Analyse Approfondie des Performances des Joueurs de Football (2013-2021)

Fossi Cedric

20 Mai 2024

Contents

1 Introduction

1.1 Contexte

Le football est l'un des sports les plus populaires au monde, et les analyses statistiques jouent un rôle crucial dans la compréhension et l'amélioration des performances des joueurs et des équipes. Les données quantitatives sur les performances des joueurs permettent aux entraîneurs, analystes, et clubs de prendre des décisions éclairées sur les tactiques, les stratégies de jeu, et le recrutement.

1.2 Objectifs

L'objectif principal de cette étude est de déterminer les tendances dans les performances des joueurs, en se concentrant spécifiquement sur les participants des cinq plus grands championnats d'Europe. Nous cherchons également à découvrir si les caractéristiques des joueurs peuvent prédire le championnat auquel ils appartiennent.

1.3 Base de Données

Les données utilisées dans cette étude proviennent de Kaggle et couvrent les années 2013 à 2021. Elles comprennent des statistiques détaillées sur les performances des joueurs à travers diverses saisons et tournois.

2 Prérequis pour Débutants en Football

2.1 Termes Clés

- **Tournoi :** Compétition organisée où plusieurs équipes jouent des matchs pour remporter un titre ou un trophée.
- **Saison :** Période pendant laquelle une série officielle de matchs de football est jouée, généralement d'été à printemps.
- **But :** Nombre de fois qu'un joueur a marqué pendant un match.
- **Passe Décisive :** Passe effectuée par un joueur qui mène directement à un but.
- **Carton Jaune/Rouge :** Pénalités données pour comportement antisportif ou jeu dangereux.
- **Duel :** Confrontation directe entre deux joueurs adverses pour le contrôle du ballon.
- **Surface de Réparation :** Zone rectangulaire située devant chaque but. Les fautes dans cette zone peuvent entraîner un penalty.

2.2 Importance des Statistiques

Les statistiques de joueurs, telles que les buts, passes décisives, et cartons, sont cruciales pour évaluer les performances individuelles et d'équipe. Elles permettent d'identifier les points forts et les faiblesses, d'optimiser les stratégies de jeu, et de prendre des décisions de recrutement basées sur des données objectives.

2.3 Compétitions Européennes

Les cinq plus grands championnats européens, souvent au cœur des analyses de performances dans le football, incluent :

- Premier League (Angleterre)
- La Liga (Espagne)
- Bundesliga (Allemagne)
- Serie A (Italie)
- Ligue 1 (France)

Ces ligues attirent certains des meilleurs talents mondiaux et sont suivies par des millions de fans à travers le monde.

3 Description des Ensembles de Données

3.1 Structure des Données

Les données se composent de quatre fichiers CSV principaux :

- **players_stats_per_season.csv :** Contient des statistiques détaillées par joueur et par saison.
- **player_id_mapping.csv :** Correspondance entre les identifiants des joueurs et leurs noms.
- **season_mapping.csv :** Relation entre les identifiants des saisons et leurs noms.
- **tournaments_mapping.csv :** Relation entre les identifiants des tournois et leurs noms.

3.2 Variables Clés

- **totalRating :** Note cumulative attribuée à un joueur sur une saison ou un tournoi.
- **countRating :** Nombre de fois qu'un joueur a été évalué.

- **appearances :** Nombre total de fois que le joueur est apparu sur le terrain.
- **accuratePasses :** Nombre total de passes réussies.
- **keyPasses :** Passes clés menant à une occasion de but.
- **shotsOnTarget :** Nombre de tirs cadrés.

4 Processus d'Importation et de Nettoyage des Données

4.1 Importation

Les données ont été importées à l'aide du package `dplyr` de R, permettant une manipulation efficace des grands ensembles de données.

4.2 Nettoyage et Fusion des Données

Une fois les données importées, il a été crucial de les fusionner pour créer un ensemble de données complet et cohérent pour l'analyse. Les étapes de nettoyage ont inclus la gestion des valeurs manquantes, l'élimination des doublons, et la conversion des types de données.

