Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 *versão eletrônica*

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

www.ibpefex.com.br / www.rbff.com.br

MACHINE LEARNING APPROACHES TO PREDICT THE MATCH RESULT: BRAZILIAN FUTSAL LEAGUE CASE

Denio Duarte[1], Jefferson Alexandre Coppini[1]

ABSTRACT

The use of machine learning approaches in sports has been grown in the last decade. Sports analytics, outcome match results, and possible player's injury are examples of machine learning applications. Accordingly, this work aims to use machine learning techniques to build models to predict FutSal National League (LNF) results (win/loss/draw) based on data collected in the first half of a match. To accomplish that, we extract the data from the LNF website, and, based on the data, we propose six new features using the concept of team strength. The data correspond to the 2016 to 2019 seasons. The models are built usimg machine learning approaches, and they are validated through an accuracy metric. We build ten models, and the predictions are organized as follows: the individual performance of each model and a voting approach (committee) based on the majority of the predicted results. The results show that the individual models get better performance when predicting a single result (e.g., home win) with 95% accuracy. On the other hand, the committee gets a better performance regarding the overall results. The win, loss, and draw results reach almost 79% accuracy.

Key words: Futsal. Supervised Machine Learning. Prediction Models.

RESUMO

Utilizando abordagens de aprendizado de máquina para prever resultados de jogos: o caso da liga nacional de futsal

O uso de técnicas de aprendizado de máquina na área esportiva cresce dia a dia. Áreas como análise esportiva, previsão de resultados e prevenção de contusões se apoiam cada mais nessas técnicas para obterem resultados mais eficazes. Neste contexto, este trabalho tem como objetivo prever os resultados de partidas de futsal da Liga Nacional de Futsal (LNF) (vencedor mandante, vencedor visitante e empate) utilizando os dados gerados durante o primeiro tempo da partida. Os dados foram extraídos do sítio da LNF e, além, dos atributos extraídos, seis novos atributos baseados na força dos times foram propostos. Os dados correspondem às temporadas de 2016 a 2019. A previsão dos resultados é feita através de modelos construídos por algoritmos de aprendizado de máquina. A validação do modelo foi feita através da acurácia dos resultados de previsão. Foram criados dez modelos de previsão e os resultados foram organizados da seguinte forma: o desempenho individual de cada modelo e um comitê de votação em que o resultado mais votado é o resultado utilizado na previsão. Resultados apontam que os modelos individuais possuem melhores desempenhos em prever um resultado específico (e.g., vitória do mandante) chegando a 95% de acurácia. Por outro lado, o comitê obteve um melhor desempenho nos resultados agrupados, chegando a quase 79% de acurácia.

Palavras-chave: Futsal. Aprendizado de Máquina Supervisionado. Modelo de Predição.

1 - Universidade Federal da Fronteira Sul, Campus Chapecó, Chapecó, Santa Catarina, Brasil.

E-mail dos autores:
duarte@uffs.edu.br
jefferson.coppini@schumann.com.br

Autor para correspondência:
Denio Duarte.
duarte@uffs.edu.br
Universidade Federal da Fronteira Sul
Campus Chapecó, Santa Catarina, Brasil.
Rodovia SC 484 - Km 02, Fronteira Sul.
CEP: 89815-899.

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 *versão eletrônica*
Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício
www.ibpefex.com.br / www.rbff.com.br

INTRODUCTION

Predicting match events in football has drawn significant attention over the past years. The prediction of individual matches' outcomes is very challenging because the amount of variables that affects the match situation is significant (Anderson and Sally, 2013).

For example, the number of fouls, number of yellow and red cards, and substitutions. Futsal matches, as a variant of football, follow the same challenges, including that substitutions are on the fly and unlimited, each team is allowed to use one timeout each period, and direct free kicks are counted as accumulated fouls (Frencken and Lemmink, 2009).

Machine learning (or data mining) has been applied to help teams and managers to make sound decisions to enhance the overall teams and players performance. This field has overgrown in the past decade, mainly because of the available data and powerful tools (Fried and Mumcu, 2016). For example, Kaggle (www.kaggle.com) provides datasets, like FIFA FutsalWorld Cup, with data about futsal matches. Besides, machine learning approaches offer techniques for learning patterns from data and, so, build models that describe the input data (Duarte and Ståhl, 2019).

