Research Article ISSN Online: 2783-4859 Computational Mathematics and Computer Modeling with Applications

Received 6 March 2025
Accepted 5 August 2025

DOI: 10.48308/CMCMA.4.1.43
AMS Subject Classification: 68-XX; 68Txx

Predicting football player features through hierarchical clustering and representative selection

Mahdi Nouraie[a], Meysam Agah[b], Changiz Eslahchi[c] and Arnold Baca[d]

This study offers a comprehensive analysis of both classic and advanced features of football players, aiming to enhance player evaluation and feature prediction. We applied a three-step methodology: first, hierarchical clustering was used to reveal the group structure of features in each position. Second, representative features were identified within clusters leveraging the correlation matrix of features, reducing the dimensionality of the dataset while retaining critical information. Third, several regression models were employed to predict other player features using the representatives. This approach was applied across the distinct positions of goalkeeper, defender, midfielder, and forward. Bootstrap resampling confirms the robustness of the results obtained, revealing consistent clusters against random data variations. The findings indicate that representative features effectively encapsulate the entire feature space for each position, allowing other features to be predicted accurately with minimal errors. This study contributes to football analytics by providing a robust method for feature selection and prediction, ultimately improving the accuracy and efficiency of player performance analysis.

Keywords: Football tactical analysis; Advanced football analysis; Feature selection; Machine learning.

1. Introduction

Football, or soccer as it is known in some parts of the world, stands as the most popular and widely followed sport globally, captivating millions of fans across continents [19]. This global appeal has transformed football into a cultural phenomenon and an important area of study within sports analytics. Football analytics focuses on the detailed examination of player and match data to optimize team strategies, enhance player performance, and ultimately increase the likelihood of success on the field. As the sport has evolved, so has the role of analytics, becoming indispensable for clubs aiming to gain a competitive edge in an environment where small details can decisively influence match outcomes.

The development of football analytics has been driven by rapid advancements in Artificial Intelligence (AI) and Machine Learning (ML) techniques. These innovations have introduced a range of advanced approaches, from statistical learning methods to deep learning algorithms, capable of uncovering complex patterns and insights previously difficult to access. Nowadays, analysts can evaluate every aspect of the game from player movement and decision-making to team formations and tactical approaches with unprecedented precision. This capability has revolutionized how teams prepare for matches, and how scouts evaluate player potential, manage player workloads, and make critical in-game decisions [7, 15, 17, 21].

In football analytics, player features are generally categorized into classic features such as shooting accuracy, passing precision, and ball control and advanced features that delve deeper into the game's nuances. The gathering of these features has been greatly enhanced by complex systems of player evaluation carried out by websites and analysis teams around the world. These entities assess each player's features on a standardized scale, after meticulously examining football matches. These scoring systems provide a quantifiable measure of player skills, which can be used to inform team strategies and player development programs [16].

[a] *Department of Statistics, Shahid Beheshti University, Tehran, Iran.*
[b] *Department of Mathematics, Shahid Beheshti University, Tehran, Iran.*
[c] *Department of Computer and Data Science, Shahid Beheshti University, Tehran, Iran.*
[d] *Centre for Sport Science and University Sports, University of Vienna, Vienna, Austria.*
[*] *Correspondence to: C. Eslahchi. Email: ch-eslahchi@sbu.ac.ir and A. Baca. Email: arnold.baca@univie.ac.at*

In recent years, the landscape of football analysis has undergone a transformative shift with the introduction of advanced features, such as Expected Goals (xG) and Expected Assists (xA). These features have revolutionized the analysis of player and team performance and have been extensively studied and applied in various contexts, including match outcome prediction, player scouting, and tactical analysis [1, 3]. For instance, a study used xG to predict match winners, showing that calculated probabilities closely align with actual results [5]. Furthermore, a study highlighted how xG could be integrated with other performance data to evaluate team improvements over a season [17].

As the volume and complexity of player data have increased, so has the need for effective methods to manage and interpret this information. Football analytics now encompasses a vast array of both classic and advanced features, leading to datasets with high dimensionality. This complexity can make it challenging to extract meaningful insights and can introduce redundancy in the data. To address these challenges, clustering algorithms and feature selection methods have become integral to football analytics. These approaches help identify key player features and groups of similar entities, such as teams or performance features, based on their characteristics. Such techniques have been widely applied not only in football but also across other fields. For example, a study demonstrated the effectiveness of clustering and feature selection in identifying critical genes and optimizing drug testing processes, further underscoring the importance of these algorithms in advancing biomedical research [13].

Clustering techniques have become essential for grouping players, teams, or game scenarios based on similar characteristics. By leveraging statistical methods and machine learning, clustering enables football analysts to uncover hidden patterns in the data. A novel clustering method was introduced using a "situational score line" as a key feature to describe and categorize seasonal team performance, revealing underlying patterns in team performance across a season [22]. Similarly, a clustering algorithm was developed to analyze team formations in football matches, providing new insights into tactical structures and how teams organize their play [14]. also clustering algorithms have significantly enhanced the analysis of football player and team performance by allowing for more nuanced categorizations and comparisons. Several studies have leveraged clustering algorithms to integrate advanced features into football analytics. For instance, Principal Component Analysis (PCA) was combined with a Gaussian clustering method to characterize professional football players, effectively reducing data dimensionality while identifying key performance indicators [18].

Although numerous clustering algorithms have been employed across diverse football analytics applications, a significant gap persists in leveraging these methods for feature selection. Many existing methods rely on complex machine learning and statistical techniques that require specialized knowledge, making them difficult for coaches and analysts who lack experience in these areas to use. In contrast, hierarchical clustering paired with dendrogram visualizations provides a clear, unified way to select important features. The tree like diagram shows how data points relate to each other in an intuitive format, allowing users of all skill levels to gain practical insights without needing expertise in machine learning or statistics.

In this study, we focus on utilizing hierarchical clustering to analyze football player features, with the goal of identifying the most significant features that define player performance. We use hierarchical clustering because it provides a visual representation of related features through dendrograms. By analyzing a comprehensive dataset from the top five European leagues over several seasons, we incorporate both classic and advanced features into our analysis. By selecting representative features from each cluster, we enhance our ability to predict other player features, providing a framework for player evaluation and performance prediction. Bootstrap resampling[4] is used to show the robustness of the results obtained against random perturbations in data. This study contributes to the growing body of research in football analytics by offering a method that integrates feature selection with predictive modeling, ultimately aiming to improve the accuracy and effectiveness of performance analysis in professional football.

2. Materials and Methods

In this section, we present a comprehensive description of the data utilized in this study, as well as the three methodological steps employed.

