MSc in Data Analytics in Football
In collaboration with
UCAM (Catholic University of Murcia)

Module 12:
Masters' Final Project

Decoding Player Roles: A Data-Driven Clustering Approach in Football

Tutor: Pablo Sanzol

Written by
Marwane Hamdani

Contents

1 Executive Summary **6**

2 Introduction **7**
2.1 Quick Overview 7
2.2 Objectives 8

3 Conceptual and Technological Architecture of the Project **9**

4 Methodologies and Techniques Employed **11**
4.1 Understanding the Business 11
4.2 Understanding the Data 11
4.3 Data Preparation 11
4.4 Score Calculation 12
4.5 Modeling 12
4.6 Evaluation 12
4.7 Visualization and Reporting 12

5 Work Development **14**
5.1 Understanding the Data 14
5.1.1 Data Source 14
5.1.2 Data Collection 14
5.1.3 Data Overview 14
5.2 Data Preparation 15
5.2.1 Data Cleaning 15
5.2.2 Feature Engineering 15
5.3 Scoring 15
5.4 Clustering 18
5.4.1 Defenders 19
5.4.2 Midfielders 19
5.4.3 Forwards 20

6 Results **21**
6.1 Clusters' Distribution . . . 21
6.1.1 Defenders . . . 21
6.1.2 Midfielders . . . 21
6.1.3 Forwards . . . 22
6.2 Analysis of Top Players in Each Cluster for Selected Teams . . . 23
6.2.1 Top Defenders by Cluster . . . 23
6.2.2 Top Midfielders by Cluster . . . 23
6.2.3 Top Forwards by Cluster . . . 24
6.2.4 Discussion of Results . . . 24
6.3 Examples of Players' Role Evolution . . . 24
6.3.1 Achraf Hakimi . . . 25
6.3.2 Kylian Mbappé . . . 25
6.3.3 Erling Haaland . . . 25
6.3.4 Joško Gvardiol . . . 26
6.3.5 İlkay Gündoğan . . . 26
6.3.6 Jude Bellingham . . . 26

7 Conclusions and Future Work **27**
7.1 Conclusions . . . 27
7.2 Future Work . . . 27

List of Figures

1 Python was the programming language used. 9
2 Google Colab is the environment used to run Python. 9
3 The workflow of the project. 13
4 FBRef: the data source. 14
5 Top 10 Players by Passing and Creativity Score 16
6 Top 10 Players by Defense Score 16
7 Top 10 Players by Possession and Dribbling Score 17
8 Top 10 Players by Shooting and Finishing Score 18
9 Distribution of Defenders Across Clusters 21
10 Distribution of Midfielders Across Clusters 22
11 Distribution of Forwards Across Clusters 22

List of Tables

1 Average Scores for Different Defenders' Clusters 19
2 Average Scores for Different Midfielders' Clusters 19
3 Average Scores for Different Forwards' Clusters 20
4 Top Defenders by Cluster for Selected Teams 23
5 Top Midfielders by Cluster for Selected Teams 23
6 Top Forwards by Cluster for Selected Teams 24
7 Role Evolution of Achraf Hakimi 25
8 Role Evolution of Kylian Mbappé 25
9 Role Evolution of Erling Haaland 25
10 Role Evolution of Joško Gvardiol 26
11 Role Evolution of İlkay Gündoğan 26
12 Role Evolution of Jude Bellingham 26

1 Executive Summary

This project presents an innovative approach to player role classification in football, leveraging advanced data analytics to move beyond traditional positional labels. By analyzing player performance data from the top five European leagues over the last three seasons, we developed a methodology to classify players into distinct roles based on their contributions across four key aspects of the game: Passing and Creativity, Defense, Possession and Dribbling, and Shooting and Finishing.

Utilizing clustering algorithms, we identified specific player roles within each positional group : defenders, midfielders, and forwards. This analysis uncovered unique roles such as Playmaking Defenders, Box-to-Box Midfielders, and Creative Forwards, providing deeper insights into the diverse ways players contribute to their teams. Additionally, the project tracked the evolution of several high-profile players by season, demonstrating how their roles have shifted over time.

The findings from this project offer practical applications for football managers, scouts, and analysts, enabling them to make data-driven decisions about player deployment, recruitment, and development. The methodology not only classifies players but also tracks their evolution, helping to identify emerging talents and optimize team strategies.

Future work could expand this analysis by incorporating data from leagues outside the top five in Europe, as well as integrating event data and positional tracking to further refine player role classification. Such enhancements would broaden the applicability of the model, providing even more granular insights into player performance across different contexts.

2 Introduction

2.1 Quick Overview

In the dynamic world of football, where strategies, formations, and player roles constantly evolve, understanding player performance beyond traditional statistics has become essential for gaining a competitive edge. This project leverages the power of data analytics to provide a nuanced perspective on player roles, offering insights that go beyond simple metrics like goals, assists, or tackles. By clustering players based on their performance metrics, we aim to uncover patterns and tendencies that can inform tactical decisions, player development, and even transfer strategies.

Football is no longer confined by rigid formations or static positions. As Luciano Spalletti famously said during his successful season, "Systems no longer exist in football, it's all about the spaces left by the opposition. You must be quick to spot them and know the right moment to strike, have the courage to start the move even when pressed." This philosophy highlights the fluid nature of modern football, where players constantly adapt to the game's demands rather than being tied to predefined positions. In this landscape, traditional labels like defender, midfielder, or forward fail to capture the full scope of a player's contributions.

