Bunn JA, Reagor MK, & Myers BJ. Creating a random forest model to determine success in women's collegiate lacrosse. J Sport Human Perf 2025; 13(1):1-9.
DOI: 10.12922/jshp.v13i1.200

JSHP
JOURNAL OF SPORT AND HUMAN PERFORMANCE

ORIGINAL RESEARCH OPEN ACCESS

CREATING A RANDOM FOREST MODEL TO DETERMINE SUCCESS IN WOMEN'S COLLEGIATE LACROSSE

Jennifer A. Bunn [1,*], Mary K. Reagor [2], and Bradley J. Myers [3]

1 College of Health Sciences, Sam Houston State University, Huntsville, TX, USA
3 Independent statistician, Apex, NC, USA
2 College of Pharmacy & Health Sciences, Campbell University, Buies Creek, NC, USA

Corresponding author: jab229@shsu.edu

ABSTRACT

Predicting the outcomes of sports is difficult due to the variation created with human performance, environmental conditions, and style of play. Linear models have proven ineffective in creating usable equations that hold true across these variations. The purpose of this study was to use a random forest model to determine the variables involved in predicting game success (wins and losses) in Division I women's collegiate lacrosse. Data from the 2013-2018 seasons (103 games) were used as training input to the basic random forest model, with the 2019 data (17 games) used as a hold-out set to test the accuracy of the model. The model was also tested with data from the other teams from the same conference. After optimization, the accuracy of the model was 88.2% using the 2019 team data and 86.0% using the conference data. The variables with the highest importance solely emphasized shots taken by the team of interest and preventing shots from being taken by the opposing team. These data can be used to help coaches design drills based on the most important variables. Because the two models were so similar in accuracy, the designed drills are likely to be transferable to teams of similar capability.

Keywords: team sport, machine learning, performance match analysis, win probability

INTRODUCTION

Predicting the outcomes of sports is challenging due to the variation created with human performance, mentality, matchups between teams, match location, environmental conditions, style of play, referees, and specific game scenarios (1,2). Traditional linear models have proven ineffective in creating usable equations that hold true across these variations (3). However, machine learning created through decision trees, random forest plots, and neural networks has improved sport prediction beyond linear modeling.

Artificial neural networks are capable of learning attributes from past or present data to learn patterns that may then be used to make predictions (4). Match results have been predicted with varying accuracy (55-83%) using artificial neural networks with football (1), rugby (5), American football (6), and basketball (7). The volleyball sporting community has utilized artificial neural networks for prediction of league standings (4) and the success of specific plays in siding out (8). Tumer and Kocer were able to predict Turkish volleyball league standings with 98% accuracy (4). Wenninger et al. compared different artificial neural networks in predicting technical and tactical behaviors in elite-level beach volleyball (8). Specifically, Wenninger et al. used gradient boosted classification tree, multi-layer perceptron, convolutional neural network, and a recurrent neural network. The results indicated that combined input variables from team data created better prediction models than position-only input, but there was little variation in prediction differences among neural networks. Artificial neural networks appear to be one of the most common and longest utilized machine learning techniques to predict match outcomes in sports, but it is not the only method that has been utilized. Neural networks are also 'black box' learning, in that it is sometimes hard to determine which variables are driving the prediction. If knowing more about which variables contribute to the prediction is desired, using a supervised learning technique such as a decision tree makes more sense.

Decision tree learning is a predictive modeling approach that uses observations about an item to create conclusions related to the item's target value. Previous literature in Australian Football (9), rugby (10), and basketball (11) have utilized this method to predict rankings and outcomes. These studies have implemented data from either individual or team statistics for predictions, with varying levels of success. Similarly, sports scientists have utilized random forests to predict outcomes in rugby (12) and American Football (13). Random forests combine the input from multiple decision trees to predict the outcome, thus random forests are theoretically superior to decision trees because they utilize more information and scenarios before predicting the outcome. Further, random forests have less bias than a single decision trees because the data are randomly selected multiple times and repeatedly resampled. Both methods utilize a percentage of the data to train or establish the algorithm and the remaining portion of the dataset to test its accuracy. Bennett et al. showed 80% accuracy in game outcome prediction using data from both teams involved in the match in elite level rugby play (12). Lock and Nettleton examined play-by-play data from the National Football League, estimating the probability of winning using a random forest method. With this type of analysis, the probability of winning fluctuated for each team throughout a game based upon the performance of athletes and coaches' decision-making.