2.1. Dataset

In this study, we explore both classic and advanced features for player analysis using the FBref dataset [6] . The dataset we used includes data from the five major European leagues (England, Italy, Spain, France, and Germany) from the 20172018 season to the 20212022 season.

The raw dataset contains some issues requiring preprocessing. The preprocessing pipeline involved several stages. First, to create a unique entry for each player, we filtered the multi-season dataset to retain only the most recent season's data for any individual. Second, we addressed missing values by systematically removing players with incomplete data in key columns. This involved the exclusion of 12 players with missing 'onxGA' values, 10 with missing 'OG' values, 52 with missing 'Cmp%' values, and 16 with missing 'Rec%' values, leading to the removal of 90 players in total. Third, we performed feature cleaning and reduction. We identified and removed 17 columns with a high incidence of missing data (a minimum of 104 missing values each). A further four columns ('Matches', '+/-', '+/-90', 'On-Off') were removed due to their redundancy in the context of our per-match normalized analysis. Finally, 23 columns were removed as they were identified to be duplicates of existing features (e.g., "Gls.1", "Ast.1", "npxG.1", etc.). This comprehensive preprocessing step resulted in a refined dataset containing 4,904 players and 71 features.

In addition to features like each player's name, age, nationality, club, and position, the remaining features represent their classic and advanced features. The features used in this research are listed below:

Player, Nation, Pos, Squad, Comp, Age, Born, Gls, Ast, PK, xG, npxG, xA, Sh/90, SoT/90, FK, np:G-xG, Cmp%, KP, 1/3, PPA, CrsPA, Prog, TB, Sw, Crs, CK, In, Out, Str, Head, Other, Int, Blocks, SCA90, PassLive, PassDead, Drib, Fld, Def, GCA90, Tkl, TklW, Def 3rd, Mid 3rd, Att 3rd, Past, Succ, ShSv, Pass, Clr, Err, Att Pen, #Pl, Megs, CPA, Mis, Dis, PPM, onG, onGA, +/-90, onxG, onxGA, xG+/-90, Fls, PKwon, PKcon, Recov, Won, Lost.

Given the large number of features, we will not introduce them in detail here. For more information about these features, please visit the website[†].

One of the features in the dataset, labeled as "MP", indicates the number of matches a player participated during the relevant season. While MP is not used directly in the clustering and subsequent processes, we divided all the player's features by MP in the preprocessing stage. This adjustment is essential because different players participate in varying numbers of matches throughout a season, which can significantly influence the values of their features. By normalizing the dataset with dividing the features to the MP, we eliminate the effect of the total number of matches played by each player, ensuring a more accurate analysis.

Finally, among the 71 features, 7 represent the identity and basic information of the players. These features were excluded in the clustering and subsequent processes.

2.2. Methodology

The algorithmic approach we used in this research comprises three principal steps which are clarified in detail in the subsections that follow. The goal of our method is to identify a small set of features that can accurately predict the remaining classic and advanced features of players for each position.

2.2.1. First step: Clustering the features of the players The first step involves clustering the features of the players. Although various clustering methods are available, we have selected a hierarchical approach for this study. This choice is motivated by the effectiveness of dendrograms in summarizing and organizing data into a hierarchy, which facilitates the examination and interpretation of clusters. In addition, this approach allows us to identify features that can be derived from other features and to explore the relationships underlying features. Hierarchical clustering also highlights the relative importance of features, thereby enhancing the interpretability of advanced features, particularly for the general football audience.

Before performing clustering, all features are normalized to z-scores by column. This normalization step is crucial because hierarchical clustering is metric-based, meaning that features with larger values could disproportionately influence the clustering.

To carry out the clustering, we calculate the correlation matrix corresponding to the features in each cluster using the Pearson correlation coefficient. The formula that is used to calculate the distance between features is as follows:

$$d_{i,j} = 1 - \text{corr}_{i,j} \tag{1}$$

where di,j and corri,j represent the distance and Pearson correlation between features i and j respectively. The minimum distance is 0, indicating perfect positive correlation, while the maximum distance is 2, indicating perfect negative correlation.

In hierarchical clustering, it is essential to compute the distance between clusters to determine how they should be merged as the algorithm progresses. we employ complete linkage method for calculating the distance between clusters because it yielded better discrimination between the clusters of this problem in our experiments.

To determine the optimal number of clusters to prune the dendrogram, we used the Kelley-Gardner-Sutcliffe (KGS) criterion [10]. The KGS criterion helps to determine the optimal number of clusters by calculating distances, applying linkage methods, and then finding a penalty value for each potential number of clusters. The optimal number of clusters is chosen as the one with the lowest penalty value, ensuring the most significant and optimal clustering structure.

After completing the clustering process, we employed the bootstrap resampling, a statistical technique to evaluate the robustness of the clustering results against random variations in data. The Bootstrap method is a powerful tool that reveals the reliability of results by accounting for sample variability. This method involves repeatedly drawing random samples from the original dataset with replacement. By doing this, we generate multiple bootstrap samples that allow us to estimate the distribution of clustering results. Analyzing these distributions allows us to assess the robustness of the clustering results against random perturbations.

For this analysis, we generated 10,000 bootstrap resamples from the dataset, with a relative sample size ranging from 0.5 to 1.5. Clustering was performed on each bootstrap resample, allowing us to test the significance of each cluster at a specified Type I error probability level. This procedure was carried out separately for each position: Forward, Midfielder, Defender, and Goalkeeper.

2.2.2. Second Step: Finding Representatives of Clusters In this step, our goal is to select representative features for each cluster obtained in the first step. These representatives are crucial for reducing the dimensionality of the dataset and identifying the most significant features for each position. To identify representatives for each cluster, we first calculate the Pearson correlation

[†]Edd Webster Github FBref Player Stats Data Engineering.ipynb

matrix of the features within each cluster. Next, we compute the row sums (or equivalently column sums) of the correlation matrix. The feature with the highest row sum value is selected as the first representative. Subsequently, features with correlation coefficients exceeding a specified threshold θ with the selected representative are excluded from further consideration in the correlation matrix. This recursive process continues until no feature remains in the cluster's correlation matrix. During this iterative process, the most correlated feature is selected as the representative in each round and features closely related to the chosen representative are excluded from further consideration. This method ensures that the selected representatives are the most informative features within each cluster.

The threshold θ is a hyperparameter in the algorithm which can take any value between 0 and 1. Increasing the threshold θ leads to a less stringent process, ensuring that only highly correlated features are discarded in each iteration. In this research we selected 0.7 which represents a balanced trade-off between eliminating redundancy and preserving informative features.