Our project addresses this gap by using data to predict a player's optimal role based on their performance across multiple aspects of the game—passing and creativity, defense, possession and dribbling, and shooting and finishing. By calculating weighted scores for each of these aspects, we capture a comprehensive view of a player's abilities. These scores, derived from a variety of detailed performance metrics, serve as the foundation for the clustering process. Furthermore, we also calculate an overall score that combines these aspects, offering a holistic assessment of a player's contribution on the field.

By utilizing advanced clustering techniques, we move beyond simple positional classifications and instead offer a more sophisticated understanding of a player's role in the team's overall strategy. This not only helps coaches and analysts in maximizing player potential but also ensures that decisions about tactics, player recruitment, and development are backed by data-driven insights. The overall score and individual aspect scores can be particularly useful in the recruitment process, as they allow scouts and managers to identify players who excel in specific areas that align with the team's needs.

The utility of this project extends to various stakeholders within a football club. Coaches can use these insights to refine tactical setups, scouts can identify recruits who fit specific playing styles, and players themselves can understand how they compare to their peers. Moreover, tracking player roles across multiple seasons enables the identification of successful transitions between roles, creating a richer, data-driven narrative around player development.

In essence, this project is not just about clustering players; it's about redefining how we view player performance in football. By integrating advanced data analytics with practical football knowledge, we aim to create a tool that enhances decision-making processes at all levels of the sport, ensuring players are deployed in roles where they can have the greatest impact.

2.2 Objectives

The specific objectives of this project include:

- Collecting and preprocessing performance data from multiple seasons.
- Defining and calculating weighted scores for key performance metrics across different roles.
- Applying clustering algorithms to group players based on their performance data.
- Evaluating the clusters to ensure they provide meaningful insights into player roles.
- Analyzing the evolution of certain high-profile players' roles across multiple seasons.

These objectives are designed to build a comprehensive framework that categorizes players not just by their positional play, but also by the underlying attributes that contribute to their effectiveness in different roles. By adopting a data-driven approach, this project aims to create clusters that accurately reflect the dynamic and evolving nature of football, enabling more precise predictions of a player's optimal role. Additionally, by tracking the evolution of players season by season, the project provides insights into how roles can shift over time, further aiding in tactical decision-making, player development, and overall team performance. This ensures that each player is utilized to their fullest potential, contributing meaningfully to the team's success.

3 Conceptual and Technological Architecture of the Project

This project was developed using Python as the primary programming language, executed within the Google Colab environment.

Figure 1: Python was the programming language used.

Figure 2: Google Colab is the environment used to run Python.

Python's versatility and rich ecosystem of libraries made it an ideal choice for this project, allowing for efficient data processing, analysis, and model development. Google Colab provided a powerful and collaborative platform to run the code, taking advantage of cloud-based computing resources.
Within the Python ecosystem, the following libraries and frameworks were utilized:

- **Pandas and NumPy**: Libraries used for data manipulation and numerical computations.


- **scikit-learn**: A machine learning library used for implementing clustering algorithms and evaluating model performance.


- **Matplotlib and Seaborn**: Libraries used for data visualization and generating insights from the data.


- **Beautiful Soup**: A library used for web scraping, allowing for the extraction of player performance data from online sources.


In this project, a variety of tools and libraries were employed to ensure efficient data processing, analysis, and visualization. Python served as the backbone of the development, providing a flexible and powerful programming environment. Within Python, the Pandas and NumPy libraries were essential for handling and manipulating large datasets, offering functions to clean, transform, and analyze the data effectively. The scikit-learn library was integral to the implementation of machine learning algorithms, particularly for clustering players based on performance metrics, allowing for the extraction of meaningful patterns and insights.

Data visualization was handled by Matplotlib and Seaborn, which provided a range of plotting capabilities to represent the data graphically, making the insights more accessible and understandable. Additionally, the Beautiful Soup library was utilized for web scraping, enabling the extraction of detailed player performance data from various online sources. These tools collectively provided a robust framework for conducting comprehensive analyses and generating valuable insights from the data.

4 Methodologies and Techniques Employed

This project follows a structured approach to data analysis, incorporating various phases essential for the development and execution of the project. The methodology is designed to address the specific challenges of analyzing and clustering football players based on their performance data, with an additional focus on tracking the evolution of player roles over multiple seasons.

4.1 Understanding the Business

The first phase involves comprehending the project objectives and requirements from a strategic perspective. The primary goal of this project was to analyze football players' performance data from FBRef for the top 5 European leagues over the last three seasons.
The objective was to identify and cluster players based on their roles on the field, moving beyond traditional positional labels to provide deeper insights into optimal player positioning and strategy. Additionally, the project aimed to track the evolution of players' roles over time, offering valuable insights into how a player's contributions and positional responsibilities may change from season to season.

4.2 Understanding the Data

In this phase, the collected datasets of the last 3 seasons were thoroughly analyzed to understand their content and structure. The datasets included various performance metrics from FBRef, a well-known website for football statistics and player data, covering top players from major European leagues over the last three seasons.

- **Exploratory Data Analysis (EDA):** Key features, distributions, and potential anomalies were identified to gain a better understanding of the data.
- **Data Quality Assessment:** This involved checking for missing values, inconsistencies, and potential outliers to ensure the integrity of the data.

4.3 Data Preparation

The data preparation phase was critical to ensuring that the data was ready for modeling. This phase included:

- **Data Cleaning:** Handling missing values, correcting inconsistencies, and ensuring the data was in a usable format.
- **Feature Engineering:** Calculating weighted scores for different performance aspects such as Passing and Creativity, Defense, Possession and Dribbling, and Shooting and Finishing.
- **Data Transformation:** Standardizing and normalizing the data to ensure comparability across different features.
- **Data Integration:** Merging data from multiple seasons and ensuring consistent formats and scales across datasets.

