Sports Analytics algorithms for performance prediction

Konstantinos Apostolou
The Data Mining and Analytics Research Group
School of Science & Technology, International Hellenic University
14th km Thessaloniki – Moudania 57001 Thermi, Greece
k.apostolou@ihu.edu.gr

Christos Tjortjis
The Data Mining and Analytics Research Group
School of Science & Technology, International Hellenic University
14th km Thessaloniki – Moudania 57001 Thermi, Greece
c.tjortjis@ihu.edu.gr

***Abstract*— Sports Analytics is an emerging research area with several applications in a variety of fields. These could be, for example, the prediction of an athlete's or a team's performance, the estimation of an athlete's talent and market value, as well as the prediction of a possible injury. Teams and coaches are increasingly willing to embed such "tools" in their training, in order to improve their tactics. This paper reviews the literature on Sports Analytics and proposes a new approach for prediction. We conducted experiments using suitable algorithms mainly on football related data, in order to predict a player's position in the field. We also accumulated data from past years, to estimate a player's goal scoring performance in the next season, as well as the number of a player's shots during each match, known to be correlated with goal scoring probability. Results are very promising, showcasing high accuracy, particularly as the predicted number of goals was very close to the actual one.**

Keywords— Sports Analytics, Prediction, data mining, classification.

I. INTRODUCTION

Sports analytics exist as a concept for many years, but a lot of steps are required in order to understand and improve team performance. It is a topic that is increasingly gaining interest recently [1], [2]. More and more teams try to use such solutions to improve performance. For the purposes of this paper we will mostly focus on football (soccer). We aim at predicting performance of individual players, based on previous seasons' data. We tried to predict results in three major fields. In these experiments, a player's position in the field could be predicted using suitable algorithms. By accumulating data from past years, we could have an estimation for a player's goal scoring performance in the next season.

Sports analytics is the application of the Machine Learning methods and implementations to sports in order to draw useful conclusions. Such conclusions may affect the performance of an individual athlete, the team as a whole for a specific game or for the whole season. It can also help teams make predictions for upcoming talents, a player's market value and the possibility of an injury. It is a field that is becoming more and more popular nowadays and it is likely to be adopted by a plethora of teams, coaches, individual athletes and companies. [3]

Furthermore, in order to be more focused, the number of a player's shots can be predicted in each match, something that has a correlation with goal scoring possibility. In order to achieve good accuracy, not only do we need to find an accurate database, but also the attributes of the database to be relevant to the research. So, sometimes we had to combine data from more than one database and create ours in order to conduct experiments. In the end, we managed to obtain promising results, encouraging us to continue our work in the future.

There are many ways that a team can use data and there are many kinds. First, they could be related to the way that the game is played (in this case football), not only by each player, but also by the team as a whole. This kind of data have to do with the players average stats, such as the number of goals that they score, the number of fouls they commit, with how many red and yellow cards they are booked, how many tackle-ins they do, how many kilometers they run during a match (the use of cameras helps for these stats) and many more.

However, it is difficult to compare all those stats in successive matches, because a player's performance depends on his opponents as well. In other words, a striker may score many goals playing versus a badly organized defense, or a bad goalkeeper. Moreover, there are data that give information about how a team manages to score a goal. For, example how many passes they achieve before they score a goal, what is the average ball possession, and the field that they occur. Obviously, it is different to keep the ball close to the opponent's goalpost. Sometimes, though, there are outliers that show that other factors play important roles as well, because teams win in football even if they have little ball possession.

So, it is required to use data that are more difficult to collect. These data have to do with players' physical condition, such as pulse rate when being calm and when sprinting, measure sweatiness and track a player's sleep. However, most of these data were impossible to collect until recently because these devices were not allowed in football games. It was in March 2015 that the use of Electronic Performance and Tracking Systems (EPTS) was allowed and gave the opportunity for sports analysts to explore other aspects of the game. There are aspects that affect a team's performance such as the weather conditions, the condition of the field, and even psychological factors, such as the fans support. Another factor is injuries which sometimes could be predicted or prevented.