We are also witnessing the growth of sports betting; the number of bookmakers, who offer the opportunity to bet on the outcome of football macthes, has grown as well thanks to the development of the world wide web (Stübinger, Mangold, Knoll, 2020).

Predicting a match outcome can help both team managers and bookmakers (and wagers). Team managers can use the prediction to either strengthen the team strategy if the team is successful in its strategy or change the strategy, otherwise. Bookmakers (and wagers) can rely on their strategy regarding the predicted result (Horvat and Job, 2020), e.g., raising or lowering the monetary rewards.

Recently, several studies have applied machine learning techniques to sports. For example, Tsunoda et al., (2017) propose a play recognition method for passes, shoots, and dribbles in Futsal by convolution neural network (CNN) using videos of multiple cameras. However, their method's accuracy is 70% at most, i.e., seven of ten predictions are right.

Imai et al., (2018) claim that to improve recognition performance, it is essential to combine multiple information sources such as videos and wearable sensors.

On the other hand, Van Haaren and Van den Broeck (2015) propose a learning task comprising the prediction of the goal difference for individual football matches. The goal difference is determined by the home score minus the away score.

Constantinou and Fenton (2017) applied Dynamic Bayesian networks to develop a model that generates accurate predictions of football teams' evolving performance. The model enables to predict, before a season starts, the total league points a team is expected to accumulate throughout the season.

Nevertheless, Flôres et al., (2019) built a statistical model to predict the effectiveness of football substitutions, and they conclude that substitutions are essential for coaches to improve team performance.

Baboota and Kaur (2019) use random forest and gradient boosting models to help the bookmaker's predictions. The proposed approach was not able to outperform the bookmakers' predictions; however, they claim that incorporating factors such as information about injured players and a key player's presence could help their approach to outperform bookmaker's predictions.

We refer the readers to (Bunker and Thabtah, 2019) and (Horvat and Job, 2020) for a review on machine learning applied to sports.

The present work intends to predict the result (win/draw/loss) of futsal match of Liga Nacional de Futsal do Brasil (LNF) based on data available for the first half of the match and historical data from the teams. We extract the data from 2016 to 2019, corresponding to 598 matches.

From this data, we extract the most informative features, and six new features are proposed. Using the built dataset, we apply machine learning approaches to predict the match result. Ten different approaches are applied, and we investigate their performance in predicting the results correctly.

Moreover, we use the ten build models to create a committee for voting a given match result.

The result with a large number of votes is chosen as a predicted result.

MATERIALS AND METHODS

Machine Learning

Machine learning is a sub-field of computer science that aims to build models from input data. The built model describes the input data to foresee the outcome of unseen data.

Therefore, the learning process receives a dataset as inputs and builds a model that predicts the classes representing the data. The learning task can be supervised or unsupervised. The former receives the data with a label for every example, the latter the examples have no label.

The supervised learning is divided into classification and regression. If the labels are discrete (classes), we work with a classification approach; otherwise, if the labels are continuous values, a regression approach must be applied.

The dataset used in this work consists of a set of examples labeled with the result of the match (i.e., win, draw, or loss), so we applied classification supervised learning approaches. Within those approaches, we choose ten different algorithms to build a predicting model: (i) Decision Tree (DT), (ii) Random Forest (RF), (iii) Ada Boost (AB), (iv) k-Nearest Neighbors (KNN), (v) Multilayer Perceptron (MLP), (vi) Gradient Boosting (GB), (vii) Support Vector Classification (SVC), (viii) Nu Support Vector Classification (NUSVC), (ix) Linear Support Vector Classification (LSVC), and (x) Gaussian Naive Bayes (GNB). We use scikit-learn API (Pedregosa et al., 2011) to implement the models.

We show the best algorithm for our proposal and a voting method: each model's prediction is used to identify the best outcome; we call that voting committee. The predictions of the models are evaluated using an evaluation metric. We choose accuracy, and it can be described as follows: given w_p, d_p, and l_p as the number of wins, draws, and losses correctly predicted, respectively; and w_r, d_r, and l_r as the real number of wins, draws, and losses, respectively. The accuracy of a model is:, that is, the number of correct class predictions over the number of the real classes (Duarte and Sthål, 2019). Note that if the accuracy is 1.0, it means that all classes are correctly predicted.

Dataset and Feature Extraction

The dataset was extracted from the LNF website (ligafutsal.com.br) using a web scraper. The web scraper extracted data from futsal matches from 2016 to 2019. The built dataset consists of 598 matches. We use the matches from 2016 to 2018 to train our models and the matches of 2019 to test them.