The purpose of this study was to use a random forest to determine the variables involved in predicting game success (wins and losses) in Division I women's collegiate

lacrosse. In this case, the observations include on-field statistics, and the target value was success as measured by a win or loss for a specific team of interest (TOI). A model was built using five seasons of data, and the results were compared to games throughout the 2019 season for the TOI and the entire conference to test if the prediction held true outside of one team. If so, this random forest could conceivably be used as a guide for lacrosse coaches within this league during training and games. If not, then perhaps machine learning prediction methods are best utilized in sport for within team evaluation and prediction.

METHOD

Study Design

This was a retrospective analysis of publicly available data. This study was approved for exempt status by the Campbell University Institutional Review Board.

Data Collection

Archived aggregate game statistics were used for analyses. These data are publicly available through www.gocamels.com. We obtained game-by-game team data for the TOI and their opponent for each game during the 2013-2019 seasons. No individual data were gathered. Rather, all data collected were composite team data as presented in the "game-by-game" stats archives. The following variables for both the TOI and opponents were recorded for the first half, second half, and whole game: shots, shots on goal, saves, turnovers, clears, ground balls, draw controls, and free position shots. A detailed description of each game variable and abbreviations used are shown in Table 1 (14).

Table 1: Definitions and abbreviations of game statistics evaluated in women's collegiate lacrosse.

Variable	Definition	Abbreviations
Shots	A ball propelled toward the goal with the offensive player's stick	Shot1, Shot2, Shot Total, O-Shot1, O-Shot2, O-Shot Total
Shots on goal	Includes only shots that score and those that were saved by the goalie	SOG1, SOG2, SOG Total, O-SOG1, O-SOG2, O-SOG Total
Saves	Any time the ball is stopped or deflected with any part of the goalie's stick or body	Save1, Save2, Save Total, O-Save1, O-Save2, O-Save Total
Turnovers	When a team in possession of the ball loses possession, both in live-ball or under certain dead-ball situations	TO1, TO2, TO Total, O-TO1, O-TO2, O-TO Total
Clears	Attempts to clear begin when a team has possession of the ball behind their own defensive restraining line and begins to transfer the ball to the offensive attack area. Teams who do this are awarded a successful clear, and teams who lose possession prior to the offensive attack area are awarded a clear attempt	Clear1, Clear2, Clear Total, O-Clear1, O-Clear2, O-Clear Total
Ground balls	When the ball changes possession during a live-ball play	GB1, GB2, GB Total, O-GB1, O-GB2, O-GB Total
Draw controls	Awarded to the team who controls the ball after taking the draw	DC1, DC2, DC Total, O-DC1, O-DC2, O-DC Total
Free position shots	When a shot is taken from the 8-m line. This shot also counts towards total shots	FPS1, FPS2, FPS Total, O-FPS1, O-FPS2, O-FPS Total

1: indicates number in the first half; 2: indicates number in the 2nd half; Total: indicates number for the whole game; O: indicates number from the opposing team

Data were gathered from teams within the same competitive conference as the TOI for the 2019 season to evaluate if the random forest would hold true for similar teams as the TOI. Data were publicly available (15) and collected from each conference game in the same way as outlined above.