4.3 Nettoyage des Données

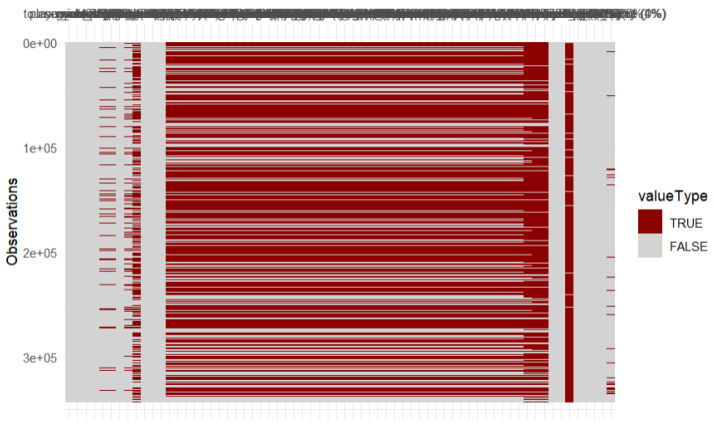


Figure 1: Distribution des valeurs manquantes dans les données

Pour nettoyer nos données, nous avons traité les NAN présents colonne par colonne. Nous avons jugé s'il était intéressant de garder chaque colonne, et si

oui, avec quelle valeur il était intéressant de remplacer les NAN. En analysant les NAN dans chaque colonne, nous avons remarqué que la colonne *rating* comportait beaucoup de NAN.

4.3.1 Méthode Initiale de Nettoyage

À première vue, il semblait plus simple de supprimer entièrement la colonne *rating*. Cependant, nous avons remarqué que lorsqu'il y avait un NAN dans *rating*, il y en avait également dans de nombreuses autres colonnes (fig 3). Nous avons donc compris qu'il était logique que lorsque les joueurs ne sont pas notés, certaines variables ne sont pas remplies. Nous avons alors décidé, dans un premier temps, de supprimer les lignes avec des NAN dans *rating*, estimant que ces observations n'étaient pas intéressantes. Cependant, cette approche nous a fait perdre beaucoup d'informations sur certains joueurs. En effet, même si un joueur n'est pas noté, certaines variables comme les buts ou les passes décisives étaient remplies. Supprimer ces lignes nous faisait donc perdre des données précieuses.

4.3.2 Méthode de Nettoyage Adoptée

Pour remédier à ce problème, nous avons opté pour une autre méthode de nettoyage :

- **Pour les variables de comptage (comme les buts, passes décisives, etc.) :** Nous avons remplacé les valeurs manquantes par 0, car une valeur manquante dans ces colonnes indique généralement qu'un événement (comme marquer un but) n'a pas eu lieu.
- **Pour les variables de pourcentage et de mesure (comme le pourcentage de passes réussies, etc.) :** Nous avons imputé les valeurs manquantes par la moyenne de ces variables pour chaque joueur, car un 0 serait trompeur. On à du cependant faire attention à remplacer les NAN par la moyenne du joueur en question.
- **Pour les variables ratings et colonnes affectées par rating :** Nous avons opté pour une imputation par la moyenne générale du joueur. Les valeurs NAN des variables de comptage ont été remplacées, cette fois-ci, non pas par 0 mais par leur moyenne, pour que cela soit cohérent avec leurs ratings.

Ensuite, après avoir fait tous ces traitements, il restait quelques rares joueurs qui avaient toujours des NAN dans *rating*. Ceux-là étaient les joueurs qui n'ont jamais été notés. Très souvent, c'étaient des joueurs qui ont joué que très peu de matchs. Nous avons donc supprimé ces lignes ainsi que les joueurs avec.

4.3.3 Contradiction des Données

Nous avons observé une contradiction entre les données des colonnes *minutesPlayed* et *goals*. Certains joueurs ont en effet enregistré des buts sans avoir

de minutes de jeu comptabilisées, ce qui n'est pas cohérent. Nous avons donc décidé de supprimer la colonne *minutesPlayed* ainsi que *scoringFrequency*, cette dernière étant calculée à partir des minutes jouées. Nous avions *Appearances*, qui fournissait la même information que *scoringFrequency*.

