# Deeper Insights for Sports Analytics

Ashwin Vaswani*[1], Rijul Ganguly*[1], Het Shah*[1], Sharan Ranjit S*[1], Shrey Pandit[1], and Samruddhi Bothara[1]

Birla Institute of Technology and Science Pilani, K. K. Birla Goa Campus

**Abstract.** Sports data has become widely available in the recent past. With the improvement of machine learning techniques, there have been attempts to make use of this data to analyze not only the outcome of individual games but also to get better insights and develop better strategies. A lot of sports leagues in the world have interrupted their seasons due to the outbreak of the COVID-19 virus, and there are significant questions among people around the world regarding the outcomes of those seasons. What if the season was not interrupted and concluded as normal? Which teams would end up winning trophies? Which players would perform the best? Which team will end their season on a high and which teams will fail to keep up with the pressure? We aim to tackle this problem and come up with a solution to these questions. In this paper, we propose **UCLData**- a dataset containing detailed information of UEFA Champions League games from the past six years. We also propose a novel autoencoder based machine learning architecture that can come up with a story on how the rest of the season will pan out.

**Keywords:** Sports Analytics · Machine Learning · Data Mining · Auto-encoder.

## 1 Introduction

Sports analytics has received great attention in the community over the past few years. While a lot of work in sports analysis is emphasized on visual [1, 2] and tactical analysis [3], there have been attempts in recent years to predict the outcome of individual games and seasons. However, most of these attempts only predict the outcome without providing insights or having internal statistics to back their results. Another major issue is the lack of large clean datasets for this task. While most of the existing datasets provide data summarising matches, there is very little focus on the little intricacies of matches which might be of interest. To tackle this, we propose a dataset called **UCLData** which consists of both match and individual statistics from Champions League matches over the past six years. We further handle dataset size issues with the help of some intuitive priors or handcrafted features which makes our model robust and realistic.

In this work, we propose a novel autoencoder based architecture that not only predicts the outcome of a game, but also predicts it's internal statistics to give a more complete picture of how a match is expected to pan out. Moreover, apart from match-wise statistics, we also provide player-wise statistics to provide details about the contribution of each player and minor details about a match which is generally ignored. We also back up our results with a detailed study of the data from previous matches.

## 2 Related Work

Most work in the area of prediction of results of sports games using machine learning aims to predict simply the results of the games, instead of running a simulation predicting all the statistics.

Kampakis *et al.*[4] uses both player and team data for cricket matches to predict the performance of various teams based on different features. Kickoff[9] uses statistical modeling and Bayesian Inference to predict the results of individual football games. Rotshtein *et al.*[5] uses seven predictive models to predict outcomes in the English Premier League and the Premiership Rugby in England.

Joseph *et al.*[6] incorporated expert knowledge into a Bayesian model, again for predicting the results of individual football matches rather than for running simulations. Huang *et al.*[7] focuses on using neural networks to predict the results of the 2006 Football World Cup and is the most similar to what we have tried to achieve in this paper. They achieved an accuracy of 76.9% on the games' results, having special difficulty in predicting draws.

Hucaljuk *et al.*[8] incorporated expert opinion into Champions League matches, but in this case, there was no increase in accuracy in their work to predict the score of the games using neural networks.

## 3 Dataset

To create a dataset from which we can derive meaningful predictions for the remaining Champions League matches, we use web scraping as a tool. The following sections details our approach.