The original dataset comprises 39 features, 38 of them describe the match, and one stores the match result with three possible values: local winning, away winning, or draw. As we are interested in the first half of a match's features, we retain 16 of them (and the label: match result). Table 1 presents the retained features.

Remark that some features are originally alphanumeric, and we have to transform them into numeric features. For example, the dataset label (Match result), which initially stored draw, home win, or away win, stores now 0, 1, or 2, respectively. We use selectKBest scikit-learn method for getting the importance of the features in the dataset to build the models. In Table 1, the features are ranked according to their importance.

Note that the first-half goals are the most important features to build the prediction model. This is expected since the goals are often a good indicator of which team will win the match. Interestingly, the features of technical time-out are placed in the top-5 features and the home team's identification.

Another interesting point is that the number of yellow and red cards does not play an essential role in building a prediction model.

The intuition behind the yellow card's unimportance is that it does not change the match's behavior, and the red card, differently from football, another player can replace a red-carded player after two minutes.

Table 1 - Features corresponding to the first half of LNF match.

Id	Feature Name	Type
1	Away team goals	number
2	Home team goals	number
3	Home team technical time-out	number
4	Away team technical time-out	number
5	Home team identification code	number
6	Number of substitution away team	number
7	Number of yellow cards home team	number
8	Number of substitution home team	number
9	Number of faults home team	number
10	Number of yellow cards away team	number
11	Number of round	number
12	Competition phase	number
13	Number of red cards home team	number
14	Away team identification code	number
15	Number of red cards away team	number
16	Number of faults away team	number
label	Match result	number

Table 2 shows an extract of the first-half dataset (the first four rows of the dataset). The columns' names correspond to the features Id of Table 1. Note that the results (column label) of those examples (rows) are home win, away win, draw, and home win, respectively. In the second row, we can see that the home team was winning in the first half, the other match results corresponding to the draw.

Table 2 - Extract from the dataset before post-processing.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	label
0	0	1	1	3	6	2	5	0	1	1	1	0	4	0	0	0
0	2	1	1	1	5	1	5	0	1	1	1	0	2	0	0	1
0	0	0	1	15	5	0	5	0	1	1	1	0	12	0	0	2
0	0	1	0	16	5	0	5	0	0	1	1	0	8	0	0	0

The number of features impacts the machine learning algorithms computational performance (Duarte and Ståhl, 2019), so we decide to use the top-5 original features to build the model.

However, we propose six new features to help the algorithms to build better models.

Table 3 - Variables used to calculate the proposed features.

Id	Description
tm	total number of matches
tw	total wins
td	total draws
dwin	# of draws becomes win
draw	# of draws from losses
gwin	# of wins from losses
wwin	# of wins from wins
kdraw	# of draws (score 0 x 0)

Given the variable shown in Table 3, the proposed features are based on three concepts: Performance: this concept measures the performance of the teams in seasons 2016, 2017, and 2018. The performance is measured according to the win and draws in all considered seasons, and it is Given as follows:

$$P=((tw \times 3 + td) \times 100) \div (tm \times 3)$$

Keep result: this concept considers the number of matches that the team keeps the result in the second half. Here we consider the draw or win result, and it is given as follows:

$$R=((wwin \times 100) \div tm) \times 3 + ((kdraw \times 100) \div tj)$$

Reaction (R): it is similar to KR, but considering that the team was losing in the first half, the result was a win or a draw. It is given as follows:

$$KR=((dwin \times 100) \div tm) + ((draw \times 100) \div tj) \times 2) + (((gwin \times 100) \div tj) \times 4)$$

The intuition behind these concepts is: (i) the number of wins is three times more important than the number of draws for the performance concept, (ii) the number of matches that a team starts winning and ends winning is three times more important than the matches in which a team keeps the draw result for the keep result concept, and (iii) the number of loss to win is weighted as four, the number of loss to win is weighted as two for the reaction concept.

These three concepts help us to build the six new features:

HomeStrength (HS) and AwayStrength (AS): these features represent both teams' strength during a given match. Their values depend on the result of the match in the first half:

Draw: HS and AS are set to P×3+KR×2+R
Home Winning: HS=KR×2 and AS=R×2
Home Losing: HS=R×2 and AS=KR×2

We conduct an empirical experiment to identify the ideal weights for the above equations. Several values were tested, and the best ones were three as the weight for P, two for KR, and 1 for R in equations when there is a draw.

