Partie 2 – Clustering

Shaina Boutebba

Novembre 2025

Introduction

Cette seconde partie du projet s'inscrit dans la continuité de la première, où un travail de nettoyage, d'analyse exploratoire et de prétraitement des données issues du jeu **FIFA** a été réalisé. L'objectif principal est ici de **segmenter les joueurs de champ** à l'aide de différentes méthodes de **clustering non supervisé** afin d'identifier des groupes homogènes selon leurs caractéristiques techniques et physiques.

Plus précisément, cette étude met en œuvre les algorithmes **K-means**, **DBSCAN** et **Ward**, tout en évaluant la qualité des regroupements à l'aide d'indices de validité tels que la silhouette, le Calinski–Harabasz et le Davies–Bouldin. Enfin, un cas pratique illustre l'application du clustering dans un contexte de recrutement, à travers la recherche d'un joueur statistiquement proche de Kylian Mbappé.

1 Préliminaire

Avant l'application des algorithmes de clustering, un travail rigoureux de **prétraitement des données** a été effectué afin de garantir la qualité, la cohérence et la pertinence statistique des informations exploitées.

Sélection et nettoyage des données

Conformément à la méthodologie définie, les **gardiens de but (G)** ont été exclus du jeu de données. Leur profil particulier (statistiques propres) aurait pu perturber l'analyse de similarité avec les joueurs de champ. Les joueurs étudiés appartiennent donc aux catégories *DEF, MID, FWD* et *SUB*.

Seules les **variables numériques pertinentes** ont été conservées. Les attributs fortement dépendants du marché (*Player Value*, *Wage*) ou des évaluations globales (*Overall Rating*) ont été exclus pour éviter tout biais. L'objectif était de fonder les regroupements uniquement sur les performances intrinsèques du joueur.

Les variables catégorielles ont été encodées :

— **Encodage ordinal** pour les catégories de salaire et de valeur (*wage_cat, value_cat*) ;
— **Encodage one-hot** pour les positions en club et en équipe nationale.

Analyse et traitement des valeurs manquantes

Une inspection initiale a montré que certaines variables contenaient jusqu'à 80 % de valeurs manquantes. Celles-ci ont été supprimées pour éviter des imputations trop spéculatives.

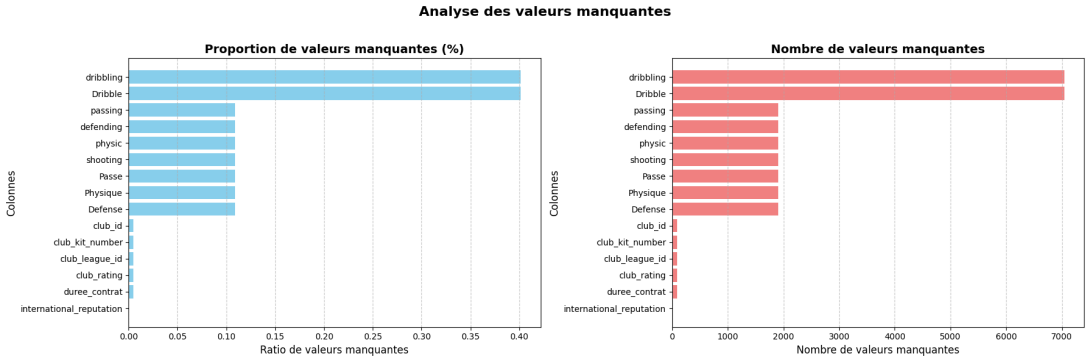

FIGURE 1 – Analyse des valeurs manquantes

Les variables restantes (*dribbling, passing, shooting, defending, physic*) ont été imputées selon deux approches :
— **Régression polynomiale** pour les variables fortement corrélées, permettant d'estimer les valeurs manquantes tout en respectant les relations entre attributs ;
— **Imputation par la médiane** pour les variables à faible corrélation, notamment *dribbling*, afin d'éviter une propagation d'erreur.