4.4 Score Calculation

The scoring phase involved assigning weighted scores to players across different aspects of their game. Each performance metric was evaluated based on its significance within its respective category—Passing and Creativity, Defense, Possession and Dribbling, and Shooting and Finishing.
These scores were then aggregated to provide a comprehensive overview of a player's contributions. The final scores were standardized using MinMaxScaler to ensure consistency across different metrics, making the data ready for clustering analysis.

4.5 Modeling

In the modeling phase, clustering algorithms were applied to the prepared data to group players based on their performance metrics. The key steps were:

- **Algorithm Selection:** The K-Means clustering algorithm was selected for its effectiveness in grouping players based on performance metrics.
- **Model Training:** The algorithm was applied to the data.
- **Model Evaluation:** The quality of the clusters was assessed to ensure they provided meaningful insights into player roles.

4.6 Evaluation

The evaluation phase focused on assessing the effectiveness of the clustering model. The key activities included:

- **Cluster Validation:** The stability and interpretability of the clusters were analyzed to ensure reliable results.
- **Comparison with Known Roles:** The clusters were compared with known player roles to validate the model's accuracy and relevance.
- **Feedback Incorporation:** Refinements were made to the model based on the evaluation results, enhancing the overall performance of the clustering process.

4.7 Visualization and Reporting

The final phase of the project involved the visualization and reporting of the analysis results. Effective visualization techniques were employed to represent the clustering outcomes, player roles, and their evolution over time.
Bar plots were used to illustrate the distribution of players across different clusters within each position, while tables were created to highlight the top players in each cluster from the most prominent European football clubs. Additionally, player evolution was tracked and presented using tables that showed the shifts in cluster membership over the three seasons analyzed.
These visualizations, coupled with thorough reporting, provided valuable insights into player performance trends, role evolution, and the effectiveness of the clustering approach. The findings were systematically documented to ensure that they could be easily interpreted and utilized by football managers, recruiters, and analysts.

This phase ensured that the results were not only analytically sound but also presented in a manner that supports informed decision-making.

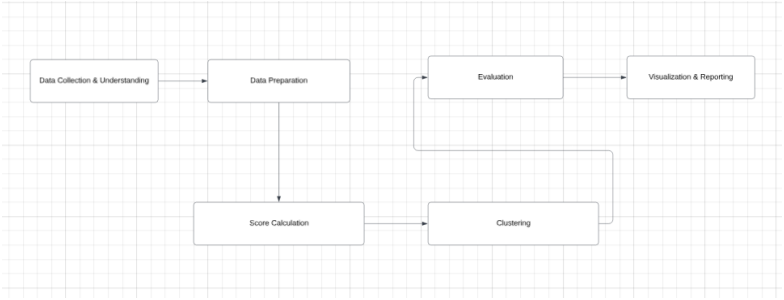

Figure 3: The workflow of the project.

5 Work Development

This section provides an in-depth look at the development of the project, focusing on the specific methodologies and techniques employed. The project was tailored to address the unique challenges and objectives of analyzing football player performance data.
Each phase of the project was meticulously executed to ensure the generation of comprehensive and meaningful insights. The development process involved a structured approach to data collection, preparation, analysis, and visualization, all aimed at understanding and categorizing player roles, as well as tracking their evolution across multiple seasons.

5.1 Understanding the Data

The next phase involved gaining a deep understanding of the data that would be used for analysis.

5.1.1 Data Source

The data is sourced from FBref.com, the ultimate hub for football analytics and insights! In collaboration with Opta, FBref provides comprehensive coverage of over 20 competitions, including the big five men's European leagues, the thrilling Champions League, and the electrifying World Cup.

Figure 4: FBRef: the data source.

5.1.2 Data Collection

The data was collected through web scraping techniques using the Beautiful Soup library in Python. The scraping process focused on capturing a wide range of performance datasets for each season, including passing, creativity, defense, possession, dribbling, shooting, and finishing.
We focused on collecting the datasets of the last 3 seasons.

5.1.3 Data Overview

The datasets for each season included player-specific metrics, such as minutes played, goals scored, assists, tackles, passes, and more. These metrics provided a comprehensive view of each player's contributions on the field, making it possible to analyze their roles in depth.

5.2 Data Preparation

Data preparation is a critical step that ensures the data is ready for analysis and modeling.

5.2.1 Data Cleaning

Our initial focus was on the 2023-24 season, the most recent dataset, where we filtered out players who had played fewer than 10 full matches to ensure the reliability of our analysis. Following this, we carried out feature engineering, which we will elaborate on in the next section.
Lastly, we identified and retained the same players across the three datasets (2022, 2023, and 2024) to enable meaningful comparisons and insights over time.

5.2.2 Feature Engineering

Feature engineering involved normalizing the statistics per 90 minutes for all players to ensure a fair comparison across different playing times. We then integrated these normalized stats into a single dataset, combining all aspects of a player's performance into one comprehensive row, rather than having separate datasets for each aspect. This approach allowed us to evaluate each player holistically, capturing the full scope of their contributions on the field.
For players who represented two clubs within a single season, we aggregated their statistics by summing metrics that are additive, such as total goals or assists, and calculating averages for percentage-based metrics, such as pass completion rate. This ensured that each player's overall performance for the season was accurately represented.

