Player Role Profiling & Style Clustering: A Case Study on Replacing Nicolò Barella

Federico Cattó Andrea Matteo Re

September 2025

Contents

Chapter 1

Introduction

1.1 Dataset

For this project, we collected player-level aggregated statistics via **web scraping** from FBref, a trusted source for advanced football metrics. Specifically, we scraped data for all players in the **Serie A, LaLiga, and Premier League** during the **2024/25 season**.

We then combined these data into a **single unified dataset** and filtered it to include only **midfielders**. This filtering allows for a more direct and coherent comparison with the profile of Nicolò Barella, focusing on players who could realistically match his role and style.

For a detailed description of the variables included in the dataset, please refer to the following document: **Variables**.

1.1.1 Methods of Data Acquisition

Data acquisition is a critical component of the *data science workflow*, serving as the **foundational step** upon which all subsequent analysis, modeling, and interpretation depend. It refers to the process of **identifying**, **collecting**, and **storing** data from various sources, ensuring it is **accurate**, **reliable**, and **suitably structured** for analysis.

In the context of **sports analytics**, robust data acquisition enables informed evaluations of *player performance*.

High-quality acquisition influences the **validity**, **depth**, and **reproducibility** of analysis, making it essential to choose the right tools and techniques.

For this project, the collection of player-level football statistics was achieved through **web scraping**, a programmatic method to extract structured data from websites. Web scraping enables the gathering of large-scale datasets

that are otherwise not easily accessible in downloadable formats, providing flexibility and precision in selecting the relevant information.

To implement this process, two widely used Python libraries were employed:

- **BeautifulSoup**: This library is ideal for parsing *static HTML* content. Its lightweight design allows for efficient extraction of structured elements from pages with fixed markup.
- **Selenium**: Dynamic web pages, often rendered using JavaScript, require browser simulation to access content. Selenium provides this functionality, enabling automated interactions with page elements to retrieve asynchronously loaded data.

1.1.2 Data Sources

The primary source of information was **FBref.com**, a comprehensive football statistics platform widely used for advanced metrics. All relevant data, including player performance statistics across multiple leagues, were systematically scraped to ensure completeness and consistency. This approach allowed the creation of a unified dataset that could support subsequent analysis focused on midfielders comparable to **Nicolò Barella**.

This combination of methods and tools ensured a robust, reproducible, and scalable data acquisition process, forming the foundation for the subsequent stages of feature engineering and clustering analysis.

1.1.3 Data Cleaning and Preparation

Following the acquisition of raw player-level statistics, a structured data cleaning and preparation pipeline was applied to ensure the reliability and interpretability of the dataset. This process was essential to transform heterogeneous scraped data into a consistent analytical resource, suitable for clustering and similarity analysis.

First, all **non-numerical columns** were excluded, as the focus of the study lies in quantitative performance metrics. In addition, variables that were not meaningful for midfielders — such as statistics specific to goalkeepers — were removed to maintain role-specific coherence.

To address missing values, all variables containing **NaN entries** were imputed with zero. This assumption is consistent with football statistics, where the absence of an event (e.g., a player recording zero tackles) is better represented by zero rather than by an undefined value.

Next, only players with a **significant amount of playing time** were retained, ensuring that the dataset reflects stable and representative performances rather than noisy observations from limited minutes.

A further refinement step involved analyzing the **pairwise correlation among variables** (see Figure 1.1). Whenever two or more variables exhibited a correlation coefficient greater than 0.9, only one was retained. This reduced redundancy, mitigated multicollinearity, and enhanced the interpretability of clustering result.

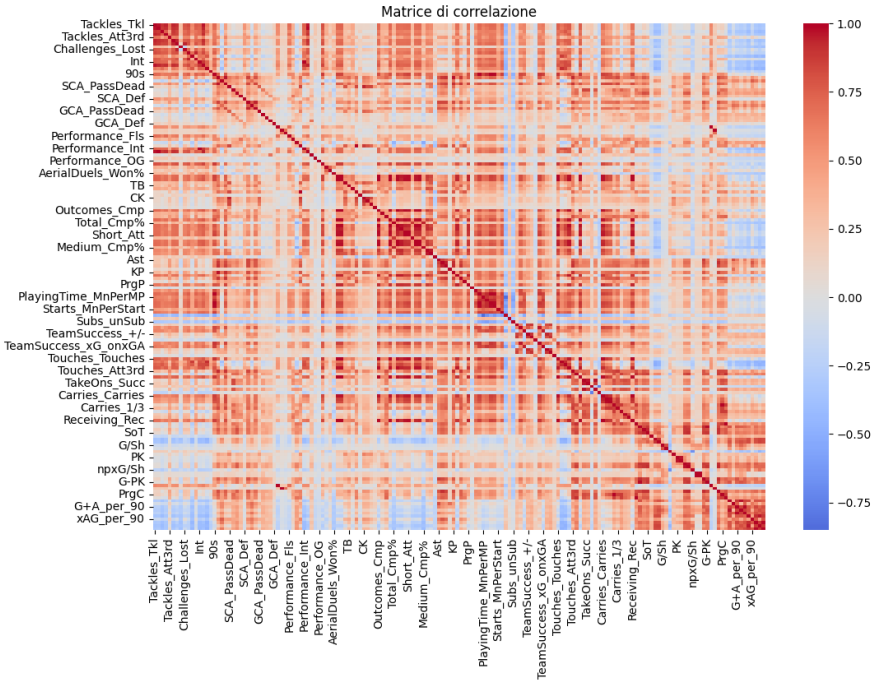

Figure 1.1: Heatmap of pairwise correlations among variables before dimensionality reduction. Highly correlated features (correlation > 0.9) were pruned to mitigate redundancy and multicollinearity.

Finally, all numerical features were **normalized** to a common scale. This step was critical for unsupervised methods, as it prevents variables with larger numerical ranges from disproportionately influencing the clustering algorithm.

Through this systematic cleaning and preparation pipeline, the dataset was transformed into a balanced, role-specific, and methodologically robust foundation for the subsequent stages of feature engineering, clustering, and similarity analysis.

1.1.4 Description of Final Dataset

After the data acquisition and cleaning pipeline, the final dataset was structured to provide a robust and interpretable foundation for subsequent analysis. The resulting dataset consists of **82 numerical variables** and covers a total of **496 midfielders** drawn from Serie A, LaLiga, and the Premier League in the 2024/25 season.

The variables capture a comprehensive spectrum of player performance, ranging from passing and possession metrics to attacking contributions and defensive actions. Each feature was selected and preprocessed to ensure that it meaningfully contributes to describing the playing style of a midfielder, while redundant or role-irrelevant variables were systematically excluded during the cleaning phase.

The choice to restrict the dataset to midfielders guarantees a **role-specific focus**, which is essential for generating valid comparisons with Nicolò Barella. Moreover, the balance between breadth (82 features) and sample size (600 players) provides sufficient statistical richness to enable effective clustering and similarity analysis without introducing excessive dimensionality.