II. BACKGROUND

It is common knowledge that more and more teams are trying to invest in some form of data analytics in order to improve their performance, by gaining even a slight advantage over their opponents. However, there are not so many teams that admit that they have implemented such methods. In addition, it is difficult for a team to reserve the needed funds in order to employ data scientists who will help them draw conclusions [4] [5].

Furthermore, it is important to point out that not only teams are interested in predicting the outcome of a game or a team's performance. People around the world who are keen on sport betting would also benefit from such predictions [6].

The first attempt of applying analytics in a sports game was conducted by T.C. Reep during the 1950s. While watching a football game at Swindon Town he was disappointed at the fact that the team could not score. So, in the second half he started to take notes about the game. He concluded that the team should slightly increase the scoring rate in order to be promoted. A manager at Brentford was fascinated by Reep's work and thus, he was hired as an adviser. His goal there was to help the local team to avoid relegation. After his arrival the team easily gathered the needed points, so they managed to remain in the division [7].

Nowadays, more sophisticated methods are used in football and in other sports in general, for data mining and decision making based on data. However, we should not forget that Reep was one of the pioneers at this field that was going to become a hot topic in our decade.

One of the greatest modern examples for data analytics usage is the German National Football Team. In summer 2014, there was the World Cup in Brazil. In the semifinals, the interest was of course on the match of Germany vs. Brazil. Germany managed to win emphatically with 7-1, something that was never achieved before against Brazil. As it was later pointed out by a German assistant coach named H. Flick, he had spent two years studying the Brazilian football players. He managed to gather a lot of data and thus his team was able to achieve such a victory. It was the first time in the World Cup History that something like that was admitted publicly, however we cannot be sure that we would have a different outcome if no such data analysis occurred.

There are also similar attempts in basketball [8]. It was not until 2005 that Israeli scientists, Gal Oz and Miky Tamir, created SportVU [9]. SportVU is a system that has the ability to track not only the ball, but also the athletes, providing data about their positioning and movements in the court. All these data that are gathered can be further analyzed through sophisticated algorithms that the company has created [10].

At first the tracking system was not real time, but this changed in 2011. At the time being, not all teams had installed this system. However, in 2014 all teams had installed it in their courts, due to the benefits that it was providing [11]. Moreover, due to the systems success, in 2016 it was decided to be extended to other sports, as well such as football. The Ligue de Football Professionnel's started the adoption followed by other institutions. All those data that this system provides could be exploited by data scientists and statisticians using machine learning algorithms in order to come up with more difficult conclusions, not easily obtained with a quick look at the data [12]. There is also an attempt from the University of Virginia. Their football team had the will to use their data in order to improve their performance. Due to the lack of funds they asked the University's engineering department for assistance [4].

In addition to the previous approaches, there is another one more sophisticated that has the advantage of using data that are produced by wearable devices. It was until recently that FIFA allowed the use of such devices during football matches (Electronic Performance and Tracking Systems – EPTS). These systems include accelerometers, gyroscopes, magnetometers [13]. While in sports like Rugby, Hockey and Cricket the use of these devices is widespread, in football it is not used so much. Of course, the fact that the use of these devices was not allowed played an important role, but even nowadays there are not so many teams that have adopted their use.

