Détection de talents et profils de jeu dans le football européen à l'aide de techniques de clustering.

Projet par : Hafid OUCOUC , Youssef BOUAMAMA, Amine ITJI

2 novembre 2025

Résumé Exécutif

Dans ce rapport, nous présentons une analyse approfondie des performances individuelles de joueurs de football issus des principaux championnats européens. L'étude comprend la préparation et la normalisation des données, une analyse exploratoire des variables de performance, puis une segmentation des joueurs à l'aide de différentes méthodes de clustering non supervisé (K-Means, DBSCAN et Clustering hiérarchique). Les résultats mettent en évidence des profils distincts selon les postes — attaquants, milieux, défenseurs — et révèlent des typologies cohérentes entre joueurs « standards » et joueurs « d'élite ». Une attention particulière a été portée à l'identification de jeunes talents de moins de 23 ans au sein des clusters les plus performants, illustrant la pertinence de l'approche pour le recrutement sportif et la détection de potentiel. Nous avons travaillé de manière collaborative et équilibrée sur l'ensemble des étapes du projet, de la préparation des données à l'interprétation finale des résultats.

Remarque : Nous avons travaillé ensemble de manière équitable sur l'ensemble du projet, sans qu'aucune personne ne soit affectée seule à une partie spécifique.

Table des matières

1 Introduction

L'analyse de la performance des joueurs de football à travers les données statistiques est devenue un pilier essentiel du recrutement moderne et de la stratégie sportive. Dans un contexte où les clubs cherchent à optimiser leurs décisions à l'aide de la donnée, le *data mining* offre des outils puissants pour comprendre les profils de joueurs, identifier les talents émergents et détecter les similarités tactiques entre individus. L'objectif de ce projet est de réaliser une segmentation avancée des joueurs professionnels à partir de leurs statistiques de performance, afin de dégager des groupes homogènes reflétant des styles de jeu ou des rôles tactiques similaires. Cette étude s'inscrit dans une approche complète mêlant préparation de données, analyse exploratoire, réduction de dimensionnalité (PCA/Kernel PCA) et clustering (K-Means, DBSCAN, Hiérarchique), avec une interprétation visuelle et statistique des résultats.

Le jeu de données exploité contient près de 2 000 joueurs issus de plusieurs championnats européens. Chaque observation correspond à un joueur unique décrit par un ensemble de variables quantitatives : buts, passes décisives, tacles, interceptions, dribbles, passes progressives, taux de réussite, et autres indicateurs dérivés du modèle *Expected Goals (xG)*. Ces données proviennent d'une extraction structurée et ont été nettoyées pour assurer la cohérence des valeurs et la comparabilité entre postes. Afin de tenir compte des différences tactiques entre rôles, les joueurs ont été regroupés selon leur catégorie de poste (défenseurs centraux, latéraux, milieux, attaquants, ailiers, etc.). Cette segmentation initiale permet de mieux contextualiser les analyses statistiques et de garantir des comparaisons homogènes.

Le rapport suit une logique analytique progressive :

- **Présentation et nettoyage du dataset** : validation de la qualité et de la cohérence des données.
- **Analyse exploratoire** : étude des distributions, corrélations et profils moyens par poste.
- **Clustering** : application et comparaison des méthodes de segmentation.
- **Visualisation tactique** : représentation graphique des profils de clusters sur le terrain.
- **Détection de jeunes talents** : identification des joueurs de moins de 23 ans appartenant aux groupes d'élite.

À travers cette démarche, le projet vise non seulement à illustrer l'intérêt du *machine learning* pour la détection de talents footballistiques, mais aussi à démontrer comment les statistiques de performance peuvent être exploitées pour construire une vision objective et structurée du jeu moderne.

2 Présentation et préparation des données

2.1 Origine et description du jeu de données

Les données exploitées proviennent d'un jeu consolidé regroupant 1 828 joueurs professionnels issus de différents championnats européens. Chaque joueur est décrit par 23 variables statistiques mesurant ses performances offensives, défensives, techniques et collectives. Les données ont été prétraitées à partir d'extractions publiques de plateformes d'analyse de performance et standardisées pour être comparables entre postes. Les principales variables concernent :

- Les indicateurs offensifs : buts, expected goals (npxG), total de tirs.
- Les variables de création : passes décisives, passes progressives, xAG (expected assisted goals).
- Les mesures techniques : dribbles réussis, taux de passes réussies, possessions progressives.
- Les statistiques défensives : tacles, interceptions, duels aériens gagnés, dégagements.

Ces mesures ont été choisies pour couvrir l'ensemble du spectre tactique du joueur : participation offensive, construction du jeu et contribution défensive.

2.2 Nettoyage et traitement des données

Un nettoyage complet a été effectué pour garantir la fiabilité des analyses :

- **Doublons supprimés** : 2 doublons détectés et retirés à partir du couple (`player_name`, `Team Name`).
- **Valeurs manquantes** : imputées via la médiane du poste pour la variable `Age`.
- **Typage des variables** : uniformisation des formats numériques (flottants) et catégorisation des postes (ex. : CB, RB, LB $\rightarrow$ Défenseurs Centraux / Latéraux).
- **Création de la variable `Categorie_Clustering`** : regroupement tactique des postes selon les rôles réels sur le terrain (attaquants, milieux, défenseurs).
- **Standardisation** : toutes les variables numériques ont été normalnées via `MinMaxScaler` pour garantir une pondération équitable dans le clustering.

2.3 Jeu de données final

Le dataset final contient 1 826 joueurs répartis en 7 grandes catégories de poste, sans valeurs manquantes. Il constitue une base solide pour les analyses exploratoires et les méthodes de segmentation à venir, garantindo une lecture homogène des performances individuelles.

3 Analyse exploratoire des données

Avant d'aborder les méthodes de clustering, une étape d'analyse exploratoire est essentielle pour comprendre la structure générale du jeu de données et valider la pertinence des variables statistiques. L'objectif est de vérifier la distribution des postes, d'étudier la composition du dataset et d'observer les tendances générales selon les rôles des joueurs.

3.1 Répartition et catégorisation des postes

L'analyse de la variable `Categorie_Clustering` montre une répartition équilibrée des rôles sur le terrain, bien que certaines catégories soient plus représentées. Les défenseurs centraux et les latéraux constituent près de 40 % de l'effectif total, confirmant l'importance de ces postes dans les effectifs européens. Les attaquants (13.5 %) et milieux offensifs (6.9 %) sont logiquement moins nombreux, traduisant la rareté des profils à vocation purement offensive.

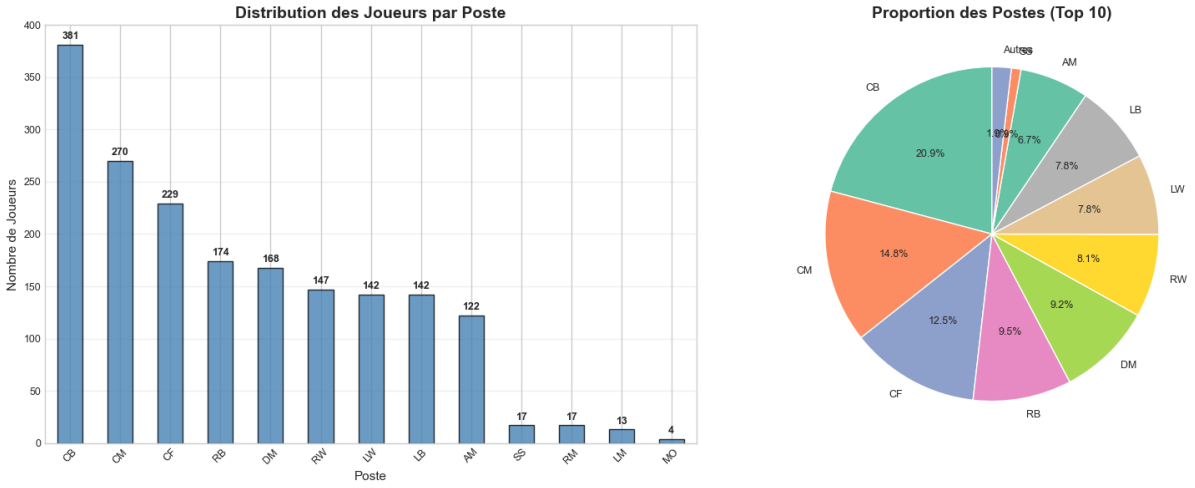

Figure 1. Répartition des joueurs par catégorie de poste (bar chart + pie chart)

Les postes initiaux ont été regroupés selon leur rôle tactique principal. L'objectif était de réduire la granularité excessive des positions et de créer des ensembles de joueurs comparables. La classification finale comprend 7 catégories :

Table 1. Catégorisation des postes pour le clustering

Catégorie de Clustering	**Postes inclus**	**Nombre total**	**% du dataset**	**Description du profil**
Défenseurs Centraux	CB	381	20.9 %	Défenseurs axiaux spécialisés dans les duels aériens et le marquage.
Latéraux	LB, RB, LWB, RWB	316	17.3 %	Joueurs de couloir polyvalents, alternant entre défense et attaque.
Milieux Défensifs	DM, CDM	168	9.2 %	Récupérateurs chargés de protéger la défense et relancer proprement.
Milieux Centraux	CM, LCM, RCM	270	14.8 %	Relayeurs équilibrés, cœur du jeu et de la transition.
Milieux Offensifs	AM, CAM, MO	126	6.9 %	Meneurs de jeu, créateurs et soutiens à l'attaque.
Ailiers	LW, RW, LF, RF, RM, LM	319	17.5 %	Joueurs excentrés, forts en dribbles et en percussion.
Attaquants	CF, ST, FW, SS	246	13.5 %	Finisseurs et attaquants de surface, souvent point d'appui de l'équipe.

