ISSN No:-2456-2165

Player Stats Analysis Using Machine Learning

Sylvester Anthony A
Department of Computer Science and Engineering
Velammal Engineering College
Chennai, India

Dr. S.L. Jayalakshmi
Department of Computer Science and Engineering
Velammal Engineering College
Chennai, India

Akash D
Department of Computer Science and Engineering
Velammal Engineering College
Chennai, India

Abstract:- A significant area that needs critical thinking to ensure a team performs well is the strategizing of a specific team. The secret to overcoming this dilemma is to use the talent of the players inside the team that can be disregarded at times. With ever growing rivalry, a talented team, with an old and obsolete plan, could have to face undesirable and bad outcomes. In this article, we have performed an experimental analysis in the field of outdoor sports for soccer. The approach considered in the current paper work focuses on deciding the lineup of a squad by measuring the abilities of the soccer players. To collect the data set in the proposed method, we created our own web scraping algorithm. To predict the best location of a player, machine learning classifiers such as Neural Network (Multilayer Perceptron), Random Forests, KNN, Naïve Bayes and Logistic Regression are used. Using various ultra-modern classifiers, the precision of the method proposed was evaluated.

Keywords:- *KNN, Naïve Bayes, ANN, Random Forests, Logistic Regression, Machine Learning, Strategy Management.*

I. INTRODUCTION

In history, there have been many occasions where the application of analytics and statistics has changed the world of sports entirely. One of them is when the FIFA World Cup 2018 used data mining by the French football team. Today we see a variety of football clubs using data analytics in different scenarios be it scouting, team strategizing etc. We used statistical and advanced analytics in our experiment to help look for the talent pool for football clubs, nations and determine the playing roster on the basis of performance statistics for individual players. Today, some specialist statistical analysis companies such as WyScout Sports and Opta Sports, statsbomb etc. are present, providing football clubs, managers, and leagues with data.

It becomes really important for players who play at any specific position, to put a very talented and strong team to do their best and give a good performance at that specific position. This identification, historically performed in the manual mode, can be performed using different prediction model used today to make the job faster and much more efficient. We have trained the machine learning models for the prediction using the data produced by Electronic Arts for the latest version of their legendary game FIFA.

In different studies and posts, it has been said that the use of FIFA edition data has different uses and many advantages that conventional historical data and datasets do not really bring. The FIFA Football Games have been giving the world of football a big, out-of-bounds and cogent scout of players around the world since 1995. We have a number ranging from 0 to 100 for each parameter that really tells how good a player is at that specific parameter. Examples of characteristics include: blocking, pace of the run, stronger foot, heading and agility. If it is noticed, then we can accept the fact that it seems almost impossible to immediately classify all the players precisely in these attributes. All of these are then gathered and nurtured by a single corporation or organisation, whose main purpose is to close the distance between the gameplay and the football team. It seeks to get as close as possible to truth, while maintaining honesty and representativeness across the dataset.

The FIFA 19 dataset used for this study and prediction offers various different statistics for about 18,000 players on about 70 different attributes, 25 of which, we say, are taken for position classification by a dimensionality reduction algorithm. These characteristics are optimum metrics to assess how well a specific player can play at that particular position, and based on the results we get, they can also provide the player's best position. To classify players into three distinct positions, namely the positions of Attacking, Midfield and Defence, we have used ultra-modern machine learning classification algorithms.

Machine learning models such as Logistic Regression, Random Forest, Naïve Bayes, KNN and Artificial Neural Networks were used after extracting the data and forming a clean dataset. For training purposes, the train data with the different labels is fed into the classifier. The test data is provided as input to the qualified classifier after the completion of the training process, whose output class needs to be predicted based on the training performed. As described above in the earlier descriptions, the performance is categorized into three positions. Different performance metrics have been measured to assess accuracy, such as precision recall and F1 score. In order to better understand the efficiency of the predictive model, the Confusion Matrix has also been generated for each of the algorithms.

The following is the structure of the remaining part of the research paper. In Section II, a short summary of the Literature Review is given. Section III provides an idea of the different approaches and helps with the paper's overall approach. Section IV provides experimental findings along

with a brief clarification. Section V provides a brief and crisp conclusion that the cloud allows to allow further changes.

II. LITERATURE REVIEW

In paper[1],The author uses a technique for assessing players based on a data-based methodology. A regression model based methodology uses a restricted input and gave results by comparing two players from two different teams. A small dataset was the basis of this method. The result was that the PDVE team had 4-7% greater power across three metrics. The conclusion notes that there is more space for changes in this basic model.

