Determinants of Success in Football: A Machine Learning Approach

Patrick Oxenham, [1] **Julio Costa,** [1] **Tom Hounsell** [2]

[1]Fulham Football Club, Motspur Park, London
[2]Burton Albion Football Club, Burton upon Trent

Youth Football | Decision Tree | K-Means Cluster | Match Outcomes

Headline

One advancing area in football research is the use of statistical modelling and machine learning algorithms to predict match outcome (1, 2, 3, 4). These techniques afford football practitioners the opportunity for deeper analysis into identifying key variables of interest during training and matches for the preparation of differing competitive situations.

Statistical modelling techniques utilise historical performance data to identify patterns and trends to predict future outcomes. Multiple methods of statistical modelling have been used to predict match outcome such as; mann-whitney u non-parametric tests (5), t-tests and discriminant analysis (6, 7), and one-way ANOVA (1, 7, 8, 9).

More recently, due to the availability of big data, machine learning algorithms have become increasingly popular due to their flexibility and ability to identify more complex patterns compared to statistical modelling techniques. These include; linear regression (10), log linear modelling (11), multinominal logistic regression (12), logistic regression (13, 14), bayesian networks (15) and decision trees (1, 9). One of the most popular classification algorithms is decision trees (16), which aims to create outputs by minimising classification error. The algorithm represents predicted outcome-based decisions (leaf nodes) from a singular partition (root node) through a process of decision nodes. Consequently, a decision tree algorithm was used in this study to represent the relationship between the chosen performance variables and match outcomes.

When focusing on the determinants of success, it is important to consider external parameters that may also influence football performance. As a result, the concept of 'situational variables' has emerged as an important aspect of performance research (17). Two prominent variables which have been researched heavily are match status, the effect on performance when 'winning', 'drawing or losing', (18, 19, 20) and quality of opposition, the effect on performance when playing against 'strong' 'balanced' or 'weaker' opponents (11, 12, 21, 22). Effective evaluation of sports performance in football requires knowledge of the aforementioned situational variables with above research evidencing the need for inclusion when analysing performance. As a result, within this study, quality of opposition and scoring first were included in the analyses.

Traditional methodologies of determining opposition level were based upon current standing (23), end of season classification (11) or defining due to differences in the end of season ranking between opposing teams (24). These methods have come in for criticism, as using end of season and current rankings fail to recognise in season momentum and personal changes over time. Therefore, to improve methodological rigour, authors now utilise distance-based machine learning algorithms such as K-Mean Clustering (1, 25, 26).

The above research provides a thorough view on the determinants of success at 1st team level, using methodologies to predict match outcomes that incorporate machine learning algorithms (4). However, there appears to be a lack of application in youth football to fully understand what indicators are important for positive sports performance.

Aim

To understand the determinants of success within the U21s Premier League 2 using a machine learning approach.

Methods

Data Sample and Variables

The data used within this study comprises of 77 key performance indicators and situational variables acquired from Stats Perform (London, United Kingdom) that show the characteristics of all U21s (n = 28) English Football Teams in the 2021-2022 Premier League 2 – Division 1 (n = 14) and Premier League 2 – Division 2 (n = 14). This resulted in a total of 364 matches included for analysis with 728 observations in total, as there are 2 teams that contest each match. All matches included within the study were from the group stage phase of the league season. The semi finals and final of the Premier League 2 – Division 2 were excluded from the study sample as different competition state could impact playing style (Gomez et al, 2013). The breakdown of observations can be seen in Table 1.

All variables used within this study relate to the value of the teams, rather than the individual value of the players and were broken down into three categories to represent the different phases of football (variables related to goal scoring (n = 21), variables related to passing (n = 20) and variables related to defending (n = 33)). The aim of the study is to use the 76 independent variables to predict the dependent variable, which is match outcome. All variables used within the study along with their operational definitions can be found on the Stats Perform glossary website (`https://www.statsperform.com/opta-event-definitions/`).