Data Analysis

Analysis was performed in Python using the random forest classifier model from Scikit-learn (16). The random forest model was formulated similarly as described by Groll et al. (17). Data from the 2013-2018 season games (103 games, 48 variables) were used as training input to the basic random forest model, with the 2019 data (17 games, same 48 variables) used as a hold-out set to test the accuracy of the model. We did this to see how well past data can be used to predict future performance for the team. During training, the random forest model parameters were further optimized using a cross-validation grid search algorithm (17). Cross-validation grid search is a way to tune the hyperparameters of the random forest algorithm for optimal performance. For a random forest, hyperparameters can include variables like the number of decision trees in the forest, the criterion used to determine the split, the number of features considered before splitting a node, and more. The grid search with cross validation took all the parameters chosen and tried out all combinations, giving the variables that produced the highest accuracy for the model. Gini importance or mean decrease in impurity (MDI) of each variable was determined in the model. Gini importance or MDI is defined as the total decrease in node impurity (weighted by the probability of reaching that node) which is approximated by the proportion of samples reaching that node and averaged over all trees of the ensemble.

The Boruta method was used to determine which variables were most important to the prediction (18). This step reduced the size of the model while maintaining accuracy. Features are selected based on whether they were better than a randomized set of all the features. Boruta iteratively removed features that were statistically less relevant than a random probe, and with each iteration, rejected variables are removed from consideration in the next iteration. The Boruta algorithm consisted of following steps:

1. Extend the information system by adding copies of all variables (the information system was always extended by at least five shadow attributes, even if the number of attributes in the original set was lower than five).
2. Shuffle the added attributes to remove their correlations with the response.
3. Run a random forest classifier on the extended information system and gather the Z scores computed.
4. Find the maximum Z score among the shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA.
5. For each attribute with undetermined importance perform a two-sided test of equality with the MZSA.
6. Deem the attributes which have importance significantly lower than MZSA as ‘unimportant’ and permanently remove them from the information system.
7. Deem the attributes which have importance significantly higher than MZSA as ‘important’.
8. Remove all shadow attributes.
9. Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

The optimized model created from the 2013-2018 data accuracy was tested with the 2019 data from 1) the TOI and 2) other conference teams. Accuracy was tested for the full season of the TOI to evaluate if the model would hold true from one season to the next. Conference teams were used to test if the model for the TOI data was generalizable to other data sets. The variables for this data set were the same as the original TOI data set but consisted of seven unique teams across 50 games. For this test, we used the model we created on the TOI data (2013-2018 data) and input the conference data as the test set.