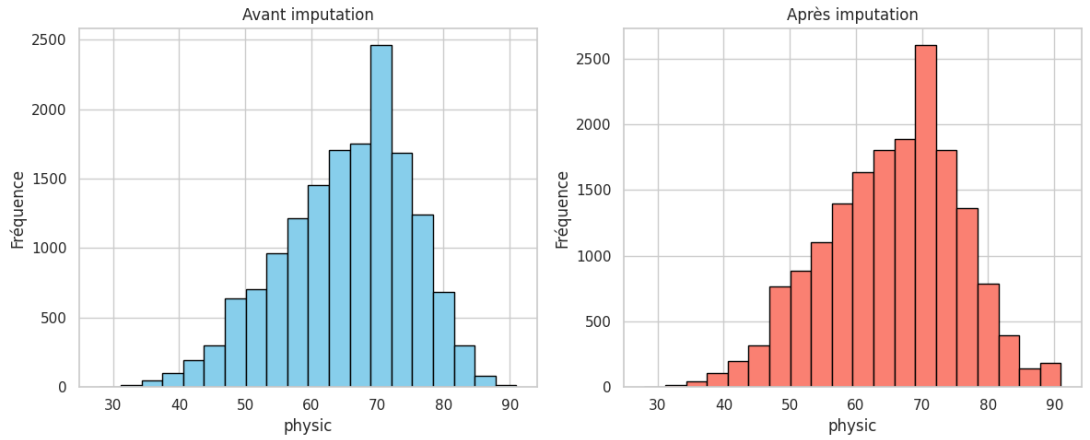

FIGURE 2 – Comparaison avant/après imputation de la variable *physic*

Cette stratégie mixte a permis de préserver la distribution initiale tout en améliorant la complétude du jeu de données.

Normalisation des variables

Avant le clustering, toutes les variables ont été **normalisées par StandardScaler**. Cette étape est essentielle pour donner à chaque variable une influence équivalente dans le calcul des distances euclidiennes utilisées par la majorité des algorithmes.

2 Partie 2 – Clustering

2.1 K-means

L'algorithme du K-means a été appliqué après normalisation, avec un maximum de 20 exécutions pour garantir la stabilité des résultats et éviter les minima locaux.

Méthode du coude – Solution C_1

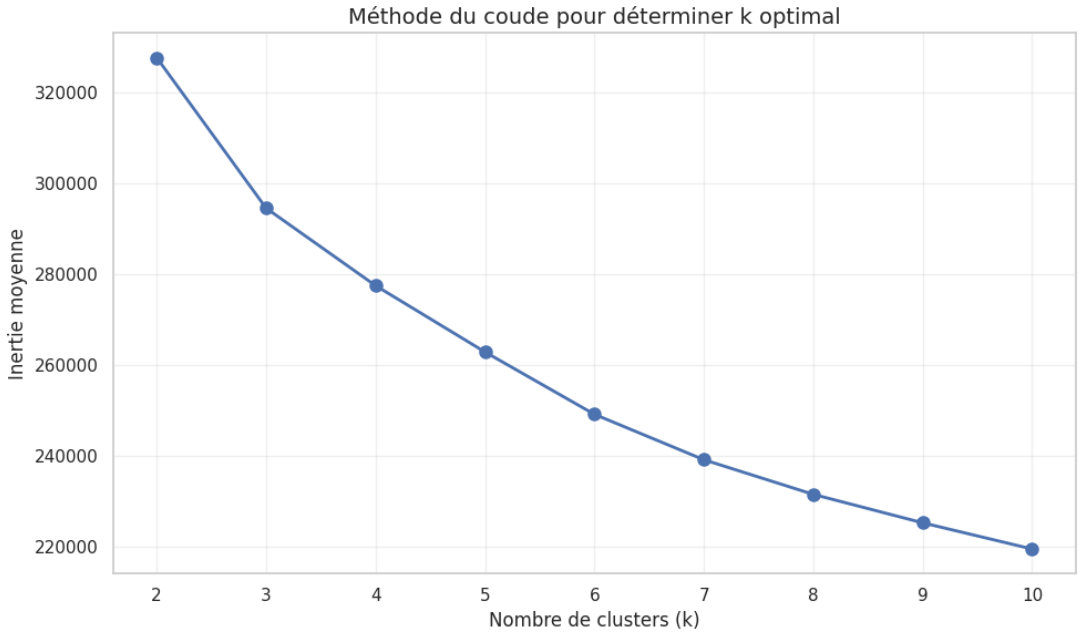

FIGURE 3 – Méthode du coude – Inertie moyenne selon k

Le coude du graphe "apparaît" pour $k = 3$, indiquant un bon compromis entre homogénéité intra-cluster et séparation inter-cluster. Cette configuration donne la solution C_1 composée de trois groupes équilibrés.