5.3 Scoring

Before clustering the players, we decided to focus on four key aspects of the game: Passing and Creativity, Defense, Dribbling and Possession Retention, and finally, Shooting and Finishing. For each aspect, we carefully selected the relevant features and assigned weights to each based on their importance to the overall performance. For example, positive contributions such as assists and goals were given higher positive weights, while negative contributions like red cards and miscontrols were given negative weights. The coefficients were determined to reflect the relative impact of each feature on a player's overall performance in that aspect.
Using these weights and coefficients, we generated a score for each player in each aspect, which was then scaled using the MinMaxScaler to produce a score between 0 and 1. This approach allowed us to objectively compare players across different roles and playing times. Below, we present the top-scoring players in each aspect for the 2023-24 season:

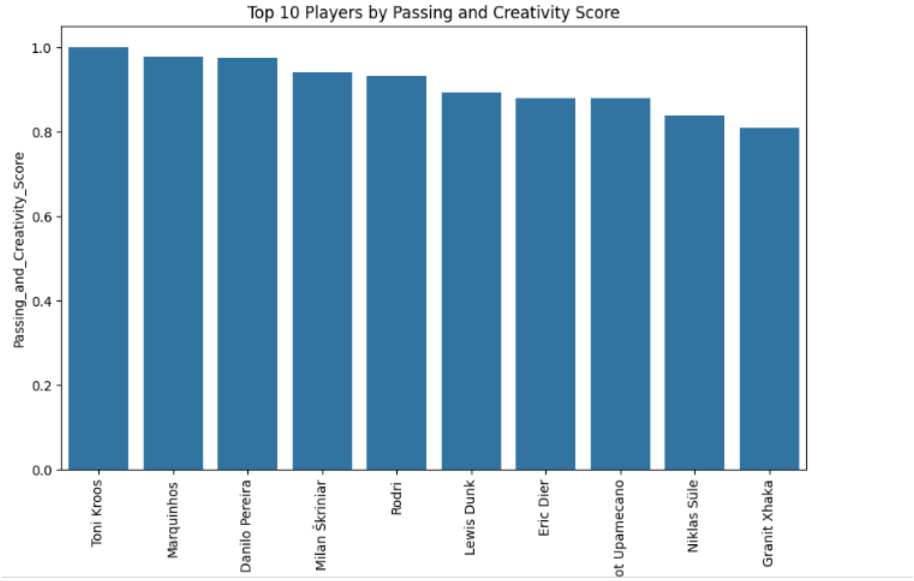

Figure 5: Top 10 Players by Passing and Creativity Score

Toni Kroos tops the list, reflecting his exceptional ability to influence the game through passing. It's noteworthy that several defenders also appear in this ranking, which highlights their role as key distributors, often taking risks to initiate plays from the back.

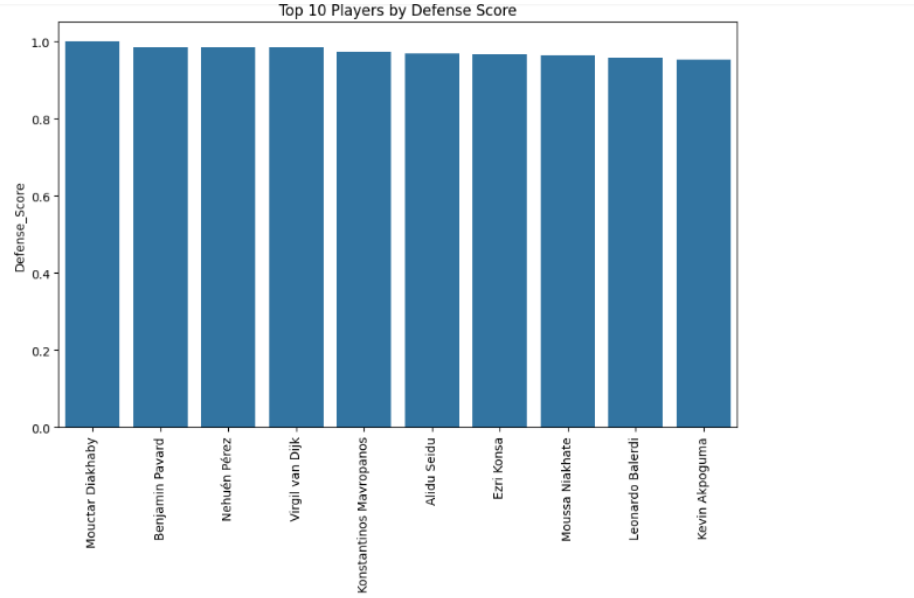

Figure 6: Top 10 Players by Defense Score

The list is dominated by well-known defenders, including Virgil van Dijk and Benjamin Pavard. However, it also features players like Mouctar Diakhaby, who may have flown under the radar but have been defensively solid throughout the season.

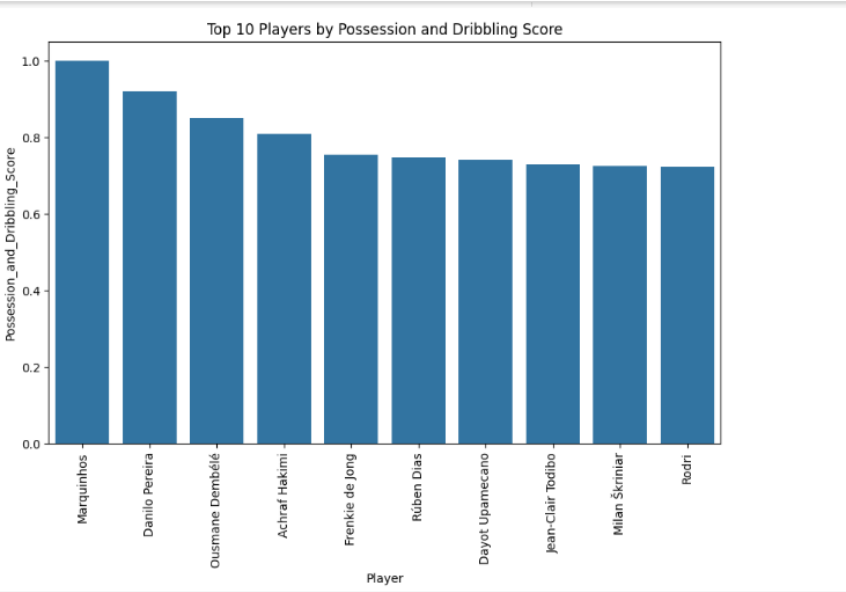

Figure 7: Top 10 Players by Possession and Dribbling Score

This ranking showcases a diverse group of players, including wingers, midfielders, defenders, and full-backs. Marquinhos and Danilo Pereira, primarily known as defenders, rank high due to their ability to carry the ball forward and maintain possession under pressure.