RESULTS

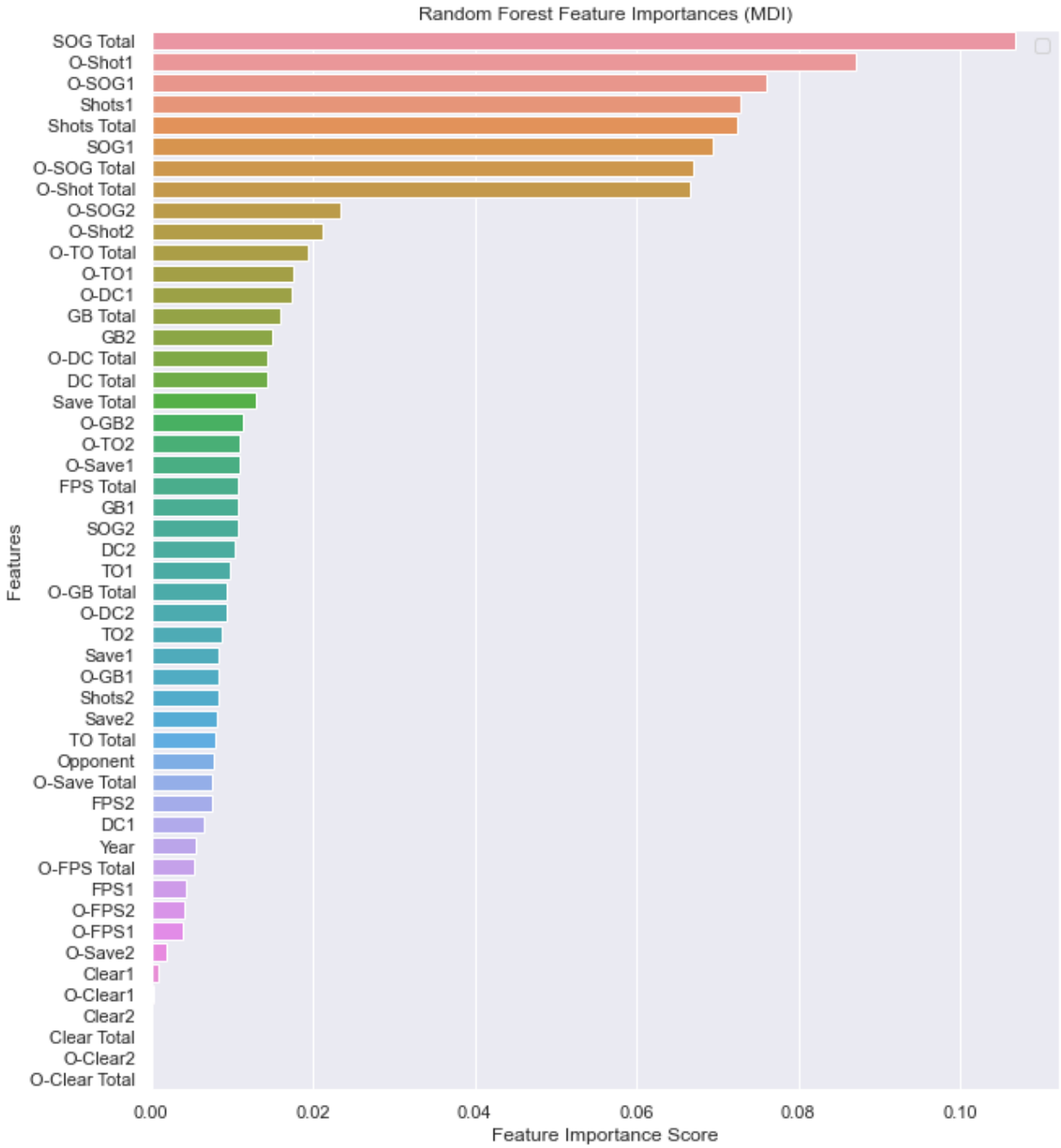
The original accuracy of the random forest model for the 2019 TOI data was 82.4%. The variable order by importance (Gini importance MDI) is shown in Figure 1 and the top 10 variables shown in Table 2. The order of variable importance was initially relatively unstable but gained stability with the optimization of the random forest and became very stable after the Boruta. After the random forest was optimized, and Boruta was used to reduce the number of variables, the accuracy of the model was 88.2% using the 2019 TOI data. The Boruta method determined that 10 variables were most important to determining a win or loss (alpha level 0.001). These 10 variables are listed in order of importance in Table 2, under the column, "TOI data after optimization." These variables were all focused on the TOI taking many shots throughout the game and preventing shots from being taken by the opponent throughout the game. After optimization, the top four variables were all related to offensive measures by the TOI and the bottom six variables are all related to shots taken by the opponent.

The extrapolation of the model with the conference data set showed a prediction accuracy measure of 86.0%. This is similar accuracy to that of the TOI data after Boruta optimization. This similarity suggests that prediction of game outcomes are likely similar between teams that play within the same league.

Table 2: Comparison of the ten most important predictor variables of the random forest model before and after the Boruta optimization. Variables related to the TOI are shown in black, and variables related to the opponent are shown in red.

Order of importance	*Original TOI data*	*TOI data after optimization*
1	SOGTot	Shot1
2	OShot1	ShotsTot
3	OSOG1	SOG1
4	Shot1	SOGTot
5	ShotsTot	OShot1
6	SOG1	OShot2
7	OSOGTot	OShotTot
8	OShotTot	OSOG1
9	OSOG2	OSOG2
10	OShot2	OSOGTot

Figure 1: Graphical representation of the MDI values for each variable utilized in the pre-optimized model. Higher MDI indicates greater importance of the variable.



DISCUSSION

This study sought to utilize a random forest model to determine the variables involved in predicting game success (i.e., wins and losses) in Division I women's collegiate lacrosse and to test the random forest model for accuracy using team data and data from conference opponents. The model proved reasonably accurate for both sets of test data.