TABLE 1 – Résultats du clustering C_1 (K-means, méthode du coude)

Cluster	Nombre de joueurs	Positions dominantes	Caractéristiques principales
0	4827	SUB, MID, DEF	Valeur modérée, forte défense
1	7260	SUB, DEF, MID	Tir et attaque supérieurs
2	5368	MID, FWD	Potentiel et passes élevés, joueurs d'élite

Indice de silhouette – Solution C_2

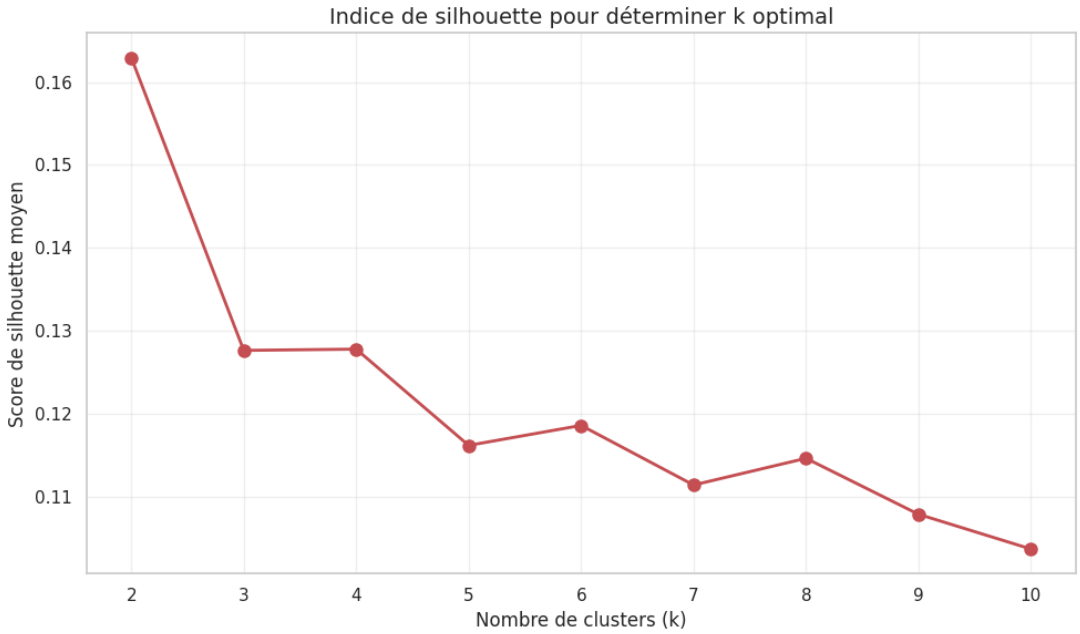

FIGURE 4 – Indice de silhouette moyen en fonction du nombre de clusters

Le score maximal est obtenu pour $k = 2$ (silhouette = 0.162), mais une valeur $k = 4$ (score environ 0.128) offre une segmentation plus fine, permettant d'identifier un cluster d'élite.

TABLE 2 – Résultats du clustering C_2 (K-means, indice de silhouette)

Cluster	**Nombre de joueurs**	**Positions dominantes**	**Caractéristiques principales**
0	6623	SUB, DEF	Finition et tir solides
1	4622	DEF, MID	Joueurs équilibrés défensifs
2	790	MID, FWD	Superstars internationales, valeur/salaire élevés
3	5420	MID, SUB	Jeunes talents en progression

Ces résultats montrent que C_2 affine C_1 en isolant un segment d'élite représentant 4.5 % des joueurs.

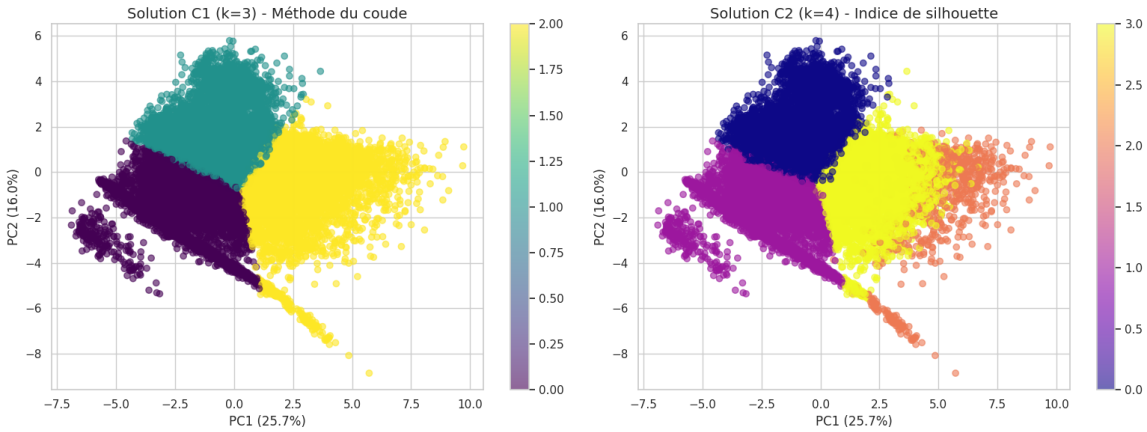

FIGURE 5 – Visualisation PCA des solutions C_1 et C_2

Remplaçant de Kylian Mbappé

La distance euclidienne normalisée indique que le joueur le plus similaire à *Kylian Mbappé* est :

Jude Victor William Bellingham (MID) — Distance = 3.607

Ce résultat est cohérent avec son appartenance au cluster d'élite (C_2), confirmant sa proximité statistique avec les joueurs de haut niveau.