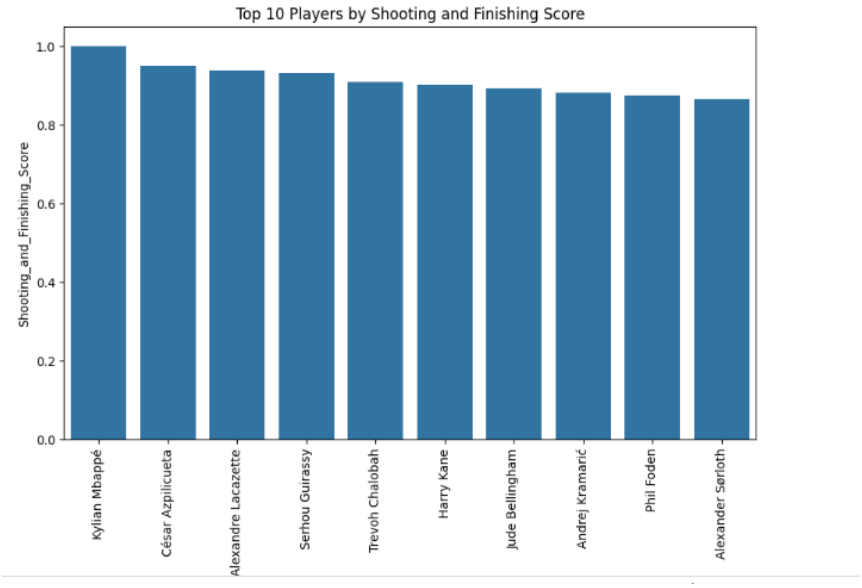

Figure 8: Top 10 Players by Shooting and Finishing Score

As expected, big names like Kylian Mbappé and Harry Kane appear in the top 10, showcasing their elite finishing ability. Interestingly, players like César Azpilicueta also make the list, indicating their efficiency in front of goal despite their primary defensive duties. Jude Bellingham's inclusion reflects his remarkable debut season at Real Madrid, where he has been a significant goal-scoring threat from midfield.

5.4 Clustering

In this section, we present the clustering results for each position: defenders, midfielders, and forwards. Each cluster represents a distinct role based on the players' performance metrics across various aspects of the game. The rationale for clustering players based on their primary position allows us to capture the diverse roles within each positional group more effectively, rather than merging all players into a single classification.
For players listed under multiple positions on FBRef, we designated their primary position based on the first listed role, ensuring consistency in the clustering process. This approach helps to maintain the integrity of the analysis, focusing on their most influential role within their teams.

5.4.1 Defenders

Cluster	Passing and Creativity	Defense	Possession and Dribbling	Shooting and Finishing
0	0.705296	0.678522	0.685330	0.324834
1	0.335333	0.702771	0.251732	0.247299
2	0.506902	0.726418	0.448801	0.286158
3	0.397642	0.625637	0.371762	0.508147

Table 1: Average Scores for Different Defenders' Clusters

Each cluster represents a unique role among defenders:

- **Cluster 0: Playmaking Defender** - This cluster is characterized by high scores in Passing and Creativity, indicating that these defenders are not only solid defensively but also play a crucial role in initiating attacks from the back. Their ability to distribute the ball effectively and contribute to building play sets them apart as playmaking defenders.
- **Cluster 1: Traditional Center-Back** - Defenders in this cluster have strong defensive metrics, with a primary focus on stopping the opposition and securing the defensive third. They have lower scores in Passing and Creativity, suggesting a more traditional role focused on defense rather than ball distribution.
- **Cluster 2: Balanced Defender** - Players in this cluster exhibit a well-rounded skill set, with decent scores across Defense, Passing and Creativity, and Possession and Dribbling. They contribute both defensively and offensively, making them versatile players who can adapt to different situations on the field.
- **Cluster 3: Attacking Full-Back** - Higher scores in Shooting and Finishing, coupled with moderate scores in Passing and Creativity, indicate that these defenders frequently join the attack, particularly from wide positions. They play a significant role in both defense and offense, often contributing to goal-scoring opportunities.

5.4.2 Midfielders

Cluster	Passing and Creativity	Defense	Possession and Dribbling	Shooting and Finishing
0	0.384597	0.597429	0.366038	0.326131
1	0.329509	0.427581	0.381887	0.584834
2	0.220109	0.529338	0.220834	0.394413
3	0.562443	0.541060	0.593102	0.451201

Table 2: Average Scores for Different Midfielders' Clusters

For midfielders, the clusters break down as follows:

- **Cluster 0: Defensive Midfielder** - High defensive scores characterize players in this cluster who prioritize protecting the backline and disrupting opposition play. Their lower scores in Passing and Creativity suggest a primary focus on defense rather than initiating attacks.