Otherwise, KR and R get the same weight, i.e., two. Note that in the case of one of the teams winning in the first half, KR or R are used since they represent the team's strength to keep the result or change it, respectively.

Another two features are built based on the goals scored:

Difference of goals of the home team (HTDiff) the away team (ATDiff):

HTDiff=GoalsHomeTeam - GoalsAwayTeam
ATDiff= HTDiff × -1 (change the positive to negative or vice-versa)

Finally, the two last new features are based on HS, AS, HTDiff, and ATDiff attributes and correspond to the difference of strength of the two teams:

Difference of strength of home team: DSH=HS × HTDiff

Difference of strength of away team: DSA=AS × ATDiff

After the selection of the attributes (the original and new ones), we split the dataset into two subsets: (i) the training set composed of 564 matches and corresponds to seasons 2016, 2017, and 2018, and (ii) the test set composed of 34 matches and corresponds to matches of season 2019. The training set is used to train the models and the test set to assess the models.

Table 4 shows the same extract as Table 2 but with the six new features: HS, AS, HTDiff, ATDiff, DSH, and DSA. Note that in the third row, the away team strength (AS) was smaller than the home team strength (HS) (66 and 172); howeverm the away team was winning in the first half of the match.

Table 4 - Extract from the dataset used to build the models.

HS	AS	HTDiff	ATDiff	DSH	DSA	1	2	3	4	5	label
278	333	0	0	0	0	0	0	1	1	3	0
172	66	2	-2	344	-132	0	2	1	1	1	1
252	328	0	0	0	0	0	0	0	1	15	2
311	202	0	0	0	0	0	0	1	0	16	0

The training phase is the more time-consuming phase: the models must be tuned (e.g., the definition of the best hyperparameters and cross-validation batches definition), the result must be analyzed, and the models might be rebuilt based on the analysis.

RESULTS

We conduct the test phase to identify the performance of the proposed model. Every algorithm was executed ten times, and we report the averages of the executions. In the test dataset, there are 20 home winnings, 11 away winnings, and three draws.

Figure 1 shows the performance of every algorithm based on the accuracy metric. Analyzing the bar plot, Random Forest (RF) and Gradient Boosting (GB) are the best models regarding the others, with a prediction rate of 79.4% for both. Observe that the bars represent the overall performance, i.e., the percentage of correct predictions of all results (win, loss, or draw) over the observed ones. Ada Boost (AB) and Gaussian Naïve Bayes (GNB) get the worst performance with 64.70% and 67.64%, respectively.

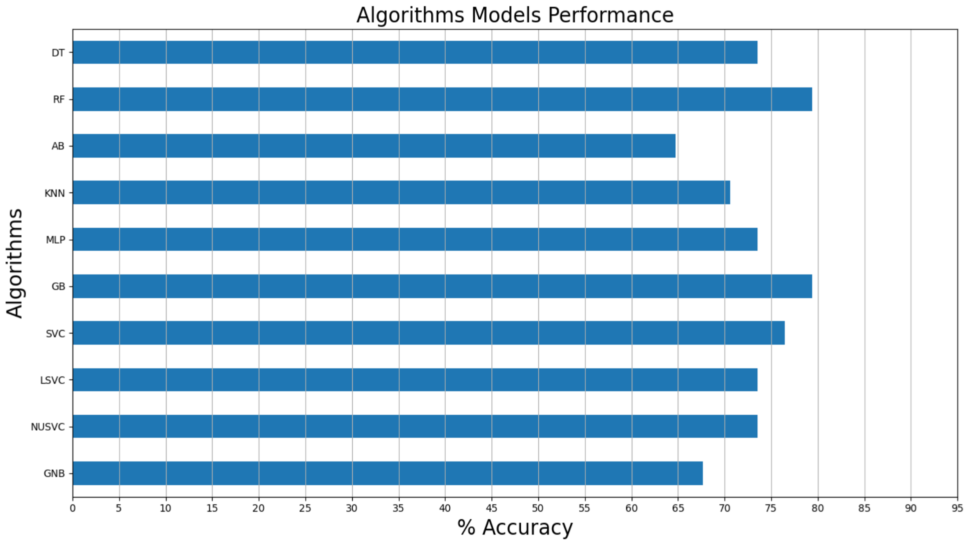

Figure 1 - Overall Models Performance.