2.2 Autres algorithmes de clustering

DBSCAN

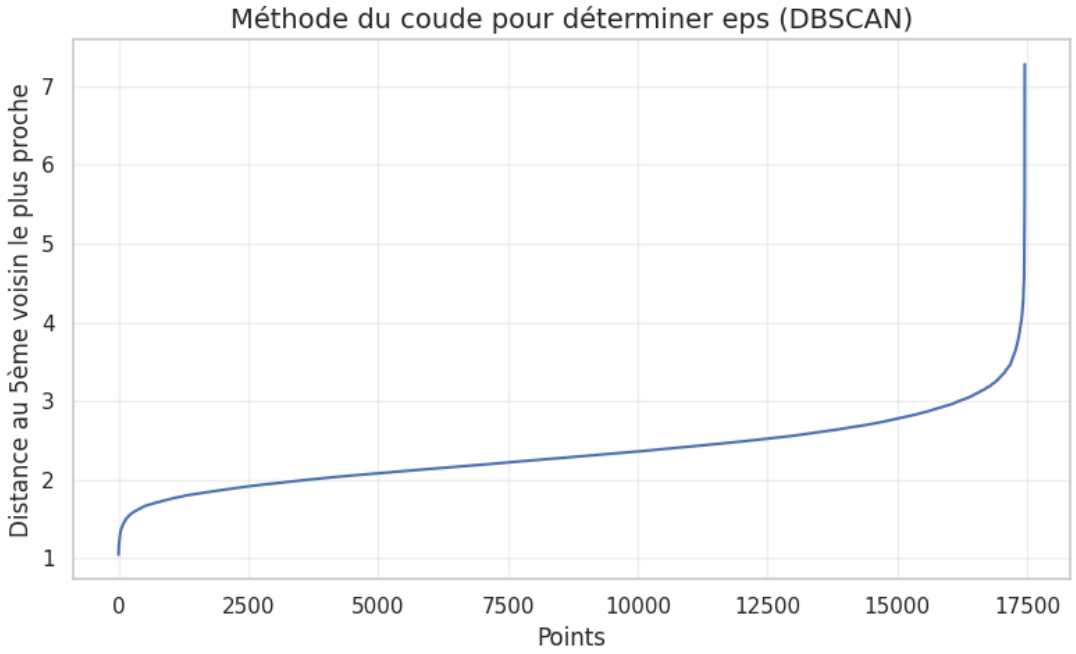

FIGURE 6 – Méthode du coude pour la sélection de ε – DBSCAN

La cassure observée vers $\varepsilon = 3.3$ indique la frontière entre les zones denses et les points isolés.

TABLE 3 – Résultats du DBSCAN selon différents ε

ε	Clusters	Points bruit	CH	DB	Silhouette
3.0	5	458	228.75	1.296	0.162
3.1	4	329	101.99	1.246	0.291
3.3	4	171	103.20	1.170	**0.392**
3.4	4	120	112.61	1.086	0.392
3.5	2	99	216.25	0.774	0.417

Le meilleur compromis se situe pour $\varepsilon = 3.4$, qui conserve 4 clusters distincts avec une structure équilibrée. Le score silhouette atteint 0.39, supérieur à celui du K-means.