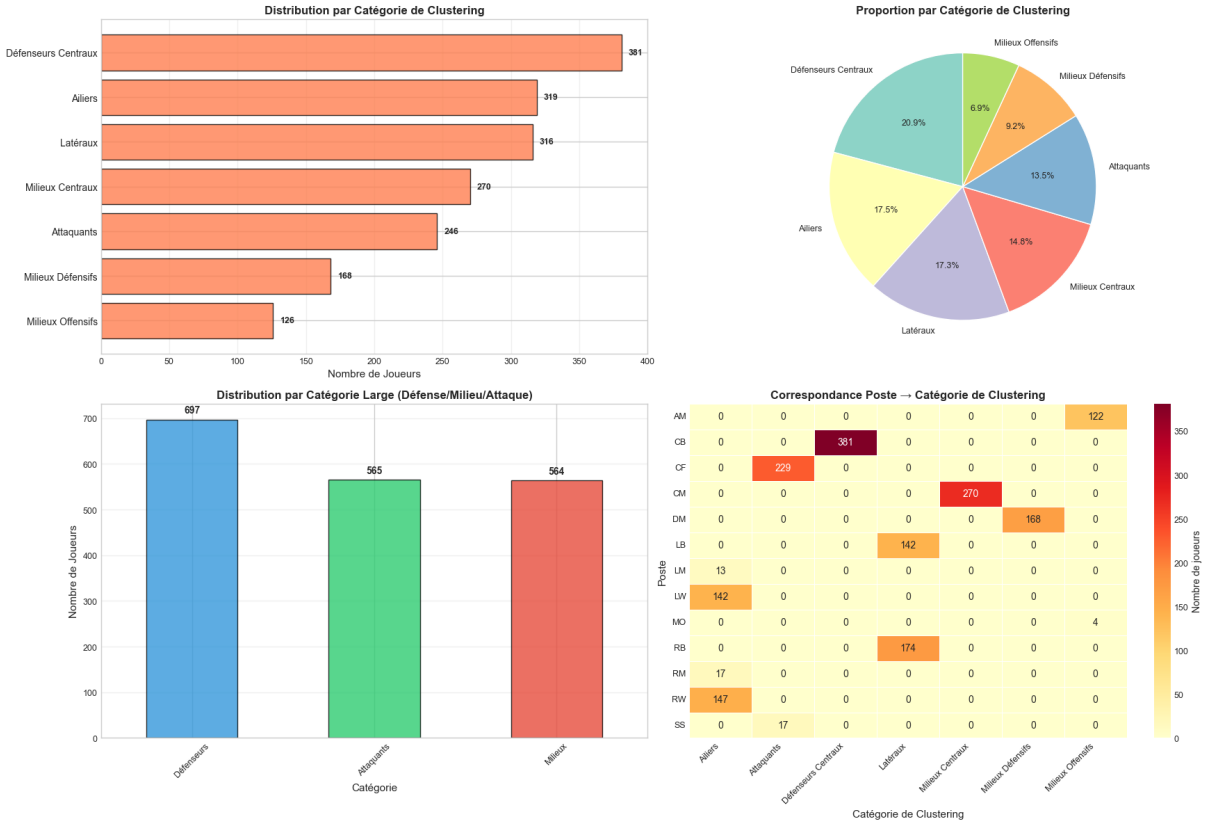

Figure 2. Répartition et catégorisation des postes (barres, secteurs et heatmap)

3.2 Analyse spécifique des RM/LM

Les milieux latéraux (RM/LM) présentent une grande similarité statistique avec les ailiers (RW/LW). L'analyse multidimensionnelle (ACP sur 5 variables offensives) confirme un fort recouvrement entre ces profils. Ces postes ont donc été fusionnés dans la catégorie "Ailiers" afin de garantir des groupes homogènes pour le clustering.

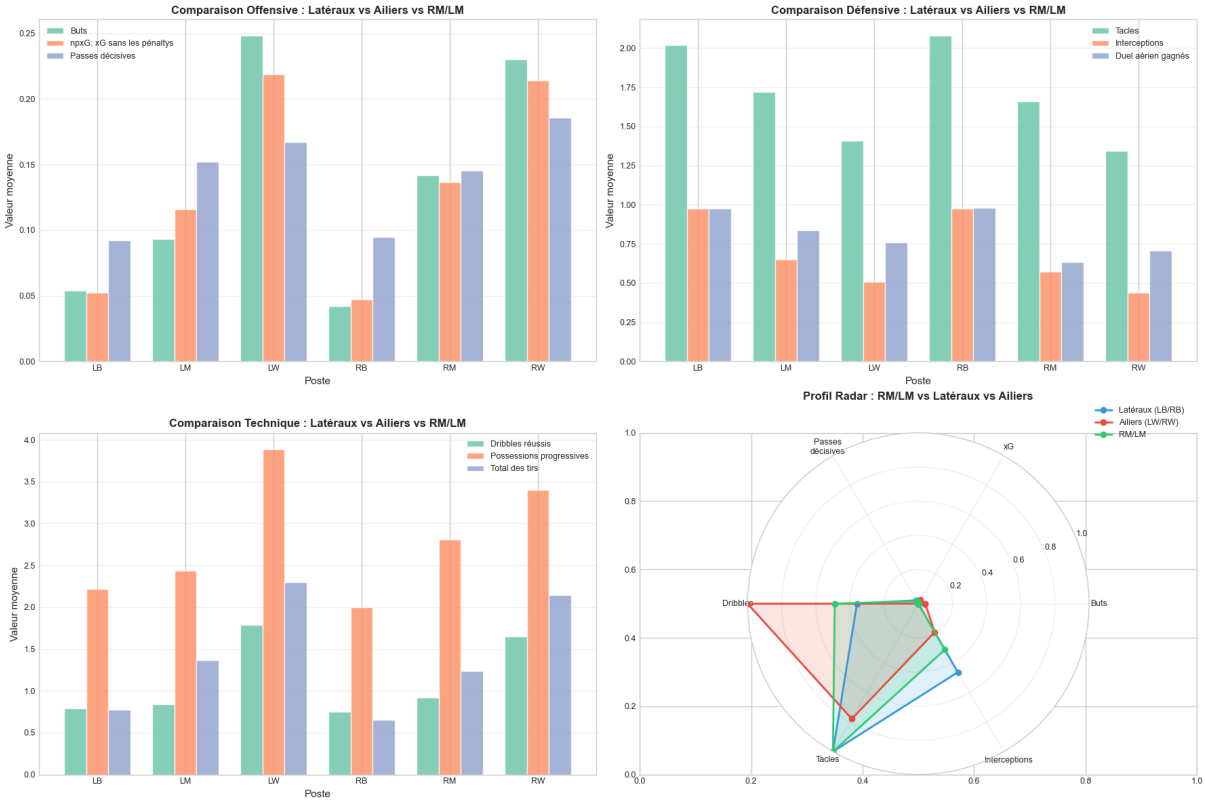

Figure 3. Analyse de similarité entre RM/LM, latéraux et ailiers

3.3 Analyse de la variable Âge

La distribution de l'âge des joueurs est centrée autour de 26.3 ans, avec une forte concentration entre 22 et 29 ans. Près de 20 % des joueurs ont moins de 23 ans, une population particulièrement intéressante dans le cadre de la détection de jeunes talents. Les ailiers et attaquants présentent les moyennes d'âge les plus faibles, traduisant la valorisation des profils explosifs et rapides.

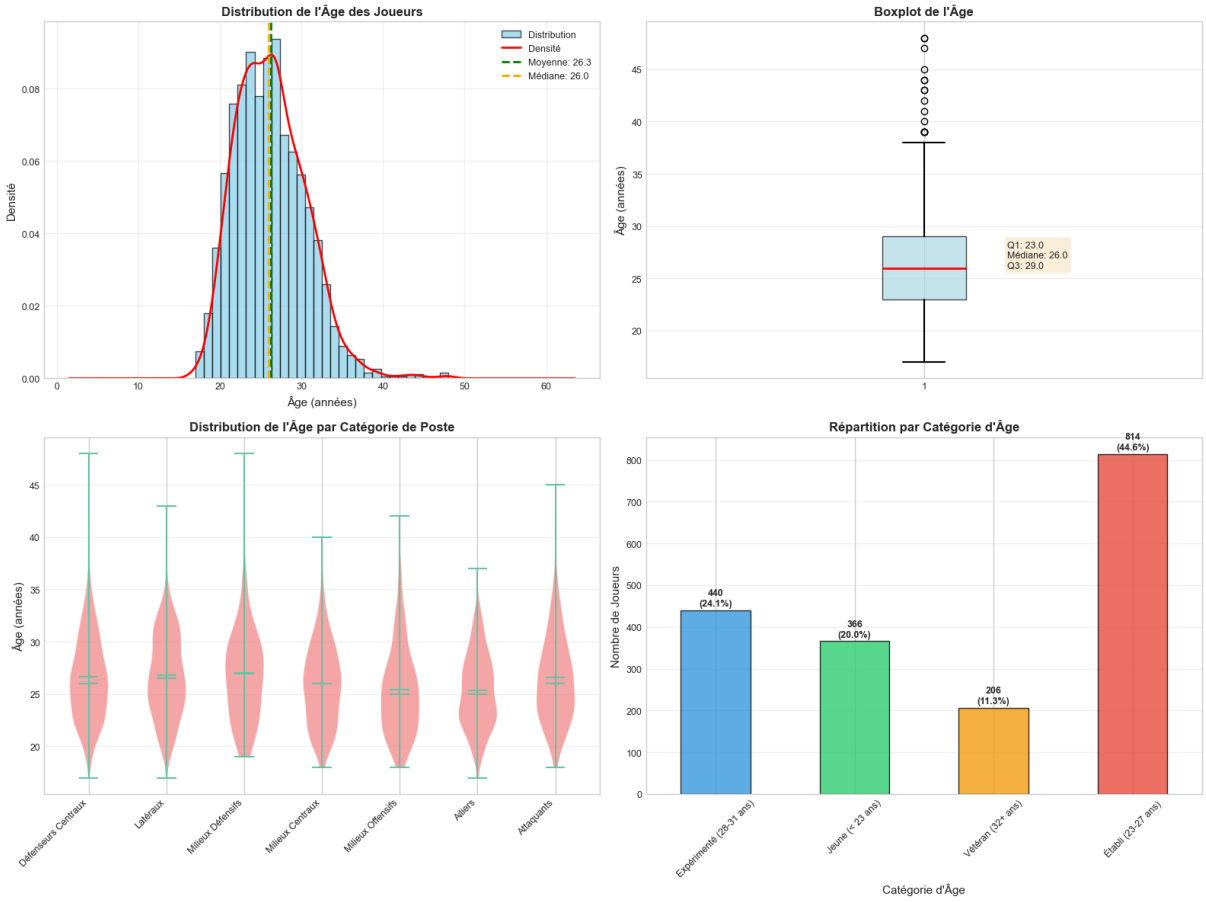

Figure 4. Distribution de l'âge globa

3.4 Profils statistiques moyens par catégorie

L'étude des moyennes normalisées sur un ensemble de variables clés (offensives, techniques et défensives) met en évidence des différences nettes de style entre postes :

- **Attaquants** : dominent sur les indicateurs buts, npxG et tirs.
- **Ailiers** : se distinguent par les dribbles réussis et possessions progressives.
- **Milieux offensifs** : excellents en passes décisives et xAG.
- **Milieux centraux** : profils équilibrés ("box-to-box"), combinant passes et tacles.
- **Milieux défensifs** : forts en interceptions et récupérations.
- **Latéraux** : hybrides, contribuant autant en passes progressives qu'en tacles.
- **Défenseurs centraux** : dominent les duels aériens et dégagements.