To strengthen the robustness and reliability of the feature selection process, an extensive sensitivity analysis was conducted by varying both the threshold parameter θ and the number of bootstrap samples. Specifically, we employed 10,000 bootstrap iterations to ensure statistical stability in the selection process. Higher numbers of bootstrap iterations were found to improve convergence, stabilizing the selection of representative features across multiple resamples. This dual-parameter analysis ensured that the final set of representatives was not overly sensitive to specific parameter choices or sampling variability. The detailed results of this analysis, including the impact of varying θ and bootstrap size, are provided in the appendix.

The representatives chosen for different positions are reported in the results section. It is noteworthy that the representatives selected for different positions demonstrate distinctiveness, highlighting the distinct feature sets relevant to each role on the field.

2.2.3. Third Step: Predicting remaining Features of the Players Using Representative Features In this step, our objective is to predict the remaining features for each player within each position using the selected representative features. To achieve this, we employ four distinct algorithms: linear regression, ridge regression, least absolute shrinkage and selection operator (LASSO) regression, and Random Forest [9, 8, 20, 2].

Given that each position involves multiple clusters, each with one or more representative features, these representatives collectively serve as predictor variables, while the remaining features are treated as response variables. For each response variable, the entire group of representatives is used as the predictor set. Models are fitted using an 80:20 train-test split ratio, and the corresponding test errors are computed and recorded. This process is repeated for each response variable, keeping the same predictor variables and train-test split ratio.

For comprehension, the three main steps of methodology have been condensed and visually represented in Figure 1.

3. Results

In this section, we present a detailed analysis of the results derived from the three steps of the methodology to understand the performance and outcomes associated with each step.

3.1. Results of the First Step

As previously discussed, the optimal number of clusters was determined using the Kelley-Gardner-Sutcliffe (KGS) criterion. Specifically, the optimal number of clusters was found to be 6 when consider all positions, 7 for goalkeepers, 6 for defenders, 7 for midfielders, and 6 for forwards.

Figure 2 illustrates the clustering results across all positions, with each cluster represented by a distinct color. However, as shown in Figure 3, many clusters are not enclosed within red rectangles, which indicate statistical significance at the 0.05 level for Type I error. This suggests that these clusters are not robust against random sample variations, emphasizing the need for position-specific cluster analyses.

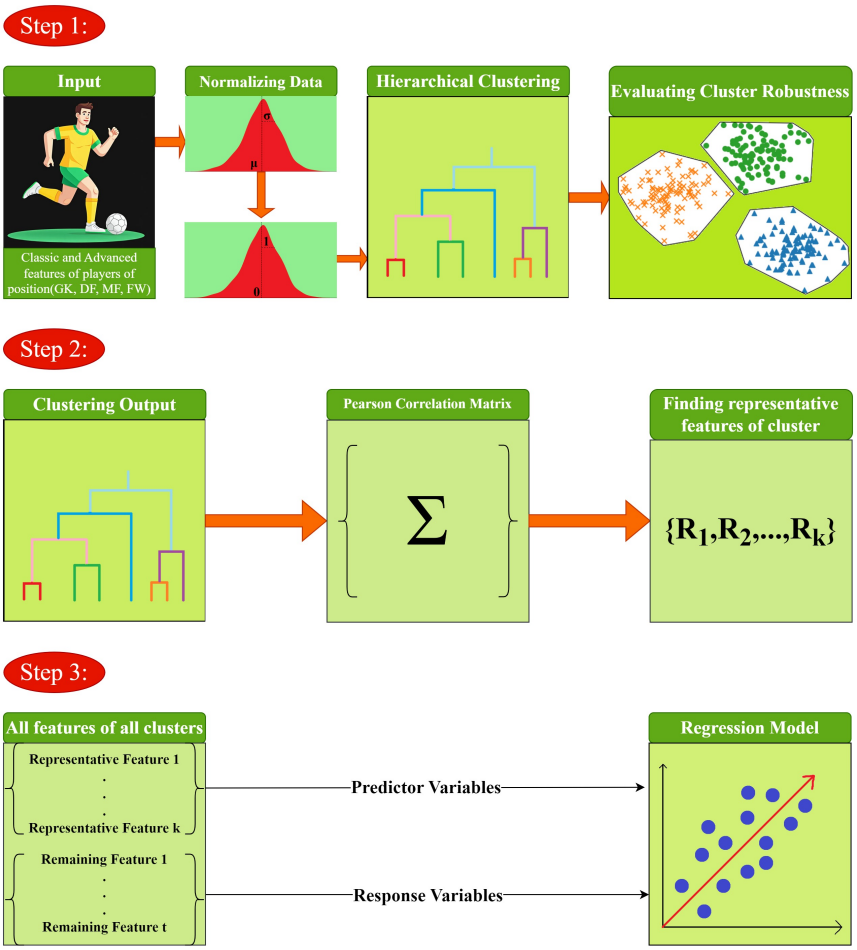
Subsequently, Figure 4 and Figure 5 present clustering and bootstrap results for the forward position. As shown in Figure 5, most clusters are enclosed within red rectangles, indicating statistically significant and robust patterns.

Clustering and bootstrap results for other positions (goalkeeper, defender, and midfielder) are provided in appendix section. These figures demonstrate the robustness of position-specific clusters, as indicated by red rectangles highlighting significant patterns across different random samples.

3.2. Results of the Second Step

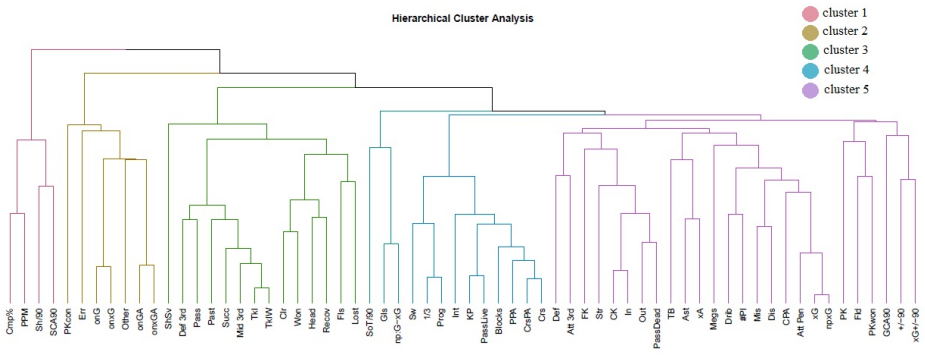
In the second step of our analysis, we aimed to identify position-specific features that define distinct playing styles and roles on the field. Summaries of the key results for each position are provided below.