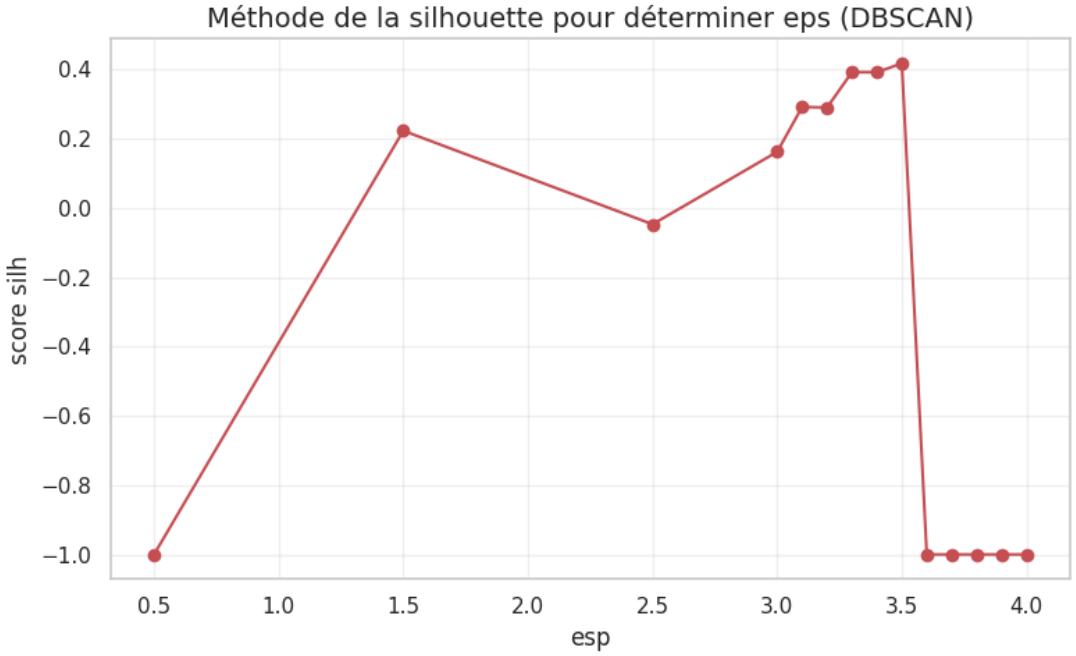

FIGURE 7 – Score de silhouette selon ε – DBSCAN

Méthode de Ward

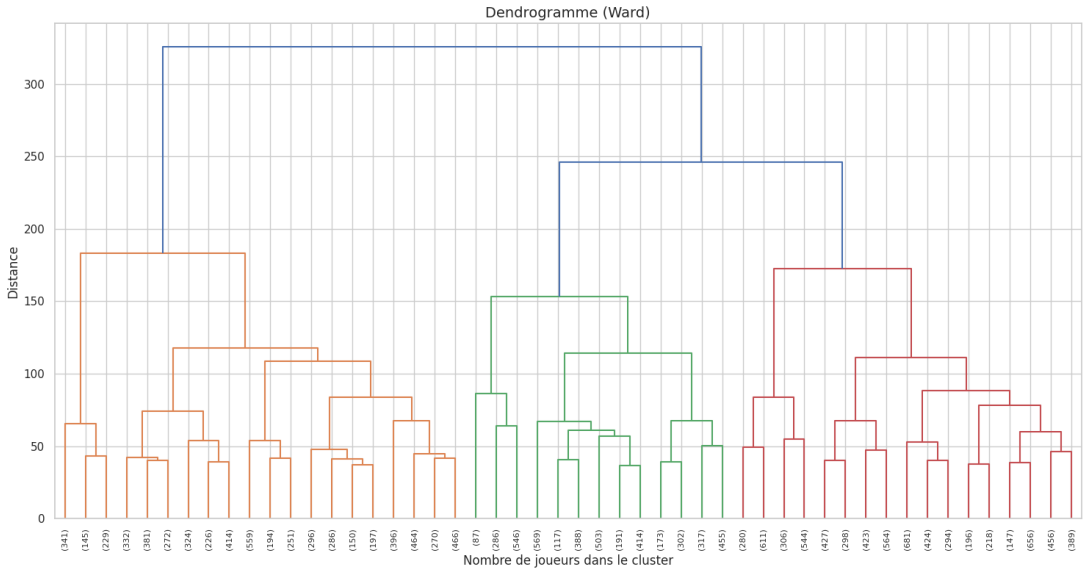

FIGURE 8 – Dendrogramme – Méthode de Ward

Nous pouvons voir un saut net dans la hauteur des fusions entre environ 200 et 300 sur l'axe des distances. Ce grand écart vertical suggère qu'en coupant le dendrogramme à ce niveau (autour de 240 de distance), on obtient trois grands clusters principaux visibles par les trois ensembles de couleurs (orange, vert, rouge). En dessous de ce seuil, les branches sont plus courtes, indiquant des regroupements cohérents et proches.

