Predicting Physiological Variables of Players that Make a Winning Football Team: A Machine Learning Approach

Alberto Cortez[1](✉), António Trigo[1], and Nuno Loureiro[2,3]

[1] Coimbra Business School Research Centre | ISCAC, Polytechnic of Coimbra, Coimbra, Portugal
alberto.v.cortez@protonmail.com, aribeiro@iscac.pt
[2] Sport Sciences School of Rio Maior, 2040-413 Rio Maior, Portugal
nunoloureiro@esdrm.ipsantarem.pt
[3] Life Quality Research Centre (CIEQV), Polytechnic Institute of Santarem, Santarem, Portugal

Abstract. As football gained popularity and importance in the modern world, it been under greater scrutiny, so the industry must have better control of the training sessions, and most importantly, the football games and their outcomes. With the purpose of identify the physiological variables of the players that most contribute to winning a football match, a study based on machine learning algorithms was conducted on a dataset of the players GPS positions during the football matches of a team from the 2nd division of the Portuguese championship. The findings reveal that the most important players' physiological variables for predicting a win are Player Load /min, Distance m/min, Distance 0.3 m/s, Acceleration 0.2 m/s with an accuracy of 79%, using the XGBoost algorithm.

Keywords: Football · Winning · Physiological variables · Predicting · Machine learning · Classification

1 Introduction

Football is a team sport seen by several authors, as one of the most popular forms of sport in the world, as referred by Oliver et al. [1]: "football is the most popular sport in the world". As its popularity increases it comes under greater scrutiny, so the industry must have better control of training sessions, and most importantly, football games.

With the purpose of identifying which physiological variables of the players most contribute to winning a football match, a study was performed using Machine Learning (ML) algorithms on a dataset with the GPS positions of the players during football matches, sampled at 10 Hz (Playertek, Catapult Innovations, Melbourne, Australia [2]). The study aimed to both to determine the athletes' physiological variables, in a team, best contribute to the team's win, and the variables that most influence the wins by player position on the field.

Regarding this, was used this information, to apply ML algorithms. Various classification algorithms were used for understanding how these variables influence the victory

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 3–15, 2021.
https://doi.org/10.1007/978-3-030-86970-0_1

in the game. For the ML process the Cross Industry Standard Process for Data Mining (CRISP-DM) [3] process was selected, since it is considered the most relevant and comprehensive approach for this kind of analysis [3].

2 Related Work

It is widely known that technology fosters learning with generalized impact on society, and in sports it is not different: "the future of the sports industry lies in the hand of technology" [4]. Because of that, "sports activities have many opportunities for intelligent systems" [5].

As it can be easy understandable, data-driven methods can effectively overcome the subjective limitations (manual analysis) of the match and offer better results for football clubs [6]. As Baboota and Kaur state, "whatever form of Artificial Intelligence (AI) is used, it is evident that football is a sport that benefits from technical integration" [7].

ML has been employed by several authors in different areas of knowledge, for example, Žemgulys et al. [8] used ML techniques to recognize the hand signals made by a basketball referee in recorded games and achieved a 95,6% of accuracy, regarding three of the hand signals.Yang [4] defends, ML offers a contemporary statistical approach where algorithms have been specifically designed to deal with imbalanced data sets and enable the modelling of interactions between a large number of variables. As Knauf et al. [9] affirm in their study, this field of research, known as ML, is a form of AI that uses algorithms to detect meaningful patterns based on positional data. ML is a relatively new concept in football, and little is known about its usefulness in identifying performance metrics that determine match outcome.

Nonetheless, there are some studies regarding ML in football, such as a study that explores injury risk [1], studies about the categorization of football players and football training sessions [4, 11], and even one study that tries to evaluate the football players regarding their market value [12]. Regarding injury risk, the study done by Oliver et al. [1] aims to understand whether the use of ML improved the ability of a neuromuscular screen to identify injury risk factors in elite male youth football players. In categorization, there are some studies. García-Aliaga et al. [11] aimed to find out, by using ML methods, whether the technical-tactical behaviours of players according to position could be identified by their statistics, without including spatial-temporal descriptors. In addition, the authors of this work wanted to check the capacity of ML to identify the most influential variables in each of the positions and to find groups of anomalous players. Another model regarding categorization is the one proposed by Yang [4] to evaluate football training using AI, through the analysis of the evaluation indexes of the running ability of athletes at different positions, showing that the evaluation efficiency is 24.12% higher than that of traditional artificial team, which proves the feasibility of this model. In [12], authors propose a novel method for estimating the value of players in the transfer market, based on the FIFA 20 dataset. The dataset was clustered using the APSO-clustering algorithm which resulted in detecting four clusters: goalkeepers, strik-

ers, defenders, and midfielders. Then, for each cluster, an automatic regression method, able to detect the relevant features, was trained, and they were able to estimate the value of players with 74% accuracy. As illustrated by those examples, ML algorithms hold the potential to provide coaches and analysts with additional information to evaluate the game [10].