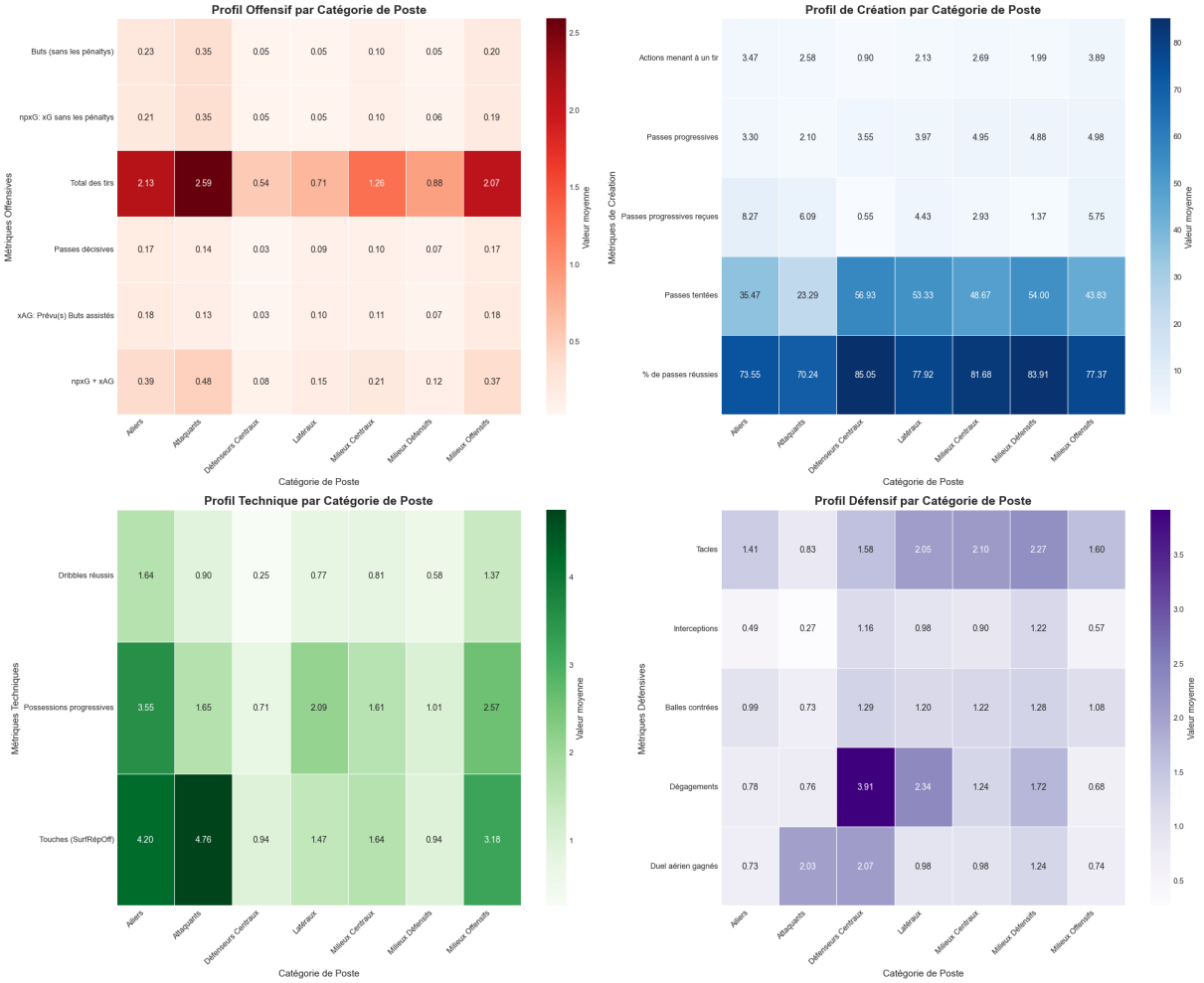

Figure 5. Profils moyens par catégorie de poste (offensif, créatif, technique et défensif)

Ces résultats confirment la cohérence tactique des catégories de poste. Chaque groupe statistique se distingue clairement par son rôle sur le terrain :

— Les **attaquants et ailiers** forment la zone offensive dominante.
— Les **milieux** équilibrent entre création et récupération.
— Les **défenseurs** assurent la stabilité défensive et la relance.

3.5 Corrélations et variables discriminantes

L'étude de la matrice de corrélation révèle plusieurs associations fortes entre indicateurs :

— npxG $\leftrightarrow$ Buts ($r = 0.84$)
— xAG $\leftrightarrow$ Passes décisives ($r = 0.77$)
— Dribbles $\leftrightarrow$ Possessions progressives ($r = 0.77$)
— Possessions $\leftrightarrow$ Passes progressives reçues ($r = 0.78$)

Ces corrélations confirment la redondance partielle entre certaines métriques, ce qui justifie l'application d'une réduction de dimension (PCA) avant le clustering. Aucune corrélation négative majeure n'a été détectée, suggérant que les dimensions offensives et défensives ne s'opposent pas mais coexistent selon les rôles. Les variables les plus discriminantes — buts, npxG, tacles, passes progressives reçues — serviront de base à la segmentation tactique.