The model obtained 88.2% accuracy on the TOI data even after reducing the number of variables from a set of 48 to a set of 10. Additionally, the optimized model predicted game outcomes with 86% accuracy for similar teams within the same competitive conference. The most important variables in the model solely emphasized shots taken by a team and shots taken by their opponent.

The results of the present model showed higher accuracy in predicting game outcome than models previously created for rugby, which used a random forest model ranging in 70-80% accuracy (12). The rugby random forest model also identified 10 performance indicators that were used in the model. Two different studies in volleyball employed the use of artificial neural networks with varying success in prediction accuracy. Tumer et al. showed 98% accuracy in predicting team league standing (4), whereas Wenninger et al. showed 59% accuracy in predicting the outcome of volleyball rallies (8). A decision tree model used with Olympic men's basketball predicted game outcome at 93% accuracy (11). The decision tree method was not only accurate, but also allowed for a team to meet specific thresholds for a given metric to help determine the game outcome. Thresholds are useful in providing coaches and teams a target to obtain to improve their chance of winning. Alternatively, random forest models do not provide these thresholds because the model is made of multiple decision trees, not just one. While random forest models generally provide a higher level of accuracy than decision trees, the drawback is that there is no output of thresholds or target metrics for the team to direct their training and goals. The random forest model created in the present study did provide good accuracy in predicting game outcome and provided performance indicators, but thresholds for these indicators would be of value for coaches.

The performance indicators from the studies in rugby and basketball (11,12) tended to favor defensive strategies. Six of the top ten variables of interest from the present study were related to opponents' shots, with the remaining four variables related to shots taken by one's own team. The simplest interpretation of these data suggest that a team take as many shots as possible, especially in the first half of play, and limit as many shots from the opponent as possible. Interestingly, all three team success studies favored defensive variables over offensive variables. Of further interest, is that player-specific stats in lacrosse are offensive leaning, with the only individual defensive statistic being caused turnovers. Because the random forest model provided a 60% emphasis on defensive concepts and limiting shots, perhaps developing a new player statistic related to this concept would be beneficial coaches in selecting starters and creating optimal offensive-defensive matchups in a game. Further, developing defensive drills to limit the number of shots taken within a given possession would also be beneficial. The results of the present study do not provide either a game threshold or shots per possession threshold and team possessions are not currently tracked within women's lacrosse. It would be beneficial to accurately track and explore these metrics in the future.

A limitation of this study was that there were no data with thresholds or values for the top 10 variables within the optimized model. This limits the application of the data for coaches and athletes. Rather, the coaches and athletes have generalized concepts for success around these variables. Addressing the values provided within specific decision trees produced in the model may be useful in providing some of this missing information, but each of the trees produced in the random forest will be different. Further, it is unknown if the accuracy of this model can be

extrapolated to teams of differing levels of play from other conferences and divisions. Future research could potentially address the use of this model beyond its current test data. Lastly, women's collegiate lacrosse has moved from playing in halves to quarters in the 2022 season. While these data are still likely useful, it may be beneficial to create a new model utilizing quarters.

Conclusions

Coaches and athletes can use the results of this study to guide and structure drills and game strategies. Understanding that 60% of game success was related to reduction of opponents' shots taken is an important consideration in assembling a practice. These data also suggest that the current individual and team game statistics may be inadequate for providing sound and detailed predictive indicators for team success. The random forest model created in this study showed similar accuracy between the TOI, whose data was used to create the model, and teams within their conference. Thus, the model retained external validity for teams within their conference, which improves the extrapolation and use of the data for improving training and the game.

Conflicts of interest: The authors certify that there are no conflicts of interest to report.

Funding: This research was partially funded by the National Association of Kinesiology in Higher Education.