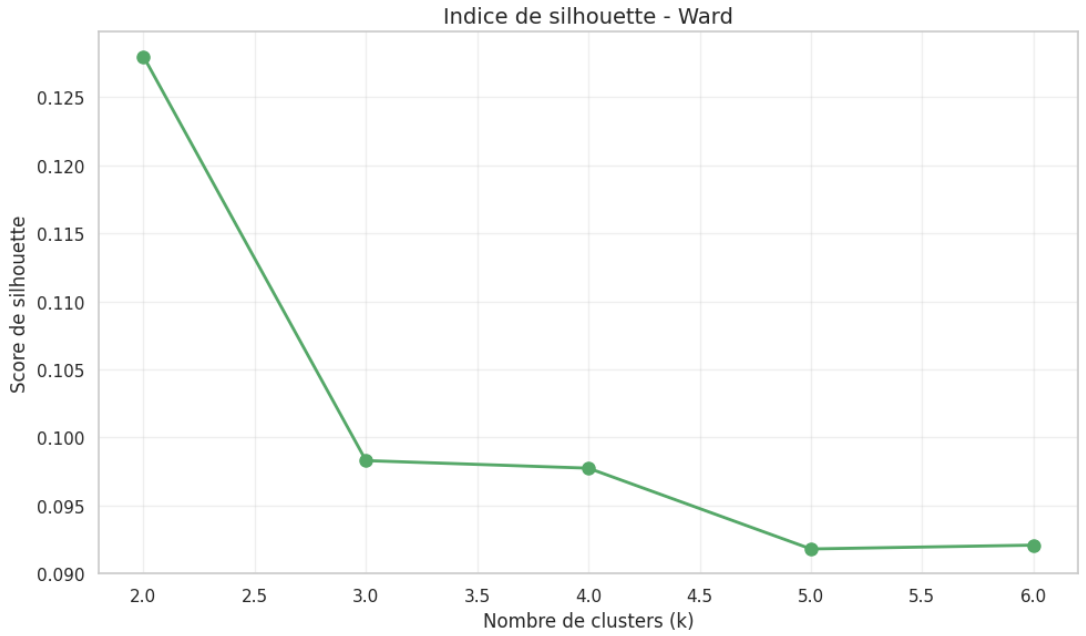

FIGURE 9 – Indice de silhouette pour différents k – Ward

TABLE 4 – Comparaison des performances des algorithmes de clustering

Algorithme	**Nb Clusters**	**Silhouette**	**Observation**
K-means (coude)	3	0.1276	Groupes équilibrés
K-means (silhouette)	4	0.1301	Cluster élite isolé
DBSCAN	4	**0.3916**	120 points bruit
Ward	3	0.1280	Structure hiérarchique claire

Résumé : D'après le résultat précédent, **le DBSCAN s'impose comme l'algorithme le plus performant** en termes de qualité de regroupement, tandis que le **K-means demeure le plus interprétable**, et la **méthode de Ward** vient corroborer la validité de la structure à trois groupes observée dans les résultats précédents.

2.3 Indices de validité du clustering

Trois indices ont été utilisés pour évaluer la qualité des partitions :

— **Silhouette Coefficient** : évalue la séparation et la compacité des clusters ;

— **Calinski–Harabasz Index (CH)** : mesure la dispersion inter- vs intra-cluster ;

— **Davies–Bouldin Index (DB)** : quantifie la similarité moyenne entre clusters.

TABLE 5 – Indices de validité du clustering – Méthode K-means

Nombre de clusters (k)	**Calinski–Harabasz (CH)**	**Davies–Bouldin (DB)**	**Silhouette (Sil)**
2	**3548.01**	2.0960	**0.1629**
3	2956.96	2.1304	0.1276
4	2450.48	**2.0352**	0.1278
5	2182.87	2.0141	0.1162
6	2035.02	2.0432	0.1186
7	1888.60	2.0780	0.1114
8	1754.27	2.0726	0.1146
9	1638.44	2.1254	0.1078

Les indices de validité montrent que les meilleures valeurs de **CH (3548.01)** et de **Silhouette (0.1629)** sont atteintes pour $k = 2$, indiquant une structure à deux grands groupes compacts et bien séparés. Cependant, cette segmentation reste trop grossière pour capturer la diversité réelle des profils de joueurs. Le **minimum de DB (2.0352)** est obtenu à $k = 4$, traduisant une amélioration de la compacité interne et une segmentation plus équilibrée. Ainsi, bien que $k = 2$ optimise les indices globaux, la solution à $k = 4$ offre le meilleur compromis entre performance statistique et interprétabilité des résultats, en distinguant notamment un groupe d'élite bien défini.