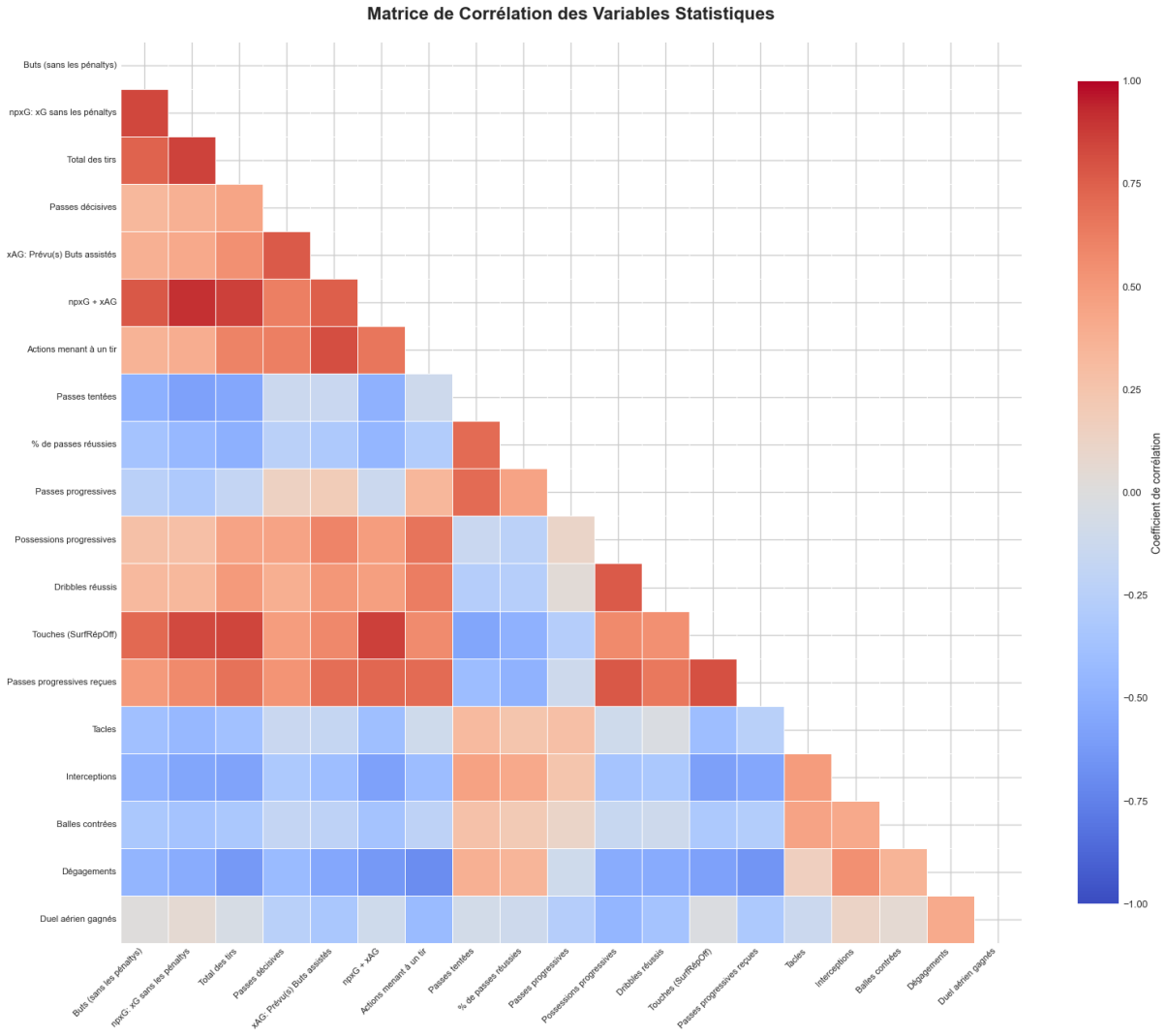

Figure 6. Matrice de corrélation des variables statistiques principales

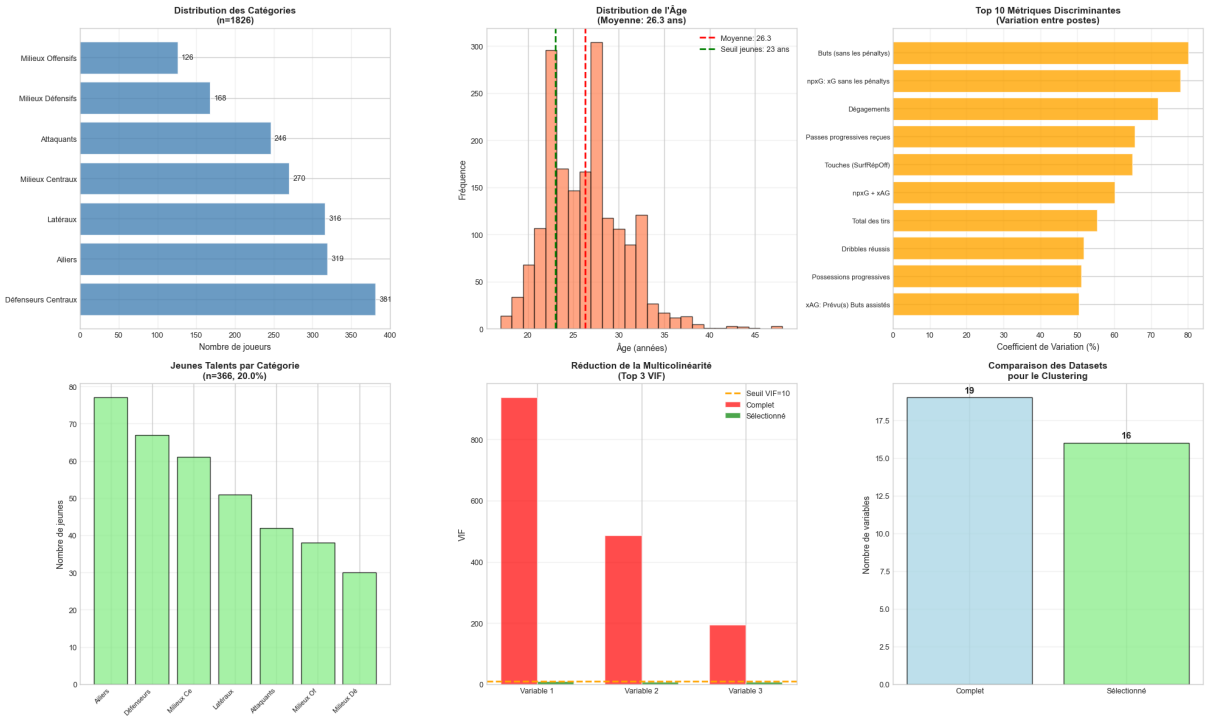

Figure 7. Graphe résumant les parties d'exploration de données.

4 Clustering et segmentation des joueurs

4.1 Méthodologie et choix des algorithmes

Avant d'appliquer les algorithmes de clustering, plusieurs étapes méthodologiques ont été nécessaires pour garantir la comparabilité des joueurs, la stabilité des modèles et la pertinence des résultats.

4.1.1 Prétraitement et normalisation des variables

Les données initiales comportaient des variables exprimées dans des unités très différentes (tacles, passes, xG, dribbles, etc.). Une étape de normalisation a donc été indispensable afin de ramener toutes les variables sur une même échelle et éviter qu'une métrique à forte amplitude ne domine les autres.

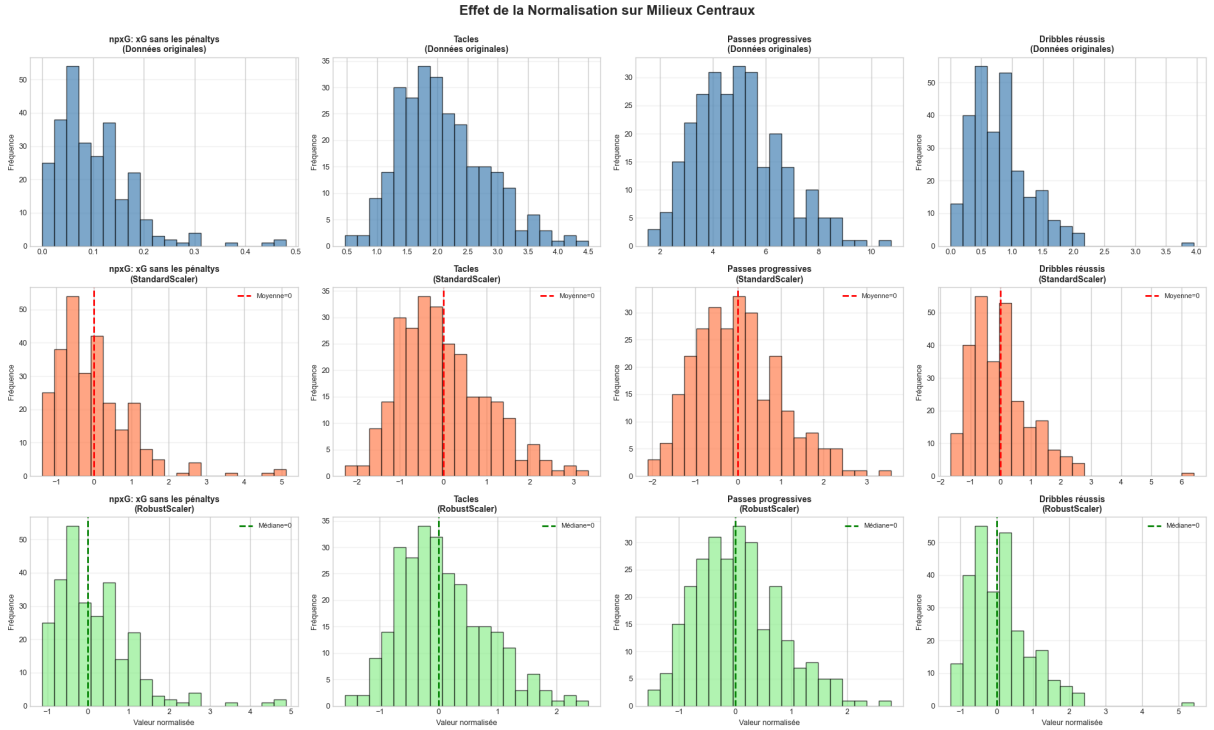

Figure 8. Effet de la normalisation sur les distributions de variables clés

Trois méthodes ont été testées :

- `StandardScaler` : centrage-réduction classique (moyenne = 0, écart-type = 1).
- `RobustScaler` : basé sur la médiane et l'écart interquartile, plus robuste aux valeurs extrêmes.
- `Aucune normalisation` : référence pour comparaison.

L'analyse montre que le `StandardScaler` produit une distribution symétrique et centrée, adaptée à des méthodes comme K-Means qui supposent des distances euclidiennes homogènes. Le `RobustScaler` conserve mieux la forme des données, mais reste légèrement moins performant pour la variance inter-catégories.

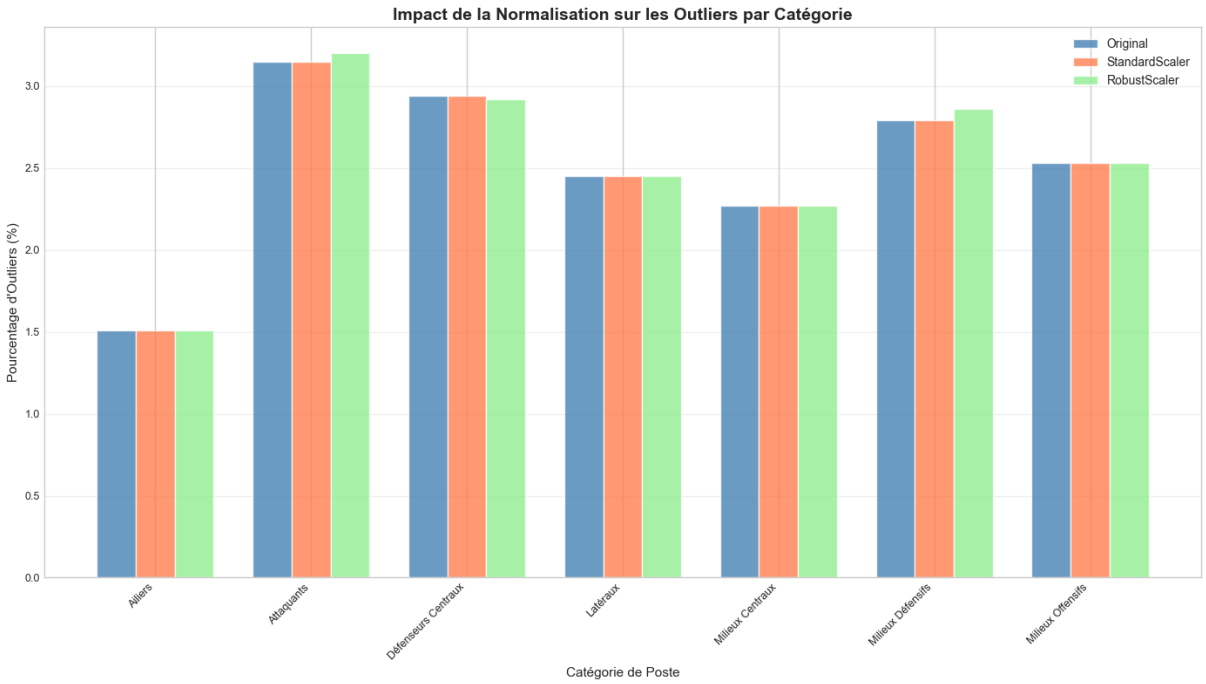

Figure 9. Impact de la normalisation sur le pourcentage d'outliers par catégorie

L'étude des valeurs extrêmes confirme la stabilité du StandardScaler : la proportion d'outliers reste quasi constante (<3,5%) dans toutes les catégories de postes, garantindo une répartition équilibrée des joueurs lors du clustering.