5 Analyse Exploratoire des Données

5.1 Statistiques Descriptives

5.1.1 Distribution des Performances (figure 4)

1. **Buts et Passes Décisives :**
   - La distribution des buts montre un maximum extrême de 232 buts, ce qui indique probablement la présence de quelques joueurs exceptionnels, tandis que la médiane à 1 suggère que la majorité des joueurs marquent peu, ce qui est finalement normal car notre jeu contient des joueurs de tous les postes (attaque, défense, milieu et gardien) ; donc pour la plupart, leur rôle n'est pas de marquer des buts.
   - Pour les passes décisives, le maximum est à 98, indiquant là aussi quelques joueurs très influents, tandis que la plupart des joueurs (médiane à 1) contribuent peu en termes de passes décisives, ce qui aussi ici est logique car les joueurs défensifs ne sont pas concernés par les passes décisives.
2. **Cartons Jaunes :**
   - Une moyenne de près de 8 cartons jaunes et une médiane à 4 montrent une répartition assez large, ce qui pourrait refléter les différents styles de jeu ou la discipline sur le terrain.
3. **Précision et Création de Jeu :**
   - **Passes Totales et Pourcentage de Passes Réussies :** Des valeurs élevées en passes totales et un pourcentage moyen de passes réussies à 75% montrent une forte tendance à la possession et à la distribution efficace du ballon, montrant que ce sport est vraiment un sport en équipe.
   - **Créations de Grosses Chances :** Une moyenne de 5,15 grandes occasions créées, mais avec un maximum à 186,89, montre que certains joueurs sont des créateurs clés, démontrant une capacité à ouvrir les défenses adverses.
4. **Défense et Tacles :**
   - **Interceptions :** Une moyenne de 50 interceptions par joueur, avec un maximum à 550, suggère que certains joueurs ont un rôle défensif très actif.

On peut donc voir qu'il y a une grande variabilité dans presque toutes les métriques, ce qui suggère des différences significatives dans les rôles des joueurs, leurs compétences et leur efficacité.

5.2 Corrélations

L'analyse des corrélations entre différentes variables est essentielle pour comprendre les relations et les interdépendances entre diverses mesures de performance des joueurs (fig 5 et 6). Voici quelques-unes des principales observations issues de notre analyse des corrélations :

- **Buts et Passes Décisives :** Il existe une corrélation positive significative entre le nombre de buts marqués et le nombre de passes décisives. Les joueurs qui marquent beaucoup ont également tendance à créer des occasions de but pour leurs coéquipiers.
- **Minutes Jouées et Apparitions :** La corrélation entre le nombre de minutes jouées et le nombre d'apparitions est très élevée, ce qui est attendu, car plus un joueur participe aux matchs, plus il accumule de temps de jeu.
- **Passes Précises et Passes Clés :** Il y a une forte corrélation entre le pourcentage de passes précises et le nombre de passes clés. Cela indique que les joueurs qui sont précis dans leurs passes sont également efficaces pour créer des occasions de but.
- **Interceptions et Tacles :** Les données montrent une corrélation positive entre le nombre d'interceptions et le nombre de tacles réussis. Les joueurs qui excellent dans la récupération du ballon tendent également à être actifs dans les tacles.
- **Cartons Jaunes et Fautes :** Comme prévu, il existe une forte corrélation entre le nombre de fautes commises et le nombre de cartons jaunes reçus. Les joueurs qui commettent plus de fautes ont plus de chances de recevoir des cartons jaunes.
- **Offsides et Buts :** Il y a une corrélation intéressante entre le nombre de hors-jeu (offsides) et le nombre de buts, indiquant que les attaquants qui prennent des risques pour marquer des buts se retrouvent souvent en position de hors-jeu.
- **Total Duels Gagnés et Fautes :** Une corrélation notable entre le total de duels gagnés et les fautes commises montre que les joueurs actifs dans les duels physiques sur le terrain ont également tendance à commettre plus de fautes.
- **Total Contest et Variables d'Attaque :** Il existe une corrélation entre le nombre total de duels disputés (total contest) et diverses variables

d'attaque telles que les dribbles réussis. Cela montre que ce sont souvent les joueurs offensives qui contest les decisions des arbitres.

Ces corrélations nous aident à identifier les attributs qui tendent à se manifester ensemble. Par exemple, un attaquant performant est non seulement un bon buteur mais aussi un créateur d'occasions. De même, un bon défenseur combine souvent des compétences en tacle et en interception. Comprendre ces relations peut guider les entraîneurs dans l'optimisation des tactiques et des formations en fonction des points forts des joueurs.