Statistical Analysis Procedures

Data was acquired from Stats Perform and exported into CSV format in preparation for data cleaning. Any blanks within the data set were identified and removed (U21 – Division 1, 22 matches were removed and U21 – Division 2, 16 matches were removed). This resulted in a final data set 326 matches and 652 observations as shown in Table 2. Basic statistical descriptors (mean and standard deviation) for all variables were calculated with respect to match outcome (win, draw or loss).

A shapiro-wilk test was carried out on all numerical variables to check for normality. However, all variables were included within the next stage of the analysis process as sample size exceeds 30. As the main goal of an ANOVA is to examine if variable effects are significant, the assumption of normality of the residuals is only required for small samples (27) as violating normality assumptions bears risk that are limited and manageable (28). A one-way ANOVA was carried out to test for statistical differences for each of the variables. Following this, Tukey HSD post hoc test was applied to the variables whose means were statistically significant to examine where the differences in match outcomes (win, draw loss) originated with a p value less than 0.05 used as a measure of statistical significance.

K-means clustering was used to group (n = 3) the games by quality of opponent (strong, balanced, weak) to examine the key performance indicators according to situational effects. Prior to K-means clustering, all independent variables were scaled to ensure the distance-based learning algorithm was not affected by bias towards certain variables. Finally, the decision tree machine learning algorithm was applied to investigate the variables that had the most influence upon match outcome. A 75/25 split was applied to partition the training and holdout data sets in which a K-Fold cross validation was used due to its acceptance across literature as an 'optimum' approach (16). A flowchart of the full analysis process can be found in Figure 1.

Basic statistical descriptors, normality testing, a one-way ANOVA with post hoc testing and K-Means cluster were carried out using SPSS. To carry out the decision tree machine learning algorithm, the 'rpart' (29) and 'rpart.plot' (30) packages in R were applied.

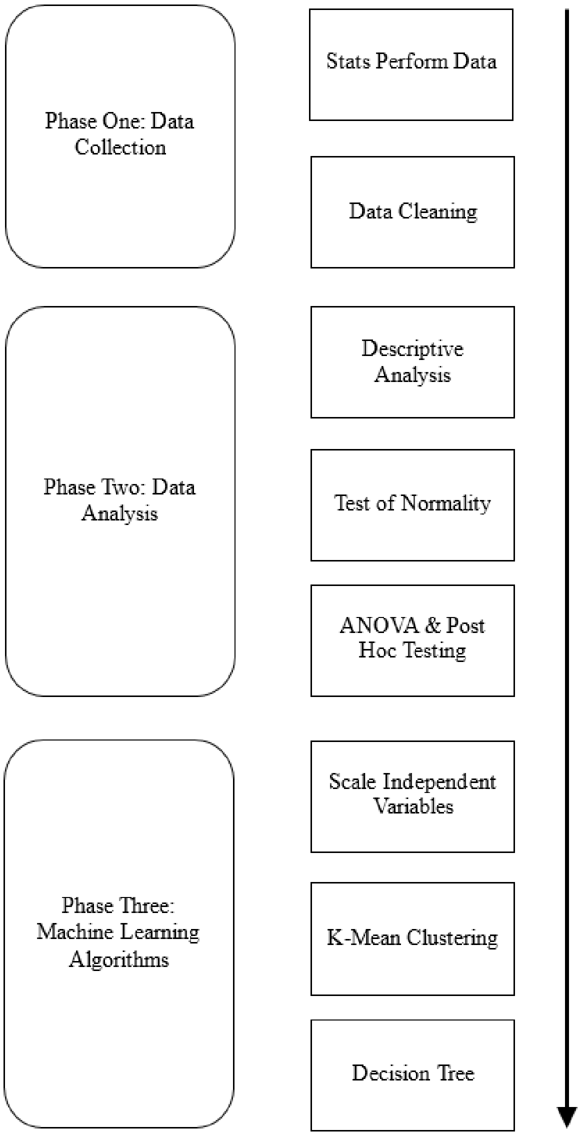

Fig. 1. Flowchart of the Analysis Process.

