ELEKTROTEHNIŠKI VESTNIK 86(4): 197-202, 2019
ORIGINAL SCIENTIFIC PAPER

Importance of the training dataset length in basketball game outcome prediction by using naive classification machine learning methods

Tomislav Horvat[1], Josip Job[2]

[1] University North, Department of Electrical Engineering, 104. brigade 3, 42000, Croatia
[2] Josip Juraj Strossmayer University in Osijek, Char of Visual Computing, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Kneza Trpimira 2B, 31000 Osijek
E-mail: tomislav.horvat@unin.hr

Abstract. The focus of the paper is on using naive machine learning algorithms for predicting the NBA game outcomes. In order to complete a convincing result, the data of nine full NBA seasons are scraped for the proposed model training and result evaluation. The aim of the paper is to present the possibilities of naive machine learning methods and to define the length of the training phase as well as of the evaluation phase to be optimal for predicting the NBA games outcome. The research serves as an initial stage in the development of a doctoral dissertation on the outcome prediction in sport. The proposed supervised classification machine learning methods is used and two possible outcomes (win or loss) are predicted. The data segmentation is used as an evaluation method for a training dataset occurring chronologically prior to the testing dataset. The best results are achieved by using a single training season and one to three evaluation seasons and all the played games during the training phase.

Keywords: basketball, classification, machine learning, NBA, outcome prediction

Vpliv dolžine testnih podatkov pri naivnem algoritmu strojnega učenja na napoved izidov košarkarske tekme

V prispevku predstavljamo naivni algoritem strojnega učenja za napovedovanje izidov v košarkarski ligi NBA. Pri razvoju algoritma smo uporabili rezultate tekem v devetih sezonah. Namen prispevka je predstaviti možnosti metod naivnega strojnega učenja ter določiti dolžino faze usposabljanja in faze ocenjevanja, ki bodo optimalne za napovedovanje izidov iger NBA. Uporabljene so predlagane metode nadzorovane klasifikacije strojnega učenja in predvidena sta dva rezultata (zmaga ali poraz). Segmentacija podatkov se uporablja pri ocenjevanju učenja pred testiranjem. Najboljši rezultati so doseženi na podlagi podatkov iz ene sezone in ene do treh ocenjevalnih sezon ter vseh odigranih iger v fazi učenja.

1 INTRODUCTION

Nowdays, the sport outcome prediction is very popular, especially in sport betting among fans and sport workers around the world. This is particularly evident for the most popular sports such as basketball, football and soccer. A lot of researchers have proposed various algorithms to predict game outcomes, but their prediction ranges vary not only from sport to sport but also from the same sports leagues and seasons. It is almost impossible to determine the boundaries of the prediction possibilities, so it is important to determine the predictions results using simple prediction methods. The possible outcome number, competitiveness of sports and thus the possibilities of predictions vary from sport to sport, therefore, satisfactory outcome prediction results depend on the type of sport, but also on the competitiveness of the competition itself. The paper presents initial prediction results based on NBA game outcome prediction and will serve as a starting point for proposing a more advanced NBA league prediction algorithm. The research will serve as an initial stage in the development of a doctoral dissertation on the outcome prediction in sport. The proposed supervised machine learning methods will be used, more precisely the classification machine learning methods in which two possible outcomes will be predicted.

Arthur Samuel, the founder of machine learning, defines machine learning as a field of computer science that gives the computer the ability to learn without being explicitly programmed [1]. A newer definition defines machine learning as a method of programming computers to optimize the performance criterion using example data or past experience [2]. There are various types of machine learning, but outcome prediction in sport is mostly used by supervised machine learning. The goal of supervised learning is to develop a predictive model that based on both the input and output data predicts future events on the previously unseen data. Sports predictions are usually treated as a classification problem by which one class is predicted [3], and rare cases are predicted by numerical values.

Received 30 May 2019
Accepted 11 July 2019

Results in paper [4] also reveal that the classification predictive schemes predict game outcomes better than the regression schemes. The types of the machine learning techniques and their short descriptions can be seen in Figure 1.