As GPS devices have become commonplace in professional football to track player performance. As some authors defend, the identification of internal and external load measures can provide answers about which variables to include in an integrated approach [13]. The amount of information about athletes collected by these GPS systems is increasing, making it difficult for users to analyse this information, so it is necessary to create analytical algorithms that effectively analyse it. Optimization methods, especially metaheuristic optimization algorithms, are powerful methods that can help the researchers to overcome the difficulties and challenges in different fields [12]. Data science has emerged as a strategic area that, supported by the great possibility of data production for analysis, allows knowledge discovery in sport science with the aim of filling some gaps that traditional statistical methods could not achieve [9]. The enormous advances in technology make it possible to process an exorbitant amount of data to draw extremely useful conclusions [5].One such area where predictive systems have gained a lot of popularity is in the prediction of football match results [5].

3 Methodology

In this work was used the CRISP-DM process, being especially useful in modelling. Laureano et al. [14] use this methodology to better model for predicting the waiting time for internments. This allowed them to identify the clinic attributes related to time of interment. Another study done in the medicine with this framework was done by Morais et al. [15] and they aim to better understand what characteristics of a new-born, brings the need of assistance for breathing at birth. This was done by analysing the characteristics of the mother and the pregnancy to prevent neonatal mortality.

The CRISP-DM framework has six phases [3]: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. This study followed these phases which are presented in this section.

For this study, a dataset was compiled with the physiological variables of football players recorded by a GPS system during the playing of football games of a Portuguese team from the 2nd Regional Division of AF Santarém in the 2018/2019 season. Besides the physiological variables, the dataset also contains the results of the games that will be the label/target variables of the study.

During the 2018/2019 the team with 28 players with an average age of 22 years old played 14 games in the first phase of the championship, 10 games in the second phase and 2 extra games in Ribatejo Cup. A total of 33748 different episodes were registered, regarding the different players, and games they played. ML algorithms were applied to these records, which results are presented in Sect. 4.

3.1 Business Understanding

The business goal of this work is the prediction of the physiologic variables that have more influence in the game winning. For this was consider, all the variables recorded by the GPS tracking system, some environment characteristics, the rate of perceived exertion (RPE) and the final score of the game. This prediction can have influence in training analysis, because the coaching staff can understand where they should apply more focus in training sessions to have a better performance on game day.

3.2 Data Understanding

Table 1 presents the variables used in this study that are present in the dataset.

Table 1. Dataset variables

Variables	Description
Game	Is the number of the game in the Championship games sequence
Final Score	Represents the points that the team won in the game, which relates to a victory, draw or loss. 0 it's for a lost game, 1 for a draw and 3 for a win
Minutes	It is the number of minutes that the players are actively in the game
RPE	It is the rate of perceived exertion, made by a numeric estimate of someone's exercise intensity. It is a way to measure how hard a person is exercising, which ranges from 1 (no exertion) to 10 (extremely hard)
Heart Rate (HR)	The maximum heart rate calculated as HRmax = 220 – age. It is calculated in absolute and %
Player Load	Calculated based on the acceleration data that are registered by the triaxial accelerometers. This variable, considered as a magnitude vector, represents the sum of the accelerations recorded in the anteroposterior, medio-lateral and vertical planes. Represented in Total and in Arbitrary Units (U.A.) per minutes
Distance (Total and m/s)	The Total distance provides a good global representation of volume of exercise (walking, running) and is also a simple way to assess individual's contribution relative to a team effort. It is divided in five different speed zones: "walking/jogging distance, 0.0 to 3.0 m/s; running speed distance, 3.0 to 4.0 m/s; high-speed running distance, 4.0 to 5.5 m/s; very high-speed running distance, 5.5 to 7.0 m/s; and sprint distance, a speed greater than 7.0 m/s [13]
Work Ratio (WRRatio)	It is used to describe footballer's activity profiles, which is divided in two categories: 1) pause if the distance is travelled at a speed < 3.0 m/s; and as 2) work if the distance is travelled at a speed > 3.0 m/s