* Equal contribution

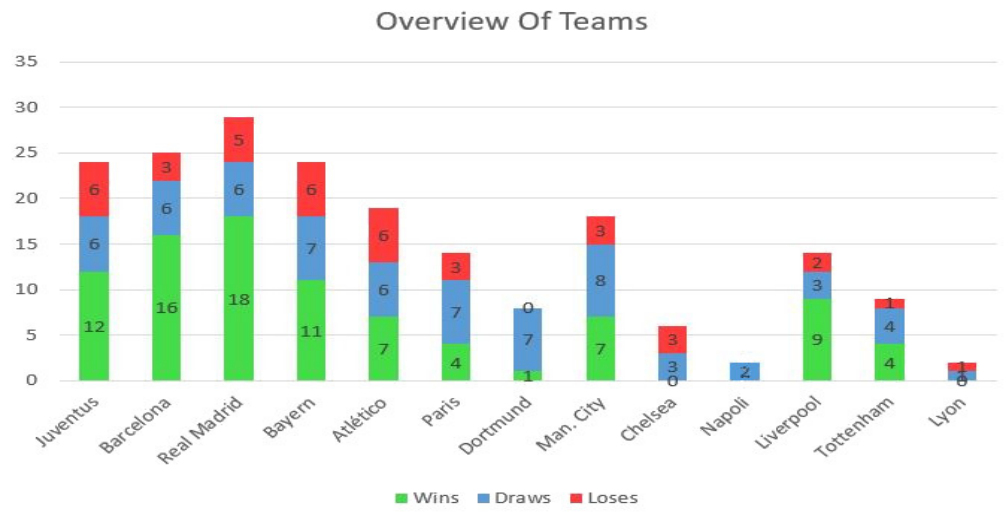



Fig. 1: Overview of Dataset

### 3.1 Data Collection

We scrape data from the official UEFA Champions League website to build our dataset. We use the data from the years 2014 to 2020. Overall we collect the data for 157 knockout stage matches. We do not collect data for group stage matches because our predictions will be on the knockout stage games of the $2019-20$ season of the Champions League, and hence we did not want the context of group stage matches misleading our model.

To scrape the data, we use the Python library Beautiful Soup ([12]), which is a library that allows us to take the data directly from the website of the relevant sites. We divide our data into two categories - team data and player data. Team data contains the statistics for the entire team playing in the match for both the sides, while player data includes the statistics for the teams' players.

To obtain the team data, we used the official UEFA website for the individual matches. However, we were unable to obtain the statistics for the players from the official website. Hence, we extracted individual player data from the FBref website[10] and the Global Sports Archive website[11]. Table 1 summarises the attributes we considered for our dataset.

| | Attributes |
|---|---|
| Team | Total Goals, total attempts, attempts on and off target, blocked shots, shots which hit the woodwork, corners, offsides, amount of possession, total passes, passing accuracy, completed passes, distance covered, number of balls recovered, tackles, clearances, blocks, yellow and red cards, fouls. |
| Individual player | Goals scored, total shots, shots on target, assists, interceptions, crosses, fouls committed, offsides, total time played |

Table 1: List of attributes for a team and an player

### 3.2 Data Preprocessing

Our data in its raw form contains numbers that vary in a wide range - from hundreds in the fields such as passes completed to only one or two in areas such as goals. Passing such fields without any preprocessing would lead to our proposed model not accurately capturing this wide range. Hence we normalize our data to the range of zero to one. This makes sure that our model doesn't give any undue importance to any fields because of scaling issues. After preprocessing, we create embeddings from our normalized data.

### 3.3 Creation of Embeddings

There are a few problems with using individual match data at all times. First, the information about matchups that happened previously cannot be captured efficiently. The argument can be explained with the help of an example. Let's say two teams A and B play against each other in year Y1. Then let us assume the two teams play against each other again in year Y2. Now, these two games are not independent as the two sides have played multiple other teams in this period and have evolved their gameplay, and thus, it's not right to directly use individual match stats without capturing this context. Another issue is regarding players switching teams,

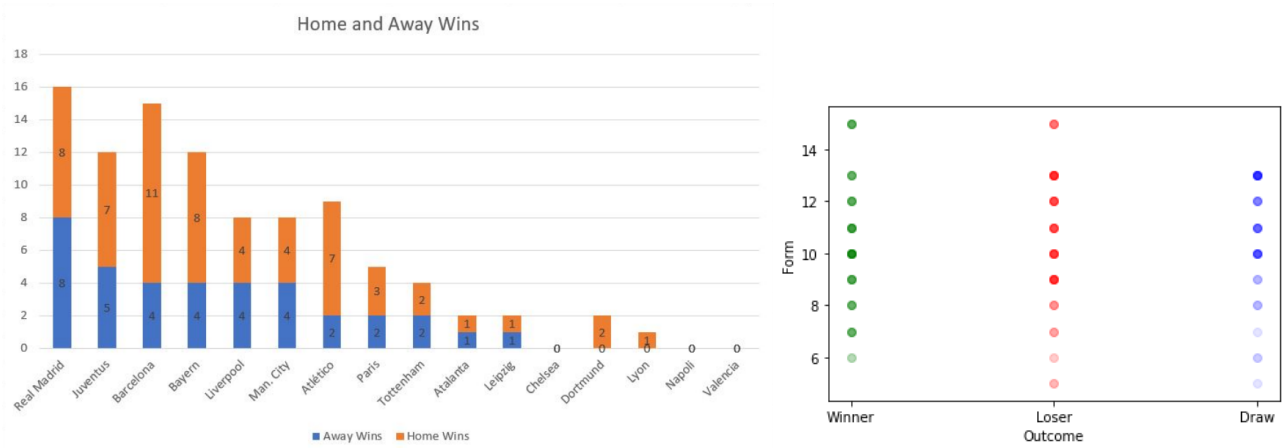
which is quite common in sports. If a player plays in team A in year Y1 and switches to team B in year Y2, we need a way to represent it so that his individual information is maintained. We solve these problems with the use of embeddings. We create embeddings for each team and each player so that when two teams are matched up, these representations can maintain the interactions with other teams and players and can preserve contextual information from previous data.