Table 1. Match and Observation Breakdown by Competition.

Age Group	League	Number of Teams	Total Number of Matches	Total Number of Observations
U21s	Premier League 2 – Division 1	14	182 (14 x13)	364
U21s	Premier League 2 – Division 2	14	182 (14 x13)	364

Table 2. Match and Observation Breakdown by Competition After Blanks Were Removed.

Age Group	League	Number of Teams	Total Number of Matches	Total Number of Observations
U21s	Premier League 2 – Division 1	14	160	320
U21s	Premier League 2 – Division 2	14	166	332

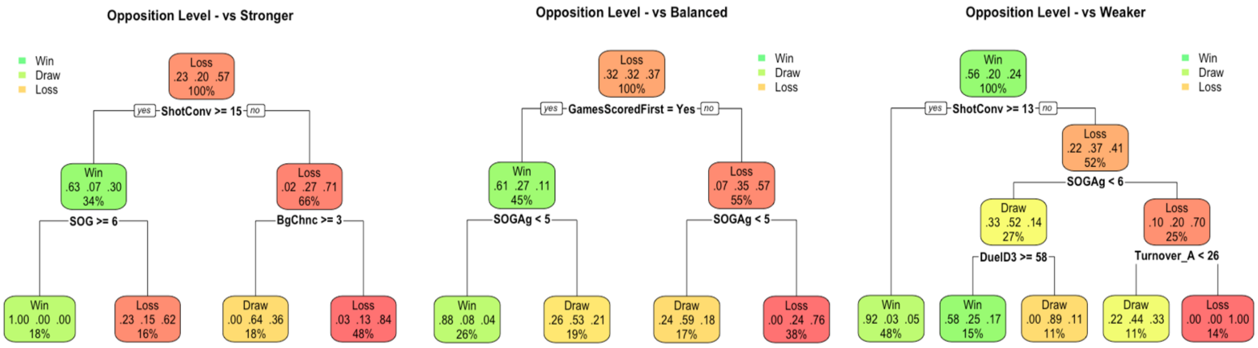

Fig. 2. Decision Tree Results for Teams v Stronger, Balanced, and Weaker Opponents in Division 1.

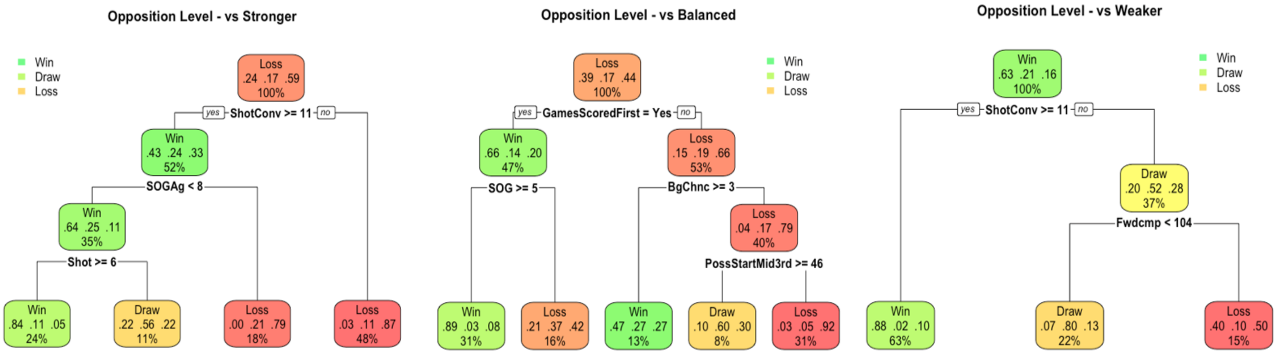

Fig. 3. Decision Tree Results for Teams v Stronger, Balanced, and Weaker Opponents in Division 2.

