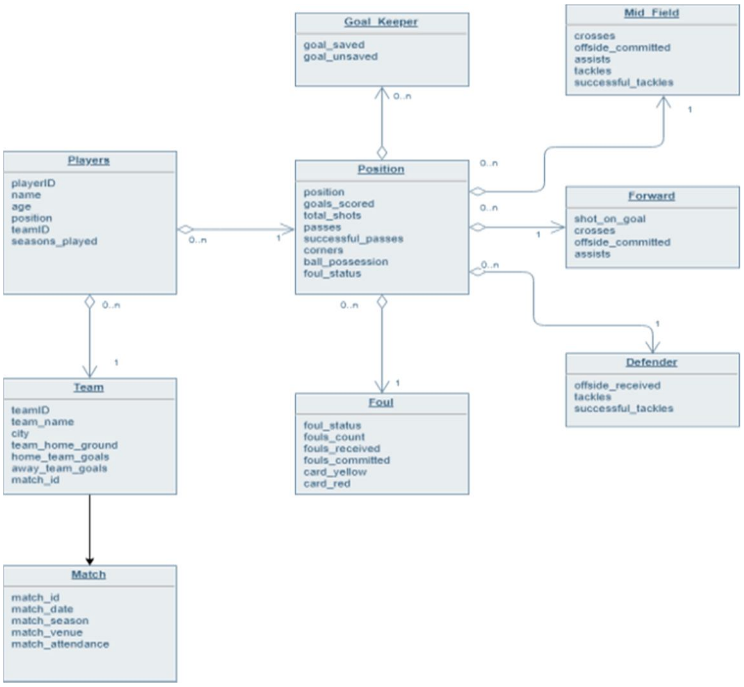
*a)* Propertie

*i)* *Variable Importance:* Random forest can be used to rank the variables in classification or regression problem.

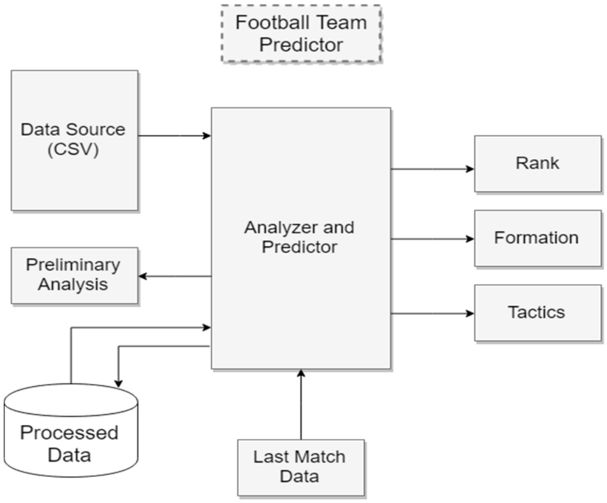
*ii)* *Relationship with Nearest Neighbour:* This can be merged and understand with knn algorithm where weighted neighborhoods are used.

## II. FIGURES AND TABLES

### *A. UML Diagram*



### *B. System Architecture(Visual Flow)*



All diagrams are created using online tool.

## III. LITERATURE SURVEY

This paper [1] the general discussion of grading the team and implementing ML algorithm to predict the team has been discussed. Establishment of collaboration, team management and grading methods has been implemented..

The paper [2] gives the idea about performance indicators affecting the team's analysis and formulating along with co-relating the performance indicators which discus reference values of the game indicating winning, losing and drawing teams in football.

The paper [3] taken a compound framework in predicting sports results – rule-based reasoning and Bayesian inference – and combined it with an in-game time-series approach for more accurate and realistic predictions. Implementation of football results predictor called FRES (Football Result Expert System) has been discussed.

In [4] the improvement in accuracy of systems using data mining techniques, AI methods and knowledge base discovery (KDD) has been discussed.

### *A. Overview of Project Modules*

1) *Team Prediction:* In this module, the user (manager/coach) decided the team as they require and the model predicts win, lose or draw on the basis of previous match data and the overall rating of the player. Random Forest classifier is used to train the model because it provides the best accuracy among many others.
2) *Overall Rating Prediction:* In this module, the overall rating of a new player is predicted using the XGBoost model trained using the player attributes data of all 11,075 players. Whenever a new player is introduced in the team having no past record, this module is used to avoid Mathew Effect on the system.
3) *Analysis Module:* Analysis such as team interactions, past lineup and outcome, and world map indicating the matches played in the countries (venues). Providing stats as well as insights for the player, team and matches.
4) *Player Comparator:* In this module, one on one Comparison of players on the basis of all the attributes to find out the best among all is done. Distinguishing the strength and weakness of player in order to establish the scope for improvement.
5) *Top teams Extractor:* In this module, the top 10 teams of individual league as well as the overall among all of Europe is extracted on the basis of goals and stats. To determine the league table and the team status in the league this module is used.

## IV. CONCLUSION AND FUTURE WORK

In this system by using random forest, we achieved an accuracy of 80.85% with win accuracy of 88%, losing accuracy 87% and draw accuracy of 62% . In a provided sample for two teams on an upcoming match the system predicted a win situation which was later verified as true after the game. On adding a new player with overall raring provided by linear regression and XGBoost with explained variant score of 0.81 and 0.96 respectively.

Future work include providing depth and in-field analysis of the player and reach and coverage according to their skills and performance indicators.

## REFERENCES

[1] Dragutin Petkovic, Kazunori Okada, Marc Sosnick, Aishwarya Iyer, Shenhaochen Zhu, Rainer Todtenhoefer, Shihong Huang, “A Machine Learning Approach for Assessment and Prediction of Teamwork Effectiveness in Software Engineering Education”, IEEE 2013.

[2] Carlos Lago-Peñas, Joaquín Lago-Ballesteros, Ezequiel Rey, “Differences in performance indicators between winning and losing teams in the UEFA Champions League”, Journal of Human Kinetics 2011.

[3] Byungho Min, Jinhyuck Kim, Chongyoun Choe, Hyeonsang Eom , R.I. (Bob) McKay, “A compound framework for sports results prediction: A football case study”, Elsewer 2018.

[4] Igiri, Chinwe Peace, Nwachukwu, Enoch Okechukwu, “An Improved Prediction System for Football a Match Result”, IOSRJEN 2014.

[5] Yiftach Nagar, “Combining Human and Machine Intelligence for Making Predictions”.

[6] COUCEIRO, Micael S., DIAS, Gonçalo, ARAÚJO, Duarte and DAVIDS, Keith , “The ARCANE Project: how an ecological dynamics framework can enhance performance assessment and prediction in football”, Sheffield Hallam University Research Archive (SHURA).

[7] Carlos Lago-Peñas, Joaquín Lago-Ballesteros, Alexandre Dellal, and Maite Gómez “Game-Related Statistics that Discriminated Winning, Drawing and Losing Teams from the Spanish Soccer League”, J Sports Sci Med. 2010.