This carefully curated dataset therefore represents not only a comprehensive statistical snapshot of top-level midfielders across three major European leagues, but also a consistent and methodologically sound basis for the application of data-driven profiling and clustering techniques.

1.2 Idea and Motivations

The transfer of Nicolò Barella, one of Inter's key midfielders, represents a significant tactical and strategic challenge for the club. Barella's possible departure creates the need to identify potential replacements who can fulfill both his statistical profile and his role within the team's playing philosophy. This scenario inspired the development of a **data-driven player profiling and clustering framework**, aimed at supporting recruitment decisions with objective insights.

The motivation behind this project is threefold:

- **Enhancing scouting efficiency:** Traditional scouting often relies on subjective assessments or limited observational data. By leveraging quantitative analysis of player statistics, the process becomes faster, more systematic, and less prone to human bias.
- **Maintaining tactical identity:** Inter has a specific style of play, and replacing a key midfielder requires understanding not only raw

performance metrics but also stylistic similarity. Clustering players based on comprehensive statistical features helps identify those who fit the tactical profile.

- **Supporting evidence-based recruitment:** Transfer market decisions carry financial and sporting risk. A framework that ranks and compares potential targets based on multi-dimensional similarity metrics provides a structured, reproducible method to reduce uncertainty.

In essence, this project seeks to bridge the gap between advanced statistical analysis and practical decision-making in football scouting. By combining data preprocessing, feature engineering, and clustering algorithms, we aim to provide actionable insights that help Inter club identify midfielders comparable to Barella, while also uncovering hidden talents who align with the team's strategic needs.

1.3 Aim of the project

The primary aim of this project is to develop a **player-role profiling and style-clustering framework** capable of identifying footballers who are statistically and stylistically similar to Nicolò Barella. By leveraging advanced data processing and unsupervised learning techniques, the framework is designed to support clubs in making informed, evidence-based decisions in the transfer market.

Specifically, the project seeks to achieve the following objectives:

- **Data preprocessing and feature engineering:** Transform raw player statistics into standardized, comparable metrics, and create meaningful features that capture key aspects of midfield performance.
- **Clustering and similarity analysis:** Apply clustering algorithms to group players with similar playing styles, and compute similarity scores to rank potential targets relative to the reference player.
- **Interactive visualization:** Provide an intuitive dashboard where users can select a reference player, explore cluster memberships, compare statistical profiles, and access similarity metrics, facilitating practical interpretation of the results.

By accomplishing these objectives, the framework not only aids in identifying potential replacements for Barella but also serves as a generalizable tool for scouting and tactical analysis. This approach ensures that recruitment decisions are grounded in objective, multi-dimensional evidence while also uncovering emerging talents that align with the club's strategic philosophy.

Chapter 2

Methodologies

In this chapter, we present the methodological framework adopted to perform player profiling and clustering. Each method is introduced by outlining its core principles, theoretical underpinnings, and mathematical formulations, followed by a critical discussion of advantages and limitations in the context of football analytics.

2.1 Dimensionality Reduction: Principal Component Analysis (PCA)

Principles

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that projects high-dimensional data onto a lower-dimensional space while retaining as much variance as possible. It is particularly effective when dealing with high dimensionality, as in our dataset.

Theory and Formulas

Given a standardized data matrix $X \in \mathbb{R}^{n \times p}$, PCA computes the eigenvalue decomposition of the covariance matrix $\Sigma = \frac{1}{n-1} X^\top X$. The principal components are obtained as:

$$Z = XW,$$

where W contains the eigenvectors associated with the largest eigenvalues of Σ.

Application in This Study

PCA was applied to the cleaned dataset in order to reduce dimensionality and mitigate multicollinearity. From the initial set of 82 variables, we retained 36 principal components, which captured the majority of the variance and served as the input features for the clustering stage.

Pros and Cons

- **Pros**: Reduces dimensionality, mitigates multicollinearity, and facilitates visualization of high-dimensional data.
- **Cons**: Assumes linear relationships, components may lack intuitive interpretability for non-technical audiences.

2.2 Clustering Framework

Principles

Clustering is an unsupervised learning approach aimed at grouping players with similar statistical profiles. By operating in the reduced PCA feature space, clustering enables us to identify meaningful role-based groupings among midfielders.

Evaluation of Optimal Cluster Number

To ensure robustness, the choice of the optimal number of clusters was guided by multiple internal validation metrics:

- **Calinski-Harabasz Index (CH)** – favors partitions with dense and well-separated clusters.
- **Davies-Bouldin Index (DBI)** – lower values indicate better clustering compactness and separation.
- **Silhouette Score** – measures cohesion and separation; higher values indicate better-defined clusters.

2.2.1 K-Means Clustering

Theory and Formulas

K-Means clustering partitions observations into k groups by minimizing within-cluster variance. The optimization problem is:

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

where μ_i is the centroid of cluster C_i.

Pros and Cons

- **Pros**: Simple, computationally efficient, scalable to large datasets.
- **Cons**: Requires pre-specification of k, sensitive to initialization, assumes spherical clusters of similar size.

2.2.2 Hierarchical Clustering

Theory and Formulas

Hierarchical clustering is a method that builds a hierarchy of clusters without requiring the number of clusters to be specified in advance. The procedure can follow two main approaches:

- **Agglomerative (bottom-up)**: Each data point starts as its own cluster, and pairs of clusters are successively merged based on a chosen linkage criterion until all points are in a single cluster.
- **Divisive (top-down)**: All data points start in a single cluster, which is recursively split into smaller clusters.

The linkage criterion defines how distances between clusters are computed. For example, in the case of Ward's method:

$$\Delta(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|\mu_i - \mu_j\|^2,$$

where C_i and C_j are clusters with n_i and n_j points, and μ_i, μ_j their respective centroids. This criterion minimizes the variance within clusters after merging.

Visualization

The results of hierarchical clustering can be effectively visualized using a **dendrogram**, which illustrates the successive merging (or splitting) of clusters at different distance thresholds. This graphical representation allows researchers to assess the natural number of clusters by identifying significant jumps in linkage distance.

Pros and Cons

- **Pros**: Does not require pre-specifying the number of clusters, produces a dendrogram for visual inspection, flexible with different linkage criteria and distance metrics.
- **Cons**: Computationally expensive for large datasets, sensitive to noise and outliers, once a merge or split is performed it cannot be undone.

2.3 Player Similarity: k-Nearest Neighbors (KNN)

Principles

To identify players with the most comparable performance profiles, we employed the **k-Nearest Neighbors (KNN)** algorithm. KNN is a non-parametric method that, given a query point, retrieves the k closest points in the feature space according to a specified distance metric.