6 Techniques de Clustering et Classification

6.1 Clustering Non Supervisé (CAH)

6.1.1 Choix du Nombre de Clusters

Pour déterminer le nombre optimal de clusters, nous nous sommes basés principalement sur deux critères :

- **Longueur des Branches du Dendrogramme :** En observant le dendrogramme, nous avons noté que conserver 5 ou 6 clusters était raisonnable car au-delà de ce nombre, les hauteurs des branches deviennent trop grandes.
- **Barplot des Hauteurs :** Nous avons également utilisé un barplot pour identifier le moment où il y a une perte considérable d'inertie inter-classe. Une grande perte d'inertie est observée lorsque l'on passe de 2 à 1 cluster, ce qui rend le choix d'un seul cluster peu intéressant. Les pertes deviennent légères lorsque l'on passe de 5 à 4 clusters, et insignifiantes de 6 à 5 clusters, justifiant ainsi notre choix de 5 clusters (fig 7).

6.1.2 Méthode Utilisée

Nous avons utilisé la méthode de Ward pour effectuer un clustering hiérarchique agglomératif (CAH). Les données ont été centrées et réduites pour accorder la même importance à chaque variable.

6.1.3 Interprétation des Groupes

```
Means_groupes <- matrix(NA, nrow=K, ncol=dim(joueur_quantitative.cr)[2])
colnames(Means_groupes) = colnames(joueur_quantitative.cr)
rownames(Means_groupes) = 1:K
for (i in 1:K) {
    Means_groupes[i,] <- colMeans(joueur_quantitative.cr[groupes.cah==i,])
}
round(Means_groupes)
```

Pour interpréter les groupes, nous avons d'abord utilisé la moyenne des caractéristiques de chaque groupe. Ensuite, nous avons réalisé une ACP pour mieux analyser nos clusters (fig 8). Enfin, nous avons examiné les joueurs de chaque groupe pour vérifier la cohérence avec la réalité.

- **Groupe 1 :** Joueurs remplaçants avec peu d'apparitions et de contributions.
- **Groupe 2 :** Joueurs polyvalents apportant de la valeur de manière équilibrée.
- **Groupe 3 :** Gardiens de but avec des statistiques spécifiques à ce poste.
- **Groupe 4 :** Milieux de terrain et défenseurs participant offensivement et défensivement.
- **Groupe 5 :** Attaquants et milieux offensifs avec des contributions élevées en buts et passes décisives.

6.2 Classification Supervisée (CART)

6.2.1 Prédiction des Championnats

Malgré nos efforts, nous n'avons pas réussi à prédire le championnat des joueurs. L'exactitude de notre modèle était d'à peine un peu plus de 0.2 (fig 9), ce qui suggère que sa performance était très faible, comparable à une sélection aléatoire. En d'autres termes, choisir au hasard le championnat d'appartenance d'un joueur aurait abouti à des résultats presque similaires. Nous avons quand même pu constater des choses pertinentes dans ces championnats (sorties R en annexes):

- **Buts et Passes Décisives:** La Premier League est le championnat avec le plus grand nombre de buts inscrits et aussi le plus de passes décisives. On peut donc dire qu'en attaque, ils détiennent les meilleurs joueurs. En revanche, la Ligue 1 est le championnat le plus bas dans ces catégories.
- **Dribbles Réussis:** Les joueurs de la Premier League ont les meilleures performances en dribbles réussis, suivis de près par LaLiga.
- **Interceptions:** La Premier League a également le plus grand nombre d'interceptions, ce qui indique une forte capacité défensive.
- **Ratings Moyens:** La Premier League et la Ligue 1 ont les ratings moyens les plus élevés, indiquant des performances globalement meilleures.
- **Duels Gagnés, Tirs Bloqués et Interceptions:** La Premier League et LaLiga ont les meilleures performances en termes de duels gagnés, tirs bloqués et interceptions.

Donc, la Premier League domine dans presque tous les domaines et peut être perçue comme le championnat le plus compétitif et performant. La Serie A et la Bundesliga sont les moins performantes.

En suite, nous avons opté pour une autre étude : prédire si un joueur était une légende, un bon joueur, un joueur moyen ou un remplaçant.

6.2.2 Prédiction de la Catégorie des Joueurs

Nous avons défini nous-mêmes ce qu'est une légende, un bon joueur, etc.