3.2.1. Position-Specific Feature Selection We identified a set of representative features that provide insights into the playing style of players. The features selected for each position are summarized in Table 1 below. Notably, for the goalkeeper position, 11 features exhibited no variance across all goalkeepers, rendering them constant. These features were excluded from the clustering



1.jpg 1.bb

Figure 1. The Three Main Steps of the Methodology. 1) The procedure normalizes both classic and advanced player features as inputs for hierarchical clustering, followed by evaluating cluster robustness. 2) Representative features are identified from each cluster using the Pearson correlation matrix derived from the clustering output. 3) The representative features are then used as predictor variables, with the remaining features serving as response variables in the regression models.

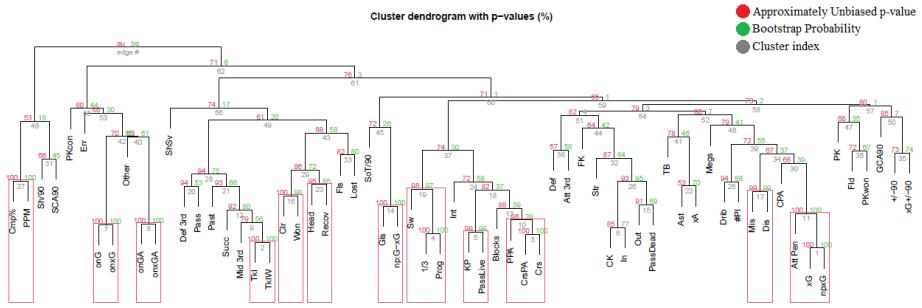


2.png 2.bb

Figure 2. Clustering features for all positions.

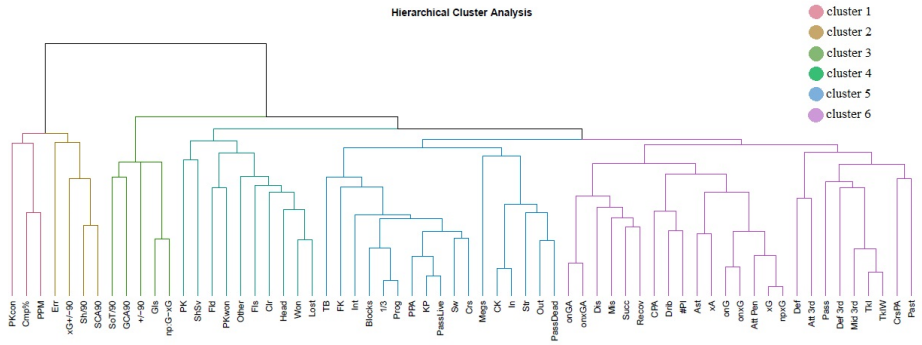
process since their lack of variability made them ineffective for distinguishing between clusters. As outlined in Table 1, some features are common across multiple positions, while others are unique to specific roles.

3.2.2. Highlights of Common Features Across Multiple Positions: Our analysis identified both common and position-specific features that contribute to the performance of players in different roles. Common features, such as Goals (Gls), Expected goals by team while on pitch (onxG), and Recoveries (Recov), are significant across multiple positions, reflecting the versatility required



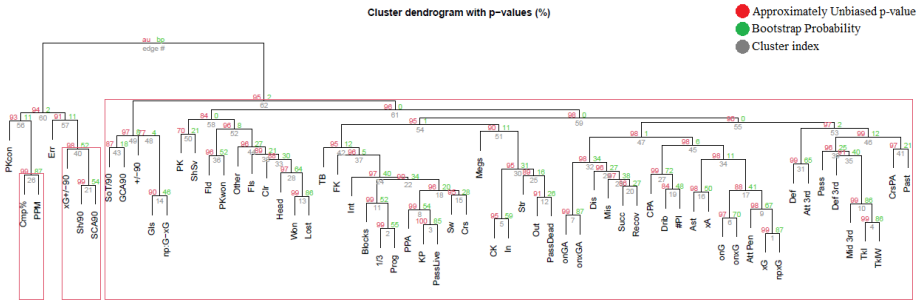
3.png 3.bb

Figure 3. The result of bootstrap on the clustering of all positions.



4.png 4.bb

Figure 4. Clustering features for the forward position.



5.png 5.bb

Figure 5. The result of bootstrap on the clustering of forward position.

Table 1. Representative Features for Each Player Position.

Position	Representative Features
Forward	Gls, onxG, Recov, Lost, Sh/90, Sw, PPM
Midfielder	Gls, KP, CK, onxG, PPA, Sh/90, Recov, xG+/-90, Drib, Tkl, Dis
Defender	Gls, PKcon, PassLive, onxG, 1/3, CK, FK, Tkl, #Pl, Cmp%, Recov
Goalkeeper	SCA90, Def 3rd, np:G-xG, PassDead, Mid 3rd, xA, Won, xG, Att Pen, PPA, Cmp%, PKcon, onxG

in modern football. Players are increasingly expected to contribute to both defensive and offensive actions, regardless of their primary position. This highlights the evolving nature of football, where even defenders and midfielders are involved in creating scoring opportunities, and forwards are expected to contribute to defensive efforts.

1. Goals (Gls): Common to Defenders, Midfielders, and Forwards. Scoring goals is a critical performance feature not only for forwards but also for midfielders and defenders who contribute to the attack in set-piece situations.
2. Expected goals by team while on pitch (onxG): Present in all four positions. This feature is vital as it reflects the quality of scoring opportunities a player creates or faces, regardless of their specific role on the field.
3. Recoveries (Recov): Found in Defenders, Midfielders, and Forwards. This shows the importance of regaining possession across various positions, highlighting a players contribution to maintaining or regaining control of the ball.

4. Pass Completion Percentage (Cmp%): Shared by Goalkeepers and Defenders. This is a crucial feature for maintaining possession and initiating play, particularly from the back.
5. Penalty Kicks Conceded (PKcon): Important for both Goalkeepers and Defenders, as it directly relates to their defensive responsibilities and can have significant impacts on match outcomes.

3.2.3. Highlights of Position-Specific Features:

1. Shot-Creating Actions per 90 minutes (SCA90) and completed dead-ball passes that lead to a shot attempt (PassDead) are unique to Goalkeepers, indicating their role in distributing the ball effectively.
2. Number of Players Dribbled Past (#Pl) and Shots from Free Kicks (FK) are key features for defenders, showcasing their ability to intercept and bypass opponents while also contributing offensively by creating goal-scoring opportunities from set-pieces.
3. Passes that directly lead to a shot (KP) and successful dribbles that lead to a shot attempt (Drib) are specific to Midfielders, highlighting their role in advancing the ball and creating scoring opportunities.
4. Points Per Match (PPM) and passes that travel more than 40 yards of the width of the field (Sw) are specific representatives of the forward position, reflecting forward's impact on team success through scoring efficiency and creating attacking opportunities with long passes.