4.1.2 Réduction de dimension par Analyse en Composantes Principales (PCA)

Après la normalisation, les 19 variables statistiques ont été projetées dans un espace réduit afin d'éliminer la redondance et de simplifier la structure de données. L'Analyse en Composantes Principales (PCA) a été appliquée séparément pour chaque catégorie de poste afin d'examiner la variance expliquée par les premières composantes.

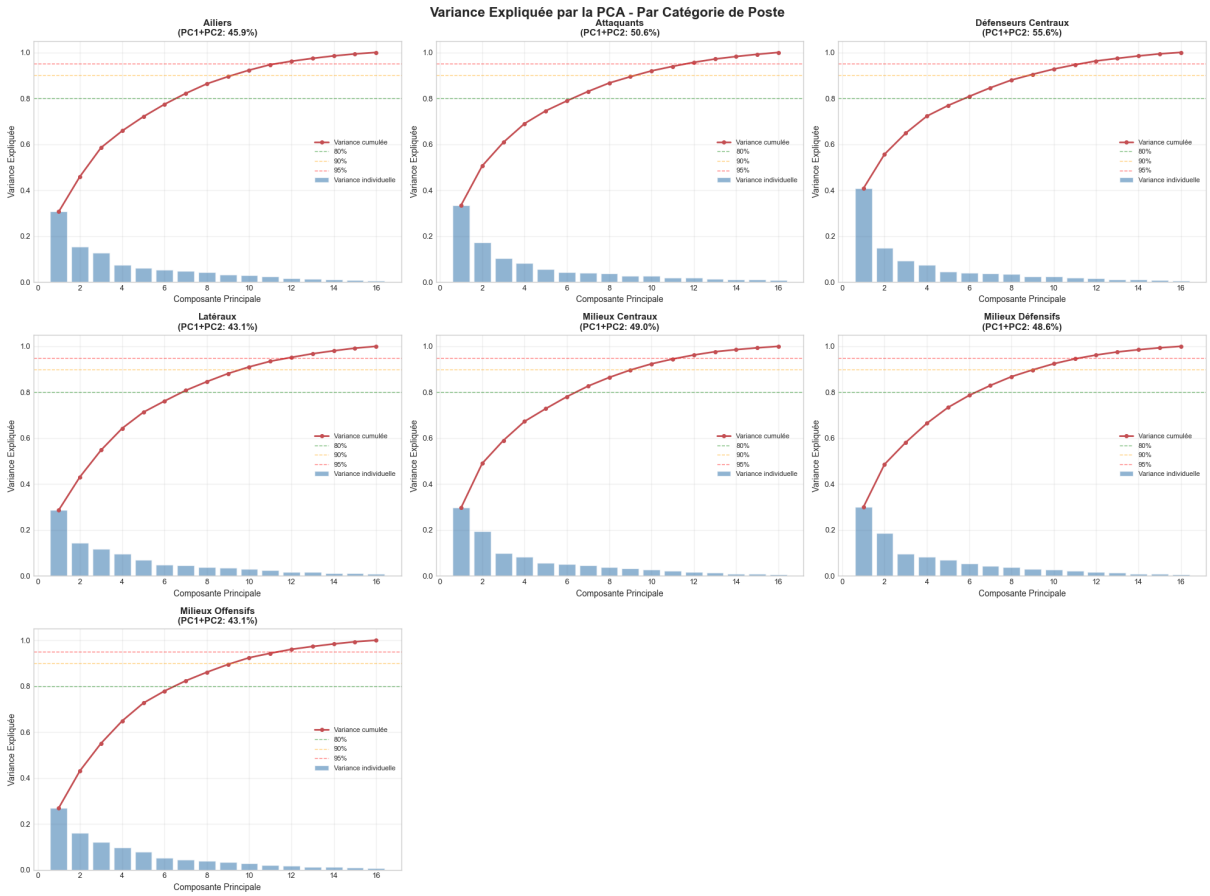

Figure 10. Variance expliquée par la PCA selon la catégorie de poste

4.1.3 Choix des algorithmes de clustering

Trois méthodes complémentaires ont été sélectionnées :

- **K-Means** : segmentation basée sur la distance euclidienne, adaptée aux données normalisées.
- **DBSCAN** : identification des groupes denses et détection des anomalies.
- **Clustering hiérarchique** : exploration des relations de similarité à différents niveaux.

Chaque approche a été testée sur les données issues de la PCA (2 à 5 composantes) afin d'obtenir une segmentation stable, interprétable et robuste aux valeurs extrêmes.

4.2 K-Means : détermination du k optimal

4.2.1 Méthode du coude et score de silhouette

Afin d'identifier le nombre optimal de clusters, deux métriques complémentaires ont été utilisées :

- **L'inertie intra-cluster (méthode du coude)**, qui mesure la compacité des groupes.
- **Le score de silhouette**, qui évalue la séparation entre clusters (valeurs proches de 1 indiquent une bonne séparation).

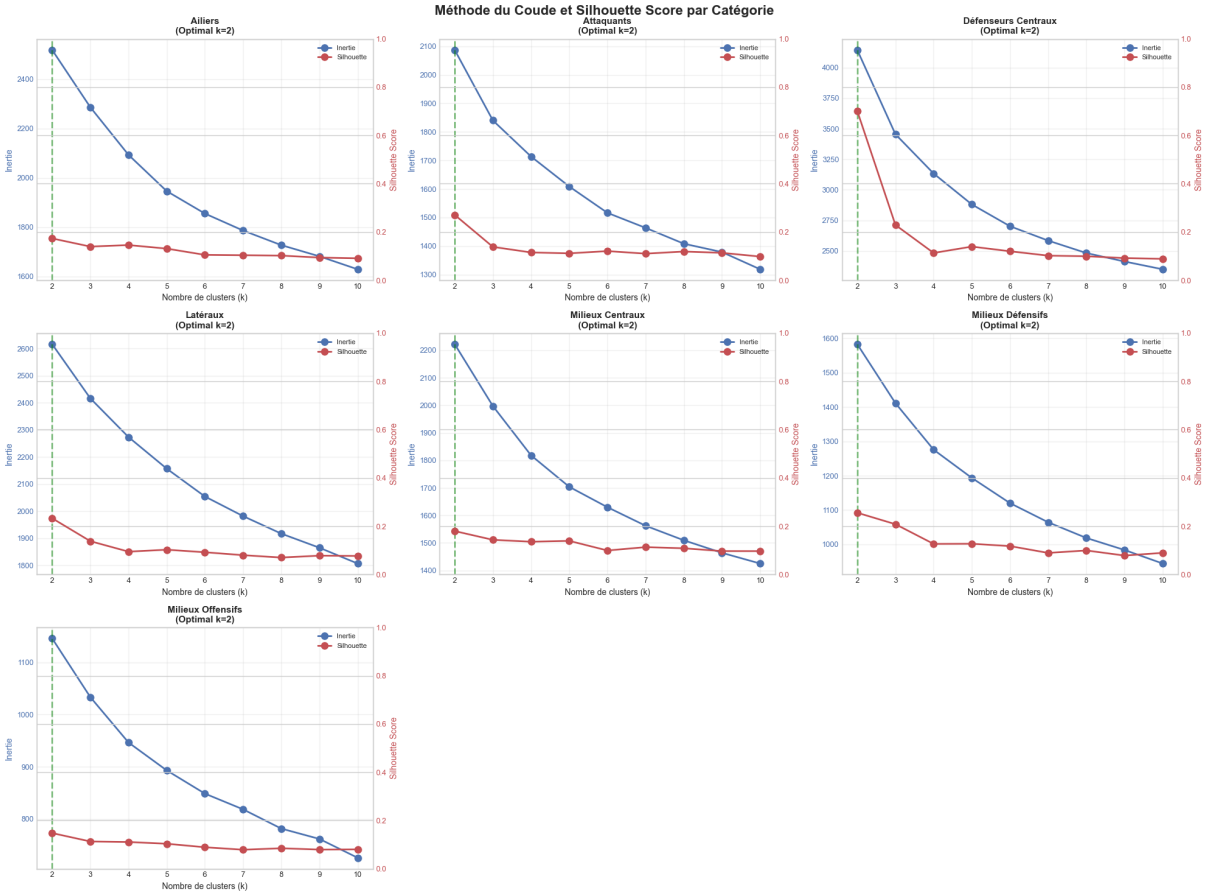

Figure 11. Méthode du coude et silhouette score par catégorie

Pour chaque catégorie de poste (défenseurs, milieux, attaquants, etc.), les deux courbes ont été comparées entre $k = 2$ et $k = 10$. On observe un point d'inflexion net dès $k = 2$ dans toutes les positions, indiquant une stabilisation de la réduction d'inertie. Simultanément, le score de silhouette est maximal à $k = 2$, confirmant la présence de deux grands profils distincts au sein de chaque catégorie. Ainsi :

- Les **défenseurs centraux** et **milieux défensifs** se divisent entre profils purement défensifs et relanceurs.
- Les **ailiers** et **latéraux** opposent les profils créatifs (offensifs) aux profils plus équilibrés (travail défensif).
- Chez les **attaquants**, la segmentation distingue les finisseurs des joueurs plus participatifs.

Cette observation est cohérente avec les tendances tactiques du football moderne, où la polyvalence crée souvent deux sous-profils dominants par poste.