III. APPROACH

Our proposed approach comprises several steps detailed here and shown in Fig. 1 as a data flow diagram. We explain the above steps more extensively below:

- Firstly, we need to find data that are relevant to our topic. One of the most complete websites that contains information about football players individually, but also about the teams, is whoscored.com. Stats about a player's appearances, how many minutes he was in a game for a whole season, the number of goals he scored, his assists and completed passes, even how many red and yellow cards he has been booked with, and many more are provided. In this website, we found the necessary data about Messi and Suarez who were the main football players for the experiments that were going to be conducted.
- Since we find the data that are related to our topic, we need a software tools to scrap them and have them in a .csv file.
- We now have a database to work with. However, it is not "clean" and includes attributes that we do not need. So, the .csv file that we obtained from the scraping tools needs to be edited and processed before we can use it for classification. We also had to eliminate missing values, so that the database could be more easily managed while conducting experiments.
- After preprocessing, we can use classification algorithms known to provide good results [13], [14], [15]. Of course, there should be a variety of classifiers that are going to be used, in order to be able to compare the results and chose the classifier that is more accurate. However, the same classifier should be used for all the players, otherwise the "method" will not be fair, and the results will be biased.

IV. EXPERIMENTAL RESULTS

We conducted three main experiments in order to predict a) a player's position b) the number of goals a player scored during a season and c) the number of shots a player attempted during a match. We detail these experiments and present the results achieved in the following subsections.

A. Player Position

The first experiment we conducted was a proof of concept, attempting to predict the position of a football player. This was handled by obtaining a database which was crawled from the website https://sofifa.com [14]. It includes data from the football game FIFA 18. The database has attributes for many players. We decided to preprocess the data to deal with several issues, like missing values, names that contain characters not compatible to the UTF-8 system, wrong data types (a column that contains integers is parsed as a string) etc. For example, some attributes were not available for all the players, because of their position on the field. For goalkeepers for instance, we had to face missing values that were not considered important and were not mentioned in the database (e.g. acceleration). However, it is obvious that we had to handle the issue of missing values otherwise the classification would not be completed. Moreover, names such as Modrić which are not UTF-8 had to be converted.