(*continued*)

Table 1. (*continued*)

Variables	Description
Acceleration (Accel.)	Categorized based upon the acceleration of the movement, which is thought to represent the "intensity" of the action. It is divided in "low intensity", 0.0 to 2.0 m/s^2; "moderate intensity", 2.0 to 4.0 m/s^2; and "high intensity", greater than 4.0 m/s^2 [13]
Deacceleration (Deacc.)	Categorized based upon the deacceleration of the movement, which is thought to represent the "intensity" of the action. It is divided in "low intensity", 0.0 to −2.0 m/s^2; "moderate intensity", −2.0 to −4.0 m/s^2; and "high intensity", greater than −4.0 m/s^2 [13]

3.3 Data Preparation

In this phase of the CRISP-DM process, the data was selected and prepared to be used by the ML models. First the dataset was scanned for null or inconsistent information.

After cleaning the data, a final transformation was done to support the two goals of the study, which created two different datasets. One dataset where the original dataset was grouped by game (computing the means for the variables), which held 676 lines and a second dataset where the original dataset was grouped by player position so that it would be possible to identify the best variables by player position on the field and was use a dataset of 10725 episodes.

3.4 Modelling

In the modelling phase, two models were created to study the importance of physiological variables of the athletes on the outcome of the football game. The first model evaluates the team as a whole, while the second model evaluates the players' positions on the field of play.

Given that the objective of our game focuses on winning the match, it does not matter to distinguish between draw and defeat, so the binary variable win/lose seems to us to be the most interesting, with draws being coded as defeat. Thus, a new label/target variable which stores the information regarding the Win of the game (1 – victory; 0 - other) was added to the original dataset. This variable was computed from the variable with the outcome of the game, "final score" (see Table 1).

3.4.1 Variables that Most Contribute to Game Victory

Several models with different features were hypothesized, both to determine which variables as a whole best contribute to the team's win and to determine which variables most influence the wins by player position on the field.

To better evaluate the variables and understanding which were the most relevant in predicting a Win, four models with different sets of variables where created (see Table 2). In the first model was chose to apply all the variables to have a better understanding how those variables all together will affect the game winning. For the second model, were

removed all the variables that had big correlation between them, and those that show high intensity for the game. Those variables that were related to high intensity were used in the third model, to understand how those variables, influence the game result. For the fourth model, were chose the variables that showed more correlation between them and victory, and the variables defended by some authors [16], like Player Load/min.

Table 2. ML models

Model	Features
Model 1	*['RPE', 'HR', '%HR', ' < 60%HR', '60-74,9%HR', '75-89,9%HR', ' > 90%HR', 'Player Load', 'Player Load.UA/min', 'Distance Total', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance. > 7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel. > 4', 'Deacc.0-2', 'Deacc.2-4', 'Deacc. > 4']*
Model 2	*['60-74,9%HR', '75-89,9%HR', 'Player Load', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Accel.0-2', 'Accel.2-4', 'Deacc.0-2', 'Deacc.2-4']*
Model 3	*[' > 90% HR', 'Distance. > 7', 'WRRatio', 'Acceler. > 4', 'Deacc. > 4']*
Model 4	*['Player Load.UA/min', 'Distance.m/min', 'Distance.0.3', 'Acceler.0.2']*

3.4.2 Physiological Variables that Most Contribute to Game Victory by Player Position

To achieve this second goal, the dataset was divided by position on the field, and was used a Recursive Feature Elimination (RFE) [17] for each position, to rank the feature variables related to winning.

The variables selected by RFE allowed to understand the different effects of a football games in the different positions, and how these variables affect winning football games. The variables are different for each position, and they are related to the specific demands. The select variables for each position are presented in Table 3 for each player position on the field.

In the scope of this work the Goalkeeper position was excluded because of the specific physiological demands regarding the positions. In this work, the following positions were used, Central Defender (CD), Full Back (FB), Midfielder (MC), Offensive Midfielder (OM), Winger (W) and Forward (F).