4.2.2 Interprétation visuelle des clusters

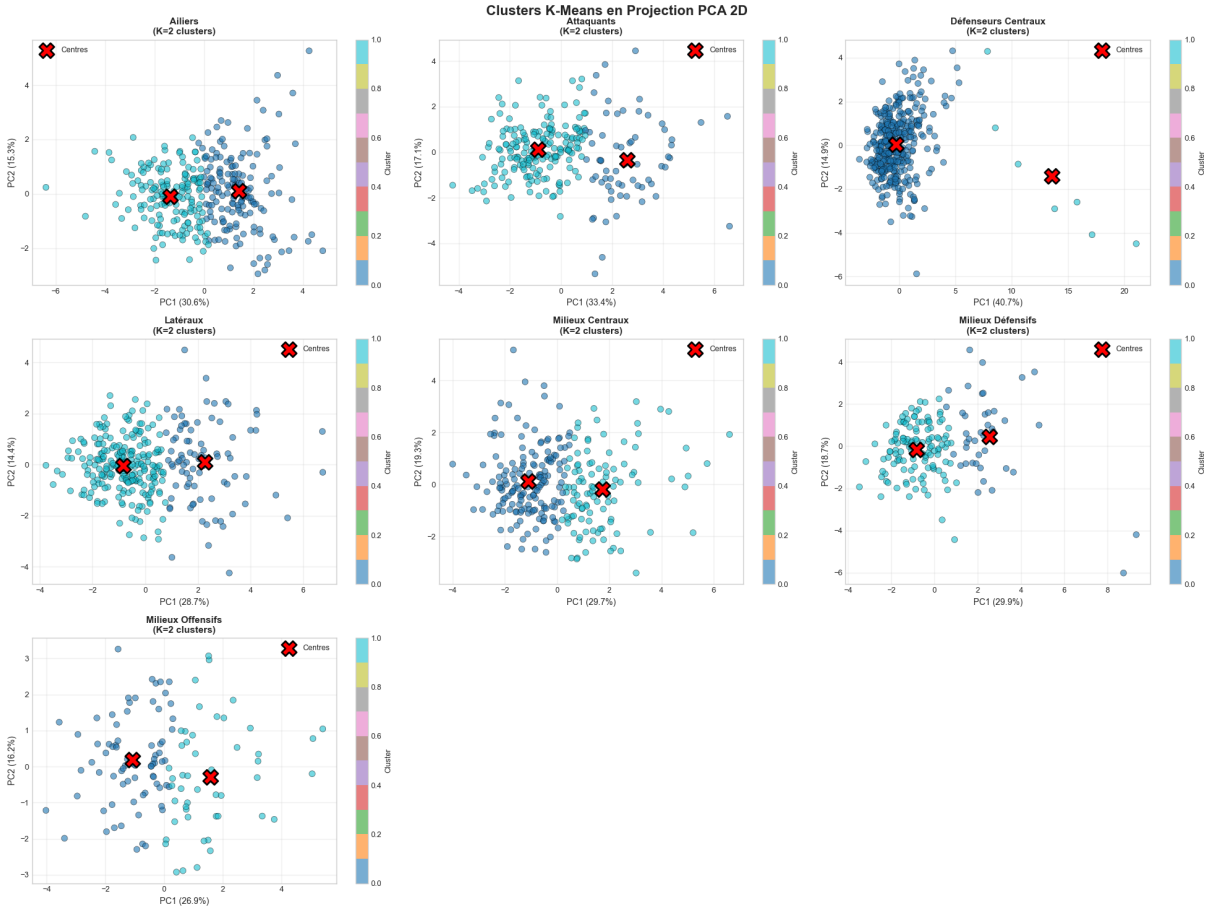

Figure 12. Visualisation PCA 2D des clusters K-Means (k=2 par poste

Les données ont été projetées dans le plan formé par les deux premières composantes principales de la PCA, permettant une interprétation intuitive des regroupements. Chaque nuage de points correspond à une catégorie de poste, et les centroïdes (croix rouges) représentent les moyennes des groupes détectés. L'analyse montre :

- Une **bonne séparation** des clusters pour la majorité des postes, notamment chez les défenseurs centraux et milieux défensifs.
- Une **superposition partielle** pour les latéraux et ailiers, où certains joueurs présentent des caractéristiques mixtes (joueurs de couloir modernes).
- Les **milieux centraux** affichent une dispersion plus large, signe d'une grande diversité de profils (box-to-box, créateurs, récupérateurs).

Ces résultats valident la pertinence du choix k=2, tout en illustrant les spécificités propres à chaque rôle. Le K-Means permet ainsi d'obtenir une première segmentation claire et interprétable avant la comparaison avec d'autres approches.

4.3 DBSCAN et Clustering hiérarchique : comparaison

4.3.1 DBSCAN – Détection de profils atypiques et analyse visuelle

L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) a été utilisé afin d'identifier les regroupements naturels de joueurs sans imposer un nombre

fixe de clusters. Contrairement à K-Means, il ne suppose pas de structure sphérique, mais se base sur la densité locale des points pour isoler à la fois des groupes homogènes et des profils atypiques (outliers).

a) Paramétrage et évaluation Les paramètres ont été calibrés pour chaque catégorie de poste :

— ϵ (epsilon) entre 1.5 et 2.0,
— MinPts = 5.

Trois indices ont servi à évaluer la cohérence des regroupements :

— Silhouette Score (séparation inter-groupes),
— Davies-Bouldin (compacité intra-cluster),
— Calinski-Harabasz (rapport dispersion inter/intra).

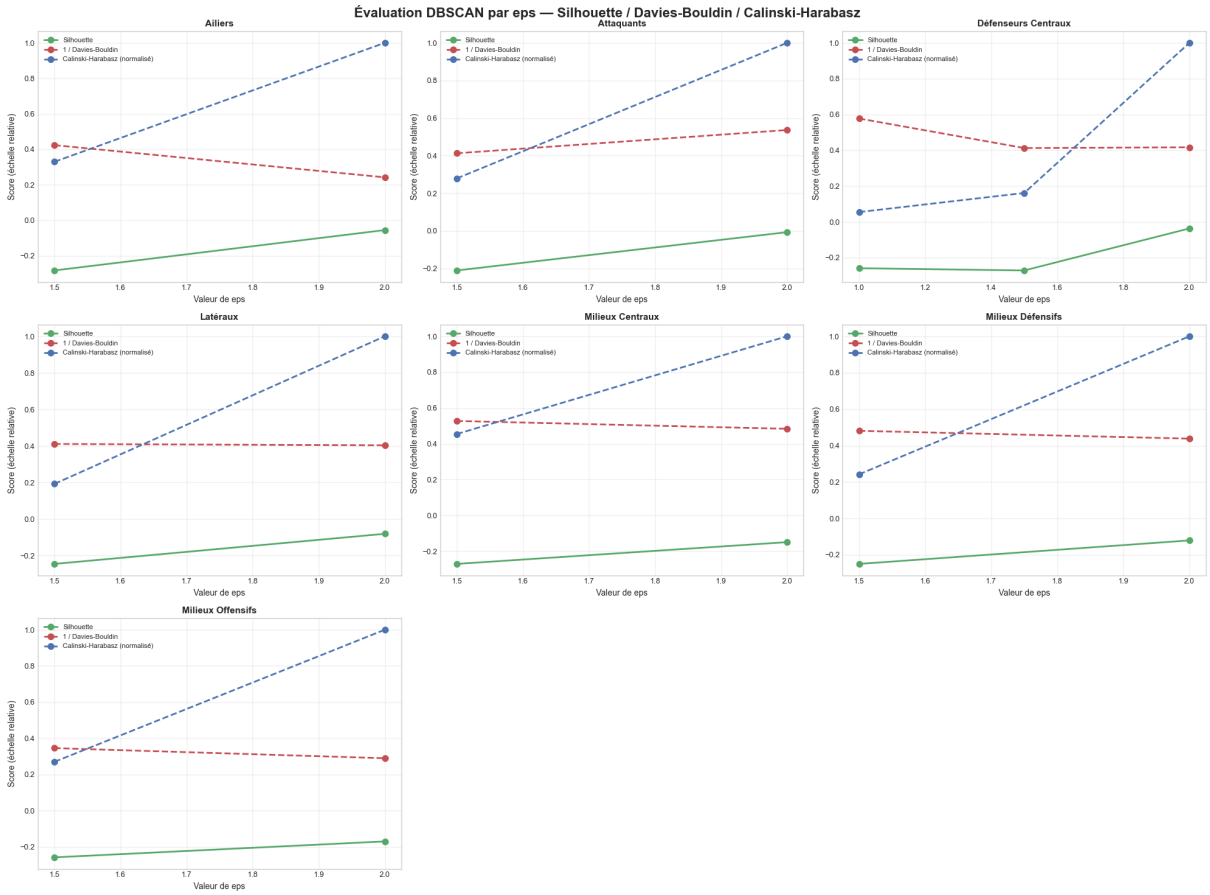

Figure 13. Évaluation DBSCAN par eps (Silhouette/Davies-Bouldin/Calinski-Harabasz)

Les courbes indiquent une stabilisation des scores pour $\epsilon \approx 1.8$, confirmant un équilibre entre compacité et séparation. Toutefois, la densité des données varie selon les postes : les attaquants et milieux centraux produisent plus de petits regroupements, tandis que les défenseurs et latéraux présentent davantage d'outliers.

Figure 14. Clusters DBSCAN en projection PCA 2D

b) Résultats globaux Les visualisations PCA montrent que DBSCAN détecte :

— 2 à 8 clusters par catégorie,

— et un fort volume d'outliers (10 à 25 %) selon la dispersion statistique.

Ces outliers correspondent souvent à des profils hybrides ou singuliers, qui combinent plusieurs qualités statistiques (ex. défenseurs bons relanceurs ou milieux à vocation offensive).

c) Interprétation par poste

— **Défenseurs centraux :** Deux groupes bien distincts apparaissent. Le premier regroupe les stoppeurs — très performants dans les duels et l'interception mais peu impliqués dans la relance. Le second regroupe les relanceurs, caractérisés par une meilleure contribution à la construction du jeu (passes progressives, longues relances). Les nombreux outliers traduisent la présence de profils intermédiaires entre ces deux extrêmes. → DBSCAN révèle ici la coexistence de deux archétypes tactiques majeurs dans le football moderne.