- **Cluster 1: Attacking Midfielder** - This cluster is marked by higher scores in Shooting and Finishing, indicating a focus on creating and converting goal-scoring opportunities. These midfielders are more involved in the offensive phase, playing a critical role in advancing the ball and taking shots.
- **Cluster 2: Holding Midfielder** - Players in this cluster have balanced defensive and possession metrics, with a particular emphasis on maintaining possession and dictating the tempo of play. They act as a bridge between defense and attack, ensuring stability and control in the midfield.
- **Cluster 3: Box-to-Box Midfielder** - Balanced across all metrics, these players are involved in both defensive and offensive phases, showcasing versatility. They are capable of contributing to both the attack and the defense, making them indispensable in a dynamic midfield role.

5.4.3 Forwards

Cluster	Passing and Creativity	Defense	Possession and Dribbling	Shooting and Finishing
0	0.119445	0.514846	0.210060	0.502545
1	0.080651	0.264397	0.155713	0.444475
2	0.106524	0.273596	0.181884	0.723758
3	0.221761	0.419692	0.460035	0.517466

Table 3: Average Scores for Different Forwards' Clusters

For forwards, the clusters are defined as:

- **Cluster 0: All-Round Forward** - Balanced scores across Shooting and Finishing, Defense, and Possession and Dribbling suggest that these forwards contribute to multiple aspects of the game. They are versatile players who can both create and score goals while also contributing defensively.
- **Cluster 1: Support Forward** - This cluster may not be the primary goal-scorers but play a crucial role in setting up offensive plays.
- **Cluster 2: Poacher** - High scores in Shooting and Finishing, coupled with lower scores in other metrics, define these players as primary goal-scorers who focus on converting opportunities in the box. They are less involved in playmaking or defensive duties.
- **Cluster 3: Creative Forward** - The best scores in both Passing and Creativity, along with solid Shooting and Finishing, indicate players who excel in both creating and finishing goal-scoring opportunities. These forwards are often key playmakers and goal-scorers in the final third.

6 Results

6.1 Clusters' Distribution

6.1.1 Defenders

The distribution of defenders across clusters highlights the diversity of roles within this position group. The "Balanced Defender" cluster has the highest representation, indicating that a significant number of defenders in modern football contribute effectively both in defense and possession. This reflects the growing importance of versatility in defensive roles, where players are expected to participate in ball progression and build-up play.

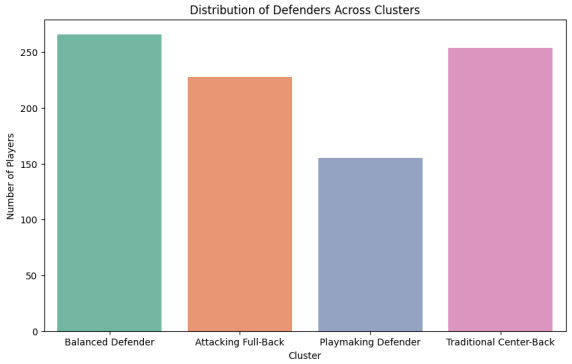

Figure 9: Distribution of Defenders Across Clusters

The "Traditional Center-Back" and "Attacking Full-Back" clusters also have substantial representation. This suggests that while traditional defensive duties remain crucial, there is a significant tactical shift towards using full-backs in more advanced, attacking roles.
Interestingly, the "Playmaking Defender" cluster, which features defenders with high passing and creativity scores, is less populated. This rarity underscores the specialized nature of defenders who can contribute creatively to their team's attacking play.

6.1.2 Midfielders

Midfielders are often seen as the heartbeat of a team, and their distribution across clusters reflects the multifaceted roles they play. The "Holding Midfielder" cluster is the most common, illustrating the importance of players who can maintain possession and control the tempo of the game.

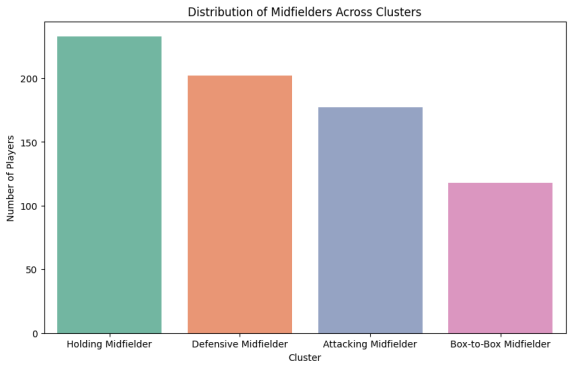

Figure 10: Distribution of Midfielders Across Clusters

"Defensive Midfielders" are also well-represented, which aligns with the tactical demand for players who can shield the defense and break up opposition attacks. The presence of "Attacking Midfielders" and "Box-to-Box Midfielders" further emphasizes the versatility required in the modern game, where midfielders are expected to contribute both offensively and defensively.

6.1.3 Forwards

The forward position has traditionally been associated with goal-scoring, but the distribution across clusters shows a more nuanced picture. "Creative Forwards" and "Support Forwards" have substantial representation, indicating that many forwards are now involved in playmaking and creating opportunities for others, rather than just finishing chances.

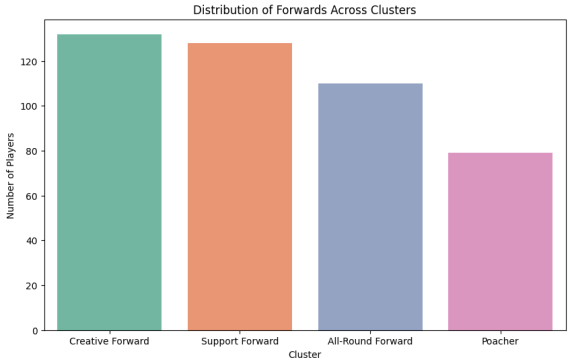

Figure 11: Distribution of Forwards Across Clusters

The "Poacher" cluster is less populated, suggesting that pure goal-scorers, who focus almost exclusively on finishing, are becoming less common. This shift may reflect the increasing tactical complexity of forward roles, where players are expected to be more involved in all phases of play. The "All-Round Forward" cluster further supports this, highlighting the demand for versatile forwards who can contribute in multiple areas.