Theory and Formulas

Given a player represented by a feature vector x, the KNN approach searches for the k players $\{x_1, \ldots, x_k\}$ that minimize the chosen distance metric $d(x, x_i)$ among all other players. The general distance function can take several forms, among which we adopted:

- **Euclidean distance:**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Manhattan distance:**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Cosine distance:**
$$d(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$
- **Mahalanobis:** $d(x, y) = \sqrt{(x - y)^T S^{-1}(x - y)}$

Application in This Study

KNN was applied in the reduced 36-dimensional PCA space to ensure computational tractability and robustness against noise. For each player, the algorithm identified the most similar peers by evaluating their proximity under multiple distance metrics (Euclidean, Manhattan, and Cosine). In particular, the methodology was used to retrieve the **Top 3 most similar players to Nicolò Barella**, providing a case study of how this approach can be leveraged for practical player comparisons.

Pros and Cons

- **Pros**: Simple to implement, interpretable, flexible to different similarity measures, and directly applicable to player profiling.
- **Cons**: Sensitive to the choice of distance metric, computationally expensive for very large datasets, and affected by the curse of dimensionality (although mitigated here by PCA).

2.3.1 Workflow Overview

To provide a clear overview of the methodological pipeline, we summarize the key steps of the project in Figure 2.1. The workflow highlights the sequential process from raw data to player similarity analysis.

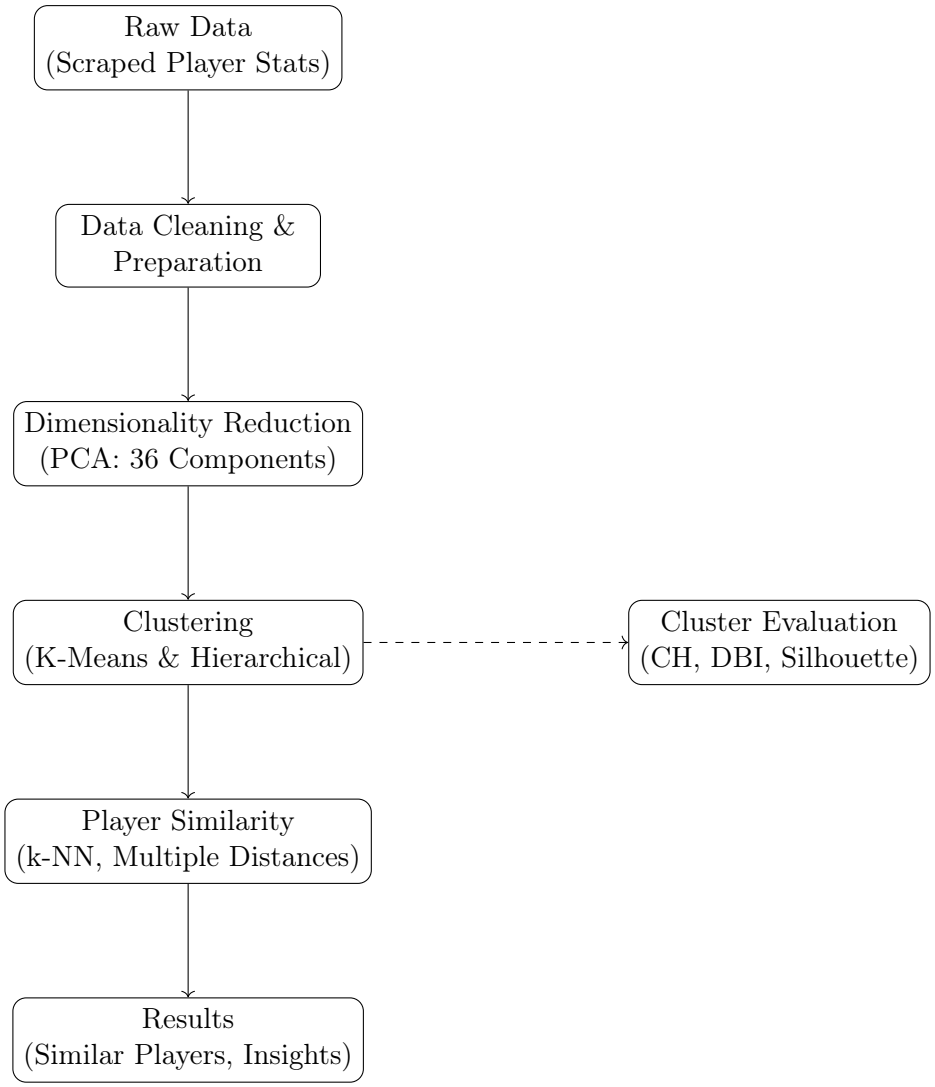

Figure 2.1: Workflow of the methodological pipeline: from data acquisition to final similarity analysis.

Chapter 3

Experiments

This chapter presents the experiments conducted to evaluate and validate the methodological framework described in the previous chapter. Each experiment is explained clearly, detailing the parameter choices and providing sufficient information for reproducibility.

3.1 Dimensionality Reduction via PCA

Objective

To reduce the original 82-dimensional feature space while retaining as much variance as possible, facilitating subsequent clustering and similarity analysis.

Procedure

- Standardized all numerical features to zero mean and unit variance.
- Applied Principal Component Analysis (PCA) on the cleaned dataset.
- Retained the first **36 principal components**, capturing the majority of variance.

Parameter Choice and Justification

The number of components was chosen based on the cumulative explained variance curve (see Figure 3.1), balancing dimensionality reduction with information retention. This ensures that the clustering algorithms operate on a compact but informative representation of player statistics.

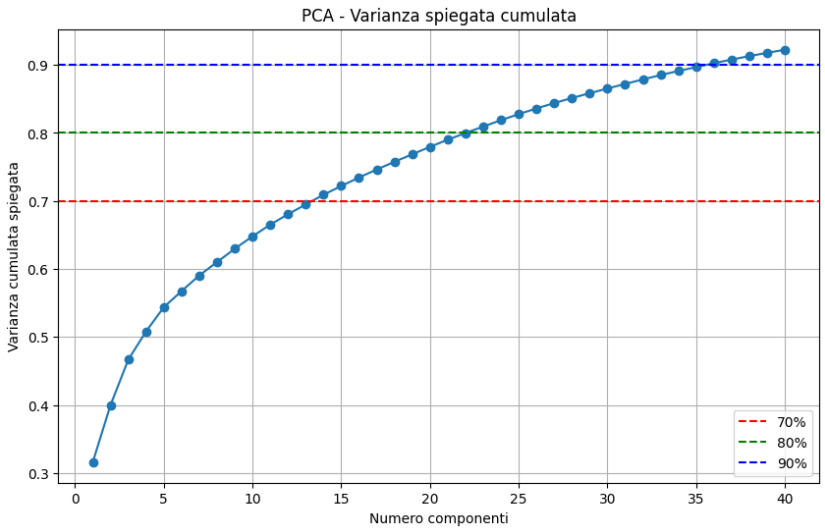

Figure 3.1: Cumulative explained variance of PCA components. The curve shows how much variance is retained as the number of components increases. The dashed horizontal lines correspond to the 70%, 80%, and 90% thresholds, which serve as practical benchmarks for dimensionality reduction. The selected cutoff (around 36 components) ensures that more than 90% of the total variance is preserved while avoiding excessive dimensionality.