After resolving these issues, given that we aimed at classifying a player according to his position, we decided that we would consolidate a variety of positions to just 4: Forward, Midfielder, Defender and Goalkeeper, since there are many positions, and many players have more than one (e.g. striker and left winger). For classification we used the well-known datamining toolkit WEKA [15]. So, given a player's attributes we try to determine their position on the field. WEKA provides the opportunity to conduct experiments and determine which attributes contribute more to the result. By selecting he most important attributes, the accuracy achieved was 81.5% with Random Forest and Sequential Minimal Optimization (SMO), using 10-fold cross-validation. It is used for the solution of the quadratic problem that occurs during the application of support vector machines. It was introduced in 1998 by J. Platt [16]. It belongs to the support vector machines family. Fig. 1 shows the relevant confusion matrix.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 24  0  9  0 |  a = FOR
  0 17  0  0 |  b = GK
  6  0 32  3 |  c = MID
  0  0  4 24 |  d = DEF
```

Figure 1- confusion matrix for position prediction

We observe that for goalkeepers there is no misclassification. The errors occur mostly in the midfielder position, due to the consolidation of positions. That means that midfielders share attributes with forwards and defenders. However, most of them are classified correctly.

B. *Number of goals per season*

The next experiment we conducted, was related to footballers' performance. Messi and Suarez of Barcelona were used as case studies to predict statistics such as the number of goals that the player has scored. In order to evaluate the process's accuracy, we used data up to season 2016-17 found in [17] to predict results for season 2017-18. The method that we followed involved firstly, to scrap data from whoscored.com and create a database. Then, we created a test set. Keeping in mind that we did not know anything about season 2017-18 we created another file to be used as a test set that looked like the training set. The only difference was that an average value for each attribute was used. More specifically, this value was estimated by accumulating all the previous seasons attributes and calculating the mean.

Then we enacted the classification process. However, we needed to decide the class attribute, the one we need a prediction for. Since Messi and Suarez are players with great scoring capabilities, the class attribute was set to be the number of goals that they score. In order to be more accurate about the results the experiments were conducted by using 4 classification algorithms obtained by sklearn libraries for python [18]. Those algorithms were: Random Forest, Logistic Regression, MLP classifier and Linear SVC.

For both players the same method was followed. Messi actually scored 34 goals during season 2017-18. The classifiers we used predicted the results shown in fig. 3. Observing fig. 3, we can determine that the best results were given by Random Forest and MLP classifier followed by Linear SVC. The worst results were provided by Logistic Regression.

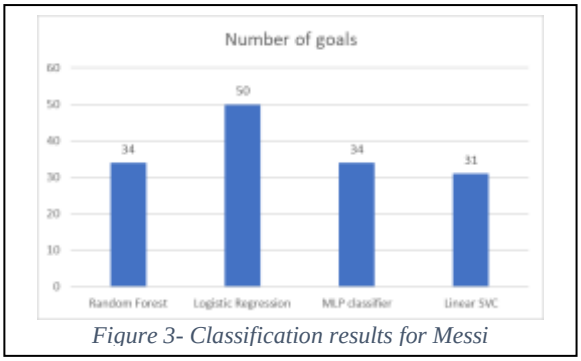

Figure 3- Classification results for Messi

The same procedure was followed for Suarez. Suarez scored 25 goals during season 2017-18. The classifiers provided the results are shown in fig. 5 We can observe that Random Forest and Logistic Regression have the same accuracy. However, the MLP classifier and Linear SVC produced the worst results by far.

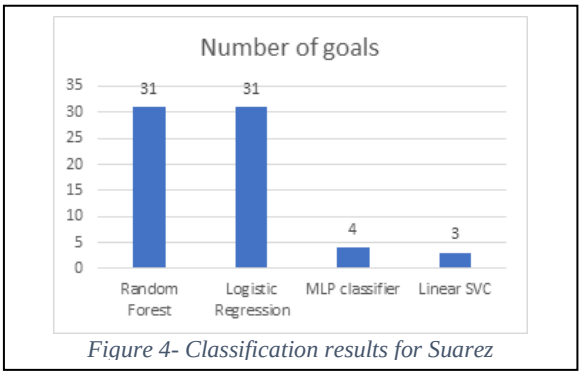

Figure 4- Classification results for Suarez

This experiment shows that we can predict quite accurately the stats that we are interested in for some players. Thus, for example, based on a player's individual performance, we could predict team performance. So, we could predict how many goals a team will score during a season, how many goals a team will concede and so on. However, the number of goals that a team scores in a season could be predicted using the team's previous years statistics, but by aggregating each player's statistics we achieve better accuracy. This is actually reasonable, considering that teams change every year and the players that are competing are not the same. We should take into account though, that a player's behavior in a game, will not be the same in every team he participates. Football and team sports in general depend not only on the behavior and performance of the teammates, but also on the opponent teams.

C. *Number of shots per match*

For the last experiment, we focused on the shots shot by a player (Messi) during a match. It was observed that Messi typically scores when he has the chance to shoot at least five times in a match. So, the idea was to predict the number of shots that Messi was going to attempt in a match. The process was the following. Firstly, data set to be accumulated for training originate from all his previous seasons. In order to create a test set we used data from understat.com, with statistics for a plethora of players and for a lot of their attributes. After completing a training and test-set and preprocessing the data, in order to eliminate missing values in cases that occurred, we used several classifiers for experiments with Random Forest providing good and accurate results.