These selected features provide insights into the unique playing patterns and responsibilities of each position. For example, the smaller number of representative features for forwards, compared to other positions, suggests that forwards exhibit more similar and consistent playing patterns. In contrast, goalkeepers, despite having fewer features due to the zero-variance issue, have a greater number of representative features, indicating a diverse range of behaviors that are effectively captured by the selected features. This diversity among goalkeepers might reflect the varying strategies teams employ in using their goalkeepers both defensively and in initiating attacks. Understanding these differences requires a deep knowledge of football analytics and the specific roles that different positions play on the field.

3.3. Results of the Third Step

In the third step of our analysis, the aim is to predict the remaining features using the representative features identified for each position in the second step. To achieve this, we applied several regression models including linear regression, ridge regression, LASSO regression, and Random Forest for different player positions. The performance of these models was assessed based on the mean squared error (MSE), with the statistical summaries of their performance detailed in Tables 2 to 5.

3.3.1. Forward Position: Regression models struggled to predict Cmp%, SCA90, and +/-90 for forwards. These features were subsequently removed from the set of response variables, leading to a final set of 54 response variables for forwards. The Linear Regression model also emerged as the best-performing model for forwards, as determined by the mean MSE (see Table 2).

Also SHapley Additive exPlanations (SHAP) [11] feature importance values for predicting key offensive performance metrics in forward players are provided in the appendix (see Table 3).

Table 2. Statistical summary of the mean squared error of linear regression, ridge regression, LASSO regression, and Random Forest algorithms for the forward position on the test data.

Model	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Linear Regression	0.01396323	0.07826028	0.1996304	0.4015492	0.5651408	2.278673
Ridge Regression	0.01591430	0.10709447	0.2330898	0.4539727	0.5815099	2.312739
LASSO Regression	0.03184951	0.25317293	0.7067298	0.7923945	1.1201457	2.377311
Random Forest	0.01359324	0.08221598	0.1876751	0.4061360	0.6058044	2.197454

Table 3. SHAP feature importance values for predicting key offensive features in forward players.

Feature	Gls	onxG	Recov	Lost	Sh.90	Sw	PPM
xG	0.3403896	0.01299458	0.002425594	0.01225014	0.69982369	0.08897313	0.06115657
xA	0.1588762	0.002613171	0.2180672	0.02968150	1.06232533	0.1666098	0.06184032
Att 3rd	0.2950519	0.005665021	0.06433792	0.02435168	0.52252373	0.4203702	0.03762546

3.3.2. Midfielder Position: The regression models for midfielders showed poor performance in predicting the features Cmp%, 1/3, SCA90, and +/-90. Due to their low predictive accuracy, these features were excluded from the response variables, resulting in a final set of 49 response variables for midfielders. The analysis revealed that the Random Forest model performed best for this position, with the lowest mean MSE among the models evaluated.

3.3.3. Defender Position: For defenders, the regression models performed well in predicting all features, and no features were excluded from the set of response variables, resulting a final set of 53 response variables. The Linear Regression model was identified as the most effective for the defender position, showing superior predictive capability based on the mean MSE.

3.3.4. Goalkeeper Position: For the goalkeeper position, 4 features FK, In, Megs, and Gls exhibited near-zero variance and were subsequently excluded from the analysis. After removing these features, the final set of response variables for goalkeepers includes 36 response variables. Among the regression models tested, the Random Forest algorithm demonstrated the best performance for this position.

Statistical summary of the mean squared error of the regression models for the midfielder, defender and goalkeeper positions on the test data provided in the appendix section.

Based on the very low mean squared error (MSE) values, we can conclude that the remaining features for each position can be predicted accurately using the representative features identified in the second step of our analysis. This high level of predictive ability demonstrates the effectiveness of the selected representative features and highlights the robustness of the clustering method in capturing the essential aspects of player performance across different positions.

4. Discussion

An additional objective of this study was to predict player positions using different sets of features. Goalkeepers were excluded from this part of the analysis due to their distinct roles, and players with multiple recorded positions were also excluded to maintain consistency. A new dataset was created by merging the FBref dataset with the SoFIFA dataset [15], resulting in a combined dataset of 1,625 players.

We employed a Random Forest model to classify players into the three positions: defender, midfielder, or forward. In our initial experiment, we used only the classic features identified as the most significant for different positions in a previous study [15], achieving an accuracy of 87.65%. When we used the advanced features identified in this study as representative features, the model's accuracy was 70.77%. Finally, combining both sets of features as predictor variables, the model achieved an accuracy of 87.35%. These results suggest that classic FIFA features are more effective than the advanced features in determining player position. The advanced features, while useful for assessing player performance, tend to evaluate aspects of a match that are not inherently tied to specific positions and can be relevant across different areas of the field. This highlights the complexity of player roles in modern football, where position-specific features may not always align with broader performance features.

This study has several limitations that warrant acknowledgment. Firstly, hierarchical clustering was employed due to its superior interpretability, particularly through dendrogram representations that allow for a clearer understanding of group relationships. While alternative clustering techniques such as k-means [12] were also tested, they did not improve the quality or clarity of the clustering outcomes. Moreover, the k-means results, which are provided in the appendix, lacked the interpretability and structural coherence offered by hierarchical methods. As a result, hierarchical clustering was preferred to support both analytical depth and visual insight into player clustering.

Secondly, the analysis was restricted to the most recent version of each players features. Expanding the analysis to include features from multiple seasons could provide a more comprehensive understanding of player performance and dynamics. Thirdly, the Pearson correlation matrix was used for both clustering and the selection of representative features. Exploring alternative methods, such as Spearmans rank correlation, may offer additional perspectives and robustness to the results. Fourthly, some features like Cmp% were more difficult to predict due to the limited scope of our feature set. These outcomes are influenced by complex contextual factors not captured in our model. Incorporating more advanced and domain-specific features would likely improve predictive accuracy for such metrics.

Additionally, in the second step of the methodology, correlations exceeding the threshold θ were excluded without considering their absolute magnitudes. Accounting for absolute values could potentially yield more informative results, as both positive and negative correlations are equally relevant to the analysis. Finally, while validating clusters against expert-defined groupings or labeled data would enhance the practical relevance of our findings, such data are not publicly available in a standardized or comprehensive format. In particular, detailed tactical roles or expert annotations at scale are scarce, making it challenging to perform direct real-world validation. However, this remains an important direction for future work as richer, domain-specific datasets become accessible. These limitations underscore areas for future research to refine and extend the findings of this study.

Furthermore, the applicability of this methodology is not confined to the leagues analyzed in this study. The proposed framework is generalizable and can be extended to other football leagues or different team sports, provided that a comprehensive set of classic and advanced performance metrics is available. Applying this method to sports like basketball, ice hockey, or cricket could yield valuable insights into their respective feature spaces, helping to identify core performance indicators unique to each domain.