Reproducibility Details

- `scikit-learn` `PCA` implementation in Python.
- Random seed set for reproducibility: `random_state=42`.

3.2 Clustering Experiments

3.2.1 Clustering Approaches

K-Means Clustering

The first step in exploring player profiles was to apply **K-Means clustering**. The goal was to identify natural groupings among midfielders based on their performance metrics, providing insights into player roles and similarities.

Optimal Number of Clusters To determine the best number of clusters, we used a combination of quantitative methods and internal validation met-

rics, including **Silhouette Score**, **Davies–Bouldin Index**, and **Calinski–Harabasz Index**. The analyses showed that:

- Silhouette and Calinski–Harabasz indicated $k = 2$ as optimal.
- Davies–Bouldin suggested $k = 3$.

Given this, clustering analysis was conducted for both $k = 2$ and $k = 3$, comparing the resulting cluster structures and their interpretability (see Figure 3.2).

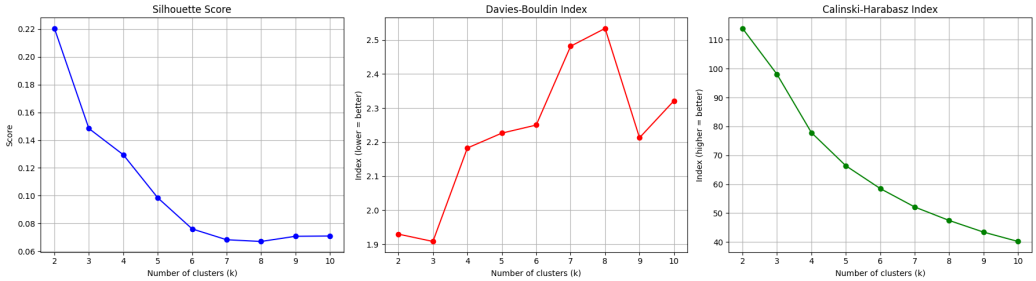

Figure 3.2: K-Means clustering evaluation for different values of k using Silhouette, Davies–Bouldin, and Calinski–Harabasz indices.

Cluster Profiling For each clustering configuration, we:

- Generated heatmaps of average profiles per cluster, providing a clear overview of feature distributions.
- Identified the 10 variables that most discriminate the clusters.
- Tabulated the values of these variables for each cluster.
- Created a PCA-reduced scatter plot of midfielders with **Barella highlighted** to visually examine cluster separation in two dimensions.

This workflow combined quantitative evaluation with intuitive visual inspection, ensuring clusters are both statistically sound and interpretable.

3.2.2 Hierarchical Clustering

To provide a complementary perspective on player groupings, **Hierarchical Clustering** was applied using the **Ward linkage method**, which minimizes total within-cluster variance and produces compact, well-separated clusters.

Dendrogram Analysis The dendrogram was initially examined to get a descriptive sense of potential cluster divisions. This suggested a possible 3-cluster solution (see Figure 3.3).

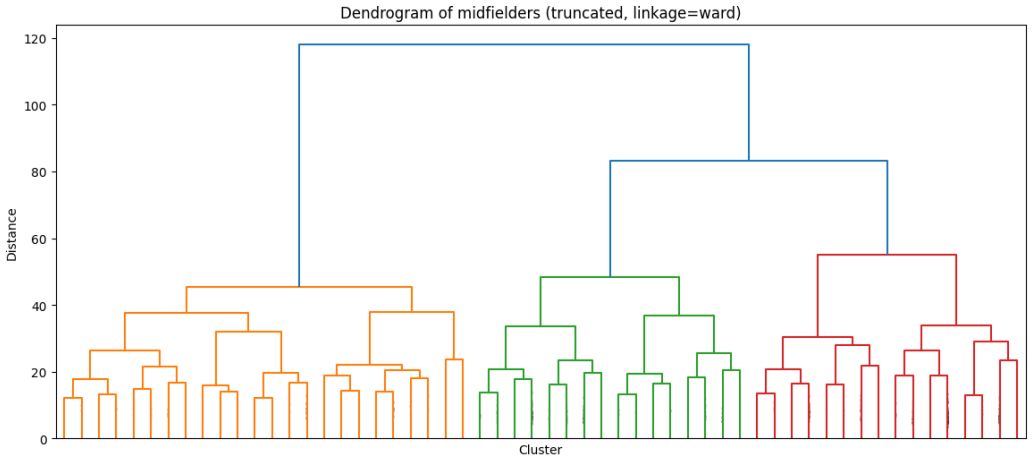

Figure 3.3: K-Means clustering evaluation for different values of k using Silhouette, Davies–Bouldin, and Calinski–Harabasz indices.

Quantitative Validation To refine the cluster selection, the hierarchical clusters were evaluated using Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. All three metrics consistently indicated $k = 2$ as the optimal solution (see Figure 3.4). Based on this quantitative evidence, the analysis focused on the 2-cluster solution, prioritizing metric-based rigor over descriptive heuristics from the dendrogram.

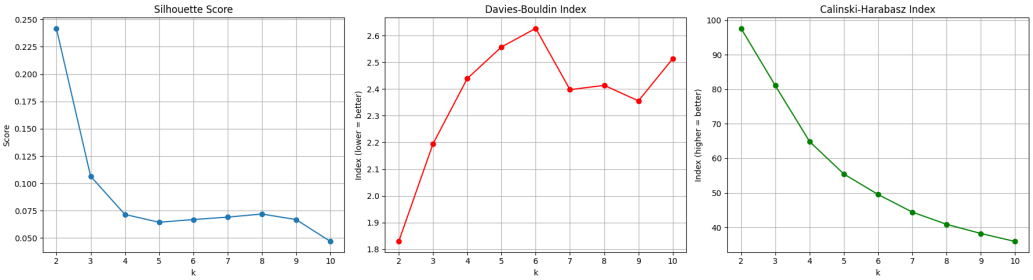

Figure 3.4: K-Means clustering evaluation for different values of k using Silhouette, Davies–Bouldin, and Calinski–Harabasz indices.

Cluster Profiling With the 2-cluster solution, the same profiling workflow as for K-Means was applied:

- Heatmaps of average profiles per cluster.
- Identification of the 10 most discriminative variables.
- Tabulated values for these variables across the two clusters.
- PCA scatter plot highlighting Barella, providing a visual perspective on the cluster structure.

This approach ensured consistency and comparability with the K-Means results, allowing a robust assessment of cluster patterns across different algorithms.

3.2.3 Player Similarity Analysis (KNN)

A K-Nearest Neighbors (KNN) analysis was conducted to explore player-level similarity, focusing on **Barella** as the reference player. Different distance metrics were used to evaluate the influence on similarity rankings:

- Euclidean Distance
- Manhattan Distance
- Cosine Distance
- Mahalanobis Distance

This allowed for a nuanced understanding of player relationships and how metric choice affects the identification of Barella's closest peers.