— **Milieux centraux / défensifs :** La structure est plus diffuse, avec plusieurs sous-groupes peu séparés. Cela illustre la variété de rôles au milieu de terrain : récupérateurs, relayeurs, meneurs bas. Les outliers représentent souvent des milieux polyvalents, capables d'assurer plusieurs fonctions selon le contexte tactique.

— **Ailiers et latéraux :** Les nuages de points sont larges et dispersés, avec beaucoup de joueurs isolés. Cela traduit une diversité stylistique importante : ailiers créatifs,

contre-attaquants rapides, ou profils hybrides impliqués défensivement. Cette variabilité rend la séparation par densité plus difficile mais souligne la richesse des profils offensifs modernes.

— **Attaquants :** DBSCAN identifie deux ou trois regroupements clairs : les finisseurs purs (hauts npxG, peu de passes) et les attaquants participatifs (fort volume de passes et dribbles réussis). Les outliers renvoient à des profils atypiques, comme les "faux neufs" ou seconds attaquants.

d) Interprétation synthétique L'algorithme DBSCAN met en lumière :

— la structure dense et compacte des profils "classiques",

— la présence marquée d'outliers, représentant les joueurs les plus singuliers du dataset,

— et la continuité tactique entre les styles de jeu, visible dans les zones intermédiaires de la PCA.

Ainsi, DBSCAN s'impose comme un outil complémentaire à K-Means : plutôt que d'imposer des frontières rigides, il révèle les zones de densité naturelle dans l'espace des caractéristiques et identifie les profils hors norme — souvent les plus intéressants dans une optique de recrutement ou d'analyse tactique avancée.

4.3.2 Clustering Hiérarchique : Structure et continuité des profils

Le clustering hiérarchique a été appliqué afin d'examiner la structure interne des données et d'observer si les joueurs se répartissent selon des niveaux hiérarchiques cohérents. Cette approche complète K-Means et DBSCAN en permettant une segmentation progressive, sans fixer a priori le nombre de groupes. Trois stratégies de linkage ont été comparées :

— **Ward**, privilégiant des groupes compacts et homogènes,

— **Average**, mesurant la similarité moyenne entre points,

— **Complete**, sensible aux points extrêmes.

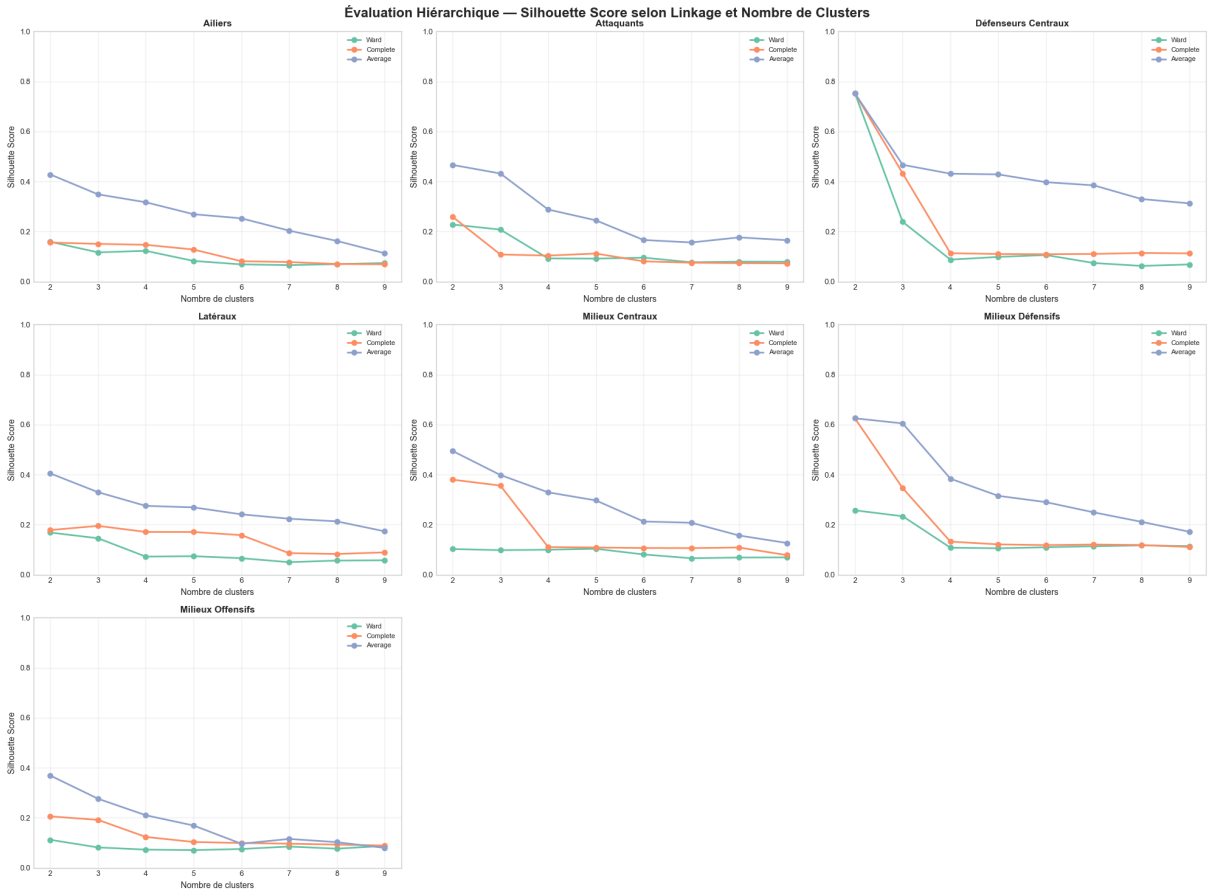

Figure 15. Évaluation hiérarchique par linkage et nombre de clusters

Les courbes de Silhouette Score ont été analysées pour $k \in [2, 9]$. Dans toutes les catégories de postes, la méthode Ward obtient les meilleurs scores de cohésion (entre 0.3 et 0.6 pour k = 2), tandis que les linkages average et complete montrent une baisse rapide au-delà de k = 3. Cette tendance indique que les profils de joueurs ne se séparent pas naturellement en nombreux sous-groupes : la structure des données reste faiblement hiérarchisée, ce qui justifie le choix d'un k = 2 pour la visualisation finale.

Figure 16. Clusters hiérarchiques en projection PCA 2D

a) Visualisation des clusters hiérarchiques

b) Interprétation par catégorie de poste

— **Défenseurs centraux (Ward, k=2)** Le nuage de points est dense et resserré, traduisant une forte homogénéité. Les deux clusters identifiés reflètent surtout un niveau d'intensité défensive différent : d'un côté des profils "stoppeurs" dominants dans les duels, de l'autre des défenseurs plus calmes mais meilleurs relanceurs. Il ne s'agit pas de rôles opposés, mais de nuances de style au sein d'un même archétype.

— **Milieux défensifs (Complete, k=2)** Légère séparation sur l'axe horizontal (PC1), correspondant à la dimension technique : les joueurs du premier groupe sont principalement des récupérateurs, tandis que ceux du second affichent une meilleure contribution à la transition offensive (passes progressives, interceptions hautes).

— **Milieux centraux (Average, k=2)** La distribution est floue, suggérant un continuum entre deux pôles : les milieux relayeurs équilibrés, et les créateurs avancés. Cette absence de frontière nette illustre la polyvalence du milieu moderne, souvent capable de s'adapter à plusieurs rôles selon la phase de jeu.

— **Ailiers et latéraux (Average, k=2)** La dispersion est importante : peu de structure interne, car ces postes regroupent une grande variété de styles (ailiers de percussion, de création ou hybrides défensifs). Les deux clusters se superposent presque totalement, traduisant une transition fluide entre les profils.

— **Milieux offensifs et attaquants (Average, k=2)** Les joueurs se répartissent de manière circulaire, sans séparation nette. Cela traduit la corrélation élevée entre les métriques offensives (tirs, xG, passes clés) et donc une segmentation faible. Les quelques variations identifiées correspondent à une opposition entre finisseurs et créateurs, mais la frontière reste poreuse.

c) Interprétation globale Le clustering hiérarchique révèle que les joueurs se distribuent de façon continue plutôt que discrète dans l'espace des performances. Il met en évidence :

— une progression hiérarchique graduelle des styles (ex. du récupérateur au relanceur),
— une homogénéité forte au sein des postes défensifs,
— et une plus grande variabilité pour les rôles offensifs, où la créativité rend la classification floue.

En somme, cette méthode met en avant la continuité des profils et complète les résultats de K-Means et DBSCAN : là où K-Means impose deux classes et DBSCAN détecte des anomalies locales, le clustering hiérarchique montre que les joueurs forment un spectre tactique fluide, sans rupture marquée entre catégories.

Synthèse des résultats et interprétations des méthodes de clustering

Table 2. Synthèse des résultats et interprétations des méthodes de clustering

Catégorie de poste	**K-Means (k=2)**	**DBSCAN**	**Clustering hiérarchique**
Défenseurs centraux	Deux profils nets : stoppeurs (duels, dégagements) vs relanceurs (passes progressives).	Deux clusters denses + nombreux outliers ; profils hybrides entre défense pure et relance.	Groupes peu séparés, homogénéité forte ; variation d'intensité plutôt que de rôle.
Milieux défensifs	Opposition entre récupérateurs purs et relayeurs plus techniques.	2–3 clusters ; profils atypiques représentant des milieux "box-to-box" ou plus offensifs.	Séparation modérée ; continuum entre profils récupérateurs et relayeurs.
Milieux centraux	Segmentation équilibrée entre profils défensifs et créatifs.	Structure diffuse ; nombreux outliers, forte diversité de rôles au milieu.	Faible hiérarchie — transition continue entre relayeurs, créateurs et récupérateurs.
Milieux offensifs	Deux sous-profils : créateurs (passes clés) vs soutiens d'attaque (tirs, xG).	Clusters peu denses, nombreux joueurs hybrides ; forte dispersion en PCA.	Répartition circulaire, segmentation faible ; continuum entre meneurs et finisseurs.
Ailiers	Distinction entre ailiers créatifs et ailiers équilibrés.	2–3 clusters peu nets, beaucoup d'outliers ; diversité stylistique élevée.	Groupes se superposant ; profils continus du dribbleur au joueur défensif.
Latéraux	Opposition entre profils offensifs et défensifs.	2 clusters peu séparés, forte proportion d'outliers ; profils hybrides.	Structure très continue, pas de rupture marquée.
Attaquants	Deux segments clairs : finisseurs vs participatifs / créateurs.	2–3 clusters + quelques outliers (faux neufs, seconds attaquants).	Groupes homogènes ; légère différenciation entre créateurs et finisseurs.
Résumé global	Structure nette et stable (k=2), utile pour typologie générale.	Détection fine des outliers et profils atypiques.	Représentation progressive des rôles, utile pour hiérarchiser les styles.