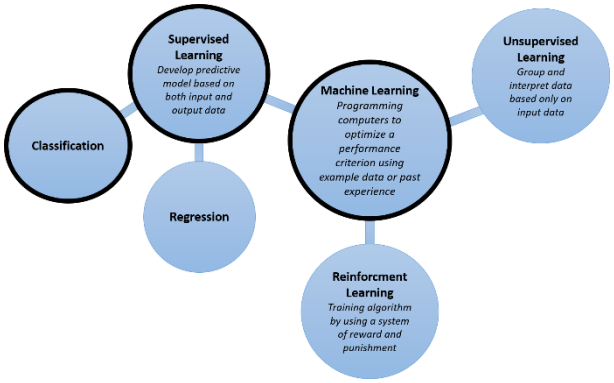

Figure 1. Types of machine learning.

The paper uses two feature extraction methods as well as two ways of defining the use of the known history. The goal of the paper is to determine the possibilities of naive learning methods and, accordingly, to propose a more advanced NBA league prediction model. Likewise, the paper also defines the best extent of the past period is the best for the NBA games outcome predicting.

As mentioned earlier, there are many researches related to the outcome prediction in sport, but the problem arises in the non-uniformity between sports and leagues. This chapter will provide a brief insight into the use of naive methods in outcome prediction in the NBA league, as well as the results of more advanced methods of machine learning primarily focused on the NBA league.

Several authors have published the results of naive machine learning methods. In paper [5], the authors use a variety of neural networks for predicting the NBA games outcomes (best result of 74.33% produced feed-forward neural network) and the obtained results are better compared to the experts' results (68.67%). In paper [6], the authors initially trained and test a variety of learning models. The best accuracy of 65.15% is achieved by a random forest method. The authors have also define two naive win prediction methods. The first prediction method is based on the greater difference between the average points scored and average points allowed per game (maximum achieved accuracy 63.5%) while the second method is based on chooses the winner based on a better win rate (maximum achieved accuracy of 60.8%). The authors also consider the accuracy of 71% achieved by the expert opinion.

The papers in which the machine learning methods are used are more numerous compared to the papers which use naive methods. In paper [7], the author propose a modelling approach based on the stacked Bayesian regressions and achieve the accuracy of 85.28%. The authors in paper [8] propose a model based on SVM with a support of a decision tree and using a CFS feature selection algorithm with the achieved accuracy of 85.25%. In paper [9], the authors propose a model based on the *k*-nearest neighbours for predicting the Euroleague games. The authors employ several models using different k and number of seasons. The best accuracy of 83.96% is achieved by using $k = 3$ for the dataset of three seasons and dataset of a single season and $k = 5$ or $k = 7$. The authors in paper [10] propose a Mixture Density Network model and achieve the maximum in-season (internal) accuracy of 86.7% and a maximum out-of-season (external) accuracy of 82%. The authors in paper [11] apply the Maximum Entropy principle to a set of features and establishe an NBA Maximum Entropy model. After that, they use a model to calculate the probability of the home team win of an upcoming game and make predictions based on this probability and achieve the accuracy of 74.40%. There is a lot of papers, but in this chapter, the papers with the highest prediction percentages are shown. Besides using different seasons, the authors use a different number of seasons, which results in a non-uniformity of the obtained results. Better results are obtained using a smaller number of seasons, which is logical with respect to the dynamics in the team roster.

The outcome prediction is also popular in other sports, especially in the most popular sports such as soccer ([3][12][13][14]), baseball ([4][15]), football ([16][17][18]), etc.

Chapter 2 depicts data preparation and introduces the used feature extraction methods and evaluation classification metric. Chapter 3 presents and explains the obtained results. Chapter 4 concludes the paper with a discussion and plans for the future work.

2 DATA AND METHODS

A sufficient amount of the relevant data is a basic condition for building a good prediction model. It is very important to well define the methods that will be used. As mentioned above, two different feature extraction methods with two ways of using historical data will be presented. This chapter will provide basic information about data acquisition and their preparing for being used for a naive machine learning algorithm. The two feature extraction methods and the two ways of using the known history will be presented. Research results and conclusions will be presented in later chapters.