Table 3. ML position models

Position	Features
CD	*['Distance.m', 'Distance.0–3', 'Distance.3–4', 'Distance. > 7']*
FB	*[' < 60.0%HR', '60–74,9%HR', 'Player Load.UA/min', 'WRRatio']*
MC	*['Player Load.UA/min', 'Distance.m/min', 'WRRatio', 'Aceler. > 4']*
OM	*[<'60.0%HR', '60–74,9%HR', 'Player Load.UA/min', 'WRRatio']*
W	*[' < 60.0%HR', '60–74,9%HR', '75–89,9%HR', 'Player Load.UA/min']*
F	*['Distance.m', 'Distance.4–5.5', 'Distance. > 7']*

3.5 Evaluation

This section presents the results obtained for the two models.

3.5.1 Variables that Most Contribute to Game Victory

Since the goal is to determine which variables contribute to the win of the game classifications algorithms were chose. Thus, six ML algorithms were selected: Naïve Bayes (NB), K-Nearest Neighbours (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Extreme Gradient Boosting (XgB).

For each ML algorithm the dataset was divided into training and testing. Training dataset was composed by 66% of all the data, the remaining was used for testing.

The performance of each ML algorithm was assessed through a confusion matrix, which shows the present results of True Positives, False Positives, True Negatives and False Negatives. With these, it was possible to calculate the sensitivity, specificity and accuracy of each algorithm, in order to evaluate their performance regarding the 4 models.

Table 4, 5 and 6, and Fig. 1 present the results of the ML algorithms for the 4 models with and without Cross-Validation (CV). For the results with CV was also calculated the Standard Deviation (StD). The best model was Model 4, which was presented with a chart to better understand the results (see Fig. 1). All the data was used for testing using 7-fold CV.

Model 1 (Table 4) has a mean accuracy of 0.476 without CV, and 0.560 with CV. The ML algorithms with the best accuracy without a CV were DT and NB with 57% and with CV were RF and SVM with 62% both.

Table 4. Results of ML algorithms for Model 1

Metrics/Algorithms	KNN	DT	RF	NB	SVM	XgB
Accuracy	**0.571**	**0.571**	0.429	**0.571**	0.286	0.429
Accuracy with CV	0.600	0.520	**0.620**	0.400	**0.620**	0.600
StD Accuracy with CV	0.330	0.140	0.080	0.330	0.280	0.220
Sensitivity	0.500	0.500	0.750	0.500	0.500	0.250
Specificity	0.333	0.667	0.667	0.667	0.667	0.667

Table 5 presents the results for Model 2 with a mean accuracy of 0.429 without CV, and 0.552 with CV. The best ML algorithm for this model regarding accuracy was KNN with a result of 57% without CV and XgB with a result of 76% with CV.

Table 5. Results of ML algorithms for Model 2

Metrics/Algorithms	KNN	DTC	RFC	NBC	SVMC	XgB
Accuracy	**0.571**	0.286	0.429	0.429	0.429	0.429
Accuracy with CV	0.600	0.620	0.400	0.450	0.480	**0.760**
StD Accuracy with CV	0.220	0.190	0.280	0.340	0.290	0.230
Sensitivity	0.500	0.000	0.750	0.250	0.250	0.500
Specificity	0.667	0.667	0.000	0.667	0.667	0.333

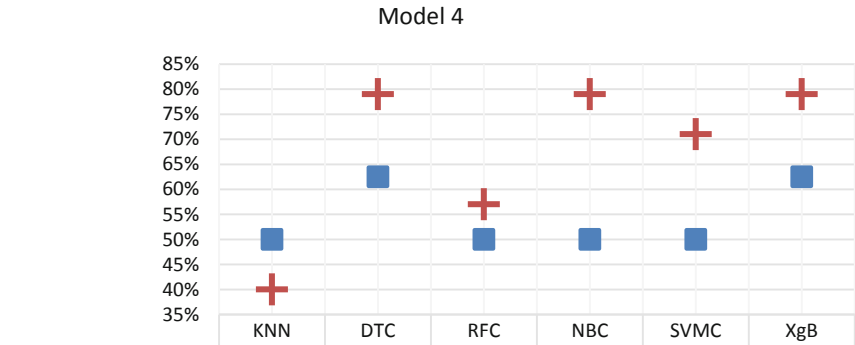
Model 3 has the worst mean results for accuracy (see Table 6) from all the models with 0.429 without CV and 0.340 with CV. The algorithms with the best accuracy without CV were RF and SVM 57% and with CV was NB with a result of 45%.