Results

Division 1

The results from ANOVA testing highlight all but two variables (Take On and Take On % Success) related to goal scoring had a significant influence upon match outcome. Only three variables related to passing (Forward Completed, Forward % Success, and Crosses) had a significant effect upon match outcome. Furthermore, there were six variables related to defending (Errors and Turnovers Leading to Goals, Number of Aerial Duels Across the Att, Mid, and Def Thirds, and Duel Success in Mid Third) which had a significant effect upon match outcome. Throughout the whole Division 1 season, 96 games were won by the team that scored first (30%) and only in 25 games (7.8%) did a team win a game when conceding first.

Division 2

The results from the ANOVA testing highlight all but four variables related to goal scoring had a significant effect upon match outcome. Interestingly, these variables were all related to take on's. There were seven variables related to passing that had a significant effect upon match outcome, these related to number of forward passes and attacking third ball retention. Finally, there were fourteen variables related to defending which had a significant effect upon match outcome. Ten of those variables were related to pressing and four of the variables related to duels in the attacking third. Throughout the whole Division 2 season studied, 109 games were won by the team that scored first (32%) and in only 28 games (8.4%) did a team win a game when conceding first.

Quality of Opposition

K means clustering analysis was applied to the quality of opposition variable to group the matches by weaker, balanced, and stronger opponents. The analysis revealed the following categories:

1. Stronger opponent if positional difference is +4 to +13 (U21s Division 1, n = 98) (U21s Division 2, n = 99)
2. Balanced opponent if positional difference is -3 to +3 (U21s Division 1, n = 124) (U21s Division 2, n = 150)
3. Weaker opponent if positional difference is -4 to -13 (U21s Division 1, n = 98) (U21s Division 2, n = 83)

Decision Tree ML Algorithm

Decision Tree classification algorithms were used to examine which of the variables had significant effects on match outcome across the three clusters in both Division 1 and Division 2 of the U21s English Premier League 2. The variables of Goals For and Against, Goals from Open and Set Play were not included within this stage of the analysis because these variables directly influence match outcome, as the main aim of a football match is to score goals to win. The below decision trees represent nodes coloured in green, yellow, and red to represent a win, draw and loss respectively with the largest proportion corresponding to the match outcome written within the node. The left side of each node represents true with the right side representing false.

When playing against stronger opponents in Division 1, it appears that shot conversion is the most important variable as it provides a 63% chance of winning if it is greater than 15. However, if it is less than 15, the chances of losing against a stronger opponent increase to 71% from 57%. If the team also have greater than 6 shots on goal as well as a shot conversion of greater than 15, chances of winning increase to 100

When playing against balanced opponents in Division 1, it appears that scoring first within the game is the most important variable as it provides a 61% chance of winning compared to a 57% chance of losing if conceding first. Furthermore, if teams have less than 5 shots on goal against, their chances of winning increase to 88%. Dependent upon the number of shots on goal against a team concedes, chances of drawing games when scoring first increase to 59% or increase losing % to 76 if conceding first.

When playing against weaker opponents in Division 1, it appears that shot conversion is the most important variable as chances of winning increase to 92% if it is greater than 13. If a team's shot conversion is less than 13, chances of losing the game are 41% with shots on goal against being less than 6 determining whether you then draw a match compared to losing.

When playing against stronger opponents in Division 2, it appears that shot conversion is the most important variable as it provides a 43% chance of winning if it is greater than 11. However, if it is less than 11, the chances of losing against a stronger opponent increase to 87% from 59%. If the team concede less than 8 shots on goal, their chances of winning increase to 64% and then to 84% if they themselves have more than 6 shots.