6.2 Analysis of Top Players in Each Cluster for Selected Teams

To better understand the clustering, we focused on analyzing the top players from prominent European clubs such as Manchester City, Paris Saint-Germain, Bayern Munich, Liverpool, Real Madrid, Inter, Arsenal, Juventus, Milan, Barcelona, and Manchester United. By isolating these top teams, we aimed to evaluate the most influential players within each cluster across different positions.

6.2.1 Top Defenders by Cluster

Player	Squad	Year	Cluster	Overall Score	Descriptive Cluster Label
Rúben Dias	Manchester City	2024	0	0.699489	Playmaking Defender
Marquinhos	Paris S-G	2024	0	0.694675	Playmaking Defender
Eric Dier	Bayern Munich	2024	0	0.684431	Playmaking Defender
Jakub Kiwior	Arsenal	2024	1	0.485566	Traditional Center-Back
Stefan de Vrij	Inter	2024	1	0.440047	Traditional Center-Back
Andrea Cambiaso	Juventus	2024	1	0.417440	Traditional Center-Back
Benjamin Pavard	Inter	2024	2	0.585269	Balanced Defender
Francesco Acerbi	Inter	2024	2	0.580059	Balanced Defender
Oleksandr Zinchenko	Arsenal	2024	2	0.569740	Balanced Defender
Andrew Robertson	Liverpool	2024	3	0.601748	Attacking Full-Back
Lucas Vázquez	Real Madrid	2024	3	0.585024	Attacking Full-Back
Raphaël Guerreiro	Bayern Munich	2024	3	0.540236	Attacking Full-Back

Table 4: Top Defenders by Cluster for Selected Teams

6.2.2 Top Midfielders by Cluster

Player	Squad	Year	Cluster	Overall Score	Descriptive Cluster Label
Fabián Ruiz Peña	Paris S-G	2024	0	0.514234	Defensive Midfielder
Oriol Romeu	Barcelona	2024	0	0.508791	Defensive Midfielder
Alexis Mac Allister	Liverpool	2024	0	0.490198	Defensive Midfielder
Jude Bellingham	Real Madrid	2024	1	0.546223	Attacking Midfielder
Kevin De Bruyne	Manchester City	2024	1	0.486756	Attacking Midfielder
Dominik Szoboszlai	Liverpool	2024	1	0.436115	Attacking Midfielder
Adrien Rabiot	Juventus	2024	2	0.462245	Holding Midfielder
Yunus Musah	Milan	2024	2	0.410853	Holding Midfielder
Ruben Loftus-Cheek	Milan	2024	2	0.392514	Holding Midfielder
Rodri	Manchester City	2024	3	0.713601	Box-to-Box Midfielder
Hakan Çalhanoğlu	Inter	2024	3	0.707259	Box-to-Box Midfielder
Toni Kroos	Real Madrid	2024	3	0.695866	Box-to-Box Midfielder

Table 5: Top Midfielders by Cluster for Selected Teams

6.2.3 Top Forwards by Cluster

Player	Squad	Year	Cluster	Overall Score	Descriptive Cluster Label
Diogo Jota	Liverpool	2024	0	0.408282	All-Round Forward
Leandro Trossard	Arsenal	2024	0	0.391279	All-Round Forward
Erling Haaland	Manchester City	2024	0	0.358870	All-Round Forward
Gabriel Jesus	Arsenal	2024	1	0.280730	Support Forward
Marcus Thuram	Inter	2024	1	0.247509	Support Forward
Joselu	Real Madrid	2024	1	0.164997	Support Forward
Kylian Mbappé	Paris S-G	2024	2	0.534289	Poacher
Harry Kane	Bayern Munich	2024	2	0.372014	Poacher
Mohamed Salah	Liverpool	2024	2	0.366760	Poacher
Phil Foden	Manchester City	2024	3	0.513440	Creative Forward
Lee Kang-in	Paris S-G	2024	3	0.480658	Creative Forward
Kingsley Coman	Bayern Munich	2024	3	0.468166	Creative Forward

Table 6: Top Forwards by Cluster for Selected Teams

6.2.4 Discussion of Results

By analyzing the top players from prominent European clubs across each cluster, we gained insights into the roles these players embody within their teams. The clustering allowed us to categorize players into specific roles, providing a more nuanced understanding of their contributions beyond traditional positional labels.

For instance, in the defender category, players like Rúben Dias and Marquinhos were identified as Playmaking Defenders, reflecting their significant involvement in building play from the back. Similarly, midfielders like Rodri and Toni Kroos were clustered as Box-to-Box Midfielders, showcasing their versatility across both defensive and offensive phases of the game. In the forwards' category, Kylian Mbappé and Mohamed Salah stood out as Poachers, highlighting their focus on scoring and finishing.

This detailed analysis, focusing on elite players from top teams, underscores the effectiveness of the clustering approach in revealing the diverse roles within football positions. The results validate the approach taken, providing actionable insights that can be used by coaches, analysts, and recruiters to optimize player deployment and recruitment strategies.

6.3 Examples of Players' Role Evolution

The following example tables illustrate the evolution of roles for selected players from 2022 to 2024. These players were chosen for their interesting role transitions over the seasons, as revealed by the clustering analysis.