Table 6. Results of ML algorithms for Model 3

Metrics/Algorithms	KNN	DTC	RFC	NBC	SVMC	XgB
Accuracy	0.429	0.286	**0.571**	0.429	**0.571**	0.286
Accuracy with CV	0.290	0.210	0.360	**0.450**	0.400	0.330
StD Accuracy with CV	0.190	0.190	0.360	0.290	0.330	0.240
Sensitivity	0.200	0.000	0.800	0.600	0.400	0.000
Specificity	1.000	1.000	0.000	0.000	1.000	1.000

The last model used in this analysis was Model 4 (Fig. 1) that held the best overall results, and the best individual ML algorithm results. This model has a mean accuracy of 0.542 without CV and 0.675 with CV. The two algorithms with the best results were

DTC and XgB with a 63% of accuracy without CV and an accuracy of 79% with CV. This model reveals the most important variables that could help predict a football game win.



	KNN	DTC	RFC	NBC	SVMC	XgB
Accuracy	50%	63%	50%	50%	50%	63%
Accuracy with CV	40%	79%	57%	79%	71%	79%
StD Accuracy with CV	8%	19%	28%	19%	19%	19%
Sensitivity	43%	50%	57%	50%	50%	50%
Specificity	100%	100%	0%	50%	50%	100%

Fig. 1. Plot of the ML algorithm results for Model 4

From the results obtained for all the four models using the ML algorithms, the algorithms that provided best mean results were SVMC and XgB (see Table 8).

Table 8. Average results of ML models with CV

Models	KNN	DTC	RFC	NBC	SVMC	XgB
Model 1	0.600	0.520	0.620	0.400	0.620	0.600
Model 2	0.600	0.620	0.400	0.450	0.480	0.760
Model 3	0.290	0.210	0.360	0.450	0.400	0.330
Model 4	0,400	0,790	0,570	0,790	0,710	0,790
Average with CV	47%	54%	49%	52%	**55%**	**62%**

Therefore, these two algorithms were chosen for the second model of the study, which objective is to detect which physiological variables most influence the game win by the player position on the field (see Sect. 3.4.2).

3.5.2 Physiological Variables that Most Contribute to Game Victory by Player Position

The best results obtained with the ML algorithms were relative to the models for the Forwards position with an accuracy of 74% for SVMC and 62% for XgB (Table 9). All the positions achieved over 50% with at least one of the algorithms.

Table 9. Players positions study with two ML algorithms and specific variables

Position	SVMC	XgB	Features
CD	66%	62%	*['Distance.m', 'Distance.0-3', 'Distance.3-4', 'Distance. > 7']*
FB	63%	61%	*[' < 60.0%HR', '60-74,9%HR', 'Player Load.UA/min', 'WRRatio']*
MC	63%	59%	*['Player Load.UA/min', 'Distance.m/min', 'WRRatio', 'Aceler. > 4']*
OM	52%	62%	*[<'60.0%HR', '60-74,9%HR', 'Player Load.UA/min', 'WRRatio']*
W	55%	56%	*[' < 60.0%HR', '60-74,9%HR', '75-89,9%HR', 'Player Load.UA/min']*
F	**74%**	**62%**	*['Distance.m', 'Distance.4-5.5', 'Distance. > 7']*

The results obtained were not very good because only one model obtained a good result that was the Forwards (F) model with a result around 70%. Nonetheless, all the results were higher than 50%, which shows that these models for the players' positions can be a contribution for the football coach.

4 Discussion

This study aimed to identify which physiological variables have the most influence on winning the game by analysing them, using ML algorithms. With that objective, several models with different features/variables were created.

These models were constructed based upon the correlation between all the variables and the understanding of the variables regarding the intensity applied in each game. For this, the first model was composed by all the variables measured, while for the construction of the second model the same set of variables was used excluding the ones with greater correlation between them and the ones used in the third model. The third model was constructed to understand if high intensity had a better association with the victory, and that was the model with the worst results. Finally, the fourth model was created with the variables defend by some authors [18] in previous studies and those that had some correlation with wining. This model was the one who got better results with an accuracy of 79% with three of the algorithms using a CV. An important finding of

this study was that the high intensity is not a key factor in game winning, which is not in agreement with the results presented in the study by Altavilla, where it is claimed that that "high intensity covered distance has traditionally been identified as a key indicator of the physical performance during the matches" [18].

This study allowed to verify that teams that kept their players more active, in the end will have better results. It is also possible to verify that the greater the distance covered in this championship, the better the proportion of victories. These results are made by the average values of the team, so with this they can only be applied to a team, not to an individual player.