## 4 Methodology

### 4.1 Handling problem of Data bias

Our data consists of matches from the last five years of Champions League games. Although we found this data sufficient to capture relationships between teams and players, there were a few issues due to imbalance. Some teams, not being Champions League regulars, had relatively fewer data points. We found that our initial results were biased towards the lower number of data points of these teams and lacked generalization. We attempted to overcome this issue with the help of prior information, which is important in the field of football analysis. We propose three additional hand-crafted features which are crucial in the context of a game. We also infer that regularisation and dropout help in solving some of these problems. We perform an ablation study and also show with examples of how the addition of each of these features helps in making our results more robust.

**Home / Away status** An important feature of Champions League knockout stages is the Home / Away concept. A fixture consists of two games out of which each game is played at the home ground of the two teams. The figure 2(a) shows some analysis on how important the location of the fixture. It can be seen that there is a




(a) Home / Away wins

(b) Outcome vs Form

Fig. 2: Home / Away wins and Outcome vs Form

general trend for most teams to perform better at home than away, which is also quite intuitive. We attempt to use this information by adding an extra flag to indicate the team is playing at home apart from our embeddings while giving input to the model.

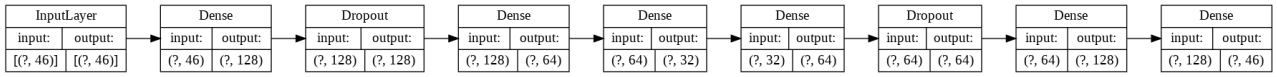
**Form Index** Another essential feature that is relevant to the context of a match is the form of the two teams playing in the fixture. It can be seen in the figure 2(b) that at lower values of the form($< 7$), teams are less likely to win whereas, in the middle range, it's difficult to comment with just form. We used the recent results of each team ( Results from the five most recent games before the fixture ) to generate a form index by giving a score of three points to a Win, one to a Draw, and zero to a Loss. This additional information helped in improving results of certain matches as a team would rather go into a game with a form of 15(five straight wins) than a form of 0(five straight losses).

**Experience** Figure 1 shows that some teams such as Real Madrid, being Champions League regulars have a lot of data points. In contrast, teams like Atalanta, who have recently started playing in the Champions League, have very few data points. Thus, the results of matches involving Atalanta was biased to the data from these limited games. This resulted in Atalanta performing exceptionally well against the odds in our initial

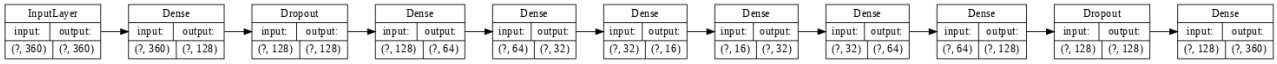
experiments. While this can be considered as a case of an "upset" or Atalanta being "dark horses," we wanted to improve our predictions further and make our results more robust. One major factor that can be crucial is a team's experience in the Champions League, because of the pressure of playing in such a high profile stage. We accumulated matches played by every team in our data to account for this experience, and it helped in solving the issue of predictions being biased because of limited data.

### 4.2 Details of the Model

The above methods helped in overcoming data issues in the pipeline. We also conducted a few experiments on the architecture to find a model that works best.