NB: Il est très important que pour les 3 premières catégories, les joueurs n'appartiennent pas au groupe 1, car ce groupe représente les remplaçants, c'est-à-dire ceux qui n'ont presque pas joué de match au cours de ces années soit parce qu'ils étaient nuls, soit parce qu'ils étaient des vétérans. Certains dans ce groupe ont donc par exemple joué 2 ou 3 bons matchs et n'ont plus jamais rejoué. Ils ne peuvent pas être considérés comme des légendes, ou même de bon ou joueurs moyen. Voici les catégories définies :

- **Légende :** Joueur avec une note de plus de 7.4 sur 10 et qui n'appartient pas au groupe 1.
- **Bon Joueur :** Joueur avec une note entre 7 et 7.4 et qui n'appartient pas au groupe 1.
- **Joueur Moyen :** Joueur avec une note inférieure à 7 et qui n'appartient pas au groupe 1.
- **Remplaçant :** Joueur appartenant au groupe 1.

6.2.3 Méthodologie

- **Ajout de la Colonne 'categorie_joueur' :** Basée sur les critères définis.
- **Séparation Train/Test :** Les données ont été divisées en échantillons d'entraînement et de test.
- **Modèle CART :** Entraîné sur les données d'entraînement et testé sur les données de test.

6.2.4 Résultats

- **Accuracy :** Le modèle a montré une très bonne accuracy, validant ainsi notre modèle.

Nous avons également utilisé un modèle Random Forest, qui s'est avéré plus performant que le modèle CART.

7 Résultats et Discussions

7.1 Interprétation des Résultats

Les clusters et classifications offrent des insights sur les caractéristiques distinctives des joueurs dans différents groupes et championnats. Par exemple, les gardiens de but et les attaquants ont des statistiques très différentes qui les rendent facilement identifiables.

7.2 Comparaison avec des Études Antérieures

Nos résultats sont cohérents avec d'autres études connues, qui montrent également la domination de la Premier League dans plusieurs domaines de performance.

7.3 Limitations

Les principales limitations de notre étude incluent la qualité des données disponibles, les valeurs manquantes, et la difficulté à prédire le championnat des joueurs en raison des similitudes entre les ligues.

8 Conclusion et Recommandations

8.1 Synthèse des Principales Trouvailles

L'analyse a révélé des tendances importantes dans les performances des joueurs, permettant de classifier efficacement les joueurs en fonction de leurs statistiques.

8.2 Implications

Les résultats de cette étude peuvent aider les clubs, les entraîneurs et les analystes à mieux comprendre les performances des joueurs et à prendre des décisions éclairées en matière de recrutement et de stratégie de jeu. Ils peuvent egalement aider les clubs et les entraîneurs à mieux comprendre les compétences requises pour exceller dans différents championnats.De plus, la capacité à prédire le niveau d'un joueur peut également aider à évaluer plus précisément la valeur marchande d'un joueur, ce qui est crucial pour les négociations de transfert et la gestion financière des clubs.

8.3 Futurs Axes de Recherche

Un axe de recherche prometteur que j'avais commencé à explorer, mais que je n'ai pas pu approfondir en raison des limitations de stockage de mon ordinateur, était d'étudier les différences entre les joueurs évoluant dans les championnats européens et ceux des championnats hors Europe. Cette analyse nécessiterait une quantité massive de données, bien au-delà de ce que mon système pouvait gérer.

L'objectif de cette recherche serait de déterminer s'il existe des profils distincts de joueurs dans ces différents contextes géographiques. En identifiant ces différences, nous pourrions éventuellement prédire à quel moment un joueur hors Europe pourrait devenir un talent de premier plan en Europe. Une telle prédiction permettrait aux clubs européens de repérer et d'acheter ces joueurs à des prix relativement bas avant qu'ils n'atteignent leur plein potentiel, offrant ainsi une opportunité d'obtenir des pépites avant qu'elles ne deviennent largement reconnues et plus coûteuses. Cette étude pourrait transformer la stratégie de recrutement des clubs européens en élargissant leur horizon de recherche et en optimisant leurs investissements dans les talents internationaux.

9 Problèmes Rencontrés

Ce projet n'a pas toujours été facile à réaliser en raison de divers défis et problèmes rencontrés.