2.1 Data acquisition

For our research purposes, the publicly available statistics of nine NBA seasons, from 2009/2010 to 2017/2018, are used. The database contains a total of 11578 games, which is more than enough to show which part of the known past is most relevant for the NBA league games outcome prediction. Using a web scraping process, structured data from the Basketball-

Reference web site are extracted, transformed and loaded into a relational database suitable for a further analysis. The process of extracting, transforming and loading is shortly called the ETL process. Due to the specificity of data retrieval, a web scraper in the scripting programming language PHP is programmed. The web scraper passes through the domain *www.basketball-reference.com*, extracts data from a page, transforms them if necessary and stores them into a relational, MySQL, database.

For the research purposes an information system called the Basketball Coach Assistant (later BCA) is built. BCA is a web application based on a relational database, built with the PHP and MySQL technologies. The application is supported by JavaScript and jQuery on the client side. The first version of the BCA information system, called *AssistantCoach*, was presented at the 2015 international conference icSports in Lisbon [19]. The proposed prediction model is shown in Figure 2.

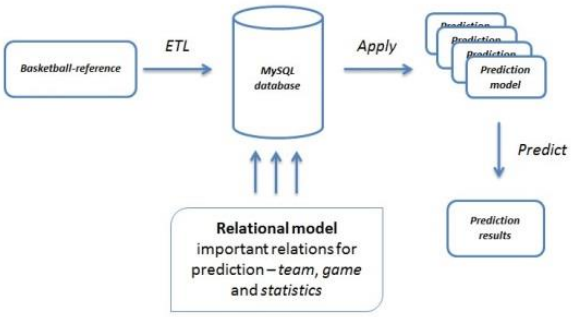

Figure 2. Proposed prediction model for the NBA games results prediction.

2.2 Feature extraction methods

Two naive feature extraction methods, as well as two ways of defining the use of the known history length, will be used in this research for the NBA game outcome prediction. The concept of the outcome prediction implies defining the winning team in the analysed game. Both feature extraction methods use a data segmentation evaluation method. In the dataset segmentation method, the input dataset is usually portioned into three different datasets: the training dataset, validation dataset and testing dataset which should be chronologically ordered. Validation datasets are not always used and their main role is to tune the final artificial intelligence model parameters. In this research, the validation dataset will not be used because there are only two features, the home and guest team win/loss ratios, and parameter tuning for a defining the final game outcome is not necessary. Using a chronologically-defined subset of the input data is recommended because sport events are not entirely independent events. The historical data can provide a very useful information and thus help in predicting future events. Figure 3 shows a graphical presentation of the data segmentation evaluation method.

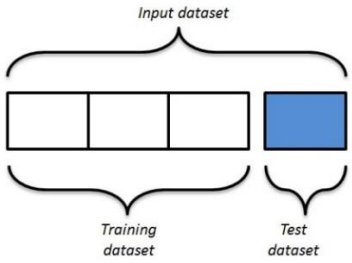

Figure 3. Data segmentation evaluation method.

The first feature extraction method includes all played games during the training period. During the evaluation phase, the win/lose ratio during the training phase is calculated and the team with a higher win percentage during training phase is pronaunced as a winner. In addition to the feature extraction methods, two ways of defining the training dataset are used. The first way of defining a training dataset includes only the games played during a training dataset, while the second way besides the training period games involves the played games during the evaluation dataset.

Second feature extraction method, as well as the first feature extraction method, defines the training and evaluation dataset. Unlike the first feature extraction method, in the second feature extraction method, the calculation of the win/loss ratio includes only the mutual games of the analysed teams. The second data preparation model also includes two ways of preparing the historical data explained in the previous sub-chapter.