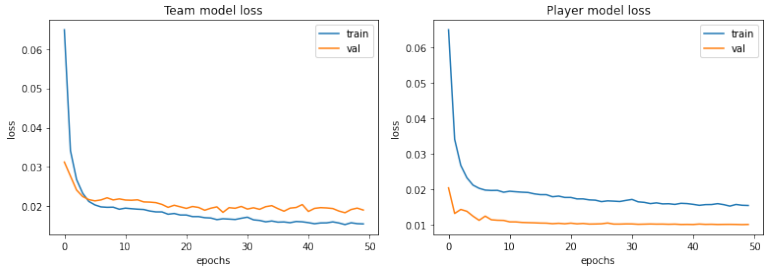

(a) Teams Model




(b) Players Model

Fig. 3: Models used

The model that gave the best results was an auto-encoder based network. The model architectures are as shown in figure 3. We took the embeddings that were created in the above steps, added Gaussian noise to them, and then took this "noisy" embedding as an input to the networks. The ground truth labels for the network were the original embedding. The reason for this is that while the model learns from the previous experiences of the team, adding Gaussian noise considers some other factors which are not consistent with the data (example a player having a lucky day/weather conditions which affect the play). So, after the training process, the model learned some important insights about the team's/player's performance, which is later helpful during the playoffs to decide the winner of a particular match. For training, the loss was taken to be **mean squared error**, and the metric that we have considered is the **root mean squared error**. The loss curves are given in figure 4. The training RMSE values for the team model were 0.1380, and for players model was 0.1127. The validation RMSE values for both the models were pretty close to the training models at 0.1379 for the team model and 0.1126 for the players' model.



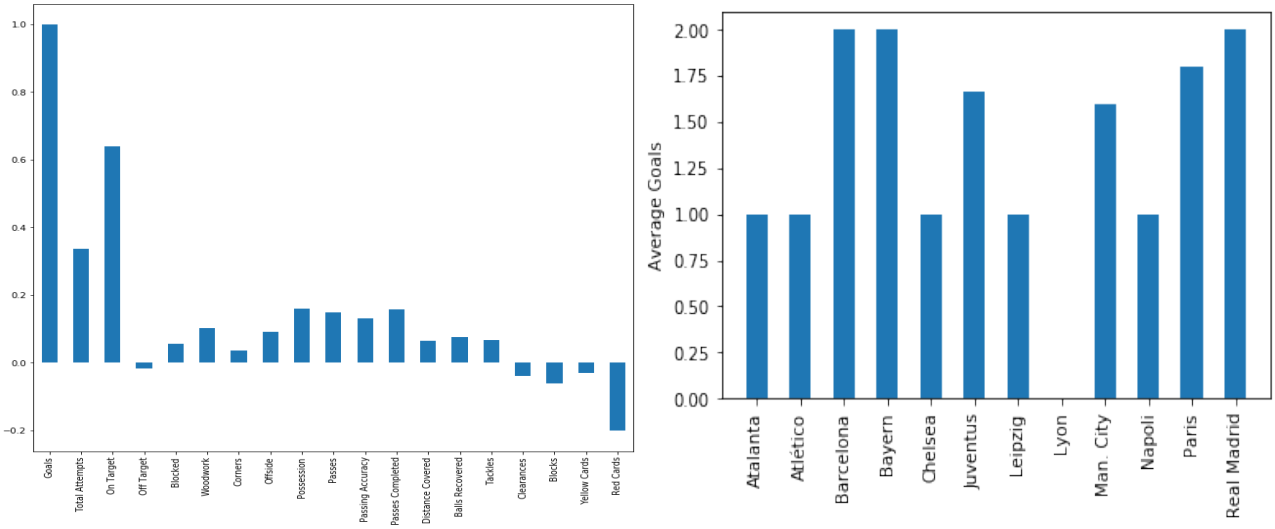

(a) Team Loss (b) Player Loss

Fig. 4: Loss curves for team and players model

## 5 Results and Observations

Figure 5 gives an overview of the simulation of the interrupted knockout stages of Champions League $2019-20$. Our model predicts both match(Total Goals, Total Passes, Possession, Blocks, Corners, etc.) and individual player statistics(Who scored the goals, Assists, Shots, Crosses, etc.) for the two teams in the fixture. The winner(team with a higher aggregate score over two legs) proceeds to the next round. In the case of a draw in the overall fixture (equal aggregate score from home/away legs), the team with the highest number of shots on target qualifies. We picked **Shots on target** as a decider, as it had the highest correlation with goals, which can be seen in Figure 6(a).