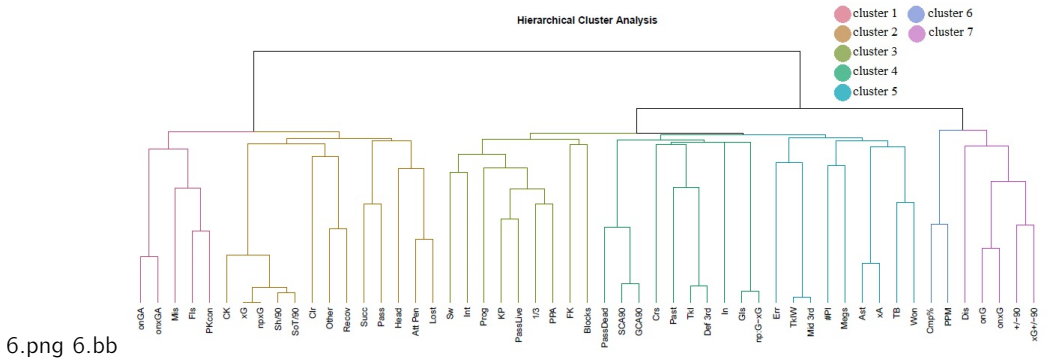
5. Conclusion

This study undertook a comprehensive analysis of both traditional and advanced features of football players to explore their utility in player evaluation and performance prediction. The findings demonstrated that these features could be robustly clustered according to players' positions, enabling the identification of representative features that encapsulate the primary information within each cluster. The results further showed that these representative features can accurately predict other player features, thereby simplifying the evaluation process while maintaining a low error margin.

Appendix

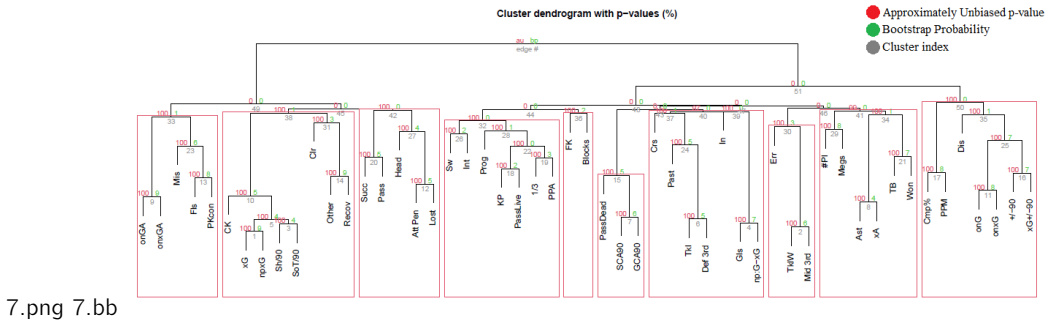
The appendix provides figures illustrating clustering results and robustness analyses, along with analysis of key parameters threshold (θ) and number of bootstrap resamples. also detailed statistical summaries of model performance across player positions. It supports and complements the main findings discussed in the paper.

Results of clustering and bootstraps for goalkeeper, defender, midfielder positions :
In this section, we present figures which are the results of clustering and robustness analysis for the goalkeeper, defender, and midfielder positions (Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11). The clustering reveals distinct groups within these positions based on players features. The robustness analysis evaluates the stability of these clusters, ensuring their consistency across different data variations and methodological approaches.



6.png 6.bb

Figure 6. Clustering features for the goalkeeper position.



7.png 7.bb

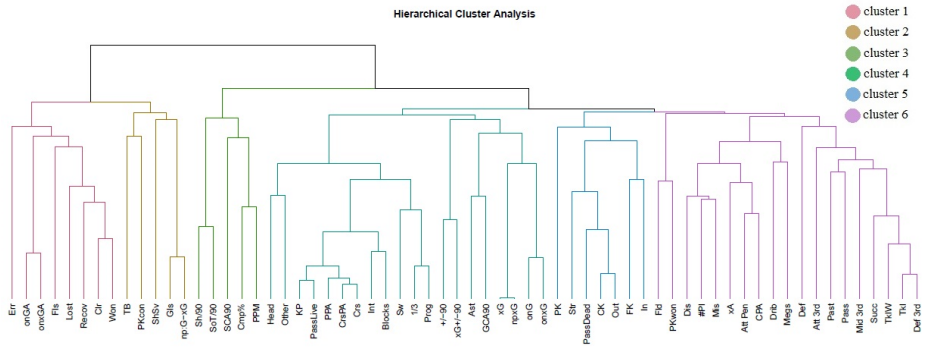
Figure 7. The result of bootstrap on the clustering of goalkeeper position.

analysis of key parameters threshold (θ) and number of bootstrap resamples:

To evaluate the robustness of the representative feature selection process, we conducted a sensitivity analysis focusing on two key parameters: the correlation threshold θ and the number of bootstrap samples. Both parameters significantly influence the stability and quality of the selected representative features.

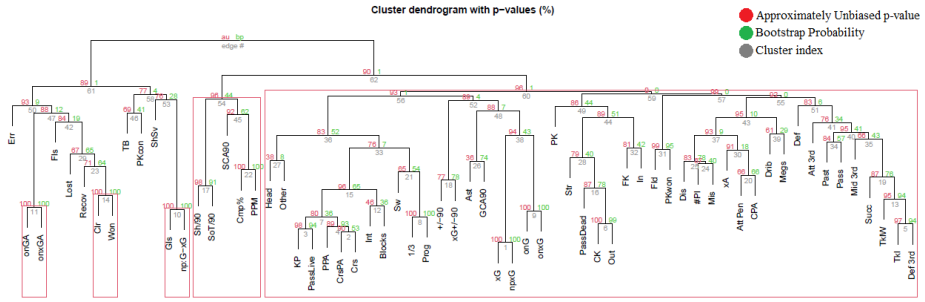
For the forward position, setting a low value of θ enforces a stricter exclusion criterion. As a result, fewer features are retained as representatives, prioritizing lower redundancy and higher distinctiveness. For example, with minimum value of θ or 0, the following features were selected:

PPM, Sh/90, Gls, Lost, Sw, onxG



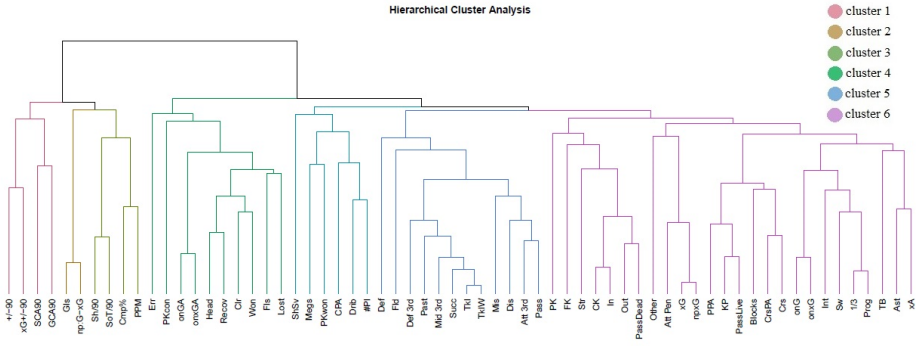
8.png 8.bb

Figure 8. Clustering features for the defender position.