2.3 Evaluating the classification metric

The outcome prediction is based on the home team where the actual result is compared to the win/loss ratio of both teams. The accuracy is a metric for evaluating the classification methods. Formally, the accuracy has the following definition:

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \qquad (1)$$

The classification often uses a confusion matrix to describe the performance of a model. Confusion matrices visualize the accuracy of a classifier by comparing the true and predicted classes. Figure 4 shows a confusion matrix for a binary classification consisting of four different combinations of the predicted and actual values.

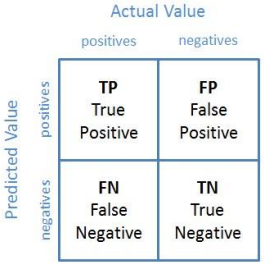

Figure 4. Confusion matrix for a binary classification.

The accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP is the true positive (an outcome where the model predicts the positive class), TN is the true negative (an outcome where the model correctly predicts the negative class), FP = false positive (an outcome where the model incorrectly predicts the positive class) and FN is the false negative (an outcome where the model incorrectly predicts the negative class). Table 1 shows the possible gamepredicted outcomes based on the current research problem.

Table 1. Possible game outcomes

True Positives (TP): Real outcome: home team wins Predicted outcome: home team wins	**False Positives (FP):** Real outcome: away team wins Predicted outcome: home team wins
False Negative (FN): Real outcome: home team wins Predicted outcome: away team wins	**True Negative (TN):** Real outcome: away team wins Predicted outcome: away team wins

Table 2. Prediction results based on all played games during a training phase.

Training dataset	Evaluation dataset	Accuracy (training period)	Accuracy (+ evaluation phase played games)
2009 – 2016	2017	54.73%	54.73%
2010 – 2016	2017	55.26%	54.57%
2011 – 2016	2017	54.95%	55.11%
2012 – 2016	2017	56.71%	56.94%
2013 – 2016	2017	57.62%	57.85%
2014 – 2016	2017	57.55%	58.00%
2015 – 2016	2017	58.77%	59.30%
2016	2017	58.84%	60.23%
2009 – 2015	2016 – 2017	56.05%	56.16%
2010 – 2015	2016 – 2017	55.97%	55.89%
2011 – 2015	2016 – 2017	57.57%	57.69%
2012 – 2015	2016 – 2017	58.26%	58.11%
2013 – 2015	2016 – 2017	58.79%	58.83%
2014 – 2015	2016 – 2017	59.90%	59.55%
2015	2016 – 2017	59.25%	60.48%
2009 – 2014	2015 – 2017	55.55%	55.40%
2010 – 2014	2015 – 2017	54.64%	55.02%
2011 – 2014	2015 – 2017	56.74%	56.50%
2012 – 2014	2015 – 2017	58.50%	58.34%
20013 – 2014	2015 – 2017	58.90%	59.45%
2014	2015 – 2017	59.41%	60.65%

3 RESULTS

This chapter gives an insight into the results of using the proposed naive machine learning methods in predicting the NBA games outcome.

As mentioned above, two naive feature extraction methods, as well as two ways of defining the use of the known history length, are used. Both feature extraction methods use the data segmentation evaluation method where part of the known history is defined as a training dataset and part as a testing dataset, where the training dataset occurrs chronologically prior to the testing dataset.

3.1 First feature extraction method

The first feature extraction method includes all the played games during the training period. During the evaluation phase, the win/lose ratio during the training phase is calculated for both teams and the team with a higher win percetage during the training phase is pronaonced as a winner. The first way of defining a training dataset includes only games played during a training dataset, while the second way besides the training period games involves the games played during the evaluation dataset. The prediction results are shown in Table 2.

Analysing Table 2, it is clear that the best results, over 60%, are achieved by using a single training season and one to three evaluation seasons and the way of defining a training dataset which includes the games played during the evaluation phase.

3.2 Second feature extraction method

The second feature extraction method, as well as the first feature extraction method, defines the training and evaluation dataset. Unlike the first feature extraction method, in the second feature extraction method the calculation of the win/loss ratio includes only the mutual games of the analysed teams. The prediction results are shown in Table 3.