Table 2 shows the performance of the algorithms by predicting each class based on the accuracy metric. In bold, we highlighted the best values. Remark that Gradient Boosting (GB) and Random Forest (RF) get an excellent result for Home Winning (95%), but they perform poorly for Away Winning (55%). However, as they get the best results for Draw and Home Winning, their overall performances are the best among the other models (see Figure 1).

Looking at Table 2, we can see that to predict Draw and Away Winning is the most challenging task.

Table 2 - Algorithms performance by class.

Model	Draw	Home	Away
DT	0.33	0.85	0.64
RF	0.67	0.95	0.55
AB	0.67	0.70	0.55
KN	0.33	0.85	0.55
MLP	0.33	0.90	0.55
GB	0.67	0.95	0.55
SVC	0.33	0.90	0.64
LSVC	0.33	0.85	0.64
NUSVC	0.67	0.85	0.55
GNB	0.33	0.75	0.64

When we apply the voting approach, i.e., the predicted result is given by the most frequent result predicted by the models, we have better results than using the individual ones.

Figure 2 shows the results separated by class and the overall result (Total). The draw is the most challenging class to predict; only half results are correctly predicted using the committee. This performance corresponds to the flip of a coin, i.e., randomly picking a Draw as a result of a match.

Using the committee instead of a specific model increases the odds of predicting the correctly Away Winning; however, it decreases the odds of predicting the Home Winning and the Draw results. For example, RF can predict correctly 95 out of 100 matches in the case of Home Winning.

This performance is useful for team managers and bookmakers. In the case of the home team performance in the first half leads to a winning (based on the model), the team manager could keep the team strategy with

95% accuracy; if the model points to a loss, some changes must be made.

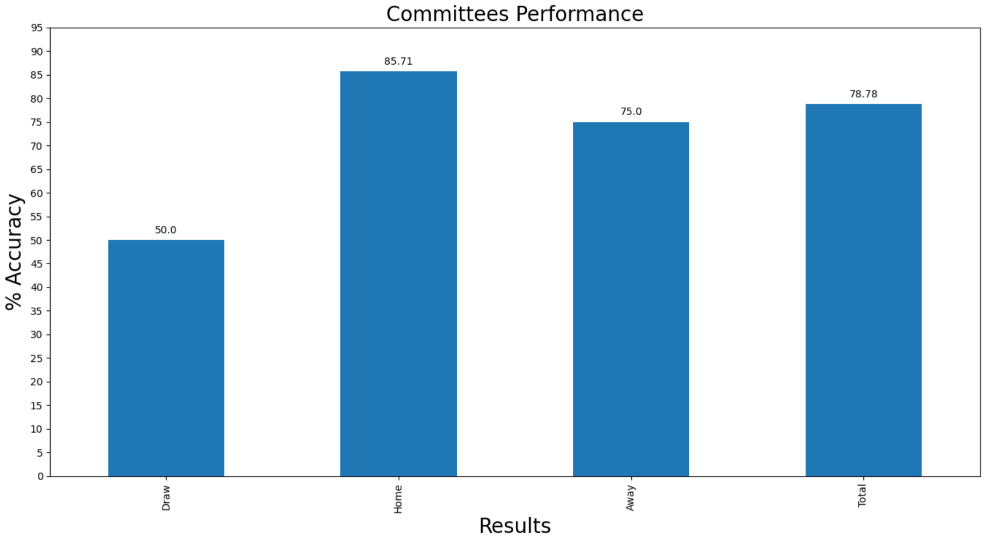

Figure 2 - Performance of the committee.

DISCUSSION

This work intends to build a model that predicts the result of a match (win/loss/draw) based on the data from the first half of the match in the Brazilian Futsal National League (LNF). To accomplish that, we extract 598 matches from seasons 2016 to 2019 directly from the LNF website. In the following, we discuss some points of this work.

Predicting Home Winning seems more straightforward than the other results. This can be explained based on the number of winnings of the home team in our dataset: more than half of the results correspond to the home winning. On the other hand, the draw is the most challenging result to predict. Maybe, it is because it is the less frequent result in LNF. For example, in Season 2018, 43 matches ended tied, i.e., 21.4%. In the same season, home winning corresponds to 52.2% of the matches, and away winning 26.4%.