When playing against balanced opponents in Division 2, it appears that scoring first within the game is the most important variable as it provides a 66% chance of winning compared to a 66% of losing if conceding first. Furthermore, if teams have more than 5 shots on goal, their chances of winning increase to 89%. If conceding first, teams would need to create more than 3 big chances to win (47% chance) or start possessions in the midfield 3rd greater than 46 times to draw (60% chance).

When playing against weaker opponents in Division 2, it appears that shot conversion is the most important variable as chances of winning increase to 88% if it is greater than 11. If a team's shot conversion is less than 11, teams have a 52% of drawing the game.

Discussion

The use of statistical modelling and machine learning algorithms to predict match outcomes are an increasing interest to football practitioners, as they afford a deeper level of analysis to identify key variables of interest in relation to successful performance. Therefore, the aim of this current research was to examine the determinants of success within the U21s Premier League 2, using a decision tree machine learning algorithm approach to establish key performance indicator influence upon match outcome.

Shot conversion (for matches against stronger and weaker opponents) and scoring first (for matches against balanced opponents) were the root nodes, inferring they were the most significant variables in predicting match outcome, regardless of the competition state (Division 1 or Division 2). The decision tree results showed that in Division 1, a shot conversion greater than 15 provided a winning probability of 0.63 against stronger opponents and a shot conversion rate greater than 13 provided a 0.92 probability against weaker opponents. In Division 2, a shot conversion rate greater than 11 provided a winning probability of 0.43 against stronger opponents compared to 0.88 against weaker opponents. These findings are supported by previous studies that have shown higher shot conversion rates influence more successful match outcomes (31, 32, 33).

When playing against balanced opponents, scoring first provided a winning probability of 0.61 in Division 1 and 0.66 in Division 2. Emphasis of scoring first is a common phenomenon throughout previous research (19, 20, 34, 35, 36) as football is a low-scoring invasion game and as such, the current findings provide further evidence of this. Previous research has shown that players adopt different tactical strategies (10) influenced by fluctuations in physical performance (18, 24) to retain possession of the ball once they have scored first. As a result, it is important applied practitioners consider this contextual variable when analysing and reporting back on team and individual performance.

Throughout the investigation there were other variables related to attacking, passing, and defending that appeared to have an influence on match outcome. Intriguingly, all these variables were related to actions in the attacking third of the pitch. These findings provide further evidence to the domain of research carried out in football confirming actions within the attacking third lead to an increased likelihood of winning games (7, 8, 13, 34, 37, 38). It would therefore be beneficial to coaches and managers to consider attacking styles of play in elite youth football to create goal scoring opportunities to increase shot conversion rates and there likelihood of scoring the first goal. Furthermore, profiling teams and players in attacking performance will allow for applied practitioners to make inferences about performance for the purposes of opposition analysis, recruitment, and talent identification.

From the 76 variables that were tested in this current investigation, there were some performance indicators that had no significant effect upon match outcome. Interestingly, these were all related to take on's. Whilst variables related to beating players in 1v1 situations has largely been ignored in previous research (1, 31, 39), further investigations are required to establish importance of these variables particularly as they underpin position specific profiles in applied practice.

The decision tree approach was chosen as it is a supervised classification machine learning algorithm that has been widely used in previous research in football at first team level (4, 40). Due to the unpredictable nature of association football, insights drawn from machine learning models should be treated with caution, particularly as no research studies have been conducted on the amount of input data needed to make a relevant prediction (4). Furthermore, the interactive nature of football suggests more frequent turnovers will happen from one team to the other (40) creating a transition imbalance. This subsequently means through counter attacks; teams are more likely to have higher quality of shots which could be influencing model selection of variable importance. Future research and application to football should test multiple models to find the best machine learning algorithm that works well for the specific contextual problem.

Practical applications

- Use of machine learning techniques to aid the identification of key performance indicators related to match outcome.
- Provides coaches and analysts a methodology to inform decisions around style of play and help prepare players for competitive scenarios during match play.
- Implementation of best practice approaches to data collection and analysis with the inclusion of contextual influence upon football performance.