REFERENCES

1. Arabzad SM, Tayebi Araghi ME, Sadi-Nezhad S, Ghofrani N. Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. Journal of Applied Research on Industrial Engineering [Internet]. 2014 Sep 1 [cited 2022 Aug 31];1(3):159–79. Available from: http://www.journal-aprie.com/article_43050.html
2. Lago-Peñas C. The role of situational variables in analysing physical performance in soccer. J Hum Kinet [Internet]. 2012 Dec [cited 2022 Aug 31];35(1):89–95. Available from: https://pubmed.ncbi.nlm.nih.gov/23487326/
3. Bynum L, Snarr R, Myers B, Bunn J. Assessment of Relationships Between External Load Metrics and Game Performance in Women's Lacrosse. Int J Exerc Sci [Internet]. 2022 [cited 2022 Aug 31];15(6):488–697. Available from: https://digitalcommons.wku.edu/ijes/vol15/iss6/8
4. Tümer AE, Koçer S. Prediction of team league's rankings in volleyball by artificial neural network method. Int J Perform Anal Sport [Internet]. 2017 [cited 2021 Mar 2];17(3):202–11. Available from: https://www.tandfonline.com/doi/abs/10.1080/24748668.2017.1331570
5. McCabe A, Trevathan J. Artificial intelligence in sports prediction. In: Proceedings - International Conference on Information Technology: New Generations, ITNG 2008. 2008. p. 1194–7.
6. Khan J. Neural Network Prediction of NFL Football Games. Joshua Kahn - PDF Free Download [Internet]. 2003 [cited 2022 Aug 31]. p. 1–19. Available from: https://docplayer.net/21763052-Neural-network-prediction-of-nfl-football-games-joshua-kahn.html
7. Ivanković Z, Racković M, Markoski B, Radosav D, Ivković M. Analysis of basketball games using neural networks. 2010 11th International Symposium on

Computational Intelligence and Informatics (CINTI). 2010.

8. Wenninger S, Link D, Lames M. Performance of machine learning models in application to beach volleyball data. undefined. 2020 Jul 1;19(1):24–36.

9. Robertson S, Woods C, Gastin P. Predicting higher selection in elite junior Australian Rules football: The influence of physical performance and anthropometric attributes. J Sci Med Sport [Internet]. 2015 Sep 1 [cited 2022 Aug 31];18(5):601–6. Available from: https://pubmed.ncbi.nlm.nih.gov/25154704/

10. Woods CT, Sinclair W, Robertson S. Explaining match outcome and ladder position in the National Rugby League using team performance indicators. J Sci Med Sport. 2017 Dec 1;20(12):1107–11.

11. Leicht AS, Gómez MA, Woods CT. Explaining Match Outcome During The Men's Basketball Tournament at The Olympic Games. J Sports Sci Med [Internet]. 2017 Dec 1 [cited 2022 Aug 31];16(4):468. Available from: /pmc/articles/PMC5721175/

12. Bennett M, Bezodis N, Shearer DA, Locke D, Kilduff LP. Descriptive conversion of performance indicators in rugby union. J Sci Med Sport [Internet]. 2019 Mar 1 [cited 2022 Aug 31];22(3):330–4. Available from: https://pubmed.ncbi.nlm.nih.gov/30146476/

13. Lock D, Nettleton D. Using random forests to estimate win probability before each play of an NFL game. J Quant Anal Sports. 2014 Jun 1;10(2):197–205.

14. McNeil K, Rhew E, Ticknor P, Col W. The Official National Collegiate Athletic Association 2020 WOMEN'S LACROSSE STATISTICIANS' MANUAL. 2020 Women's Lacrosse Statisticians' Manual National Collegiate Athletic Association; 2020 p. 1–9.

15. https://bigsouthsports.com/stats.aspx?path=wlax&year=2019#results [Internet]. 2019 Women's Lacrosse Overall Statistics.

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. Journal of Machine Learning Research [Internet]. 2011 [cited 2022 Sep 1];12(85):2825–30. Available from: http://scikit-learn.sourceforge.net.

17. Groll A, Ley C, Schauberger G, Van Eetvelde H. Prediction of the FIFA World Cup 2018 - A random forest approach with an emphasis on estimated team ability parameters. 2018 Jun 8 [cited 2022 Nov 1]; Available from: https://arxiv.org/abs/1806.03208v3

18. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw [Internet]. 2010 Sep 16 [cited 2022 Aug 31];36(11):1–13. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v036i11