3.2.4 Methodological Notes

All analyses were performed using a consistent preprocessing pipeline, including scaling and normalization of performance features. Standard Python libraries such as `scikit-learn`, `pandas`, `matplotlib`, and `seaborn` were employed, ensuring reproducibility and methodological rigor.

Chapter 4

Results

This chapter presents the empirical findings of the study, structured around cluster characterization, player positioning within clusters, and similarity analysis. The results are supported by both tabular evidence and visualizations to facilitate interpretability.

To identify the features most responsible for separating the clusters, variable importance scores were computed based on variance contribution across groups.

4.1 K-Means Clustering ($k = 2$)

The K-Means solution with $k = 2$ reveals a clear dichotomy in midfielder performance profiles. As shown in Figure 4.1, Cluster 0 groups players with consistently below-average contributions across offensive and defensive dimensions, including shots, shot-creating actions, progressive carries, and defensive interventions. This pattern suggests a more conservative or limited role in both attacking buildup and defensive transitions. Conversely, Cluster 1 aggregates players with above-average performance across nearly all metrics, indicative of dynamic midfielders actively engaged in ball progression, chance creation, and defensive support. Among the variables, shot-creating actions and carries display the most pronounced differences, underscoring their importance in discriminating between high- and low-activity profiles.

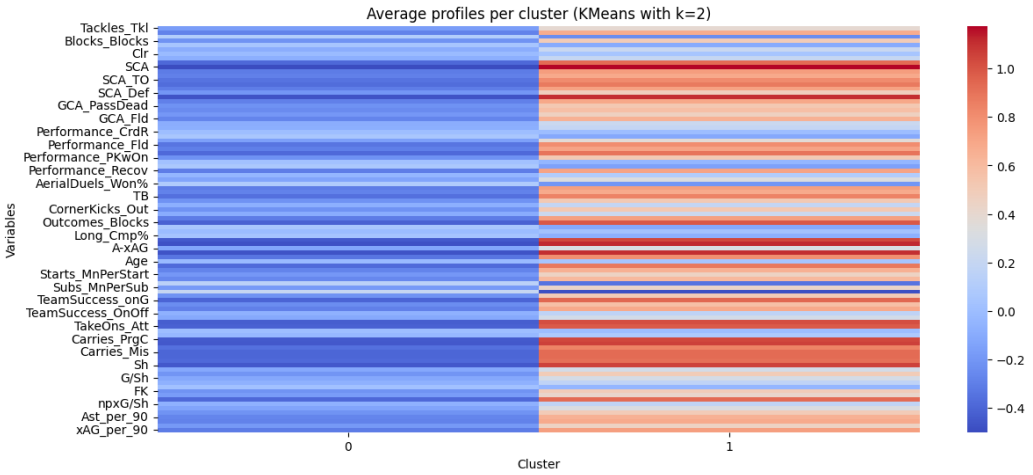

Figure 4.1: Heatmap of average normalized profiles for K-Means clusters ($k = 2$). Blue shades indicate below-average values, red shades above-average values. Cluster 0 corresponds to midfielders with limited involvement across most metrics, while Cluster 1 groups high-activity midfielders with strong offensive, defensive, and transitional contributions.

Table 4.1 lists the ten variables with the highest importance scores in separating the two clusters. Offensive and creative indicators dominate, with Shot-Creating Actions, Expected Assisted Goals, Passes into the Penalty Area, and Goal-Creating Actions emerging as the most influential. Additionally measures such as progressive carries and assists further reinforce the offensive orientation of the clustering. Defensive variables, by contrast, contribute only marginally, indicating that the two-cluster structure is primarily driven by differences in attacking involvement.

Table 4.1: Top 10 variables contributing to K-Means cluster differentiation ($k = 2$). Importance scores quantify the relative contribution of each feature to the separation between clusters.

Variable	**Importance Score**
SCA	1.675471
xAG	1.592847
PPA	1.581742
GCA	1.578727
Carries_1/3	1.515068
Sh	1.506547
Carries_PrgC	1.480255
Ast	1.480020
Touches_AttPen	1.455790
Outcomes_Blocks	1.410329

The normalized cluster means for these key variables are reported in Table 4.2. Cluster 1 records consistently positive scores, reflecting strong offensive playmaking, chance creation, and progressive involvement. In contrast, Cluster 0 shows systematically negative values, associated with below-average performance in final-third penetration and creative contributions. This confirms that the $k = 2$ solution effectively distinguishes midfielders along an offensive–defensive spectrum.

Table 4.2: Cluster means (normalized) for K-Means ($k = 2$) on the top 10 discriminative variables. Positive values indicate above-average performance, negative values below-average performance.

Variable	**Cluster 0**	**Cluster 1**
SCA	-0.499939	1.175532
xAG	-0.475285	1.117562
PPA	-0.471971	1.109770
GCA	-0.471072	1.107655
Carries_1/3	-0.452077	1.062991
Sh	-0.449534	1.057013
Carries_PrgC	-0.441689	1.038566
Ast	-0.441619	1.038401
Touches_AttPen	-0.434389	1.021401
Outcomes_Blocks	-0.420824	0.989505

The clustering structure is further illustrated in the two-dimensional PCA projection (Figure 4.2), which preserves the majority of variance in the original feature space. The two clusters appear reasonably distinct, though some overlap occurs in the central region, representing hybrid profiles. Nicolò Barella, highlighted in black, lies within Cluster 1 in the upper-right quadrant, reflecting his high involvement in chance creation and offensive progression. This position aligns with the interpretation of Cluster 1 as the group of dynamic, attack-oriented midfielders, whereas Cluster 0 predominantly includes more defensive or support-oriented players.

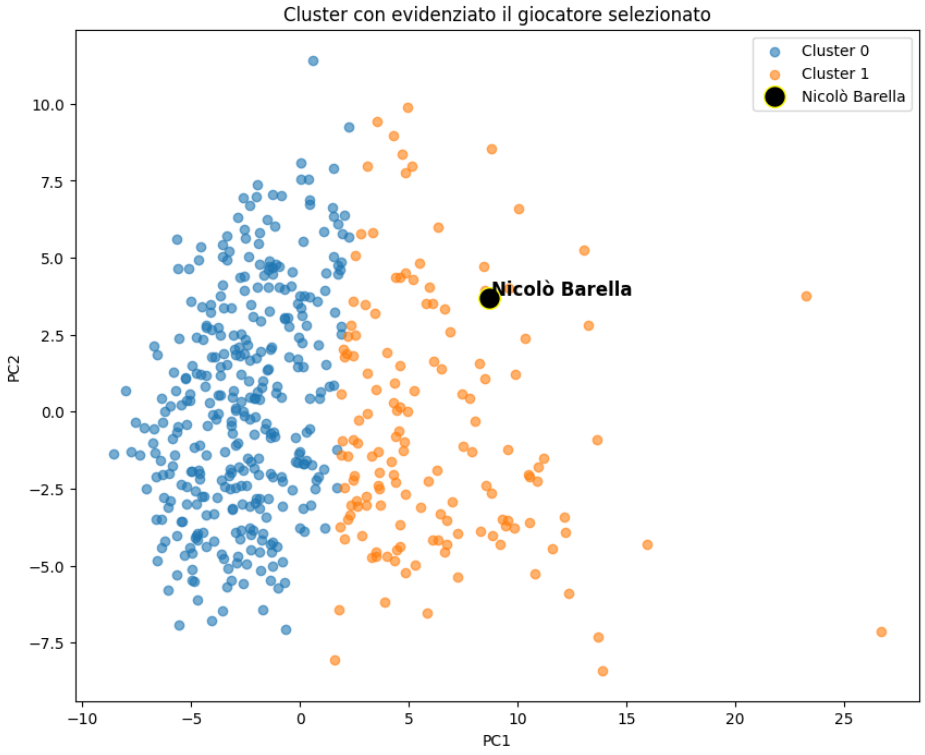

Figure 4.2: PCA projection of midfielders colored by K-Means cluster assignment ($k = 2$). Nicolò Barella is highlighted in black.