TABLE 6 – Indices de validité du clustering – Méthode hiérarchique de Ward

Nombre de clusters (k)	**Calinski–Harabasz (CH)**	**Davies–Bouldin (DB)**	**Silhouette (Sil)**
2	**2706.85**	2.3396	**0.1280**
3	2333.63	2.3595	0.0983
4	1974.62	**2.2039**	0.0977
5	1791.81	2.2886	0.0918
6	1648.47	2.2925	0.0921
7	1487.46	2.5562	0.0672
8	1371.34	2.5792	0.0605
9	1283.80	2.5970	0.0579

La méthode hiérarchique de Ward présente les meilleures valeurs du **Calinski–Harabasz (2706.85)** et du **Silhouette (0.1280)** pour $k = 2$, ce qui indique une séparation nette entre deux grands groupes de joueurs. Cependant, le **score minimal du Davies–Bouldin (2.2039)** est atteint à $k = 4$, suggérant une compacité interne supérieure et une structure plus fine des regroupements. Ainsi, bien que $k = 2$ reste optimal selon les indices globaux, la configuration à $k = 4$ offre un **meilleur équilibre entre cohésion et interprétation**, permettant d'identifier des segments plus différenciés de joueurs (défensifs, offensifs, et d'élite).

Les deux méthodes confirment une tendance similaire : les meilleurs indices (CH et Silhouette) sont obtenus pour un faible nombre de clusters ($k = 2$), traduisant une structure globale claire mais peu détaillée. En revanche, les solutions à $k = 3$ ou $k = 4$ offrent un **meilleur équilibre entre la qualité statistique et l'interprétation des profils**, permettant de distinguer plusieurs niveaux de performance chez les joueurs. Ainsi, le **K-means à 4 clusters** apparaît comme la solution la plus pertinente pour une analyse à la fois robuste et explicative.

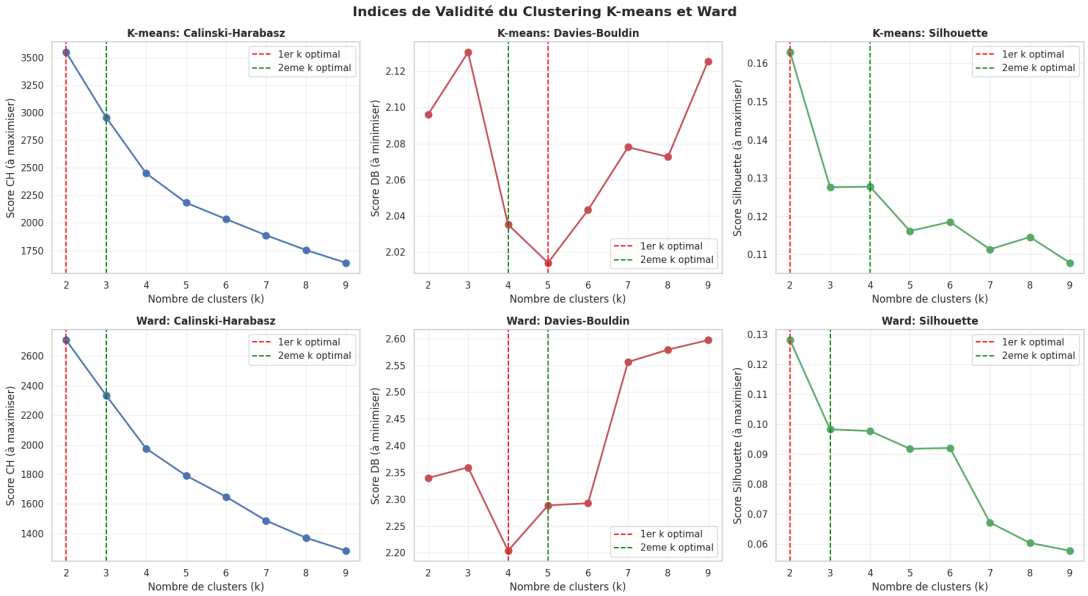

FIGURE 10 – Indices de validité du clustering K-means et Ward

TABLE 7 – Indices de validité du clustering – Algorithme DBSCAN (corrigé)

ε	min_samples	Clusters	Bruit	CH	DB	Silhouette
2.5	3	34	2184	25.63	1.4639	-0.1923
2.5	5	17	2615	28.01	1.3309	-0.0471
2.5	10	2	3577	2.70	1.0814	0.0775
3.0	3	7	397	167.72	1.4680	0.1705
3.0	5	5	458	228.75	1.2965	0.1621
3.0	10	4	601	**256.27**	1.2947	0.2715
3.1	3	6	283	73.90	1.5148	0.2628
3.1	5	4	329	101.99	1.2464	0.2913
3.1	10	3	417	113.20	1.3574	0.2968
3.2	3	4	218	130.22	1.2959	0.2899
3.2	5	4	239	117.60	1.2786	0.2898
3.2	10	3	302	127.56	1.4116	0.2933
3.3	3	3	159	165.81	1.1704	0.3975
3.3	5	4	171	103.20	1.1704	0.3920
3.3	10	2	224	202.16	0.7587	**0.4186**
3.4	3	3	115	177.28	0.7684	0.3976
3.4	5	4	120	112.61	1.0862	0.3916
3.4	10	3	153	131.59	**0.7251**	0.3972
3.5	3	2	91	215.71	0.7742	0.4169
3.5	5	2	99	216.25	0.7737	0.4173
3.5	10	3	119	142.46	0.7396	0.3958