In paper[2],the author says about player performance prediction in the game of football. The paper has given a proven model accuracy of 74.34 % which is then succeeded by the second linear regression model in the architecture with an accuracy score of 91%. This second model has a say on the future market value of the players on the measures of the overall performance value which is predicted by the first model.

In paper[3], A game of football simulation, where the different strategies are used to determine the best formation. A soccer game simulator is included in this article. The interaction algorithms of virtual soccer players here create an interaction where context-free grammar is used to perform the various simulations. The technique for strategic culls is Nash equilibrium: the strategy profiles that suit this equilibrium set up efficient game cumulations. Player and team profiles are drawn from records of the best tournaments of the past. Simulations are performed with 4-3-3, 4-4-2 or 5-3-2, midfielder-forward formation of defenders, average profile players as well as categorical profiles, respectively, so that the perks of the simulator can be calculated.

In paper[4], To strategize the squad, a machine learning approach is used. Different input features such as speed agility were used by the machine learning approach to determine the best play 11. In this model, various parameters needed for player picking are evaluated in four major divisions using a neural network model and these major categories, including the technique of the player, the celerity of the player, the physical status of the player and the resistance of the player. The neural network approach was acclimatized to create a model for player picking.

In paper[5], To strategize the team where the multiple regression algorithm was used, a machine learning approach was used. The accuracies of the different classification algorithms were also shown to demonstrate the best algorithm that matches the data well. It can be noted from the different algorithms used that the Neural Network (Multilayer Perceptron) performed the best with an accuracy of 79.01 percent and an F1 score of 0.787, taking the results into account. With an F1 score of 0.739 and an accuracy of 74.07 percent, Random Forests also performed well. With an accuracy and F1 score of 71.92 percent and 0.69 respectively, Logistic Regression gave results.

In paper[6], A multilayer perceptron neural network was used to predict a football (soccer) player's price by using real-time data on more than 15,000 players such as the video game FIFA 2017. By experimenting with different activation functions, number of neurons and layers, learning rate and decline, stochastic gradient descent based on Nesterov momentum, L2 regularization and early stoppage, the network has been optimized. There is a parallel exploration of different aspects of neural network training and a detailed look at their trade-offs. For 119 pricing divisions, the final model offers a top accuracy of 87.2 percent, and takes any footballer below 6.32 percent of the original price on average.

We come to the conclusion, from the above papers, that all these techniques provide a way to predict different outcomes. What we do in this paper is take the best practices and try to raise the efficiency by using these above-mentioned papers.

III. METHODOLOGY

Before we even go deep, the ultimate aim of our approach is to assign players the optimum position based on their set of skills. In this case, three performance groups are pre-decided: attack, midfield, and defence. The entire flow of the process can be noted in Fig. 1. The method followed in this is very modular in its organisation.

The method involves with the scraping of website data with statistics of about 18,000 players with around 70 features. On the website, the players who play are present. The attributes are statistics and personal information. For further analysis of the data in the csv format, the extracted dataset is processed.

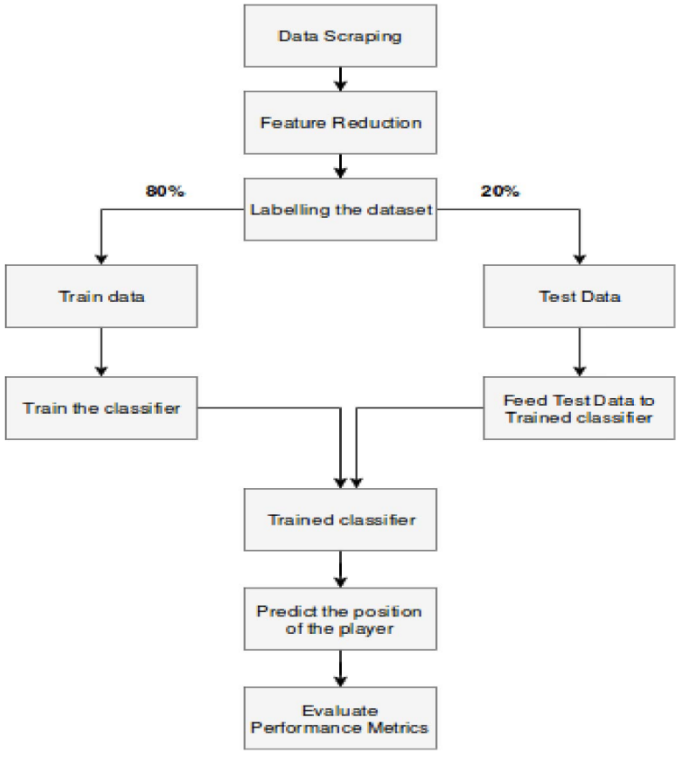

Fig 1

ISSN No:-2456-2165

The next move is to decrease the attributes in the dataset that are available. Since there are several attributes that are not important to deducting our performance, we may drop them. Thus, the selection of 25 related attributes is made using Principle Component Analysis(PCA) to improve the accuracy of the model by providing quality data to the classifier. Attributes such as personal data, for example, are irrelevant to the training of classifiers and can therefore be avoided for study.