4.2 K-Means Clustering ($k = 3$)

To further explore the clustering structure, a three-cluster solution was evaluated. The heatmap in Figure 4.3 highlights three distinct midfielder profiles, capturing stylistic differences in both offensive and defensive contributions.

Cluster 0 is characterized by strong positive values across playmaking and progression metrics such as Shot-Creating Actions, progressive carries, and Expected Assisted Goals. Players in this group emerge as creative, attack-oriented midfielders, heavily involved in chance creation and offensive transitions.

Cluster 1 displays consistently negative standardized scores across most variables, indicative of midfielders with limited involvement in both offensive and defensive actions. This cluster likely corresponds to peripheral or supporting roles with comparatively lower impact on overall team performance.

Cluster 2 presents a more defensive profile, with positive contributions in tackles and blocks but lower values in creative and final-third variables. These players appear to specialize in ball recovery and defensive stability, complementing more attacking teammates.

Taken together, the three-cluster solution differentiates midfielders along tactical dimensions: creative playmakers (Cluster 0), defensive stoppers (Cluster 2), and less active/supporting profiles (Cluster 1).

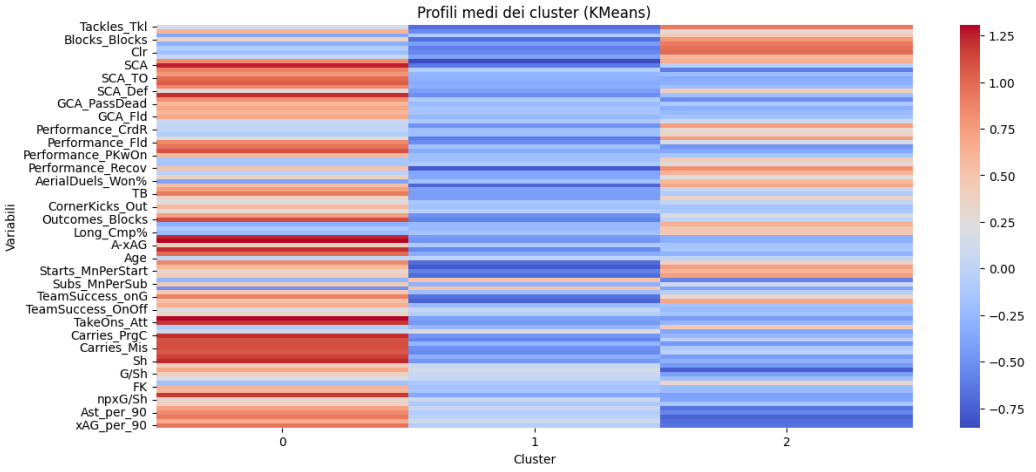

Figure 4.3: Average standardized profiles of midfielders identified by K-Means with $k = 3$. Red shades indicate above-average values, blue shades below-average values. Cluster 0 corresponds to creative midfielders, Cluster 1 to low-activity players, and Cluster 2 to defensively oriented midfielders.

Table 4.3 reports the top ten variables driving cluster differentiation. Creative and offensive metrics dominate, with Shot-Creating Actions, Expected Assisted Goals, Passes into the Penalty Area, and Goal-Creating Actions emerging as the strongest discriminators. Additional variables such as Touches in the attacking penalty area, progressive carries, and Assists further underline the offensive dimension of the clustering. Interestingly, Minutes played also appears among the top drivers, suggesting that consistent playing time contributes to distinguishing regular starters from more marginal players.

Table 4.3: Top 10 variables contributing to K-Means cluster differentiation ($k = 3$). Importance scores reflect the relative contribution of each feature to the clustering solution.

Variable	**Importance Score**
SCA	1.878182
xAG	1.789342
PPA	1.747677
GCA	1.718970
90s	1.701515
Touches_AttPen	1.696840
Carries_PrgC	1.688104
Outcomes_Blocks	1.686635
Sh	1.685878
Ast	1.674457

The normalized cluster means (Table 4.4) further clarify these patterns. Cluster 0 consistently records high positive values across offensive metrics, including chance creation, progressive carries, shots, and assists, identifying it as the most attack-oriented group. Cluster 1 shows negative scores across all dimensions, representing midfielders with limited involvement in creative or defensive actions. Cluster 2 presents moderate values, with positive scores for minutes played, indicating regular starters with balanced but less extreme profiles compared to the other groups.

Table 4.4: Cluster means (normalized) for K-Means ($k = 3$) on the top 10 discriminative variables. Positive values indicate above-average performance, negative values below-average performance.

Variable	**Cluster 0**	**Cluster 1**	**Cluster 2**
SCA	1.255359	-0.622823	-0.045684
xAG	1.308043	-0.481299	-0.307024
PPA	1.208304	-0.539373	-0.136972
GCA	1.210619	-0.508351	-0.186834
90s	0.847851	-0.853663	0.639152
Touches_AttPen	1.300956	-0.395884	-0.433488
Carries_PrgC	1.218808	-0.469295	-0.253849
Outcomes_Blocks	1.115155	-0.571480	-0.012395
Sh	1.236563	-0.449315	-0.299040
Ast	1.238806	-0.435650	-0.321987

The PCA projection (Figure 4.4) confirms the separation among clusters in a reduced two-dimensional space. The clusters are well distinguished, with Cluster 0 occupying the region associated with strong creative involvement, Cluster 1 positioned in the low-activity area, and Cluster 2 distributed in a more balanced region.

Nicolò Barella, highlighted in black within Cluster 0 (green), is located among the creative group, reflecting his high involvement in both chance creation and progression. His placement underscores his profile as a versatile midfielder, combining playmaking qualities with defensive contributions, and aligning him with the most dynamic and offensively impactful players.