- **Volume de Données :** Au début, j'avais énormément de données, ce qui rendait le clustering non supervisé impossible. J'ai dû réduire les données tout en conservant suffisamment d'informations pertinentes. J'ai essayé de supprimer les joueurs de categorie moins de 21 ans, mais cela n'a pas suffi. Finalement, j'ai décidé de ne garder que les joueurs des cinq grands championnats, ce qui a permis de maintenir un échantillon représentatif.
- **Prédiction des Championnats :** Un autre gros problème a été de ne pas réussir à prédire le championnat de chaque joueur. Les différences entre les championnats ne sont pas suffisamment marquées pour permettre une prédiction précise. Les joueurs changent souvent de championnat au cours de leur carrière, ce qui complique encore la tâche.
- **Perte de Données :** Le plus gros problème a été la perte de l'ensemble des données à cause d'une panne d'ordinateur. Cela m'a servi de leçon importante sur l'importance de sauvegarder régulièrement les travaux dans une plateforme comme GitHub, afin de pouvoir les récupérer même en cas de perte de données.
- **Sélection des Colonnes :** Choisir les colonnes les plus pertinentes à garder a été un défi majeur. J'ai dû faire des choix subjectifs pour réduire les colonnes initiales, en évaluant une par une leur pertinence.

10 Parties Préférées du Projet

- **Analyse en Composantes Principales (ACP) :** Voir apparaître sur les graphes des individus des joueurs que je connais et admire était fascinant. Par exemple, Lionel Messi se distinguait clairement comme une valeur extrême, ce qui était cohérent avec ses performances exceptionnelles.

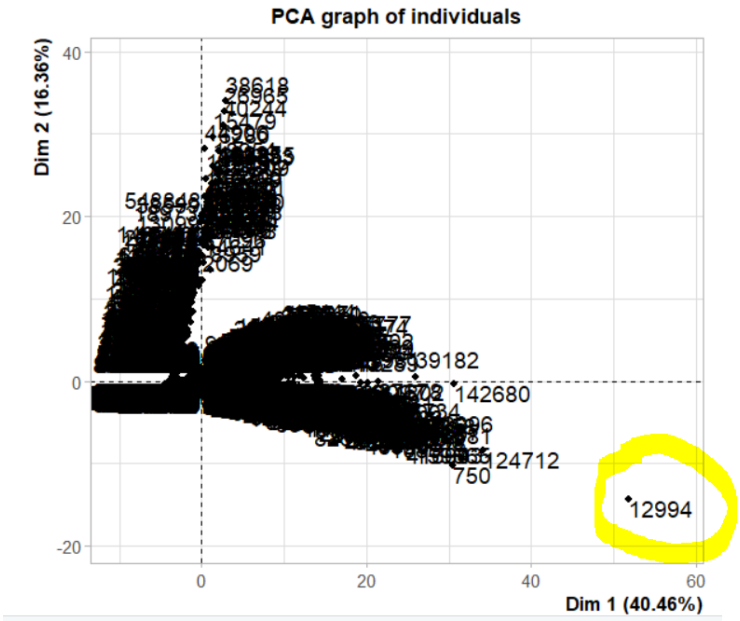


Figure 2: Représentation de Messi parmi les autres joueurs

- **Clustering :** La création et l'interprétation des clusters ont été une partie extrêmement enrichissante et fascinante de ce projet.

  Un moment particulièrement marquant a été lorsque j'ai réalisé neuf clusters distincts. Parmi ces neuf clusters, certains contenaient uniquement les légendes de l'ère 2013-2021, des joueurs qui ont dominé cette période avec des performances exceptionnelles. Il y avait également un groupe qui regroupait des remplaçants, et un autre, que j'ai surnommé les "remplaçants de luxe", composé de joueurs qui, chaque fois qu'ils entraient en jeu, faisaient une différence significative.

  Voici quelques exemples des clusters identifiés :

  - **Légendes de 2013-2021 :** Ce groupe comprenait des joueurs qui ont marqué l'histoire du football par leurs performances remarquables durant cette période. Ces joueurs étaient les piliers de leurs équipes et étaient souvent ceux qui faisaient la différence dans les moments cruciaux.
  - **Remplaçants de luxe :** Ces joueurs n'étaient pas toujours titulaires mais avaient un impact énorme lorsqu'ils étaient appelés à entrer sur le terrain. Ils étaient capables de changer le cours d'un match grâce à leur talent et leur détermination.
  - **Attaquants spécialisés dans la finition :** Ce cluster regroupait des joueurs qui excellaient dans la conversion des occasions en buts.