Fig. 5: Overview of Simulation




(a) Correlation with goals

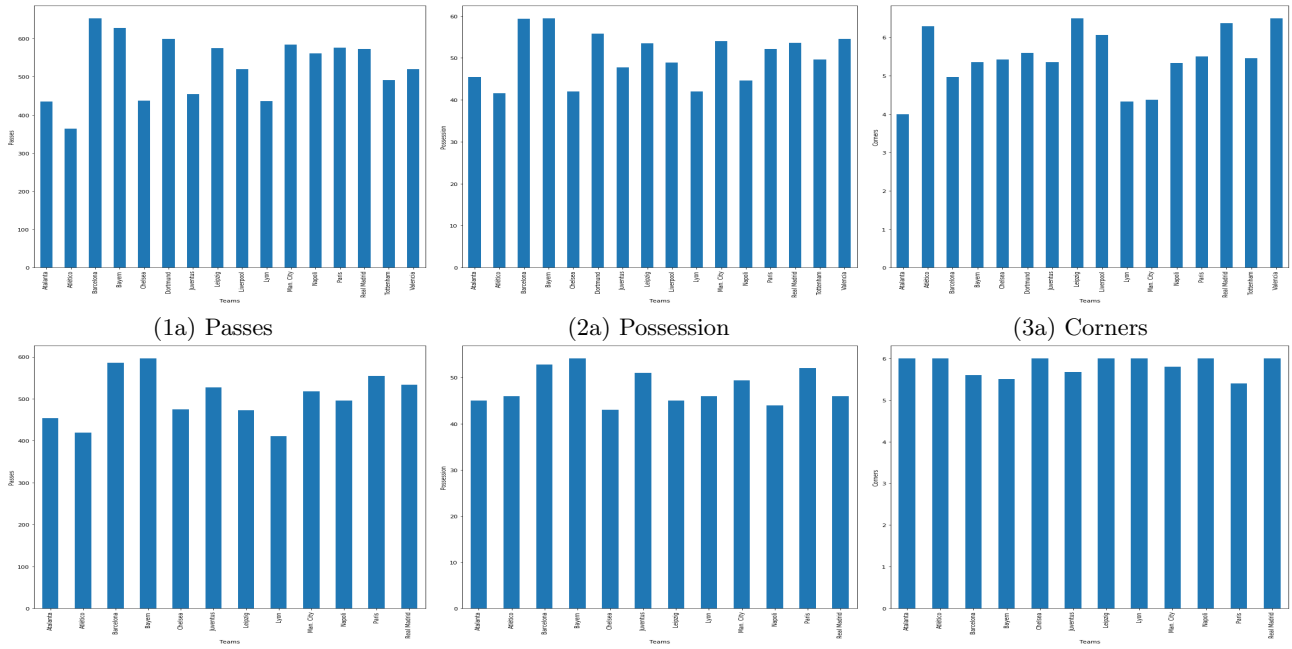
(b) Average Goals

Fig. 6: Correlation with goals and Average goals per team

The first simulation was between Bayern Munich and Chelsea(2nd Leg). Bayern Munich comprehensively beat Chelsea in the first leg fixture, which was conducted before the season was interrupted. Bayern entered the game with a form of five wins in each of its previous five games, whereas Chelsea had mixed results recently. The odds favored Bayern to win this tie, which is also backed up by our results. Bayern beat Chelsea comfortably with a scoreline of $2-1$ dominating the possession(57%) and total passing(597) stats. These stats are also backed up, as our data shows that Bayern Munich is one of the best teams in Europe with respect to passing and possession stats, which can be seen in figure 7(1a) and figure 7(2a). The goal scorers for Bayern were Robert Lewandowski and Boateng. Jorginho was the lone scorer for Chelsea. Our analysis shows Lewandowski to be one of the most prolific goal scorers in Europe over the past few years, which is backed up by these results.

Another similar result was found in the simulation of the game between Barcelona and Napoli. Again, Barcelona being European giants and one of the best passers in Europe dominated the passing(571) and possession(56%) stats and won with a scoreline of $2-1$ at home with Rakitic scoring for Barcelona. Rakitic has a good record of scoring in champions league knockouts, which is an interesting observation. Also, Barcelona has a great home record, as can be seen in Figure 1, which is also backed up by our results.

In another match, Paris (PSG) beat Atlético by two goals to one in both fixtures. Our analysis shows that Paris, being a team with a good scoring record (from Figure 6b), have a tendency to perform better against more defensive teams like Atlético. Also, Cavani is one of the most prolific scorers, scored in the fixture, validating the results of our model. Another big fixture was the game between Juventus and Man. city in which Ronaldo scored one goal, and Dybala scored two goals. However, their efforts went in vain, as Laporte scored two headed goals off corners, and Jesus scored one to take Manchester City to the semi-finals against Paris. In the crucial semi-finals, Paris being the in-form team, beat Manchester City by outwitting them in terms of both possession(58%) and passing stats, in which Cavani and Peredes scored. The fixture at Manchester City's home ground was even in terms of possession(50%) and passing statistics which can be explained by Man. City's strong record at home



(1a) Passes (2a) Possession (3a) Corners

(1b) Simulation of Passes (2b) Simulation of Possession (3b) Simulation of Corners

Fig. 7: Important statistics

in which they lost only 3 out of 18 games as seen in Figure 1. These results validate our argument that our model can learn interactions between features.