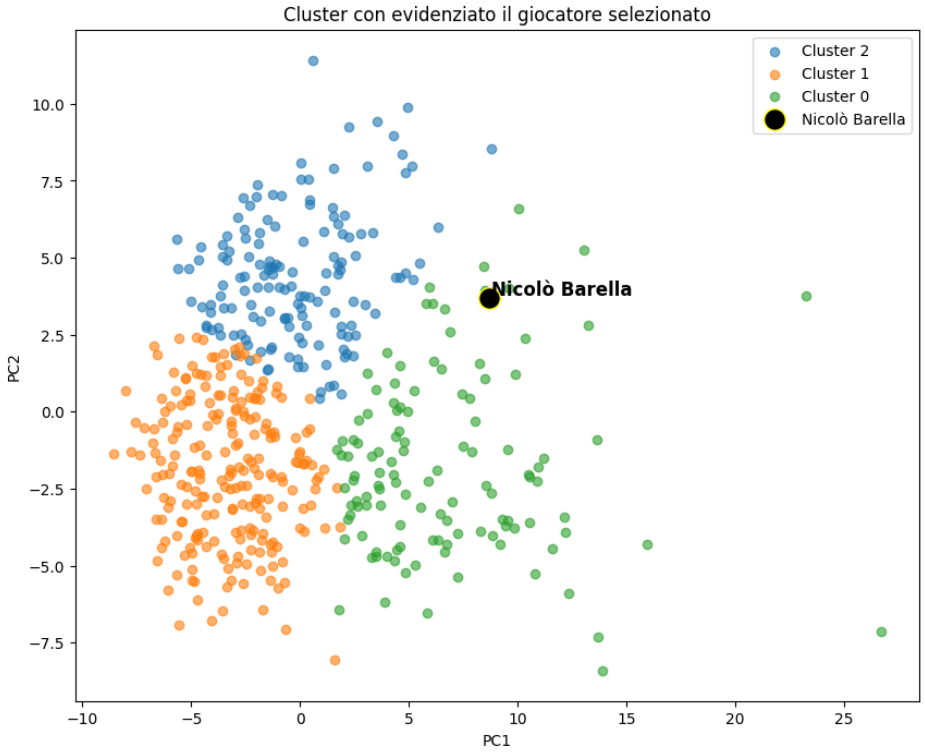

Figure 4.4: PCA projection of midfielders colored by K-Means cluster assignment ($k = 3$). Nicolò Barella is highlighted in black.

4.3 Hierarchical Clustering

The hierarchical clustering analysis confirmed the presence of two main groups of midfielders, consistent with the results obtained through K-Means. The heatmap in Figure 4.5 illustrates the average standardized profiles of the two clusters. Cluster 0 (left) is associated with lower values across nearly all performance variables, indicating players with more conservative or less impactful statistical profiles. Cluster 1 (right), by contrast, shows substantially higher scores in Expected Assisted Goals, Shot-Creating Actions, Assists, Progressive Carries, and Touches in the attacking penalty area. This cluster represents creative, attack-oriented midfielders with an active role in chance creation and offensive transitions.

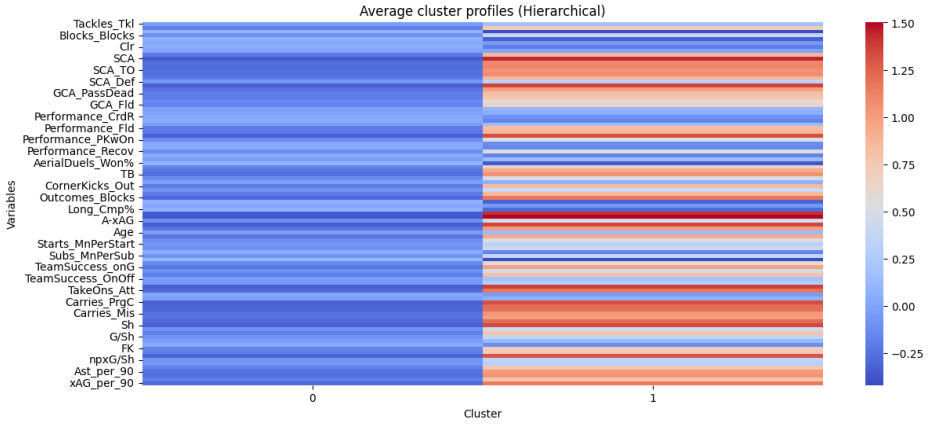

Figure 4.5: Average normalized profiles of midfielders identified by hierarchical clustering. Cluster 0 corresponds to players with limited offensive involvement, while Cluster 1 groups creative and attack-oriented midfielders.

Table 4.5 lists the ten variables that most strongly contributed to the separation of clusters. Creativity and offensive involvement emerge as the dominant dimensions, with Expected Assisted Goals, Shot-Creating Actions, and Assists ranking highest. Additional variables such as Touches in the penalty area, Goal-Creating Actions, and Passes into the penalty area further emphasize the attacking nature of the differentiation. Measures of offensive output (Shots, Progressive Carries) and complementary metrics (Crossing performance, Expected Goals) add nuance, highlighting how the model captures both playmaking and finishing contributions.

Table 4.5: Top 10 variables contributing to hierarchical clustering differentiation. Importance scores reflect each variable's relative discriminative power.

Variable	**Importance Score**
xAG	1.868350
SCA	1.790078
Ast	1.769550
Touches_AttPen	1.700685
GCA	1.699249
PPA	1.690340
Sh	1.663390
Carries_PrgC	1.652577
Performance_Crs	1.651029
xG	1.625131

The standardized cluster means reported in Table 4.6 confirm this interpretation. Cluster 1 exhibits consistently positive values across all discriminative variables, identifying midfielders with strong contributions in chance creation, assists, and offensive productivity. Cluster 0, on the other hand, shows systematically negative scores, corresponding to players with lower involvement in creative or attacking actions, likely fulfilling more supportive or defensively oriented roles.

Table 4.6: Cluster means (normalized) for hierarchical clustering on the top 10 discriminative variables. Positive values indicate above-average performance; negative values reflect below-average performance relative to the dataset mean.

Variable	**Cluster 0**	**Cluster 1**
xAG	-0.365383	1.502967
SCA	-0.350076	1.440002
Ast	-0.346061	1.423489
Touches_AttPen	-0.332594	1.368091
GCA	-0.332313	1.366936
PPA	-0.330571	1.359770
Sh	-0.325300	1.338090
Carries_PrgC	-0.323186	1.329392
Performance_Crs	-0.322883	1.328147
xG	-0.317818	1.307313

The PCA projection in Figure 4.6 further illustrates the two-cluster separation in a reduced two-dimensional space. The clusters are clearly distinguished, with Cluster 1 concentrated in the high-activity region of the plot. Nicolò Barella, highlighted in black, is positioned within Cluster 1 (orange), confirming his profile as a creative midfielder actively involved in offensive play. This placement aligns with his strong contributions to chance creation and attacking progression, contrasting with the more defensively oriented players grouped in Cluster 0 (blue).