In the second part of study the physiological variables of the players by player position on the field were analysed, this has revealed that the physiological variables that contribute to winning games vary according to the players' positions on the field and differ from the team as a whole. Some of these variables were identified and confirmed in different studies, as the key variables, for the positions on the field. As was seen in this study, distance covered at max speed (*distance.* > 7) by Forwards (F) was one of the variables that has a bigger impact/influence in winning. Almulla et al. [19] and Baptista et al. [20] defended that the distance covered by forwarders at very high speed from the winning team was higher than the losing one. The variables chosen for the Forward position presented the best results in predicting a win in this context, with a 70% in SVMC and 62% in XgB.

One of the variables chosen to apply in the Central Midfielders (CM) *was distance covered per minutes,* which is backed by Borghi et al. and Altavilla et al. [18, 21] which refer that Central Midfielders (CM) have the highest value in total distance covered during games. In the same study Borghi et al. [21] affirm that Central Back (CB) covered the shorter distances during match, and this is important because, *Distance 0.3*, was one of the variables chosen to be used in this study for this position.

Another variable that was chosen by the RFE and was the *Player Load U.A./min* for the Full Back (FB) and Winger (W) but was not selected for the Central Back (CB). This was corroborated in a study done by Baptista et al., where they defended that Central Back (CB) had less turns per match than the Full Back (FB) and Winger (W) [20]. The same author also defended that FB covered more high intensity and sprinting distance than CB during the matches [20]. This was also revealed in our study because *Work-Ratio* was one of the variables selected for the Full Back (FB). The variables selected for the Full Back (FB). The variables selected for the FB position presented the second-best results in predicting a win in this context, with a 63% in SVMC and 61% in XgB.

5 Conclusions

Modern football competition has the characteristics of fierce confrontation, long duration, intensity of the game, large amount of exercise, and high technical and tactical requirements [5]. Because of these characteristics, it can be easily understandable that football is a highly complex sport. Selecting and evaluating properly the team players is an important and difficult task in multiplayer sports such as football that has many different aspects to be considered [5].

Therefore, studies in this area may play an important role in terms of evaluating the potential performance of teams and players. It is important to stress out that teams

should be analysed as a whole in order to identify important variables that are correlated with a match winning outcome. The individual data, for instance, can be very important for predicting players performance and injuries with time, not in the scope of this work.

The results obtained in this study can be used by the coaching staff of a football team as they help to prepare the team, as a whole, in relation to the most important variables, also helping to identify by player's position on the pitch which are the most important variables for the players' performance. Bottom line, the use of ML applied to training and team selection can represent a major step forward in the evolution of the sport, providing possibly critical information to the training staff that will allow in the end to help to better understand the demands of the positions on the field.

This study was conducted in only one team, being also a reduced sample. This represents an important limitation to the study, but at the same time can serve as steppingstone to other studies to help a better understanding of the game.

In the future, it is intended to further analyse different teams and therefore improve our sample set. This will enable to get a more complete picture of what the most important variables are and correlate them with the tactical decisions of the teams. In addition, it is intended to perform this analysis in different competition contexts and verify if these variables still maintain the same importance, if some of the contextual variables changed.

Without a doubt that in the future AI will contribute to improve the results of football teams. Still, it is important to understand that when a human element is involved in sport, there will always be unpredictability and uncertainty that makes it fascinating and surprising for its viewers [22].