The dataset has a column which indicates the preferred position of the player. Then a total of 14 positions are mapped and positioned in the 3 groups already pre-decided. Except for the attribute that has the preferred role to ensure accuracy, data normalization is carried out on all data that is part of the dataset features. Thus, the value of each function varies from 0 to 1. After cleaning the dataset, 80 percent of the data is allocated randomly to train the classifier and the remaining 20 percent is given to the testing level.

The machine learning models used in this approach are Neural Network (Multilayer Perceptron), Random Forests, and Logistic Regression. Using GridSearchCV, the optimal neural network output is chosen based on the alpha value and the number of hidden layers. The optimal number of hidden layers and the alpha value are 20 and 0.001, respectively. In Neural Network training, these parameters are used. In the case of logistic regression, Multinomial Logistic Regression (MLR) is used as the dependent groups are multi-classes. In Random Woods, default parameters are used.

After training the model, test data will be fed in and the trained classifier will be loaded from the classifier. It defines the optimal output class, the optimum position of the player in this case. For the analysis, the results of the test step are then taken into account for analyses.

The performance of the classifier is measured on the basis of certain performance parameters, such as F1 score, accuracy, recall and accuracy. The primary metric is the F1 score. For visualization, a confusion matrix is also plotted as shown in Figures 2,3,4 and 5.

IV. RESULTS

The dataset forms the basis of our outcome. It is the basis of our prediction of results where, when we do the data scrapping, a dataset containing more than 18,000 items is first collected. Then we take the various algorithms for machine learning and we run them to test performance metrics.

With our primary being F1 score calculation and the remaining being the secondary, the performance indicators of a classifier include F1 score, precision, accuracy and recall. The F1 score takes account of both accuracy and recall, and is the harmonic mean of both measurements. These metrics can be defined, mathematically, as below:

- Precision: Precision is defined as the fraction of relevant instances among all retrieved instances.

$$P = \frac{TP}{TP + FP}$$

where, FP denotes for False Positive and TP denotes True Positive.

- Recall: Sometimes referred to as 'sensitivity, is the fraction of retrieved instances among all relevant instances.

$$R = \frac{TP}{TP + FN}$$

where, FN stands for False Negative and TP stands for True Positive.

- F1 score: Is the harmonic mean of precision and recall. It is better to use this as a primary metric because it considers both precision and recall as one value.

$$F1 = 2 * (\frac{P * R}{P + R})$$

where, R is the Recall and P stands for Precision.