Ils étaient moins impliqués dans la création d'occasions mais étaient des finisseurs redoutables.

- **Milieux de terrain créatifs :** Ces joueurs étaient responsables de la construction du jeu et de la création d'occasions de but. Ils possédaient une vision du jeu exceptionnelle et étaient souvent les meneurs de leur équipe.
- **Défenseurs et milieux défensifs :** Ce groupe se composait de joueurs dont le rôle principal était de protéger leur but et de récupérer le ballon. Leur contribution offensive était limitée, mais ils excellaient dans les interceptions et les tacles.

Malheureusement, un problème technique a entraîné la perte de ces données, ce qui m'empêche de présenter les résultats complets. Cependant, cette expérience m'a permis de découvrir la puissance du clustering. Les résultats étaient souvent très proches de la réalité du football, confirmant ainsi l'efficacité de cette méthode pour analyser et comprendre les performances des joueurs.

11 Références

- Sources de Données : *https://www.kaggle.com/datasets/sarangpurandare/83k-football-players-103-stats-per-season*

12 Annexes

12.1 Sortir R

```
> library(MASS)
> lda(tournament_name_most_frequent ~., data=df2)
Call:
lda(tournament_name_most_frequent ~ ., data = df2)

Prior probabilities of groups:
    Bundesliga         LaLiga        Ligue 1 Premier League        Serie A
     0.1742143      0.2167607      0.2254633      0.1772764      0.2062853

Group means:
                  goals  assists yellowCards matchesStarted appearances   rating bigChances
Bundesliga     5.723404 3.759482    6.791859       41.48289    54.87512 6.559312          5
LaLiga         5.318959 3.433457    9.470632       42.08104    55.84981 6.608294          5
Ligue 1        4.645461 2.863474    6.293781       37.41887    50.21730 6.626955          4
Premier League 6.807273 4.299091    7.901818       50.77182    66.64091 6.651184          5
Serie A        5.668750 3.428125    9.118750       43.78281    57.23906 6.568445          4
               accuratePassesPercentage accurateOwnHalfPasses accurateOppositionHalfPasses a
```

Bundesliga	73.53512	652.4987	608.8615	
LaLiga	74.18480	626.8402	676.3216	
Ligue 1	75.90892	605.1639	581.2365	
Premier League	75.13324	752.7377	857.6918	
Serie A	76.88736	672.3002	695.4629	

	keyPasses	successfulDribbles	successfulDribblesPercentage	interceptions	accu
Bundesliga	34.70749	30.47525	41.53446	50.91506	
LaLiga	33.88461	33.89770	46.66864	50.78892	
Ligue 1	30.78783	33.73860	48.43363	45.49940	
Premier League	42.86407	42.72613	48.58967	56.85549	
Serie A	37.11300	32.84682	46.65109	49.06314	

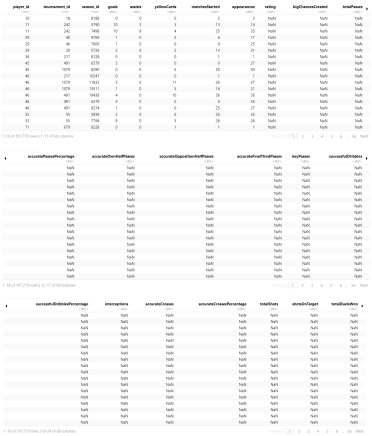
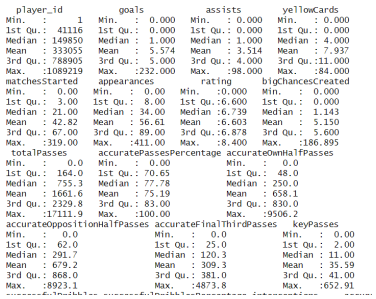