Conflict of Interest

The authors report there are no competing interests to declare.

Data Availability

Data available on request.

References

1. Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. International Journal of Performance Analysis in Sport, 19(6), 930-941.

2. Chmait, N., & Westerbeek, H. (2021). Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists. Frontiers in Sports and Active Living, 3, 682287.

3. Bunker, R., & Susnjak, T. (2022). The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review. Journal of Artificial Intelligence Research, 73, 1285-1322.

4. Rico-Gonzalez, M., Pino-Ortega, J., Mendez, A., Clemente, F.M., & Baca, A. (2023). Machine Learning Application in Soccer: A Systematic Review. Biology of Sport, 40(1), 249-263.

5. Jones, P.D., James, N., & Mellalieu, S.D. (2004). Possession as a performance indicator in soccer. International Journal of Performance Analysis in Sport, 4(1), 98-102.

6. Lago-Penas, C., Lago-Ballesteros, J., Dellal, A., & Gomez, M. (2010). Game-related statistics that discriminated winning, drawing, and losing teams from the Spanish soccer league. Journal of Sports Science & Medicine, 9(2), 288-293.

7. Lago-Penas, C., Lago-Ballesteros, J., & Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. Journal of Human Kinetics, 27(1), 135-146.

8. Lago-Ballesteros, J., & Lago-Penas, C. (2010). Performance in team sports: Identifying the keys to success in soccer. Journal of Human Kinetics, 25(2010), 85-91.

9. Parim, C., Gunes, M.S., Buyuklu, A.H., & Yildiz, D. (2021). Prediction of match outcomes with multivariate statistical methods for the group stage in the UEFA Champions League. Journal of Human Kinetics, 79(1), 197-209.

10. Lago, C., & Martin, R. (2007). Determinants of possession of the ball in soccer. Journal of Sports Sciences, 25(9), 969-974.

11. Taylor, J.B., Mellalieu, S.D., James, N., & Shearer, D.A. (2008). The influence of match location, quality of opposition, and match status on technical performance in professional association football. Journal of Sports Sciences, 26(9), 885-895.

12. Almeida, C.H., Ferreira, A.P., & Volossovitch, A. (2014). Effects of match location, match status and quality of opposition on regaining possession in UEFA Champions League. Journal of Human Kinetics, 41(1), 203-214.

13. Harrop, K., & Nevill, A. (2014). Performance Indicators that predict success in an English professional League One soccer team. International Journal of Performance Analysis in Sport, 14(3), 907-920.

14. Liu, H., Gomez, M.A., Lago-Penas, C., & Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. Journal of Sports Sciences, 33(12), 1205-1213.

15. Joseph, A., Fenton, N.E., & Neil, M., (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), 544-553.

16. Filiz, E. (2022). Evaluation of Match Results of Five Successful Football Clubs with Ensemble Learning Algorithms. Research Quarterly for Exercise and Sport, 1-10.

17. Gomez, M.A., Lago-Penas, C., & Pollard, R. (2013). Situational Variables. In McGarry, T., O'Donoghue, P., & Sampaio, J., Routledge Handbook of Sports Performance Analysis, (pp.259-269).

18. Bloomfield, J.R., Polman, R.C.J., & O'Donoghue, P.G. (2005). Effects of score-line on team strategies in FA Premier League Soccer. Journal of Sports Sciences, 23(2), 192-193.

19. Pratas, J.M., Volossovitch, A., & Carita, A.I. (2016). The effect of performance indicators on the time the first goal is scored in football matches. International Journal of Performance Analysis in Sport, 16(1), 347-354.

20. Liu, T., Garcia-de-Alcaraz, A., Wang, H., Hu, P., & Chen, Q. (2021). Impact of scoring first on match outcome in the Chinese Football Super League. Frontiers in Psychology, 12, 662708.

21. Lago, C. (2009). The influence of match location, quality of opposition, and match status on possession strategies in professional association football. Journal of Sports Sciences, 27(13), 1463-1469.