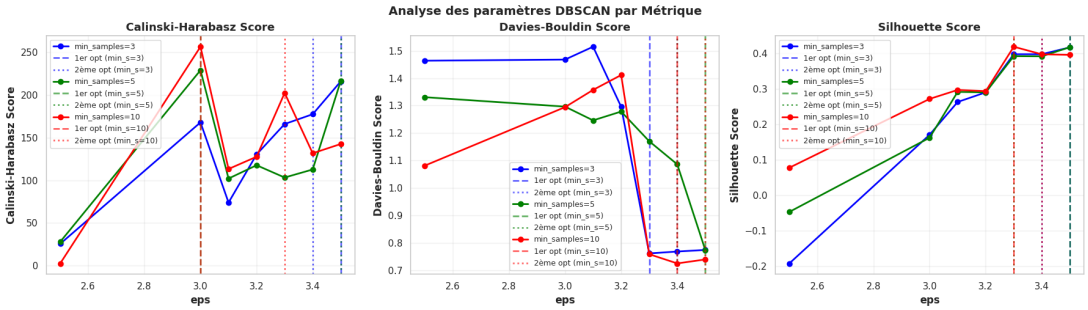

FIGURE 11 – Indices de validité du clustering DBSCAN

L'analyse des indices de validité pour l'algorithme **DBSCAN** met en évidence trois configurations selon les critères retenus :

- Le **score de silhouette maximal** atteint **0.4186** pour $\boldsymbol{\varepsilon = 3.3}$ et $\boldsymbol{min\_samples = 10}$, indiquant une excellente cohésion intra-cluster et une séparation assez nette entre groupes.
- L'**indice de Calinski–Harabasz (CH)** atteint sa valeur la plus élevée (**256.27**) pour $\boldsymbol{\varepsilon = 3.0}$ et $\boldsymbol{min\_samples = 10}$, traduisant une forte dispersion inter-groupes.
- Le **Davies–Bouldin (DB)** présente sa valeur minimale (**0.7251**) pour $\boldsymbol{\varepsilon = 3.4}$ et $\boldsymbol{min\_samples = 10}$, confirmant une bonne compacité des clusters pour cette configuration.

Ces résultats montrent que, selon le critère choisi, la configuration optimale diffère légèrement. Le maximum du CH privilégie une séparation marquée mais au prix d'un nombre plus important de points considérés comme bruit, tandis que la meilleure silhouette ($\varepsilon = 3.3$, $min\_samples = 10$) offre le meilleur équilibre entre **compacité interne** et **clarté structurelle**, ce qui en fait la configuration la plus cohérente d'un point de vue statistique.

Cependant, dans une optique d'interprétation pratique des profils de joueurs, la configuration $\boldsymbol{\varepsilon = 3.4}$ et $\boldsymbol{min\_samples = 5}$ s'avère être le **meilleur compromis** : elle conserve quatre clusters distincts, une proportion modérée de bruit (environ 120 points), et un score de silhouette encore élevé (**0.3916**). Cette solution permet de segmenter les joueurs en groupes cohérents et exploitables pour une analyse qualitative, tout en limitant la perte d'information liée à l'exclusion des outliers.

Conclusion

L'algorithme DBSCAN se distingue nettement du *K-means* et de la méthode de *Ward* par sa capacité à **détecter des structures non sphériques** et à **isoler les profils atypiques**. Les résultats empiriques confirment sa robustesse : la configuration optimale ($\varepsilon = 3.3$, $min\_samples = 10$) atteint un **score de silhouette de 0.4186**, le plus élevé parmi l'ensemble des modèles testés, tandis que les indices CH et DB affichent également

des valeurs satisfaisantes, témoignant d'une bonne séparation entre les groupes.

En pratique, la configuration légèrement ajustée ($\varepsilon = 3.4$, $min\_samples = 5$) offre une segmentation plus stable et plus interprétable, adaptée à des applications concrètes telles que l'analyse de performance ou le recrutement. Ainsi, le **DBSCAN** s'impose comme l'algorithme le plus performant et le plus pertinent pour représenter la diversité réelle des profils de joueurs FIFA, distinguant clairement les **joueurs d'élite**, les **profils intermédiaires** et les **joueurs atypiques**.