The other semi-final was a very close fixture between Bayern Munich and Barcelona. Both teams, being two of the favorites to win the competition, dominated the stats at their home, thus establishing their strong home record. The match ended in a draw with Bayern decided as winners on the basis of the highest number of shots on target (as per our chosen method). Another interesting observation was that our model could not decide the winner in this fixture over both legs, which is expected since Bayern and Barcelona were the favorites to win the competition.

The final was played between Bayern Munich and Paris, where Bayern Munich emerged victoriously. Few exciting observations from this simulation are discussed as follows: Paris, having never won the champions league, failed to perform in the final match. It was interesting that our model was able to capture this information through the means of interactions between the input embeddings. Lewandowski scored two goals for Bayern Munich, also establishing the substantial contribution of Lewandowski's goals to Bayern's success. Bayern Munich had the highest blocks per game in the simulations, which can be explained by Manuel Neuer's brilliant performances over the last few years. Finally, the results of our model are also backed up by the fact that Bayern Munich is one of the strongest teams in the competition, and had the best form leading up to the knockout stages.

Our model can easily be extended to predict other features such as time of goal. However, the extension is trivial and we leave this for future work. In order to verify the robustness of our model, we present some visualizations in Fig 7. We show the distributions of Passes, Possession, and corners in the training data and the distributions in predictions of our simulation. It can be seen that Barcelona and Bayern lead most of these stats in the training plots, and similar distributions can be seen in the simulations. It is evident from the plots in Figure 7 that our model is robust and can capture the information and interaction among features very well.

## 6 Conclusion

Inspired by the recent focus on sports analytics, and the curiosity among the community on how the current seasons would have concluded, we conducted a simulation to find out how the rest of the season would complete. We present **UCLData**, which contains data from the UCL games between the seasons 2014-2020 and propose a novel architecture that can efficiently capture the information and interactions among this data and make robust predictions on how individual matches of the season will pan out. We also propose solutions to handle some common problems related to data bias. Finally, we predict the results of the remaining champions league games and thus predict the results of this year's champions league. The code and data will be made publicly available in the hope that our work can encourage further experimentation in this field.

## References


1. Voeikov, R., Falaleev, N., Baikulov, R. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. arXiv e-prints arXiv:2004.09927.
2. H. Shih, "A Survey of Content-Aware Video Analysis for Sports," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 5, pp. 1212-1231, May 2018, doi: 10.1109/TCSVT.2017.2655624.
3. Rein R, Memmert D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. Springerplus. 2016;5(1):1410. Published 2016 Aug 24. doi:10.1186/s40064-016-3108-2
4. Kampakis, Stylianos and Thomas, William: Using machine learning to predict the outcome of english county twenty over cricket matches. arXiv preprint arXiv:1511.05837 **2**(5), 99–110 (2015)
5. Rotshtein, Alexander and Posner, M. and Rakityanskaya, A.: Football Predictions Based on a Fuzzy Model with Genetic and Neural Tuning. Cybernetics and Systems Analysis https://doi.org/10.1007/s10559-005-0098-4 (2005)
6. Joseph, Adrian and Fenton, Norman and Neil, Martin: Predicting football results using Bayesian nets and other machine learning techniques. Knowl.-Based Syst. https://doi.org/10.1016/j.knosys.2006.04.011 (2006)
7. Huang, Kou-Yuan and Chang, Wen-Lung: A neural network method for prediction of 2006 World Cup Football Game. Proceedings of the International Joint Conference on Neural Networks. https://doi.org/10.1109/IJCNN.2010.5596458 (2010)
8. Hucaljuk, Josip and Rakipovic, Alen: Predicting football scores using machine learning techniques. (2011)
9. INDY Lab Kickoff.ai: predicting football matches (2016).
10. Fbref website. https://fbref.com/en/
11. Global Sports Archive https://globalsportsarchive.com/
12. Richardson, Leonard: Beautiful soup documentation. April (2007)