Table 3. Prediction results based only on mutual games during a training phase.

Training dataset	Evaluation dataset	Accuracy (training period)	Accuracy (+ evaluation phase played games)
2009 – 2016	2017	51.98%	52.97%
2010 – 2016	2017	53.73%	53.58%
2011 – 2016	2017	56.05%	57.38%
2012 – 2016	2017	54.95%	55.56%
2013 – 2016	2017	55.56%	55.79%
2014 – 2016	2017	55.72%	56.33%
2015 – 2016	2017	57.01	57.39%
2016	2017	59.30%	58.99%
2009 – 2015	2016 – 2017	53.61%	55.70%
2010 – 2015	2016 – 2017	54.41%	56.12%
2011 – 2015	2016 – 2017	56.05%	57.38%
2012 – 2015	2016 – 2017	56.62%	58.15%
2013 – 2015	2016 – 2017	57.34%	58.34%
2014 – 2015	2016 – 2017	58.11%	58.91%
2015	2016 – 2017	57.80%	59.60%
2009 – 2014	2015 – 2017	54.76%	57.02%
2010 – 2014	2015 – 2017	54.81%	57.23%
2011 – 2014	2015 – 2017	55.93%	58.62%
2012 – 2014	2015 – 2017	57.33%	59.54%
20013 – 2014	2015 – 2017	57.94%	59.89%
2014	2015 – 2017	58.90%	60.27%

Generally, the results of using only mutual games achieve worse results compared to the feature extraction method that includes all played games during a training phase. As with the feature extraction method that includes all matches played during the training phase, the best result is achieved by using a single training season and three seasons during the evaluation phase.

3.3 Discussion

The aim of the paper is to present the possibilities of naive machine learning methods in predicting outcomes in the NBA league and to define the length of the training phase, as well as the evaluation phase, assessed as the best for predicting outcomes. By analysing Table 2 and Table 3 accompanied by Figure 5 and Figure 6, it is clear that reduction in the training phase length usually provides better results.

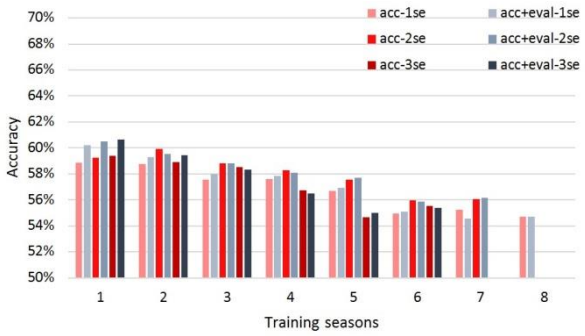

Figure 5. Graphical presentation of the prediction results based on all played games during a training phase.

Figure 5 shows the obtained machine learning results in different training seasons. It is clearly seen that by increasing the training seasons, the accuracy of the proposed prediction algorithm decreases. As expected, a slightly better prediction result provides the model in which the training phase includes the games played during the evaluation phase because the current season results give more information about the teams state.

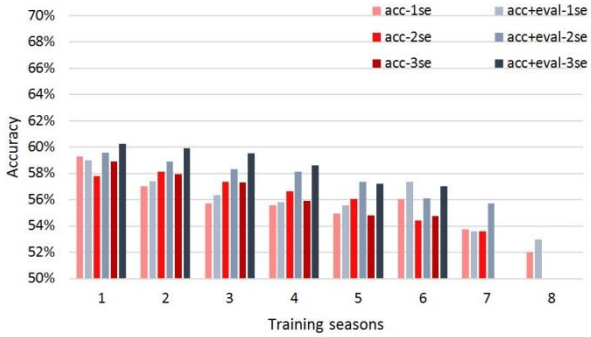

Figure 6. Graphical presentation of prediction results based only on mutual games during a training phase.

As in Figure 5, Figure 6 also shows that by increasing the training seasons, the accuracy of the proposed prediction algorithm based only on mutual games decreases. The figures given bellow show the accuracy graphs sorted by the training seasons and the evaluation season number.