6.3.1 Achraf Hakimi

Player	Year	Cluster	Descriptive Cluster Label
Achraf Hakimi	2022	3	Attacking Full-Back
Achraf Hakimi	2023	3	Attacking Full-Back
Achraf Hakimi	2024	0	Playmaking Defender

Table 7: Role Evolution of Achraf Hakimi

Overview: Achraf Hakimi consistently played as an Attacking Full-Back in 2022 and 2023, reflecting his involvement in the offensive phase and his contributions in wide positions. However, in 2024, under Luis Enrique he transitioned to a Playmaking Defender, indicating a shift towards more involvement in building play from deeper positions, contributing more to his team's passing and playmaking.

6.3.2 Kylian Mbappé

Player	Year	Cluster	Descriptive Cluster Label
Kylian Mbappé	2022	3	Creative Forward
Kylian Mbappé	2023	3	Creative Forward
Kylian Mbappé	2024	2	Poacher

Table 8: Role Evolution of Kylian Mbappé

Overview: Kylian Mbappé started as a Creative Forward in 2022 and 2023, showcasing his ability to create chances and contribute to the build-up play. By 2024, his role evolved into that of a Poacher, focusing more on finishing and capitalizing on scoring opportunities, marking a shift towards a more goal-focused role.

6.3.3 Erling Haaland

Player	Year	Cluster	Descriptive Cluster Label
Erling Haaland	2022	2	Poacher
Erling Haaland	2023	2	Poacher
Erling Haaland	2024	0	All-Round Forward

Table 9: Role Evolution of Erling Haaland

Overview: Erling Haaland was classified as a Poacher in 2022 and 2023, focusing heavily on goal-scoring. In 2024, he evolved into an All-Round Forward, indicating a broader contribution to his team's play, involving more in passing, creativity, and possibly even defensive duties, reflecting a more versatile role.

6.3.4 Joško Gvardiol

Player	Year	Cluster	Descriptive Cluster Label
Joško Gvardiol	2022	0	Playmaking Defender
Joško Gvardiol	2023	0	Playmaking Defender
Joško Gvardiol	2024	3	Attacking Full-Back

Table 10: Role Evolution of Joško Gvardiol

Overview: Joško Gvardiol was consistently categorized as a Playmaking Defender in 2022 and 2023, highlighting his ability to contribute to the build-up play from the back. By 2024, his role shifted to an Attacking Full-Back, indicating a move towards more offensive contributions from wide areas, involving more overlaps, crosses and suprisinly goals.

6.3.5 İlkay Gündoğan

Player	Year	Cluster	Descriptive Cluster Label
İlkay Gündoğan	2022	3	Box-to-Box Midfielder
İlkay Gündoğan	2023	1	Attacking Midfielder
İlkay Gündoğan	2024	3	Box-to-Box Midfielder

Table 11: Role Evolution of İlkay Gündoğan

Overview: İlkay Gündoğan showed flexibility in his role over the years. He started as a Box-to-Box Midfielder in 2022, transitioned to an Attacking Midfielder in 2023, focusing more on offensive play, and returned to a Box-to-Box role in 2024, indicating a balanced contribution to both defense and attack.

6.3.6 Jude Bellingham

Player	Year	Cluster	Descriptive Cluster Label
Jude Bellingham	2022	0	Defensive Midfielder
Jude Bellingham	2023	3	Box-to-Box Midfielder
Jude Bellingham	2024	1	Attacking Midfielder

Table 12: Role Evolution of Jude Bellingham

Overview: Jude Bellingham's evolution is a testament to his growing influence on the pitch. Starting as a Defensive Midfielder in 2022, he transitioned to a Box-to-Box Midfielder in 2023, showcasing his versatility, and by 2024, he had become an Attacking Midfielder, focusing more on creating and finishing scoring opportunities.

7 Conclusions and Future Work

7.1 Conclusions

This project has demonstrated an effective method for classifying football players based on their performance metrics, not only providing a snapshot of their current roles but also tracking their evolution over multiple seasons. By clustering players into distinct roles, we offer valuable insights that can aid managers and recruiters in making informed decisions regarding player utilization, development, and transfer strategies.
The ability to track a player's role evolution across different seasons is particularly useful for identifying emerging talents, understanding player development, and making data-driven decisions that align with team strategies.

7.2 Future Work

While this project has achieved significant results, there are several avenues for further enhancement:

- **Expanding Data Scope:** To improve the robustness and generalizability of the clustering methodology, future work could include expanding the dataset to cover leagues outside the top 5 European leagues. This would provide a more comprehensive view of global football talent and allow for the identification of emerging players in less prominent leagues.
- **Incorporating Event Data:** Integrating event data, such as key passes, tackles, and dribbles, could refine the clustering process by adding a layer of context to the performance metrics. Event data would allow us to classify players not only based on their statistical output but also on their contributions during critical moments of the game.
- **Adding Positional Data:** Including positional data could further enhance the analysis by allowing us to classify players based on their specific positions on the pitch during different phases of the game. This would provide a deeper understanding of a player's role within various tactical setups and could lead to more nuanced classifications, such as distinguishing between a defensive midfielder who drops deep to support the defense and one who presses high up the pitch.

By addressing these areas in future work, the project can be expanded to offer even more detailed and actionable insights, further supporting the decision-making processes of football managers, scouts, and analysts.

References

[1] American Soccer Analysis. Defining Roles: How Every Player Contributes to Goals. https://www.americansocceranalysis.com/home/2020/8/3/defining-roles-how-every-player-contributes-to-goals.

[2] The New York Times - The Athletic. Player Roles. https://www.nytimes.com/athletic/3473297/2022/08/10/player-roles-the-athletic.

[3] Tony El Habr. Dimensionality Reduction and Clustering. https://tonyelhabr.netlify.app/posts/dimensionality-reduction-and-clustering/.

[4] FBRef. https://fbref.com.