22. Taylor, J.B., Mellalieu, S.D., James, N., & Barter, P. (2010). Situation variable effects and tactical performance in professional association football. International Journal of Performance Analysis in Sport, 10(3), 255-269.

23. Grant, A.G., Williams, A.M., & Reilly, T. (1999). Analysis of the successful and unsuccessful teams in the 1998 World Cup. Journal of Sports Sciences, 17, 827.

24. Lago, C., Casais, L., Dominguez, E., & Sampaio, J. (2010). The effects of situational variables on distance covered at various speeds in elite soccer. European Journal of Sport Science, 10(2), 103-109.

25. Kong, L., Zhang, T., Zhou, C., Gomez, M. A., Hu, Y., & Zhang, S. (2022). The evaluation of playing styles integrating with contextual variables in professional soccer. Frontiers in Psychology, 13.

26. Shelly, Z., Burch, R.F., Tian, W., Strawderman, L., Piroli, A., & Bichey, C. (2020). Using K-means clustering to create training groups for elite American football student-athletes based on game demands. International Journal of Kinesiology and Sports Science, 8(2), 47-63.

27. Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual review of public health, 23(1), 151-169.

28. Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. Behavior Research Methods, 53(6), 2576-2590.

29. Therneau, T., Atkinson, B., & Ripley, B. (2022). Recursive Partitioning and Regression Trees. R Package Version 4.1.19.

30. Milborrow, S. (2022). Plot 'rpart' Models: An Enhanced Version of plot.rpart'. R Package Version 3.1.1

31. Kite, C.S., & Nevill, A. (2017). The predictors and determinants of inter-seasonal success in a professional soccer team. Journal of human kinetics, 58(1), 157-167.

32. Varley, M.C., Gregson, W., McMillan, K., Bonanno, D., Stafford, K., Modonutti, M., & Di Salvo, V. (2017). Physical and technical performance of elite youth soccer players during international tournaments: influence of playing position and team success and opponent quality. Science and Medicine in Football, 1(1), 18-29.

33. Mitrotasios, M., Gonzalez-Rodenas, J., Armatas, V., & Aranda, R. (2019). The creation of goal scoring opportunities in professional soccer. tactical differences between spanish la liga, english premier league, german bundesliga and italian serie A. International Journal of Performance Analysis in Sport, 19(3), 452-465.

34. Armatas, V., Yiannakos, A., Papadopoulu, S., & Skoufas, D. (2009). Evaluation of goals scored in top ranking soccer matches: Greek "super league" 2006-2007. Serbian Journal of Sports Sciences, 3(1), 39-43.

35. Garcia-Rubio, J., Gomez, M.A., Lago-Penas, C., & Ibanez, J.S. (2015). Effect of match venue, scoring first and quality of opposition on match outcome in the uefa champions league. International Journal of Performance Analysis in Sport, 15(2), 527-539.

36. Lago-Peñas, C., Gómez-Ruano, M., Megías-Navarro, D., & Pollard, R. (2016). Home advantage in football: Examining the effect of scoring first on match outcome in the five major european leagues. International Journal of Performance Analysis in Sport, 16(2), 411- 421.

37. Ruiz-Ruiz, C., Fradua, L., Fernandez-Garcia, C. and Zubillaga, A. (2011), Analysis of entries into the penalty area as a performance indicator in soccer. European Journal of Sport Science, 1(1), 1-8.

38. Castellano, I, Casamichana, D. and Lago, C. (2012), The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. Journal of Human Kinetics, 31, 139-147.

39. Merlin, M., Cunha, S.A., Moura, F.A., Torres, R.D.S., Gonçalves, B., & Sampaio, J. (2020). Exploring the determinants of success in different clusters of ball possession sequences in soccer. Research in Sports Medicine, 28(3), 339-350.

40. Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. International Journal of Sports Science & Coaching, 14(6), 798-817.