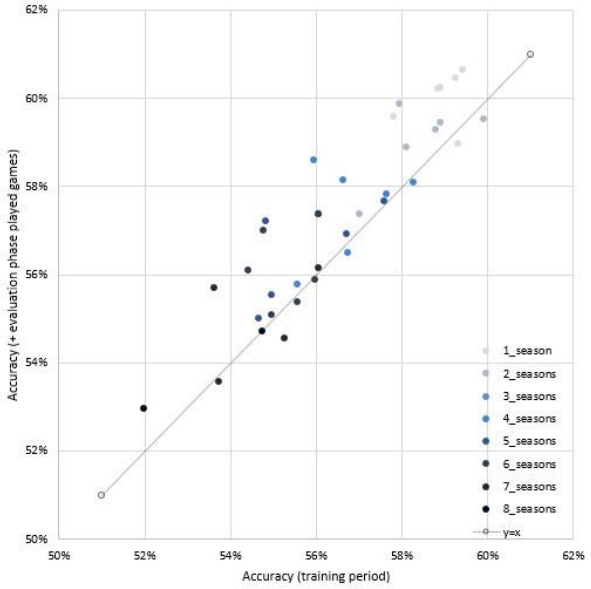

Figure 7. Graphical presentation of prediction results based on different season numbers and both ways of using the known history.

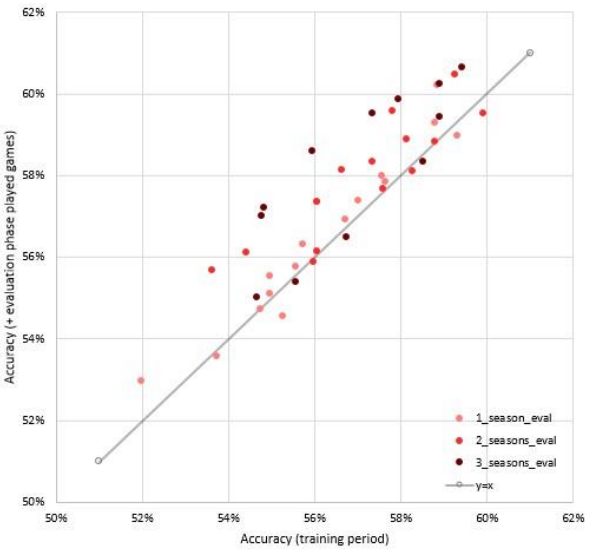

Figure 8. Graphical presentation of the prediction results based on a different evaluation season number and both ways of using the known history.

Figure 7 and Figure 8 confirm the above mentioned theses. Figure 7 clearly shows that the proposed algorithm accuracy generally decreases by increasing the training seasons and shows better results by using the played games during the evaluation phase. Figure 8 provides the prediction results of both methods for a different season number in the evaluation dataset. It can be seen that the second method, the one which uses the results of the evaluation period, shows better results

with the longer evaluation period included, i.e. the results for two seasons of evaluation are improved over the single season results, and an improvement over the two seasons of evaluation can be seen if three evaluation seasons are used.

4 CONCLUSION

The aim of the paper is to define how long the past is optimal in the NBA games outcome predicting and to define the minimal prediction capabilities /boundaries of simple/naive machine learning algorithms. Besides the ICT knowledge, the outcome prediction includes also an advanced observed-process understanding.

Two feature extraction methods and two ways of defining the history are used in this research. The first feature extraction method include all the played games during a training period, while the second feature extraction include only the mutual games of the analysed teams. The first way of defining the known history includes only the games during the training phase, while the second way of defining the known history includes also the games played during evaluation phase. Slightly better prediction results are achieved by the first extraction method which includes all the games during training phase. Better results are also achieved by using the played games of the evaluation phase during the training phase. Generally speaking, the best results are achieved by using a single training season. The length of the evaluation phase does not prove to be as important as the length of the training phase, but it shows differences between two methods if a longer evaluation period is used. The main aim of the paper is to define how long the past is optimal in the NBA games outcome predicting. The best results are achieved by using a single training season and by involving the games played in the evaluation phase into the training phase.