Figure 3: Impact des valeurs manquantes (Rating)



```
   player_id           goals             assists          yellowCards     
Min.   :      1   Min.   :  0.000   Min.   : 0.000   Min.   : 0.000  
1st Qu.:  41116   1st Qu.:  0.000   1st Qu.: 0.000   1st Qu.: 0.000  
Median : 149850   Median :  1.000   Median : 1.000   Median : 4.000  
Mean   : 333055   Mean   :  5.574   Mean   : 3.514   Mean   : 7.937  
3rd Qu.: 788905   3rd Qu.:  5.000   3rd Qu.: 4.000   3rd Qu.:11.000  
Max.   :1089219   Max.   :232.000   Max.   :98.000   Max.   :84.000  
matchesStarted    appearances         rating        bigChancesCreated
Min.   :  0.00   Min.   :  0.00   Min.   :0.000   Min.   :  0.000  
1st Qu.:  3.00   1st Qu.:  8.00   1st Qu.:6.600   1st Qu.:  0.000  
Median : 21.00   Median : 34.00   Median :6.739   Median :  1.143  
Mean   : 42.82   Mean   : 56.61   Mean   :6.603   Mean   :  5.150  
3rd Qu.: 67.00   3rd Qu.: 89.00   3rd Qu.:6.878   3rd Qu.:  5.600  
Max.   :319.00   Max.   :411.00   Max.   :8.400   Max.   :186.895  
 totalPasses      accuratePassesPercentage accurateOwnHalfPasses
Min.   :   0.0   Min.   :  0.00           Min.   :   0.0        
1st Qu.:  164.0   1st Qu.: 70.65           1st Qu.:  48.0        
Median :  755.3   Median : 77.78           Median : 250.0        
Mean   : 1661.6   Mean   : 75.19           Mean   : 658.1        
3rd Qu.: 2329.8   3rd Qu.: 83.00           3rd Qu.: 830.0        
Max.   :17111.9   Max.   :100.00           Max.   :9506.2        
accurateOppositionHalfPasses accurateFinalThirdPasses   keyPasses      
Min.   :   0.0               Min.   :   0.0            Min.   :  0.00  
1st Qu.:  62.0               1st Qu.:  25.0            1st Qu.:  2.00  
Median : 291.7               Median : 120.3            Median : 11.00  
Mean   : 679.2               Mean   : 309.3            Mean   : 35.59  
3rd Qu.: 868.0               3rd Qu.: 381.0            3rd Qu.: 41.00  
Max.   :8923.1               Max.   :4873.8            Max.   :652.91  
```

Figure 4: Descriptions statistiques des joueurs

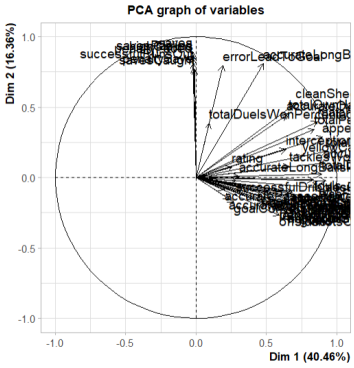


Figure 5: Graphique des Variables

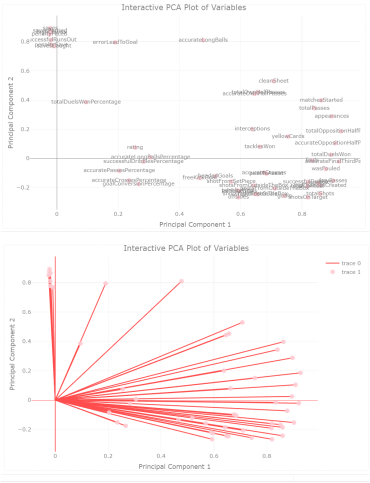


Figure 6: Graphique des Variables Interactif

Figure 7: Barplot CAH

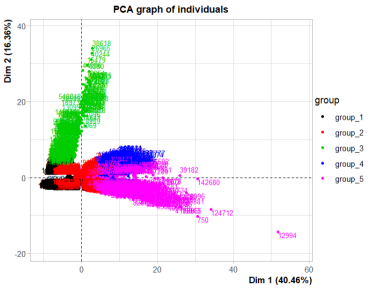


Figure 8: Interprétation des Clusters via ACP

```
## Accuracy
```{r}
accuracy_cart = mean(class_cart == data.test$tournament_name_most_frequent)
accuracy_cart
```
```

```
[1] 0.2481869
```

Figure 9: Accuracy du Modèle pour la prédiction du championnat