5 Analyses des résultats

L'objectif de cette étape est d'identifier, à partir des statistiques de la saison 2024/2025, 10 joueurs correspondant à un besoin de recrutement spécifique. Cette approche permet d'analyser les profils détectés par le clustering et de vérifier, à travers différents cas pratiques, la pertinence du modèle dans la recherche de joueurs aux styles de jeu comparables.

5.1 Cas 1- Recherche d'un joueur similaire au style de jeu proche de Kylian Mbappé

Table 3. Joueurs similaires à Kylian Mbappé

player_name	Position	Team Name	Age	Cluster	Distance
Mohamed Salah	RW	Liverpool	32.0	1.0	2.584870
Vinicius Júnior	LW	Real Madrid	24.0	1.0	2.749357
Donyell Malen	RW	Dortmund	25.0	1.0	3.257669
Luis Díaz	LW	Liverpool	27.0	1.0	3.288566
Rodrygo	RW	Real Madrid	23.0	1.0	3.330317
Rafael Leão	LW	Milan	25.0	1.0	3.404504
Ángel Correa	RW	Atlético Madrid	29.0	1.0	3.573733
Estevão	RW	Palmeiras	17.0	1.0	3.593480
Dennis Man	RW	Parma	26.0	1.0	3.692224
Khvicha Kvaratskhelia	LW	Napoli	23.0	1.0	3.726603

Le modèle a permis d'identifier des joueurs au style de jeu proche de Kylian Mbappé, caractérisés par leur vitesse, explosivité et efficacité offensive. Même sans inclure directement la vitesse comme variable, le clustering a regroupé des profils similaires tels que Vinicius Júnior, Mohamed Salah et Rafael Leão, tous issus de grands clubs européens. Cela montre que ces qualités sont captées indirectement par des variables comme les dribbles réussis, les progressions avec ballon ou le npxG, révélant des caractéristiques cachées du style de jeu.

5.2 Cas 2 - Recherche d'un jeune talent au profil similaire à Cole Palmer

Table 4. Jeunes talents similaires à Cole Palmer

player_name	Position	Team Name	Age	Cluster	Distance
Xavi Simons	AM	RB Leipzig	21.0	1.0	2.743216
Rayan Cherki	AM	Lyon	21.0	1.0	3.514274
Enzo Millot	AM	Stuttgart	22.0	1.0	3.650679
Jamal Musiala	AM	Bayern Munich	21.0	1.0	4.023606
Paul Nebel	AM	Mainz 05	22.0	0.0	4.340916
Tommaso Baldanzi	AM	Roma	21.0	1.0	4.365559
Jude Bellingham	AM	Real Madrid	21.0	1.0	4.380154
Alberto Moleiro	AM	Las Palmas	21.0	1.0	4.401434
Can Uzun	AM	Eint Frankfurt	18.0	1.0	4.421754
Mohamed Nassoh	AM	PSV	22.0	1.0	4.525554

Les résultats montrent plusieurs jeunes milieux offensifs très performants, comme Jamal Musiala, Jude Bellingham et Xavi Simons, déjà au haut niveau dans les grands championnats européens. Le modèle a aussi trouvé Rayan Cherki, un jeune talent de l'Olympique Lyonnais, dont le jeu créatif et percutant ressemble beaucoup à celui de Cole Palmer. Depuis, il a rejoint Manchester City, ce qui confirme que le modèle peut repérer tôt des jeunes joueurs prometteurs.

5.3 Cas 3 – Recherche d'une équipe complète par poste (hors gardien)

Avec la formation 4-3-3, nous avons identifié pour chaque poste les meilleurs jeunes talents afin de construire une équipe type (dream team). Parmi les résultats, on retrouve de grands espoirs déjà repérés par de grands clubs, tels que Michael Olise, Pedri, Harvey Elliott ou Arnau Martínez. Cependant, le modèle met également en évidence des joueurs encore présents dans de plus petits clubs, mais qui pourraient devenir de futures stars si leur niveau continue à progresser.

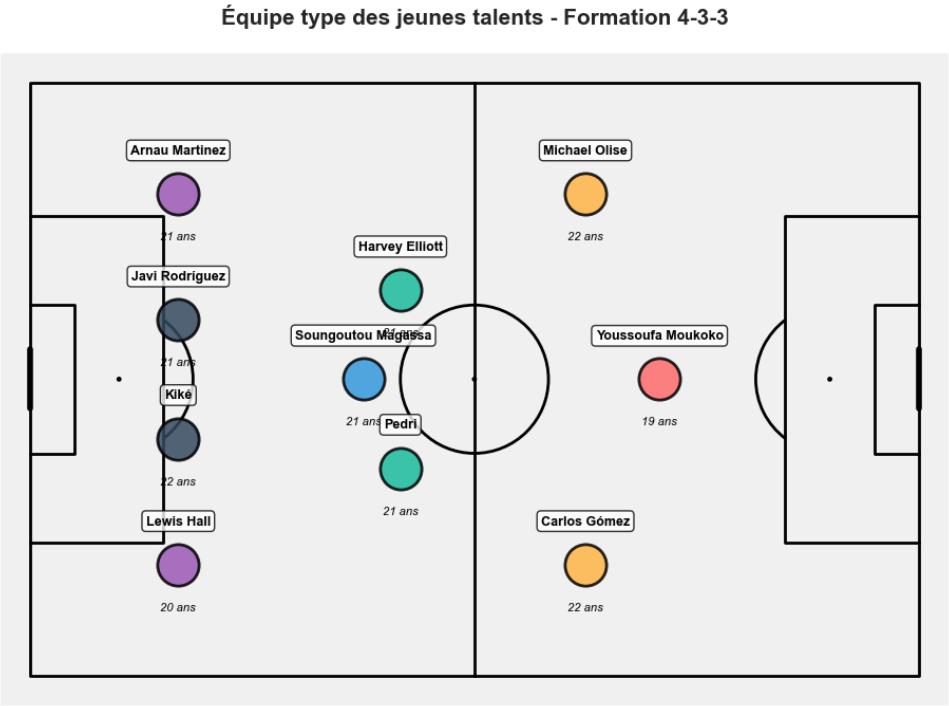

Figure 17. Équipe type des jeunes talents - Formation 4-3-3

5.4 Conclusion de nos résultats

Comme nous avons pu le voir dans les cas précédents, notre modèle de clustering offre de nombreuses applications dans la recherche et la comparaison de profils de joueurs.

- **Identifier des profils atypiques** : par exemple, des défenseurs avec un fort taux de buts ou des milieux défensifs à vocation offensive, selon le style de jeu recherché par l'entraîneur.
- **Remplacer un joueur blessé** par un profil statistiquement proche, capable de reproduire un rôle ou un impact similaire dans l'équipe.
- **Trouver des joueurs équivalents** en termes de performance et de style, mais moins coûteux en raison de leur âge, popularité ou championnat, ce qui constitue un atout pour la rentabilité du recrutement.
- **Détecter de futurs talents** avant leur explosion médiatique, afin de repérer précocement les stars de demain avant qu'elles ne soient valorisées par d'autres clubs.

Ainsi, le clustering s'impose comme un outil puissant d'aide à la décision, capable d'orienter le recrutement, la politique de transfert et la gestion stratégique des effectifs.

6 Conclusion

Cette étude de clustering, réalisée sur les joueurs de la saison 2024/2025, a permis de regrouper les footballeurs selon leurs profils statistiques et de mettre en évidence des styles de jeu distincts à partir des données de performance. Après normalisation et réduction de dimension (PCA), les joueurs ont été répartis en catégories tactiques cohérentes : défenseurs centraux, latéraux, milieux défensifs, milieux centraux, milieux offensifs, ailiers et attaquants. La comparaison entre clusters a révélé des profils contrastés :

- les **attaquants d'élite** se distinguent par leurs valeurs élevées de npxG et de tirs cadrés ;
- les **ailiers** par leurs dribbles réussis et leurs progressions avec ballon ;
- les **milieux offensifs** par leurs passes clés et xAG élevés ;
- tandis que les **défenseurs centraux** dominent sur les tacles, interceptions et duels aériens.

Parmi les trois méthodes testées — K-Means, DBSCAN et clustering hiérarchique —, le K-Means s'est montré le plus adapté. Il a permis une segmentation claire et stable, notamment en révélant des sous-groupes distincts ("élite" vs "standard") au sein de chaque poste. À l'inverse, DBSCAN a montré des difficultés liées à la densité variable entre postes, et le clustering hiérarchique produisait des regroupements moins précis.

Les applications concrètes — recherche de joueurs similaires (Mbappé, Palmer), détection de jeunes talents — ont démontré la pertinence du modèle. Le clustering a su capturer des caractéristiques implicites comme la vitesse, la créativité ou l'impact offensif, bien qu'elles ne soient pas directement incluses dans les données. Ainsi, le K-Means s'impose comme la méthode la plus efficace et lisible pour ce type d'analyse. Ce travail montre le potentiel du machine learning dans le football moderne, où les décisions de recrutement peuvent s'appuyer sur des données objectives. L'ajout futur d'informations contextuelles (trajectoires GPS, intensité, positionnement) et de modèles prédictifs pourrait encore renforcer la précision et la valeur opérationnelle du système.