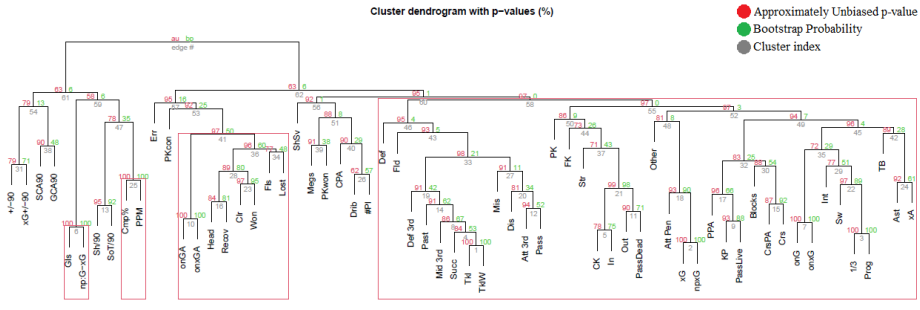
9.png 9.bb

Figure 9. The result of bootstrap on the clustering of defender position.



10.png 10.bb

Figure 10. Clustering features for the midfielder position.



11.png 11.bb

Figure 11. The result of bootstrap on the clustering of midfielder position.

Conversely, the maximum value of θ or 1 leads to a more permissive selection process, allowing highly correlated features to remain. This results in a substantially larger and more redundant set of representative features. When a high θ was applied, the representatives included the following:

PKcon, Cmp%, PPM, Err, xG+/–90, Sh/90, SCA90, SoT/90, GCA90, +/-90, Gls, np:G–xG, PK, ShSv, Fld, PKwon, Other, Fls, Clr, Head, Won, Lost, TB, FK, Int, Blocks, 1/3, Prog, PPA, KP, PassLive, Sw, Crs, Megs, CK, In, Str, Out, PassDead, onGA, onxGA, Dis, Mis, Succ, Recov, CPA, Drib, #Pl, Ast, xA, onG, onxG, Att Pen, xG, npxG, Def, Att 3rd, Pass, Def 3rd, Mid 3rd, Tkl, TklW, CrsPA, Past.

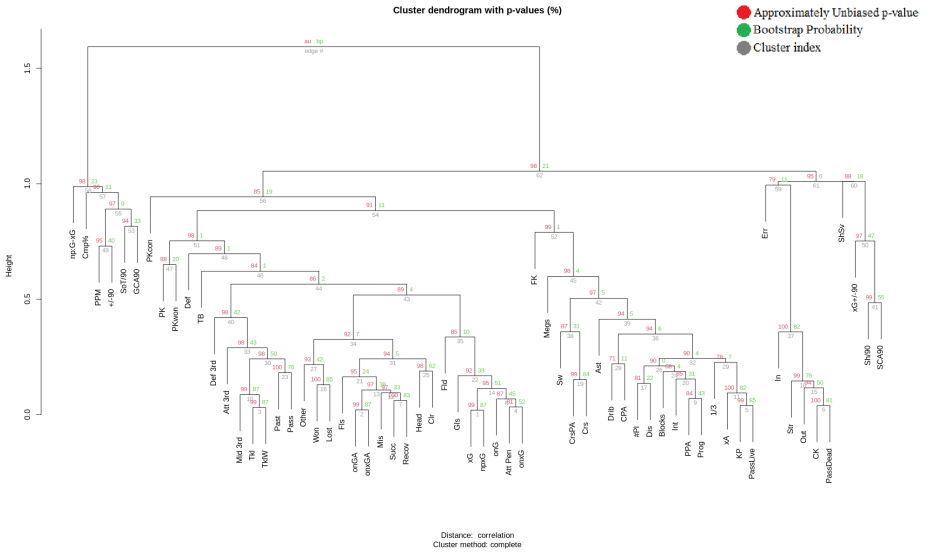
While high θ values retain a broader set of features, they risk over-representing redundant information. On the other hand, very low θ values may omit informative features. Empirical evaluation suggests that a moderate threshold, specifically $\theta = 0.7$, offers the most balanced outcome retaining informative yet non-redundant features and resulting in stronger downstream performance. At this value, the selected representatives for the forward position were:

Gls, onxG, Recov, Lost, Sh/90, Sw, PPM

These features demonstrated improved predictive capacity when used in subsequent modeling, as discussed in the main results section.

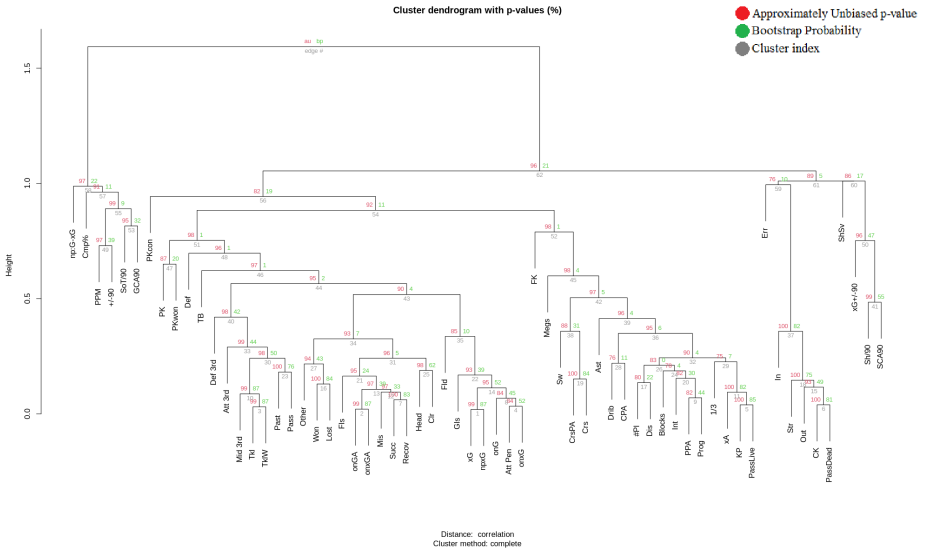
We investigated the impact of varying the number of bootstrap samples used during the representative selection process. The analysis revealed that increasing the number of bootstraps led to greater stability and convergence in the resulting feature sets. Beyond a certain threshold, further increases in bootstrap size yielded negligible differences, indicating convergence in the selection outcome and confirming the robustness of the method.