The features HomeStrength (HS), AwayStrength (AS) Difference of goals of the home team (HTDiff), Difference of goals of the away team (ATDiff), and the difference of strength of the home and ways teams (DSH and DSA) that we proposed played a crucial role to build better models. Using only the features extracted from the matches, the performance of the models was inferior.

However, as stated by (Bunker and Thabtah, 2019) and (Horvat and Job, 2020), features regarding the player individuality, team strategy, and the coaches' strength can improve the model's prediction. This brings another problem: the lack of available data to feed machine learning algorithms. It is a paradox since there are many data on the web, but they are stored in Html pages, and the format of the pages changes frequently.

We claim that the overall results of this work are satisfactory. We can predict any match result with almost 79% confidence (based on the committee result).

That can be useful for those who want to change or keep the flow of the match. It also could be useful for bookmakers and wagers since the sports-betting market is growing every year.

CONCLUSION

The literature shows that machine learning is a compelling approach in the sports analytics area. It has been used to predict from match results to a player's future position during a match.

Accordingly, in this work, we apply ten machine learning algorithms to build models predicting a futsal match result based on the first half.

Based on this work's findings, we conclude that predicting the winning of the home team is more straightforward than predicting the away winning or draw.

This can be explained because there are more examples of home winning than the other results, and, so, the algorithm can better generalize the data to build the model.

Our work has produced promising results towards the prediction of match results, and the new proposed features played a crucial role in building the models.

However, some threats to validity must be pointed out: the limited amount of the data (three seasons for training and one for testing), the weight used in the equations, the hyperparameters applied on the machine learning algorithms, and the study of other factors that can influence a match result.

We believe that those threats are mitigated since the work's overall results are satisfactory, and we test the models in real matches not seen during the training step.

Moreover, we can suggest future directions for this work: (i) gather more data from the LNF website, (ii) put experts in futsal in the loop, (iii) propose more extra-match features (as presented in the Section Discussion), and (iv) apply deep-learning as another machine learning technique.

REFERENCES

1-Anderson, C.; Sally, D.The numbers game: Why everything you know about football is wrong. Penguin UK. 2013.

2-Baboota, R.; Kaur, H. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. Vol. 35. Núm. 2. 2019. p.741-755.

3-Bunker, R.P.; Thabtah, F. A machine learning framework for sport result prediction. Applied computing and informatics. Vol. 15. Núm. 1. 2019. p. 27-33.

4-Constantinou, A.; Fenton, N. Towards smart-data: Improving predictive accuracy in long-term football team performance. Knowledge-Based Systems. Vol. 124. 2017. p 93-104.

5-Duarte, D.; Ståhl, N. Machine learning: a concise overview. In: DATA Science in Practice. Springer. 2019. p. 27-58.

6-Frencken, W. G. P.; Lemmink, K. A. P. M. Team kinematics of small-sided soccer games. In: Reilly, T.; Korkusuz, F. editors. Science and Football VI. New York. Routledge. 2009. p. 161-166.

7-Fried, G.; Mumcu, C.; editors. Sport analytics: A data-driven approach to sport business and management. Taylor & Francis. 2016.

8-Flôres, F. S.; Santos, D. L.; Carlson, G. R.; Gelain, E. Z. What Can Coaches Do? The Relationship Between Substitution and Results of Professional Football Matches. Revista Brasileira de Futsal e Futebol. São Paulo. Vol. 11. Núm. 43. 2019. p. 215-222.

9-Horvat, T.; Job, J. The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews. Vol. 10. Núm. 5. 2020. p.e1380.

10-Imai, T.; Uchiyama, A.; Magome, T.; Higashino, T. Play recognition using soccer tracking data based on machine learning. InInternational Conference on Network-Based Information Systems. Springer. Cham. 2018. p. 875-884.

11-Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. Scikit-learn: Machine learning in Python. The Journal of machine Learning research. Vol. 12. 2011. p. 2825-2830.

12-Stübinger, J.; Mangold, B.; Knoll, J. Machine learning in football betting: Prediction of match results based on player characteristics. Applied Sciences. Vol. 10. Núm. 1. 2020. p.46.

13-Tsunoda, T.; Komori, Y.; Matsugu, M.; Harada, T. Football action recognition using hierar-chical lstm. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Work shops (CVPRW). 2017. p. 155-163.

14-Van Haaren, J.; Van den Broeck, G. Relational learning for football-related predictions. In Latest advances in inductive logic programming. 2015. p. 237-244.

Recebido para publicação em 07/12/2020
Aceito em 10/03/2021