REFERENCES

[1] Samuel A.L., "Some Studies on Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development*, 3(3), pp. 210-229, 1959.

[2] Alpydin E., "Introduction to Machine Learning: Second Edition", *Cambridge, Massachusetts, London, England*, 2010.

[3] Prasetio D., Harlili D., "Predicting football match results with logistic regression", *International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, Penang, Malaysia, 2016.

[4] Soto Valero C., "Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods", *International Journal of Computer Science in Sport*, 15(2), pp. 91 – 112, 2016.

[5] Loeffelholz B., Bednar E., Bauer K.W., "Predicting NBA Games Using Neural Networks", *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[6] Lin J., Short L., Sundaresan V., "*Predicting National Basketball Association Winners*", Final Project, 2014.

[7] Lam M.W.Y., "One-Match-Ahead Forecasting in Two-Team Sports with Stacked Bayesian Regressions", *Journal of Artificial Intelligence and Soft Computing Research*, 8(3), pp. 159-171., 2018.

[8] Ping-Feng P., Lan-Hung C., Kuo-Ping L., "Analyzing basketball games by a support vector machines with decision tree model", *Neural Computing & Applications*, 28(12), pp. 4159-4167, 2017.

[9] Horvat T., Job J., Medved V., "Prediction of Euroleague games based on supervised classification algorithm k-nearest neighbours", *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support: K-BioS*, pp 203-207, Seville, Spain, 2018.

[10] Ganguly S., Frank N., "The Problem with Win Probability", *MIT Sloan Sports Analytics Conference*, 2018.

[11] Cheng G., Zhang Z., Kyebambe M.N., Kimburgwe N., "Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle", *Entropy*, 18(12), 2016.

[12] Igiri C.P., Nwachukwu E.O., "An Improved Prediction System for Football a Match Result", *IOSR Journal of Engineering, 4(12)*, pp. 12-20, 2014.

[13] Tax N., Joustra Y., "Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach", *Transactions on Knowledge and Data Engineering*, 10 (10), pp. 1–13, 2015.

[14] Zaveri N., Shah U., Tiwari S., Shinde P., Kumar T.L., "Prediciton of Football Match Score and Decision Making Process", *International Journal on Recent and Innovation Trends in Computing and Communication*, 6(2), pp. 162-165, 2018.

[15] Elfrink T., "Predicting the outcomes of MLB games with a machine learning approach", *Business Analytics Research Paper*, 2018.

[16] Delen D., Cogdell D., Kasap N., "A comparative analysis of data mining methods in predicting NCAA bowl outcomes", *International Journal of Forecasting*, 28(2), pp. 543 – 552, 2012.

[17] Blaikie A.D., David J.A., Abud G.J., Pasteur R.D., "NFL & NCAA Football Prediction using Artificial Neural network", https://www.semanticscholar.org/paper/NFL-%26-NCAA-Football-Prediction-using-Artificial-Blaikie-Abud/5207ead80e566abd29bf2c171143fa6473c28b6a, 2011.

[18] McCabe A., Travathan J., "Artificial Intelligence in Sports Prediction", *Fifth International Conference on Information Technology: New Generations*, 2008.

[19] Horvat T., Havaš L., Medved V., "Web Application for Support in Basketball Game Analysis", *icSports 2015*, Lisboa, 225-231, 2015.

Tomislav Horvat received his B.Sc. and M.Sc. degrees from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 2008 and 2018, respectively. Currently, he is a teaching assistant at the Department for Electrical Engineering at University North and is a Ph.D. student at the Faculty of Electrical Engineering, Computer Science and Information Technology in Osijek.

Josip Job received his B.Sc. and Ph.D. degrees from the Faculty of Electrical Engineering, Osijek, Croatia, in 2003 and 2010, respectively. Currently, he is an associate professor at the Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Department of Computer Science, Chair of Visual Computing. His research interests are in data visualization, machine learning and Internet of Things.