Please refer to Figure 12 and Figure 13 for detailed visualization of bootstrap stability results.



12.png 12.bb

Figure 12. The result of bootstrap on the clustering of forward position with 1000 bootstrap samples.



13.png 13.bb

Figure 13. The result of bootstrap on the clustering of forward position with 5000 bootstrap samples.

Results of statistical summary of regression models

In this section, we present tables which are results of statistical summary of the mean squared error of linear regression, ridge regression, LASSO regression, and Random Forest algorithms for the goalkeeper, defender, and midfielder positions (see Table 4, Table 5 and Table 6).

Table 4. Statistical summary of the mean squared error of linear regression, ridge regression, LASSO regression, and Random Forest algorithms for the goalkeeper position on the test data.

Model	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Linear Regression	0.00000000	0.00062690	0.00323775	0.1815835	0.06465402	2.700236
Ridge Regression	0.00000051	0.00035200	0.00391465	0.1790926	0.09998129	1.967476
LASSO Regression	0.00000051	0.00035418	0.00471001	0.2758714	0.23186958	2.074257
Random Forest	0.00000466	0.00017094	0.00170912	0.1448819	0.09338135	1.480467

Table 5. Statistical summary of the mean squared error of linear regression, ridge regression, LASSO regression, and Random Forest algorithms for the defender position on the test data.

Model	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Linear Regression	0.00910106	0.07154439	0.1759135	0.3084827	0.3090469	1.949695
Ridge Regression	0.01208780	0.15093635	0.2281311	0.3815073	0.4131106	1.973462
LASSO Regression	0.01547973	0.32605410	0.7076287	0.7863818	1.0125143	2.102995
Random Forest	0.00966391	0.08557124	0.1840530	0.3154740	0.3468538	1.602468

Table 6. Statistical summary of the mean squared error of linear regression, ridge regression, LASSO regression, and Random Forest algorithms for the midfielder position on the test data.

Model	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Linear Regression	0.01653887	0.09223787	0.1739805	0.2802615	0.4036778	1.179507
Ridge Regression	0.02170337	0.13250925	0.2332162	0.3104135	0.3623688	1.120241
LASSO Regression	0.02971558	0.43370329	0.6651402	0.7250956	1.0359088	1.872893
Random Forest	0.01562708	0.09555345	0.1942241	0.2762252	0.3735727	1.430120

Results of statistical summary of regression models with K-means for forward position

In this section results of statistical summary of the mean squared error of regression models with the K-means method for forward position are provided to show results for alternative clustering methods.

Table 7. Statistical summary of the mean squared error of linear regression, ridge regression, LASSO regression, and Random Forest algorithms for the forward position on the test data and k-means method.

Model	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Linear Regression	0.022689575	0.2192772	0.3309027	0.4708728	0.5595186	3.794340
Ridge Regression	0.005429106	0.2842394	0.3477696	0.5280368	0.5794328	3.818310
LASSO Regression	0.004407625	0.6864038	0.9987054	1.0343783	1.2418085	3.830700
Random Forest	0.045374693	0.2172367	0.3070128	0.4769509	0.5295468	3.191591

Research dataset

The dataset utilized in this research, FBref, is publicly available for both viewing and download from https://github.com/eddwebster/football_analytics/tree/master/data/fbref/raw/outfield

References

1. K. Apostolou and C. Tjortjis. Sports analytics algorithms for performance prediction. In *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4. IEEE, 2019.
2. L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
3. T. Decroos and J. Davis. Interpretable prediction of goals in soccer. In *the AAAI-20 workshop on artificial intelligence in team sports.*, Hilton Midtown, New York, NY, USA, 2019.
4. B. Efron. Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*. Springer New York, New York, NY, 1992.
5. H. Eggels, R. V. Elk, and M. Pechenizkiy. Explaining soccer match outcomes with goal scoring opportunities predictive analytics. In *3rd Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA 2016). CEUR-WS.org*, 2016.
6. FBref. Player standard stats. https://fbref.com.
7. A. Garca-Aliaga, M. Marquina, J. Coteron, A. Rodrguez-Gonzlez, and S. Luengo-Sanchez. In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, 16(1):148–157, 2021.
8. A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
9. G. James, D. Witten, T. Hastie, and R. Tibshirani. Linear regression: Multiple linear regression. In *An Introduction to Statistical Learning: With Applications in R*, pages 71–82. Springer, 2013.
10. L. A. Kelley, S. P. Gardner, and M. J. Sutcliffe. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Engineering, Design and Selection*, 9(11):1063–1065, 1996.
11. S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
12. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, University of California Press, 1967.
13. F. Y. Meybodi and C. Eslahchi. Predicting anti-cancer drug response by finding optimal subset of drugs. *Bioinformatics*, 37(23):4509–4516, 2021.
14. T. Narizuka and Y. Yamazaki. Clustering algorithm for formations in football games. *Scientific reports*, 9(1):13172, 2019.
15. M. Nouraie and C. Eslahchi. Positioning soccer players for success: A data-driven machine learning approach. *Computational Mathematics and Computer Modeling with Applications (CMCMA)*, 2(1):24–33, 2023.
16. M. Nouraie, C. Eslahchi, and A. Baca. Intelligent team formation and player selection: a data-driven approach for football coaches. *Applied Intelligence*, 53(24):30250–30265, 2023.
17. V. C. Pantzalis and C. Tjortjis. Sports analytics for football league table and player performance prediction. In *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8. IEEE, 2020.
18. C. Soto-Valero. A Gaussian mixture clustering model for characterizing football players using the EA sports' FIFA video game system.[Modelo basado en agrupamiento de mixturas Gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos FIFA de EA sports]. *RICYDE. Revista Internacional de Ciencias del Deporte*, 13(49):244–259, 2017.
19. M. Tavana, F. Azizi, F. Azizi, and M. Behzadian. A fuzzy inference system with application to player selection and team formation in multi-player sports. *port Management Review*, 16(1):97–110, 2013.
20. R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.
21. O. Uzochukwu and P. Enyindah. A machine learning application for football players selection. *International Journal of Engineering Research & Technology*, 4(10):459–465, 2015.
22. C. P. Wibowo. Clustering seasonal performances of soccer teams based on situational score line. *Communications in Science and Technology*, 1(1):1–6, 2016.