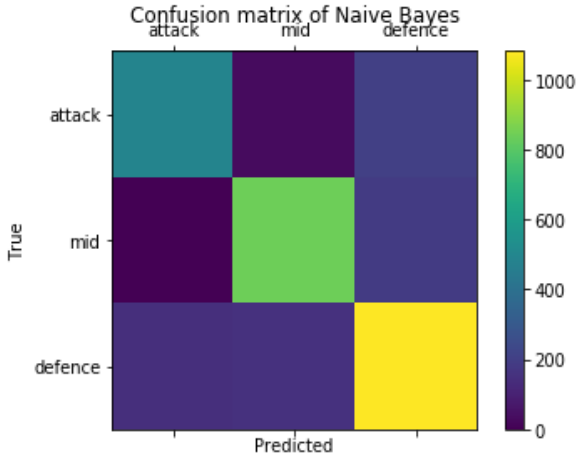

Fig 2

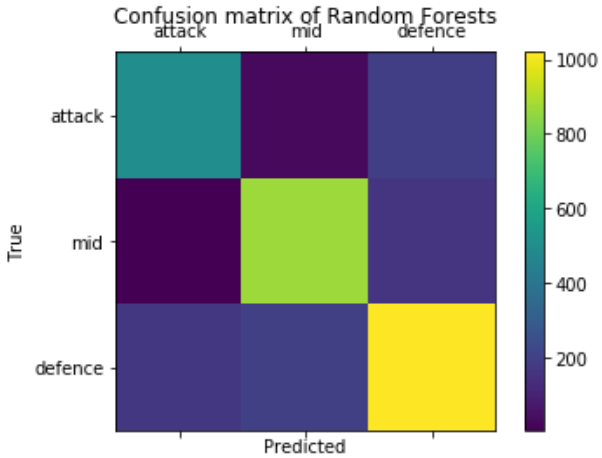

Fig 3

ISSN No:-2456-2165

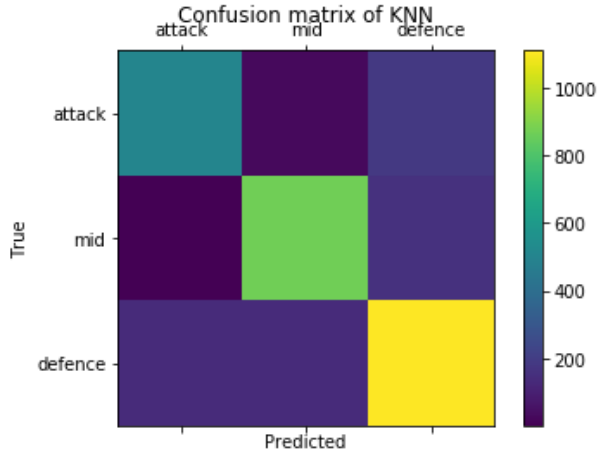

Fig 4

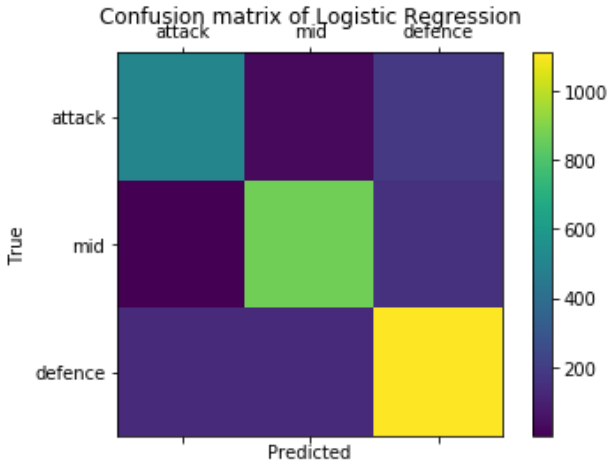

Fig 5

TABLE 1

Performance Metrics				
Model	**Precision**	**Recall**	**F1 score**	**Accuracy**
Neural Network	0.82	0.84	0.85	83.66%
Logistic Regression	0.79	0.79	0.79	79%
Random Forests	0.76	0.76	0.76	76%
KNN	0.77	0.77	0.77	77%
Naïve Bayes	0.77	0.77	0.77	77%

We may determine the results from the table. Table 1 shows that the Neural Network (Multilayer Perceptron) performed best with an accuracy of 83.66 percent and an F1 score of 0.85, With an F1 score of 0.76 and an accuracy of 76.07 percent, Random Forests also performed well, regarding the metrics taken into account. On close observation of the Confusion matrix, it can be shown that the performance for the midfield and defense positions has been correctly predicted to a large extent by all the 5 classifiers. However, in contrast to other positions, the accuracy with which the attack position has been predicted is much less. The low precision in the attack position is primarily due to the fact that as opposed to mid and defense, less players play in that position in most of the formations. Since players can play in more than one position, the prediction accuracy of all three models is not that great.

V. CONCLUSION

In this experiment, machine learning techniques were used to achieve an efficient analysis of team strategizing. Neural Network, Random Forests, KNN, Naïve Bayes and Logistic Regression are the models used in our paper. The outcomes given will give us a degree of human precision in predicting the role of the player based on his ability. The Neural Network is shown to have worked best. In this untouched area, further study may lead to increased precision. In the future, to explore and make optimum use of the available talent, this strategy can be applied to other areas such as education, industry, etc.

REFERENCES

[1]. R. Stanojevic and L. Gyarmati, "Towards Data-Driven Football Player Assessment," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016, pp. 167-172.
[2]. R. Pariath, S. Shah, A. Surve and J. Mittal, "Player Performance Prediction in Football Game," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1148-1153.
[3]. M. Alvarado and J. Tellez-Giron, "Computer football: Plays, players and strategies choices," in IEEE Latin America Transactions, vol. 16, no. 5, pp. 1485-1492, May 2018.
[4]. Q. Wang, Z. Xu and Z. Wu, "An Analysis of Football Player Transfer Problems Based on Real Options," 2010 International Conference on Management and Service Science, Wuhan, 2010, pp. 1-3.
[5]. V. Rao and A. Shrivastava, "Team strategizing using a machine learning approach," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 1032-1035.
[6]. Pricing Football Players using Neural Networks,Sourya Dey