We will focus on an example for the football match that Barcelona played against Real Sociedad on 15 September 2018. Random Forest predicted the number of Messi's shots in this match to be 2.133. By rounding that number, we can say that Messi is going to shoot twice. Messi, in this match, actually shot twice and also, he did not score any goals. The same technique could be applied for other Barcelona matches with Messi as an example and other players as well. By accumulating statistics for more players, we could have an overall estimation for the team's performance, such as other players that may score and also the clean sheets of the goalkeeper and the defense in general. As a result, we could predict the outcome of a game taking into consideration such factors, instead of just previous results. In such case though, we will have to focus on other players as well and take into account different attributes. The experiment will be different for a defender for example.

V. DISCUSSION

As highlighted by the experimental results discussed previously, good accuracy was achieved for all three tasks, especially for player position and number of goals per season. Random Forest appeared to perform very well across the board, something verified by other benchmarking efforts [19], [20]. Clearly further work is required to check how generalizable are these results and verify threats to validity originating from data selection, preprocessing and algorithmic decisions. For the purposes of this work, we focused on specific athletes. It is understandable, that these approaches may not be generalizable to all athletes, because they might have different characteristics. We must treat every athlete as an individual, in order to achieve good results. The reason that the chosen players were Messi and Suarez is that they are more consistent at their performance. They also play for Barcelona for a few years, so they have developed a certain style. Results would probably be different if it as Messi's first year in Barcelona. It would also be interesting to have the results of individual football players (for example) merged and studied as a whole team, in order to predict the team's performance.

VI. COCNLUSION

Sports analytics is expected to be an important part of a team's performance in the future [1]. It is anticipated that there is going to be a huge amount of data, which, combined with appropriate exploitation methods could improve algorithmic accuracy. Teams are increasingly embedding such practices to their inventory for improving their game. Barcelona for instance, has a whole team of data scientists for data analytics in order to study their game [21].

However, all parts that are interested in this field should be careful. Due to the fact that it is a quite new field, it is expected that mistakes

are going to be made. The data should not be easily accessible because numerous problems might occur, such as issues with betting companies, or wrong ways of athlete training leading to injuries as a result. [22]

In this paper, we explored 3 basic aspects of sports analytics in football:

The first, aimed at classifying a footballer in the position that he is more suitable, by using statistics about their style of game. We achieved good results, taking into consideration that some positions in the field are highly correlated, for example an “extreme” and a midfielder.

The second aspect was the prediction of the number of goals that two footballers (Messi and Suarez) were going to score in a season. By using appropriate algorithms, we achieved high accuracy, as the predicted number was very close to the actual one.

Finally, the third aspect was the prediction of the number of shots that a player (Messi) was going to shoot during a specific match. By accumulating appropriate data and building suitable training and test sets for the classifiers, we achieved a good prediction. The number of shots is correlated with the goal scoring possibility. Messi for example, is expected to score if he attempts at least five shots in a game [23]. Another useful extension is to include other types of classifiers known to give good results in different application domains [19], [20], [24].

In the future we plan to extend the above experiments to more teams and players and determine the differences that occur between them and the top-class players that were used for this case study. Moreover, whilst in this paper our focus was on individuals, it is important to point out that we could attempt to predict a whole’s team performance, something that is significantly more difficult to estimate [25], [26].