References

1. Oliver, J.L., Ayala, F., de Ste Croix, M.B.A., Lloyd, R.S., Myer, G.D., Read, P.J.: Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. J. Sci. Med. Sport **23**(11), 1044–1048 (2020). https://doi.org/10.1016/j.jsams.2020.04.021
2. Catapult Innovations, "Playertek" (2021). https://www.playertek.com/gb/. Accessed 12 Apr 2021
3. Jaggia, S., Kelly, A., Lertwachara, K., Chen, L.: Applying the CRISP-DM framework for teaching business analytics. Decis. Sci. J. Innov. Educ. **18**(4), 612–634 (2020). https://doi.org/10.1111/dsji.12222
4. Yang, Y.: Evaluation model of soccer training technology based on artificial intelligence. J. Phys. Conf. Ser. **1648**, 042085 (2020)
5. Maanijou, R., Mirroshandel, S.A.: Introducing an expert system for prediction of soccer player ranking using ensemble learning. Neural Comput. Appl. **31**(12), 9157–9174 (2019). https://doi.org/10.1007/s00521-019-04036-9
6. Kusmakar, S., Shelyag, S., Zhu, Y., Dwyer, D., Gastin, P., Angelova, M.: Machine learning enabled team performance analysis in the dynamical environment of soccer. IEEE Access **8**, 90266–90279 (2020). https://doi.org/10.1109/ACCESS.2020.2992025
7. Baboota, R., Kaur, H.: Predictive analysis and modelling football results using machine learning approach for English Premier League. Int. J. Forecast. **35**(2), 741–755 (2019). https://doi.org/10.1016/j.ijforecast.2018.01.003
8. Žemgulys, J., Raudonis, V., Maskeliūnas, R., Damaševičius, R.: Recognition of basketball referee signals from real-time videos. J. Ambient. Intell. Humaniz. Comput. **11**(3), 979–991 (2019). https://doi.org/10.1007/s12652-019-01209-1

9. Knauf, K., Memmert, D., Brefeld, U.: Spatio-temporal convolution kernels. Mach. Learn. **102**(2), 247–273 (2015). https://doi.org/10.1007/s10994-015-5520-1
10. Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., Meyer, T.: Machine learning in men's professional football: current applications and future directions for improving attacking play. Int. J. Sports Sci. Coach. **14**(6), 798–817 (2019). https://doi.org/10.1177/1747954119879350
11. García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., Luengo-Sánchez, S.: In-game behaviour analysis of football players using machine learning techniques based on player statistics. Int. J. Sports Sci. Coach. (2020). https://doi.org/10.1177/1747954120959762
12. Behravan, I., Razavi, S.M.: A novel machine learning method for estimating football players' value in the transfer market. Soft. Comput. **25**(3), 2499–2511 (2020). https://doi.org/10.1007/s00500-020-05319-3
13. Miguel, M., Oliveira, R., Loureiro, N., García-Rubio, J., Ibáñez, S.J.: Load measures in training/match monitoring in soccer: a systematic review. Int. J. Environ. Res. Public Health **18**(5), 2721 (2021)
14. Laureano, R.M.S., Caetano, N., Cortez, P.: Previsão de tempos de internamento num hospital português: Aplicação da metodologia CRISP-DM. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao (13), 83–98 (2014). https://doi.org/10.4304/risti.13.83-98
15. Morais, A., Peixoto, H., Coimbra, C., Abelha, A., Machado, J.: Predicting the need of neonatal resuscitation using data mining. Procedia Comput. Sci. **113**, 571–576 (2017). https://doi.org/10.1016/j.procs.2017.08.287
16. Soto, P.R., Nieto, D.C., Suarez, A.D., Ortega, J.P.: Player load and metabolic power dynamics as load quantifiers in soccer. J. Hum. Kinetics **69**, 259 (2019). https://doi.org/10.2478/hukin-2018-0072
17. Arndt, C., Brefeld, U.: Predicting the future performance of soccer players. Stat. Anal. Data Min. **9**(5), 373–382 (2016). https://doi.org/10.1002/sam.11321
18. Altavilla, G., Riela, L., di Tore, A.P., Raiola, G.: The physical effort required from professional football players in different playing positions. J. Phys. Educ. Sport **17**(3), 2007–2012 (2017). https://doi.org/10.7752/jpes.2017.03200
19. Almulla, J., Alam, T.: Machine learning models reveal key performance metrics of football players to win matches in qatar stars league. IEEE Access **8**(December), 213695–213705 (2020). https://doi.org/10.1109/ACCESS.2020.3038601
20. Baptista, I., Johansen, D., Seabra, A., Pettersen, S.A.: Position specific player load during matchplay in a professional football club. PLoS ONE **13**(5), 1–11 (2018). https://doi.org/10.1371/journal.pone.0198115
21. Borghi, S., Colombo, D., la Torre, A., Banfi, G., Bonato, M., Vitale, J.A.: Differences in GPS variables according to playing formations and playing positions in U19 male soccer players. Res. Sports Med. **29**(3), 225–239 (2020). https://doi.org/10.1080/15438627.2020.1815201
22. Rathi, K., Somani, P., Koul, A.V., Manu, K.S.: Applications of artificial intelligence in the game of football: the global perspective. Res. World **11**(2), 18–29 (2020). https://doi.org/10.18843/rwjasc/v11i2/03