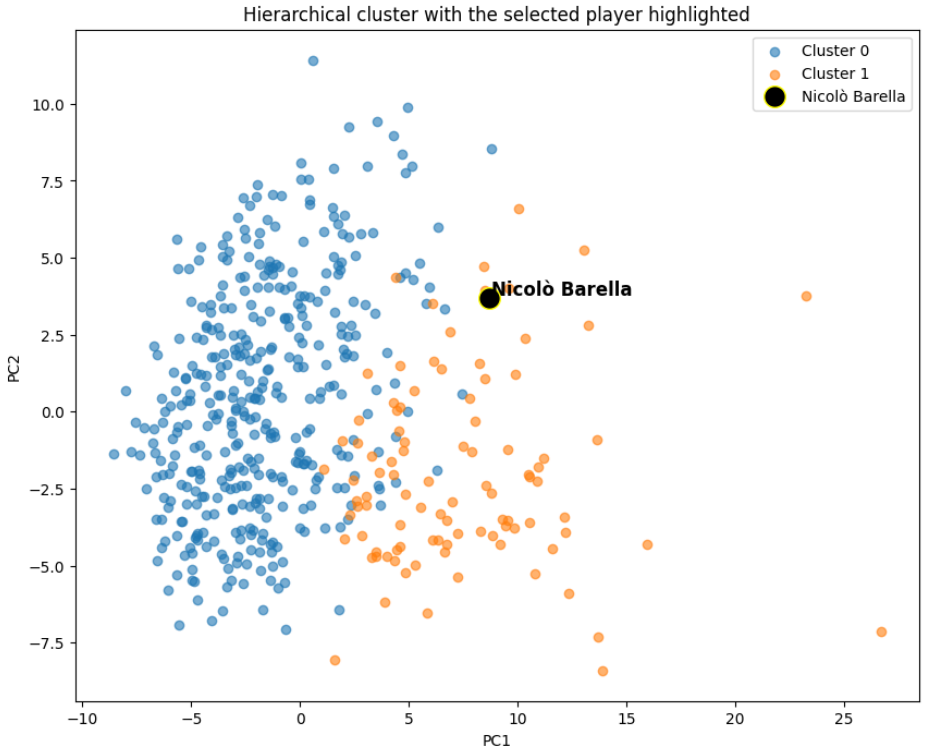

Figure 4.6: PCA projection of midfielders colored by hierarchical clustering assignment. Nicolò Barella is highlighted in black within Cluster 1.

4.4 Player Similarity Analysis (KNN)

To complement the clustering results, a k-Nearest Neighbors (KNN) analysis was conducted to identify the players most similar to Nicolò Barella in terms of their statistical profile. Four distance metrics were tested—Euclidean, Manhattan, Cosine, and Mahalanobis—to account for different notions of similarity and to ensure robustness of the results. The three closest players identified under each metric are summarized in Table 4.7, while an interactive radar chart provides a visual comparison of their performance profiles.

Table 4.7: Top three players most similar to Nicolò Barella according to different distance metrics. Lower scores indicate higher similarity.

Player	**Euclidean**	**Manhattan**	**Cosine**	**Mahalanobis**
Rodrigo De Paul	6.040	29.548	0.133	5.407
Bruno Guimarães	8.328	38.835	0.240	-
Enzo Fernández	9.092	-	-	-
Pedri	-	-	0.220	-
Martin Ødegaard	-	41.783	-	-
Luka Sučić	-	-	-	6.823
Javier Guerra	-	-	-	6.843

Across all metrics, Rodrigo De Paul consistently emerges as the closest match, reflecting strong parallels with Barella in creativity, ball progression, and overall midfield dynamism. Bruno Guimarães also appears frequently among the nearest neighbors, reinforcing his comparable profile as a versatile box-to-box midfielder.

Other players surface depending on the distance metric applied:

- **Euclidean distance** highlights Enzo Fernández, whose profile resonates in terms of passing volume and advanced positioning.
- **Cosine distance** identifies Pedri, capturing similarities in creative involvement and distribution.
- **Manhattan distance** emphasizes Martin Ødegaard, reflecting overlap in chance creation and attacking link-up play.
- **Mahalanobis distance**, which accounts for variable correlations, introduces Luka Sučić and Javier Guerra, suggesting statistically coherent but less conventional parallels.

Overall, the KNN analysis confirms that Barella's statistical fingerprint aligns most strongly with versatile and creative midfielders who combine progression, chance creation, and defensive work rate. Among all comparisons, Rodrigo De Paul consistently emerges as the closest match across multiple distance metrics, making him the most statistically similar player to Barella in the dataset. His profile mirrors Barella's blend of creativity, dynamism, and two-way contribution, positioning him as the most credible candidate to replicate Barella's role within Inter's tactical setup. This strong alignment highlights De Paul not only as a stylistic analogue but also as a realistic replacement option, should Barella's departure materialize.

Chapter 5

Conclusions

This project set out to address a concrete and strategically relevant problem: identifying potential replacements for Nicolò Barella through a data-driven framework that combines statistical profiling, clustering techniques, and similarity analysis. The results demonstrate how advanced analytics can complement traditional scouting methods by providing objective, reproducible, and interpretable insights into player performance and stylistic roles.

From a methodological standpoint, the application of clustering algorithms revealed consistent and interpretable structures within the dataset of midfielders. Both K-Means and Hierarchical Clustering converged toward the identification of two main player archetypes:

- Creative and attack-oriented midfielders, characterized by strong contributions in chance creation, assists, and progressive actions;
- Supportive or defensively oriented midfielders, with lower involvement in direct offensive metrics.

This dichotomy reflects a meaningful segmentation of midfield roles and aligns with tactical distinctions observed in modern football. The consistency across different clustering techniques strengthens the validity of the findings.

The subsequent similarity analysis using k-Nearest Neighbors provided an additional layer of actionable insight. By comparing Barella's statistical profile with those of other midfielders across multiple distance metrics, the analysis consistently identified Rodrigo De Paul and Bruno Guimarães as the closest stylistic matches, with further parallels emerging for players such as Enzo Fernández, Pedri, and Martin Ødegaard. These results situate Barella within a reference group of versatile, high-impact midfielders who balance progression, creativity, and work rate—attributes central to his role at Inter.

Taken together, these findings demonstrate the value of integrating unsupervised learning with targeted similarity measures. Clustering offers a global view of role-based player segmentation, while KNN provides player-specific recommendations tailored to a chosen reference profile. This dual approach ensures that recruitment decisions are supported both at the macro level (identifying stylistic archetypes) and the micro level (pinpointing individual targets).

Beyond the immediate case study, the framework developed here has broader applicability. By adapting the preprocessing pipeline and feature set, the methodology can be extended to other positions, leagues, or tactical contexts. The interactive visualization component further enhances the practical utility of the framework, allowing scouts, analysts, and decision-makers to explore player similarities dynamically.

In conclusion, the project demonstrates how data science can bridge the gap between statistical analysis and football decision-making. By grounding recruitment strategies in quantitative evidence, clubs can reduce uncertainty, maintain tactical identity, and uncover undervalued talents. While statistical models cannot replace expert judgment, they provide a rigorous and scalable foundation on which scouting and recruitment processes can be built.