Furthermore, it is our goal to use and obtain data from wearable devices and conduct experiments there as well. However, it is difficult to obtain these data. Even sport associations for football and basketball do not allow the use of these devices at great extent, probably because they want to protect the sports, but also the athletes. So, at first the data will be obtained by training sessions and not in real matches. Teams are not expected to provide easily such data.

REFERENCES

[1] "Beyond Moneyball: The future of sports analytics | Analytics Magazine," [Online]. Available: http://analytics-magazine.org/beyond-moneyball-the-future-of-sports-analytics/. [Accessed 10 March 2019].

[2] B. Gerrard, "Moneyball and the Role of Sports Analytics: A Decision-Theoretic Perspective," *N. American Society for Sport Management Conf.*, p. 108–109, (NASSM 2016).

[3] W. Tichy, "Changing the Game: “Dr. Dave” Schrader," 2016.

[4] J. Corscadden, R. Eastman, R. Echelberger, C. Hagan, C. Kipp, E. Magnusson, G. Muller, S. Adams, J. Valeiras and W. T. Scherer, "Developing Analytical Tools to Impact U.Va.Football Performance," in *Systems and Information Engineering Design Symposium (SIEDS)*, 2018.

[5] E. Papalexakis and K. Pelechrinis, "tHoops: A Multi-Aspect Analytical Framework Spatio-Temporal Basketball Data," 2017.

[6] D. Miljković, L. Gajić, A. Kovačević and Z. Konjović, "The use of data mining for basketball matches outcomes prediction," *SIISY 2010 - 8th IEEE Int. Symp. Intell. Syst. Informatics*, pp. 309-312.

[7] J. Wilson, Inverting the Pyramid: The History of Football Tactics, Orion, 2009, pp. 138-144.

[8] J. F. Drazan, A. K. Loya, B. D. Horne and R. Eglash, "From Sports to Science : Using Basketball Analytics to Broaden the Appeal of Math and Science Among Youth," *MIT-Sloan Sport. Anal. Conf.*, p. 1–16, 2017.

[9] Z. McCann, "Player tracking transforming NBA analytics - Tech - ESPN Playbook- ESPN," 19 May 2012. [Online]. Available: http://www.espn.com/blog/playbook/tech/post/_/id/492/492. [Accessed 10 March 2019].

[10] "Competitive fire helps Kirk Lacob make his own name with Warriors," 20 June 2015. [Online]. Available: https://www.sfgate.com/warriors/article/Competitive-fire-helps-Kirk-Lacob-make-his-own-6339796.php. [Accessed 10 March 2019].

[11] "New age of NBA analytics: Advantage or overload? - The Boston Globe," 30 March 2014. [Online]. Available: https://www.bostonglobe.com/sports/2014/03/29/new-age-nba-analytics-advantage-overload/1gAim4yKYXGUQ2CTAe7iCO/story.html. [Accessed 10 March 2019].

[12] "Stats LLC and NBA to make STATS SportVU Player Tracking data available to more fans than ever before - NBA.com: NBA Communications," 19 January 2016. [Online]. Available: https://pr.nba.com/stats-llc-nba-sportvu-player-tracking-data/. [Accessed 10 March 2019].

[13] J. Fernández, D. Medina, A. Gomez, M. Arias and R. Gavaldà, "From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data," in *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.

[14] "Sofifa.com," [Online]. Available: https://sofifa.com. [Accessed 10 March 2019].

[15] E. Frank, M. A. Hall, I. H. Witten and C. J. Pal, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Fourth ed., Morgan Kaufmann, 2016.

[16] J. Platt, *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, 1998.

[17] "Whoscored.com," [Online]. Available: https://www.whoscored.com. [Accessed 10 March 2019].

[18] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. C. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, 2013.

[19] P. Tzirakis and C. Tjortjis, "T3C: improving a decision tree classification algorithm’s interval splits on continuous attributes," *Advances in Data Analysis and Classification 11 (2)*, pp. 353-370, 2017.

[20] V. A. Tatsis, C. Tjortjis and P. Tzirakis, "Evaluating data mining algorithms using molecular dynamics trajectories," *Int'l Journal of Data Mining and Bioinformatics 8 (2)*, pp. 169-187, 2013.

[21] "FC Barcelona: Pioneering a New Field of Analytics - MIT Sloan Analytics Conference," 23 February 2018. [Online]. Available: http://www.sloansportsconference.com/content/fc-barcelona-pioneering-new-field-analytics/. [Accessed 10 March 2019].

[22] "Big Data Analytics, Machine Learning, Artificial Intelligence," 12 December 2017. [Online]. Available: http://tanukamandal.com/2017/12/12/sports-analytics-changed-play/. [Accessed 5 May 2019].

[23] "Lionel Messi - History," [Online]. Available: https://www.whoscored.com/Players/11119/History/Lionel-Messi. [Accessed 10 March 2019].

[24] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan and J. Keane, "Comparing data mining methods with logistic regression in childhood obesity prediction," *Information Systems Frontiers 11 (4)*, pp. 449-460, 2009.

[25] H. Manner, "Modeling and forecasting the outcomes of NBA basketball games," *Journal of Quantitative Analysis in Sports*, vol. 12, no. 1, 2016.

[26] Y. Yang, *“Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics”, PhD Thesis*, University of California at Berkeley, 2015.